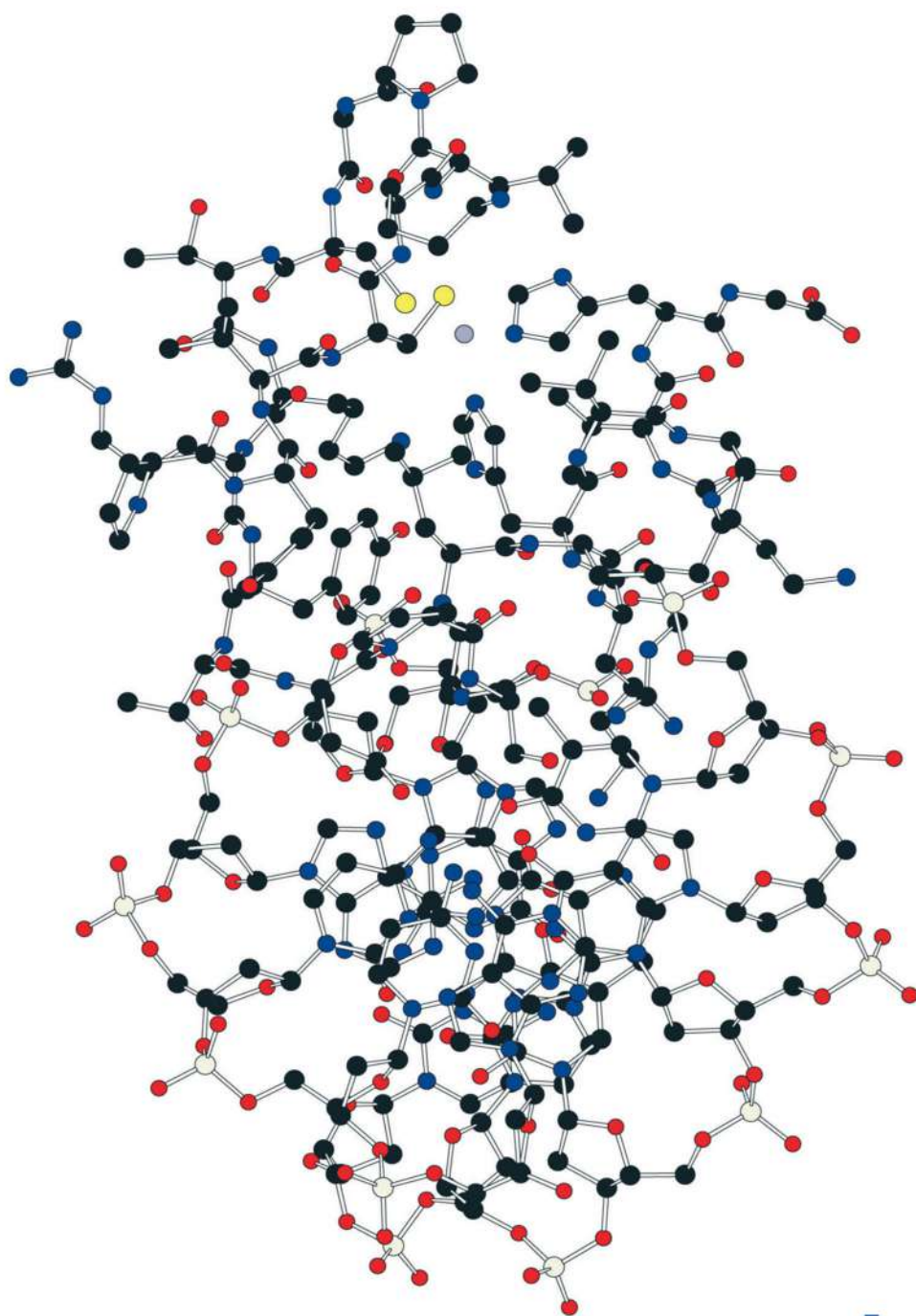


SECOND EDITION

Structure in Protein Chemistry



Jack Kyte

Structure in Protein Chemistry

*“to place before mankind the common sense of the subject,
in terms so plain and firm as to command their assent”*

Thomas Jefferson
Letter to Henry Lee, 1825

Structure in Protein Chemistry

Second Edition

Jack Kyte

Emeritus Professor of Chemistry
University of California at San Diego

Vice President	Denise Schanck
Senior Editor	Robert L. Rogers
Associate Editor	Summers Scholl
Senior Publisher UK	Jackie Harbor
Production Editor	Simon Hill
Copyeditor	Heather Whirlow Cammarn
Cover Designer	Aktiv
Typesetter	Phoenix Photosetting
Printer	RR Donnelley

© 2007 by Garland Science, a member of the Taylor & Francis Group, LLC

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

No part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

ISBN 0 8153 3867 8

Library of Congress Cataloging-in-Publication Data

Kyte, Jack.

Structure in protein chemistry / Jack Kyte. -- 2nd ed.

p. cm.

ISBN 0-8153-3867-8

1. Proteins--Structure. 2. Proteins--Analysis. I. Title.

QP551.K98 2006

572'.633--dc22

2006024536

Published in 2007 by Garland Science, a member of the Taylor & Francis Group, LLC,
270 Madison Avenue, New York, NY 10016, USA and
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN, UK.

Printed in the United States of America on acid-free paper.

10 9 8 7 6 5 4 3 2 1

Table of Contents

Preface	vii
Stereo Drawings	xi
NEWT	xiii
ExPASy	xiii
Protein Data Bank	xiii
1. Purification	1
<i>Partition into Stationary Phases and Chromatography</i>	2
<i>Assay</i>	13
<i>Purification of a Protein</i>	20
<i>Molecular Charge</i>	32
<i>Electrophoresis</i>	36
<i>Criteria of Purity</i>	45
<i>Heterogeneity</i>	47
<i>Crystallization</i>	49
2. Electronic Structure	55
<i>π and σ</i>	55
<i>Acids and Bases</i>	62
<i>Tautomers</i>	69
<i>Amino Acids</i>	74
3. Sequences of Polymers	85
<i>Sequencing of Polypeptides</i>	85
<i>Cloning, Sequencing, Expressing, and Mutating of Deoxyribonucleic Acids</i>	95
<i>Posttranslational Modification</i>	113
<i>Oligosaccharides of Glycoproteins</i>	126
4. Crystallographic Molecular Models.	149
<i>Maps of Electron Density</i>	149
<i>The Molecular Model</i>	162
<i>Refinement</i>	172
5. Noncovalent Forces	189
<i>Water</i>	190
<i>Standard States and Units of Concentration</i>	196
<i>Ionic Interactions</i>	199
<i>The Hydrogen Bond</i>	204
<i>Intramolecular and Intermolecular Processes: Molecularity and Approximation</i>	222
<i>The Hydrophobic Effect</i>	230
<i>Hydropathy</i>	241

6. Atomic Details	251
<i>Secondary Structure of the Polypeptide Backbone</i>	251
<i>Stereochemistry of the Side Chains</i>	267
<i>Hydropathy of the Side Chains</i>	272
<i>Packing of the Side Chains</i>	277
<i>Water</i>	290
<i>Ionic Interactions</i>	300
<i>Hydrogen Bonds</i>	306
<i>Association of Proteins with Nucleic Acid</i>	314
<i>Metalloproteins</i>	326
7. Evolution	345
<i>Molecular Phylogeny from Amino Acid Sequence</i>	346
<i>Molecular Phylogeny from Tertiary Structure</i>	362
<i>Domains</i>	376
<i>Molecular Taxonomy</i>	392
8. Counting Polypeptides	407
<i>Molar Mass</i>	408
<i>Electrophoresis on Gels of Polyacrylamide Cast in Solutions of Dodecyl Sulfate</i>	421
<i>Sieving</i>	423
<i>Cataloguing Polypeptides</i>	431
<i>Cross-Linking</i>	439
9. Symmetry	451
<i>Rotational and Screw Axes of Symmetry</i>	451
<i>Space Groups</i>	456
<i>Oligomeric Proteins</i>	466
<i>Isometric Oligomeric Proteins</i>	485
<i>Helical Polymeric Proteins</i>	499
<i>Heterologous Oligomeric Proteins</i>	508
10. Chemical Probes of Structure	529
<i>Covalent Modification</i>	529
11. Immunochemical Probes of Structure	555
12. Physical Measurements of Structure	573
<i>Shape</i>	573
<i>Absorption and Emission of Light</i>	592
<i>Nuclear Magnetic Resonance</i>	612
<i>Exchange of Protons</i>	640
<i>Electron Paramagnetic Resonance</i>	645
13. Folding and Assembly	659
<i>Thermodynamics of Folding</i>	659
<i>Kinetics of Folding</i>	688
<i>Assembly of Oligomeric Proteins</i>	710
<i>Assembly of Helical Polymeric Proteins</i>	717
14. Membranes	743
<i>The Bilayer</i>	745
<i>The Proteins</i>	763
<i>The Fluid Mosaic</i>	807
Index	839

Preface

Structure in Protein Chemistry is designed for a senior undergraduate or graduate course covering the structures of proteins and biophysical chemistry. The course created by this textbook is intended to bridge the gap between the research literature and the courses in introductory chemistry and biochemistry that the student has already taken. There are suggested readings at the end of each section. In these selected publications, the concepts just discussed in that section are applied in an experimental setting. There are also more than 4800 citations within the text itself that should direct the student to the scientific literature. The format of the book is intended to resemble that of a biochemical journal to ease the transition. At the completion of the course, the student should be equipped to take charge of his own education by critically reading the biochemical literature on his own. To do this he must be able to understand the experiments performed and be able to reach the same conclusions as do the authors of each publication or to realize that the authors are mistaken in their interpretations. It is my intention to develop in the student the ability to draw his own conclusions from only the experimental results. To this end, there are problems after most of the sections to reinforce the concepts that have just been presented in the text. These problems are usually based on actual experimental results, which are to be evaluated by the student, ideally in the absence of assistance or misdirection from the authors of the publications from which the results were taken.

Refined crystallographic molecular models provide most of our knowledge of the structures of proteins. Their importance and validity are self evident, and they provide the foundation on which almost all of the other experimental observations in the field must rest. They also create, in the imagination of the chemist, a reliable abstract image of what the structure of a protein consists—its atomic details, its folded polypeptide backbone, its α helices and β structure, the packing of its secondary structure, its globular or elongated shape, its irregular surface, its hydration, and the symmetric arrangement of its subunits. This abstract image of a molecule of generic protein is synthesized by her imagination from all of the particular crystallographic molecular models she has viewed. Its fully developed mental existence permits her to understand the molecular basis of all of the other physical and chemical observations that are made of proteins and thus what these observa-

tions actually mean. The abstract image also permits her to understand more clearly the evolution of proteins, the folding of proteins, and the assembly of oligomeric and polymeric proteins. Consequently, crystallographic molecular models of proteins must be discussed as soon as possible and as comprehensibly as possible in any successful presentation of the biophysical chemistry of proteins.

Structure in Protein Chemistry begins with descriptions of how proteins are purified to provide the student with an understanding of where the proteins themselves and their crystals come from. To permit him to recognize intimately the polypeptide that folds to produce the crystallographic map of electron density, the electronic and atomic details of its covalent bonds are then described, and the methods for elucidating its sequence of amino acids and defining its posttranslational modifications are explained. A comprehensive presentation of the methods of crystallography, which permits the student to understand critically its strengths and weaknesses, and a thermodynamic discussion of the properties of noncovalent forces—ionic interactions, hydrogen bonding, and the hydrophobic effect—as they are expressed in aqueous solution are a prelude to an exhaustive description of the atomic details of the structures of proteins as observed in crystallographic molecular models. The resulting understanding of their molecular structures at the atomic level and the noncovalent forces that produce those structures forms the basis for discussions of the evolution of proteins, of the symmetry of the oligomeric and polymeric associations that produce them, and of the chemical, mathematical, and physical basis of the techniques used to study their structures such as image reconstruction, nuclear magnetic resonance spectroscopy, proton exchange, optical spectroscopy, electrophoresis, covalent cross-linking, chemical modification, immunochemistry, hydrodynamics, and the scattering of light, X-radiation, and neutrons. The application of these procedures to the study of the folding of polypeptides and the assembly of oligomers and helical polymers is then described. Finally, biological membranes and the structures of their proteins are discussed.

To present a comprehensive view of the biophysical chemistry of proteins, this text combines concepts of bonding and chemical reactivity, descriptions of macromolecular structure, principles of thermodynamics, and

explanations of biophysical methods and their results. The concepts of bonding and chemical reactivity are presented in standard structural drawings of individual molecules or chemical reactions in which electronic and mechanistic aspects are emphasized as they are in courses in organic chemistry. The descriptions of macromolecular structure are illustrated with stereo images of crystallographic molecular models that are drawn by the author so that details appropriate to the particular points made in the text are emphasized by choosing the appropriate views of the structures. The principles of chemical thermodynamics are applied in relationships among the equilibrium constants and fundamental state functions. The explanations of biophysical methods rely on the mathematical equations defining the physical properties being measured. The results of the experiments themselves are found in graphs and tables derived from the experimental literature. It is this combination of chemical drawings, stereo images, mathematical equations, graphs, and tables that makes this book both unique and comprehensive. It also places severe demands on the student. She must have a firm background in physics, mathematics, analytical chemistry, organic chemistry, and physical chemistry to understand the material. In the broadest sense the intention of the course is to educate protein chemists. A protein chemist should be able to evaluate critically the results of any of the methods applied to the study of proteins.

The foregoing describes both the First Edition and the Second Edition of *Structure in Protein Chemistry* but the Second Edition is a major revision of the first. All of the sections in each of the chapters in the Second Edition of *Structure in Protein Chemistry* have been updated extensively to include the relevant observations and new discoveries in the field that have been made since the First Edition was written.

The significant progress that has been made since that time has required that some sections of the book be completely rewritten. For example, because of the explosion of knowledge in the area of protein folding, the section on the kinetics of folding has been completely redone. Likewise, there has been a dramatic increase in the number of crystallographic molecular models of oligomeric proteins so that examples are now available of all of the point groups for the symmetric assembly of asymmetric objects. As a result, because oligomeric proteins and isometric oligomeric proteins can now be discussed more systematically, the sections covering their structures have also been completely reorganized and rewritten, and stereo drawings of crystallographic molecular models of proteins representing each point group are included.

Completely new sections have also been added to the book. A new section on the structural details of the interactions between proteins and nucleic acids has been added, in part to recognize the significant progress that has been made in this area. The explosion of new

crystallographic molecular models over the last two decades has included many with heterologous oligomeric associations where few were available at the time that the First Edition was being written. Consequently, a new section discussing oligomeric proteins that are constructed heterologously has been added. As part of this section, the major classes of these proteins are discussed, including proteins involved in cellular control, motility, the cytoskeleton, the extracellular matrix, cellular adhesion, and cell-cell interactions. There is also a completely new section on the roles of metallic cations in the structures of proteins.

There are other instances in which major advances have led to extensive additions to the text. Descriptions and drawings of the crystallographic molecular models of representatives of the various classes of integral membrane-bound proteins, which were mostly unavailable for the First Edition, have been added. There is now a comprehensive description of mass spectrometry and its application to the direct sequencing of proteins, the elucidation of the structures of posttranslational modifications, and the determination of the molar masses of proteins. The section on sequencing and modifying DNA has been extensively expanded to include developments in this rapidly advancing area. The number of posttranslational modifications included in the section covering this topic has been significantly increased, a reflection of the new discoveries in this area. In particular, the recently elucidated role of inteins in the posttranslational rearrangements of the polypeptide backbone is described. There is a new discussion of the results of crystallographic molecular models of atomic resolution (Bragg spacing less than 0.1 nm) because many of these have also become available since the First Edition was written. The section on hydrogen bonding in proteins has been significantly improved by including the results of double mutant cycles, a procedure that has been developed since the First Edition was written. How the most widely used algorithms for searching data banks of amino acid sequences work is described. There is a new, detailed discussion on how an icosahedral assembly is expanded by incorporating segments of a hexagonal array, which is the strategy that viruses have used to increase the size of their coats. The use of physical measurements of a protein in solution to adjust its crystallographic molecular model, also a new development, is now discussed in the context of comprehensive descriptions of the techniques that are used to make these adjustments. For example, scattering curves from solutions of a protein are now used to adjust its crystallographic molecular model to the structure that it assumes when it is in solution.

In several instances, descriptions of procedures have been made more comprehensive to improve the student's understanding. The section on nuclear magnetic resonance has been significantly updated to describe the improvements that have been made in this

field since the First Edition, but the physical basis and the techniques of nuclear magnetic resonance spectroscopy itself are now more comprehensively discussed so that a more complete understanding of the method is gained. The limited description of electron paramagnetic resonance spectroscopy in the First Edition has been expanded to create a new section in which examples of its recent use are presented. The use of image reconstruction and cryo-electron microscopy to produce structures of helical polymeric proteins and membrane-bound proteins is more comprehensively discussed than it was in the First Edition.

All of these changes together have created a text that is not only an update but also a significant expansion of the First Edition.

It is a pleasure to thank everyone who has helped me in the preparation of this book. First and foremost I thank my wife Francey. She has entered into the computer in the proper places and the proper order all of the almost impossible to follow changes and insertions that were haphazardly written in pencil and red pen over the typescript of the First Edition or written out in my hand as inserts on sheets of scrap paper, while at the same time correcting my spelling, grammar, and punctuation. Without her assistance, it would have taken me at least an additional year to finish the job less successfully. Daniel Louvard was kind enough to provide me with an office at the Institut Curie and access to its library in the years 1995–1996 and 2000–2001 so that I could pursue the project while away from La Jolla. I would also like to thank Heather Whirlow Cammarn, my copyeditor, who converted the manuscript into the style of the American

Chemical Society and tied up all of the many loose ends with acumen. I would again like to thank all of the reviewers of the First Edition because much of their assistance has been carried into the Second Edition. Russell Doolittle and Harvey Itano read large portions of the manuscript of the First Edition and provided excellent suggestions. Individual sections of the manuscript of the First Edition were reviewed critically by Frank Huennekens, Bruno Zimm, Charles Perrin, Steven Clarke, Ajit Varki, David Matthews, John Edsall, Cyrus Chothia, Arthur Lesk, David DeRosier, Nigel Unwin, Stephen Harrison, Fred Hartman, John Simon, George Fortes, Rachel Klevit, Ken Dill, Robert Baldwin, Howard Shachman, Dennis Haydon, and Guido Guidotti. I would like to thank all of the reviewers of the Second Edition. Individual sections of the manuscript of the Second Edition were reviewed by Larry Cummings, Martin Webb, Iain Nicholl, Jeffrey Carbeck, Lloyd Waxman, Partho Ghosh, Charles Perrin, Kenneth Walsh, Tama Hasson, Steven Clarke, Ajit Varki, Brian Matthews, the late Carl-Ivar Brändén, Dave Matthews, Ken Dill, V. Adrian Parsegian, Michael Page, Malcolm MacArthur, Patrick Argos, Stephen Harrison, William Trogler, Russell Doolittle, Henryk Eisenberg, Pierre Goloubinoff, Robert Fletterick, Georg Schulz, Michael Rossmann, Ron Milligan, Fred Hartman, Donald Engelman, David Johnson, Walter Englander, C. Nick Pace, Franz Schmid, Arshad Desai, Stephen White, and Douglas Rees. Each of them provided detailed criticism, many helpful comments, and reassurance.

Jack Kyte



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Stereo Drawings

Almost all of the stereo drawings of crystallographic molecular models included in *Structure in Protein Chemistry* were produced by the program Molscript created by Per J. Kraulis. If you have the time and enjoy working on a computer, you should learn how to use the program, which is described at <http://www.avatar.se/molscript/doc/molscript.html>. It is now standard practice to publish drawings of crystallographic molecular model in this format. To appreciate the results of crystallographic studies, one must be able to view these images. Although a few individuals can view them effortlessly by crossing their eyes, the rest of us need a stereo viewer. The stereoviewer that I use and have recommended for my students is the PEAK™ Pocket Stereo Viewer with 2× magnification (124 mm legs). Suppliers of this viewer can be found using Google. It has been my experience that a student who has never viewed a stereo drawing before

will usually complain that although everyone else can learn to use one of these viewers, he cannot. It is also my experience that everyone learns to use one. When I have put a question on an examination such Problem 4.5, where one is asked to write down the sequence of the protein by examining a drawing of a crystallographic molecular model that she has never seen before, everyone in the class gets at least 90% of the sequence correct, which would have been impossible unless everyone was able to see the image in stereo. It is essential that anyone interested in the structures of proteins learn to view drawings of crystallographic molecular models in stereo. The drawings in this text have been placed vertically rather than in their usual horizontal orientation and each has been placed on the outside edge of a page. This has been done to allow each image to be spread as flat as possible for the best viewing.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

NEWT

There are hundreds of proteins discussed in the text of *Structure in Protein Chemistry*. Each of these proteins is present in many different species of organisms, but usually the details that are being discussed are specific enough that the protein from only one of these many species is described, even though what is described would fit the protein from any one of these species. Furthermore, a particular protein from a particular species is always used in a particular experiment. The names of both the protein being discussed and the species from which it was derived are usually stated in the text. It turns out that protein chemists, because they realize that the same protein from different species of organisms is basically the same, don't really care from what species of organism the protein comes, and a remarkably large collection of species of organisms are used as sources for proteins. Usually, in a particular

investigation, one particular species is chosen as a source for a particular protein for a particular reason known only to the investigator. The practical result of these diverse choices is that the names of hundreds of species of organisms are used in this book. I have chosen to name each of them with the usual Latin names of their genus and species, without explaining what the species are because I wanted to make the point that it doesn't matter where a protein comes from. The names *Escherichia coli* and *Saccharomyces cerevisiae* and the adjectives murine, equine, bovine, canine, and human are probably already familiar to you but very few of the other names will be. Even though there is no need to know, if you would like to know to what the name of a genus and species refer, go to <http://www.ebi.ac.uk/newt/display>, and enter the name of the species.

ExPASy

Hundreds of thousands of proteins from hundreds of different species of organisms have been sequenced. The sequences of their amino acids are tabulated in large data banks. The most easily used of the data banks is the Swiss-Prot/TrEMBL at <http://www.expasy.org/>. You

should become familiar with this site on the web, not only for the sequences it makes available but also for the free programs that are available at the site to analyze those sequences.

Protein Data Bank

The Protein Data Bank at <http://betastaging.rcsb.org/pdb/Welcome.do> contains the atomic coordinates of most of the crystallographic molecular models that have been constructed. At the moment there are 36,000 separate molecular models entered in the data bank. You should look at some of the lists of the full coordinates to get a feeling for what such a file contains. Enter the name of a protein for which there is a stereo drawing in the text of this book, click on the name of one of the molecular

models that are then listed, choose "Download Files", and then choose "PDB File". The atoms are listed by the name of the amino acid, the position of that amino acid in the sequence of the protein, and their locations within that amino acid by using the abbreviations given in Figure 4.14. The list is that of the *x*, *y*, and *z* coordinates of each atom in Ångströms. Each file constitutes the raw data on which the molscript program operates.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Chapter 1

Purification

The living world that teems around us, the world of species, individual organisms, organs, tissues, and cells, can be viewed as the manifestation of a vast fluid array of protein molecules, each appearing and disappearing in the proper place at the proper time. This array of protein molecules is the outcome of a long history. Each protein within the array is itself the product of evolution by natural selection, which has had more than two billion years and much of the surface of the earth to explore, by random, irrational trial and error, strategies with which to accomplish the function of that protein. There are several consequences of this fact. First, chemical principles in addition to those of which we are aware have been discovered and exploited. Second, completely different chemical mechanisms often have been applied haphazardly to achieve similar purposes. Third, there are puzzling features that are inefficient, useless, or meaningless. Fourth, the result of this process does not resemble anything the human mind would have designed, even if it were aware of all of the available chemical strategies. One consequence of these facts is that argument by exclusion is useless because it cannot be assumed that the mechanism by which a biological problem was solved is only one or more of the mechanisms of which we can conceive.

One fruitful approach in our attempt to understand life has been to study, individually or in small groups, the proteins that produce it to gain insight into the role of each one in the overall scheme. An argument could be made that a cell does seem to be no more than the sum of its parts and that a significant understanding of how it accomplishes its purpose can be gained by studying those parts individually. Because the proteins are the parts of a cell that perform almost all of the chemical and structural transformations that occur within it, they have attracted the most attention.

The most dynamic region in a living organism is the cytoplasm of the cells or cell from which it is made. About 20–30% of the total mass of cytoplasm is protein dissolved in a solution the solvent of which is water. The cytoplasm is enclosed within a thin, fragile, continuous membrane. About 60–80% of the dry weight of this membrane is protein dissolved in a solution, the solvent of which is lipid. This membrane is surrounded and supported by a tough protective integument of polysaccharide; polysaccharide and protein; or polysaccharide, lipid, and protein. Organelles, enclosed within their own

membranes, are often scattered through the cytoplasm. In a eukaryotic cell the largest of these is the nucleus, containing most of the nucleic acid in the cell.

The strategy that has been applied most frequently to the study of proteins is to identify a particular biological feature of a living organism and then purify the protein or proteins responsible for it. Typically, when a complex, beautiful, intricately organized biological specimen, such as a tissue or a suspension of cells, is submitted to the first step in any purification procedure, it is immediately sundered beyond recognition and becomes a nondescript jumble of its organelles and broken fragments of its membranes and their integuments suspended in an aqueous solution of proteins, nucleic acids, metabolites, and salts. This event is referred to as homogenization. It is usually accompanied by the dilution of the proteins in the initial specimen by addition of a buffered aqueous solution. Following the homogenization, insoluble fragments are removed by centrifugation to produce a clear solution, the protein concentration of which is 1–10%. This solution contains most of the proteins that were once the living cytoplasm of the specimen. It is from this solution that particular proteins can be isolated. The purification of a protein is the separation of that protein from all of the others in a homogenate. A particular protein must be purified before its molecular structure can be studied.

Usually, the only interest that one has in a particular protein arises from its participation in some process of biological importance. It might be an enzyme responsible for catalyzing a particular reaction; it might be a structural protein creating the macroscopic shape of the cell; it might be a protein that binds a hormone or neurotransmitter; or it might be a protein that binds to DNA and controls its transcription. To distinguish one protein from the others in a complex mixture, an assay for the protein of interest, based on its particular function, is required.

The most widely used procedure for purifying proteins is chromatography. This technique separates molecules of protein by differences in the rate at which they move along a cylinder of a porous solid phase as a liquid phase percolates through it. If the solid phase is properly chosen, each protein travels through the cylinder at a different rate and each emerges in the solution coming out of the cylinder at a different time. In this way, one can be separated from the others. In order to distinguish the

protein of interest from the others as they emerge from the chromatographic column, the assay for that protein is used. As the protein becomes purified, the preparation displays greater and greater activity in the specific assay for a given amount of total protein.

Once the protein has been purified, analytical methods must be used to demonstrate that only one protein is the major component in the final preparation and that this protein is responsible for the biological function of interest. The analytical procedure most suited to this demonstration is electrophoresis. Electrophoresis separates proteins by both their charge and their shape, and if used with discontinuous stable boundaries, electrophoresis can have high resolution.

Once a protein of known function has been purified to homogeneity, it can be crystallized. As in organic chemistry, crystallization is a way of harvesting a particular substance in a highly purified form. Ideally, every protein that was purified would be crystallized and stored in this form, as are organic molecules. In this form, each suspension of crystals would represent a pure chemical compound. In practice, because crystals are often difficult to make and yields in crystallizations are poor, purified proteins are usually left in solution or precipitated for storage. It is these solutions, precipitates, or suspensions of crystals that are the raw material for studies of the structures and functions of the proteins they contain. The purpose of this chapter is to describe how a particular protein is purified from a complex mixture of proteins such as the homogenate of either a tissue or a suspension of cells.

Adsorption to stationary phases and chromatography are the bases for both the purification of proteins and many of the assays used to identify particular proteins, so these processes will be considered first.

Partition into Stationary Phases and Chromatography

The goal of any procedure used to purify a particular substance from a complex mixture is to separate that substance from all of the other components in the mixture. When adsorption or chromatography is used for this purpose, differences in the preferences of solutes in a solution for another phase are exploited. The simplest example of such a strategy is an affinity adsorbent. Suppose a small molecule that could be tightly bound by only one particular protein in a solution was covalently attached to a solid surface. By binding them specifically, this affinity adsorbent would collect molecules of only that one protein on its surface. The rest of the molecules of protein in the solution could be washed away, and the molecules of the desired protein could then be released. Unfortunately, such highly specific adsorbents are not usually available, so small differences in affinity among proteins or among other molecules in a solution for a

separate phase are amplified by the process of chromatography.

When a chemical substance A, which will be referred to as the solute, is added to a vessel containing two immiscible phases and the system is allowed to come to equilibrium, the solute A will distribute between the two phases in a characteristic manner. The solute can be an inorganic ion, a small organic molecule, a protein, a nucleic acid, a polysaccharide, or any other similar substance. The two phases can be, for example, two immiscible liquids, a liquid and a solid, or a gas and a liquid; the only requirement is that those two phases be brought into sufficient contact to permit the **distribution of solute A** between them to reach equilibrium and that they then be separated in some way that does not redistribute the solute. The simplest examples are a two-phase, solvent-solvent extraction or the suspension of some finely divided solid in a liquid followed by its removal from the liquid by filtration.

After the equilibration and separation of the two phases, the moles of solute A in each of them can be determined. In the cases that are generally encountered, at least one of the phases is a fluid that can be freed entirely of the other phase. This fluid will be arbitrarily called the **mobile phase**. In the special case when a protein is solute A, the mobile phase is invariably an aqueous solution of moderate ionic strength buffered at a specific pH. In any situation, however, the molar concentration of solute A in the mobile phase can be readily measured. The second phase, arbitrarily referred to as the **stationary phase**, can be an immiscible liquid, a solid, or a solid in which a liquid is entrapped. Because of the peculiarities of this stationary phase, the best way to express the concentration of solute A that has become physically associated with the stationary phase, $[A]_s$, is in moles (liter of bed)⁻¹, where the volume of the bed is the volume filled by the stationary phase when it has settled.*

Three general types of behavior¹ have been observed in such a partition (Figure 1-1). The simplest behavior, type A, occurs when the concentration of the solute A in the stationary phase increases in direct proportion to its concentration in the mobile phase. This type of behavior is encountered in solvent-solvent extractions or in chromatography by molecular exclusion. In the latter example, it results from the fact that the stationary phase is nothing more than trapped, and thereby immobilized, mobile phase. Behavior of type B (Figure 1-1) is encountered when the stationary phase saturates with solute A. It results from the presence of only a finite number of sites on the stationary phase that are all equivalent in their individual affinities for solute A

* Concentrations in moles per liter of bed are indicated by primed notation. Concentrations in moles per liter of stationary phase or moles per liter of mobile phase are in the usual unprimed notation.

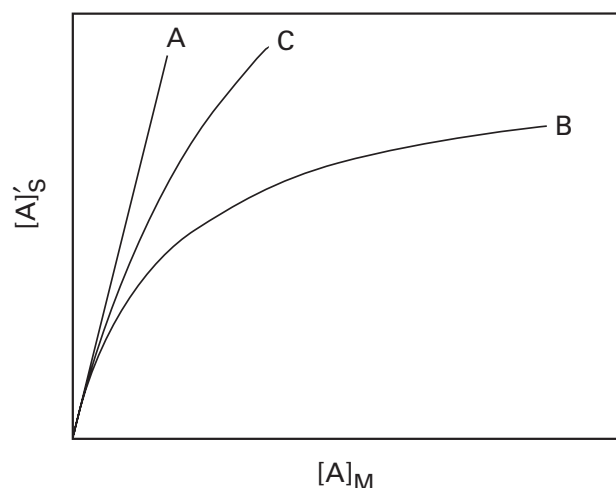


Figure 1-1: Partition of a solute between two phases referred to as the stationary phase, S, and the mobile phase, M. The concentration of solute A at equilibrium in the stationary phase in units of moles (liter of bed)⁻¹, $[A]_S$, is presented as a function of its molar concentration in the mobile phase in units of moles (liter of fluid)⁻¹, $[A]_M$, for three types of behavior designated A, B, and C.

and that are distributed over the stationary phase so that they do not interact with each other at saturation. Behavior of this type is encouraged by choosing microscopically uniform stationary phases. It is advantageous because at low concentrations of solute A the partition of solute closely approximates the direct proportionality of behavior of type A. Stationary phases showing this type of behavior are highly uniform ion-exchange resins or uniform, inert matrices to which molecules of a small organic compound displaying an affinity for solute A have been randomly and sparsely attached. In more heterogeneous stationary phases, specific sites with which molecules of solute A associate are composed and distributed in such a way that they have an array of different affinities. This means that the small number of sites with high affinities for solute A are occupied first, followed by those with lower and lower affinities sequentially. This produces behavior of type C (Figure 1-1), which is unpredictable and not uniform. Examples of a stationary phase of this type are crude hydroxylapatite or a matrix to which a polyclonal immunoglobulin has been attached.

All three of these examples are to some extent idealized descriptions of actual behavior. The deviation from ideal behavior that is the most important to the present discussion, however, is that observed during the physical adsorption of molecules of protein onto the surfaces of a solid phase.² In this circumstance, although apparently ideal behavior of type B is observed at short times, the fraction of the adsorbed protein actually in equilibrium with the protein in solution decreases with time as the amount of irreversibly bound protein increases. It is believed that in such instances molecules of protein are denatured at the interface and that the interactions of these denatured molecules with the surface are much stronger than those of the undenatured

molecules. This essentially **irreversible adsorption** of protein to the surfaces of a solid phase probably occurs in any chromatographic separation and is experienced as a less than theoretical yield of the protein collected from the chromatographic system. Often this loss is inconsequential or tolerable. Because this process is a slow one,² a consequential loss of protein due to irreversible adsorption upon chromatography can be decreased by decreasing the time during which the protein is in contact with the solid phase. Such loss of protein can also be avoided by choosing a solid phase such as agarose or polyacrylamide that is less prone to producing interfacial denaturation. It also helps to use the same solid phase repeatedly because the sites at which irreversible adsorption occur become saturated over several uses. This strategy is inappropriate, however, if the sites at which irreversible adsorption occurs are the very sites upon which the desired reversible adsorption occurs, and repeated use gradually poisons the system.

The earliest use of distributions of solutes between two phases was **selective adsorption**. Selective adsorption is a technique in which conditions are sought that promote the almost complete confinement of the substance of interest to one phase while other, unwanted substances distribute into the other phase and can thus be discarded. When a protein is being isolated in this way, the ionic strength, pH, temperature, and choice of stationary phase is varied until conditions are found that permit the protein of interest to distribute almost completely into one of the two phases while the maximum amount of the other, undesired proteins distribute into the other. An example of this strategy is one of the steps in the purification of the protein fumarate hydratase.³ Calcium phosphate gel was added to a crude mixture of proteins containing fumarate hydratase dissolved in 0.1 M sodium acetate, pH 5.2. All of the fumarate hydratase (>95%) associated tightly with the calcium phosphate gel. After the gel was washed, the adsorbed fumarate hydratase was then eluted in 97% yield with 5% $(\text{NH}_4)_2\text{SO}_4$ and 0.1 M sodium phosphate, pH 7.3, even though only 20% of the original protein remained in the final solution.

Selective adsorption is a rather unsophisticated use of the distribution of a solute between a mobile and a stationary phase. It can be remarkably improved upon by operating at concentrations of solutes low enough that $[A]_S$ is directly proportional to the molar concentration of A in the mobile phase (behavior of type A) and by causing the mobile phase to move slowly through or across the stationary phase. This process is known as chromatography.

Chromatography is the process by which solutes are separated from one another on the basis of differences in the rate at which they pass across a bed of stationary phase through which a liquid mobile phase is continuously flowing. A chromatographic system is designed so that the mobile phase passes by the

stationary phase in such a way that the contact between the two phases is maximized and equilibration of the solute between them is encouraged. Examples are **paper chromatography**, in which the liquid mobile phase moves down the paper while flowing among the cellulose fibers that form the stationary phase; **thin-layer chromatography**, in which the liquid mobile phase creeps up a thin layer of the solid, dry stationary phase drawn by the capillary force arising from its movement between finely divided particles; **column chromatography**, in which the fluid mobile phase percolates through a finely divided, solid stationary phase compacted in a cylinder; and **gas-liquid chromatography**, a type of column chromatography in which a gas containing the solutes is passed through a finely divided solid phase coated with a liquid of low volatility. All of these are examples of zonal chromatography.

Zonal chromatography is chromatography in which the mixture of solutes to be separated is introduced in a thin zone at one end of the bed of stationary phase and the mobile phase is then set in motion. The molecules of solute in the mixture meander through the system, drawn forward by the movement of the mobile phase but retarded by the stationary phase in which each spends a certain fraction of its time. The fraction of the time each solute spends in the stationary phase is determined by its affinity for the stationary phase, and this is determined by its bulk distribution behavior (Figure 1-1). Since the molecules of each solute spend a different fraction of their time in the immobility of the stationary phase, each solute moves through the system at a different rate and the components of the mixture are isolated one from the other into separate zones, which are also referred to as **peaks** or **bands**. The separated solutes are collected either by dividing the stationary phase itself and extracting them, as in paper chromatography, thin-layer chromatography, or countercurrent distribution chromatography, or by continuously collecting the mobile phase as it emerges at the opposite end of the bed of the chromatographic system, as in column chromatography or gas-liquid chromatography. Any visual display of the distribution over the field of the chromatographic system of one or more of the substances being separated is referred to as a **chromatogram**.

The important properties of the chromatogram are the relative mobilities of the solutes, the widths of the peaks of the concentrations of the solutes at their half heights, and the resolution of those peaks one from the other. The **relative mobility**, $R_{f,A}$, of a particular solute A is either (1) the distance that the peak of its distribution has traveled through the system divided by the distance traveled by the mobile phase or (2) the total volume of the mobile phase in the bed of the system, referred to as the **void volume**, V_0 , divided by the total volume that has passed through the system before the peak of the distribution of solute A emerges, referred to as its **elution volume**, $V_{e,A}$. Definitions 1 and 2 are two different ways to

define the same parameter. The **width** of the distribution of solute A at **half height**, $w_{1/2,A}$, is the width, in units of distance for definition 1 or volume for definition 2, between the two points at which the concentration of solute A is half its maximum concentration at the peak. The **resolution**, R_{AB} , between two solutes is a measure of the completeness with which they are separated, a property that increases as the difference in their relative mobility increases and decreases as their widths at half height increase. The larger the differences in the various $R_{f,i}$ and the smaller the various $w_{1/2,i}$, the more successful will be the separation of the different solutes i . Expressions for $R_{f,A}$, $w_{1/2,A}$, and R_{AB} as functions of parameters that can be manipulated are of value in the understanding and design of chromatographic separations.

There are two approaches to describing the phenomenon of chromatography in theoretical terms.⁴ It can be treated as the continuous process that it is, and differential equations can be formulated to describe the differential changes in solute positions and concentrations with time. These differential equations, however, do not have simple solutions, nor do they lead to an intuitive understanding of the process. The alternative approach is based on the concept of the theoretical plate, which was developed originally to describe the separation performed by a fractional distillation column.^{5,6} Although this is a discontinuous model for a continuous process, the treatment is formulated in terms of an easily understood mechanism and does provide, in at least one case, that of countercurrent distribution chromatography, an exact solution to the problem. Martin and Synge⁷ were the first to apply this model to the process of chromatography.

Suppose that a chromatographic separation always operates at concentrations of solute A such that the amount associated with the stationary phase and the mobile phase in the chromatographic system is a linear function of its concentration in the mobile phase (behavior of type A, Figure 1-1). If so, at equilibrium

$$[A]'_S = \alpha'_A [A]'_M \quad (1-1)$$

where $[A]'_M$ is the concentration of solute A in the mobile phase in units of moles (liter of bed)⁻¹, where the **volume of the bed**, V_T , is the volume filled by the stationary and mobile phases together as they are packed into the chromatographic system; and where α'_A is a **partition coefficient**. The units for $[A]'_S$ are, as defined earlier, moles (liter of bed)⁻¹.

The bed of the chromatographic system is formally divided into a series of equivalent theoretical plates. A set of **theoretical plates** is a set of contiguous compartments of equal volume formed by a set of evenly spaced planes passing through the bed normal to the direction in which the mobile phase flows. The **height equivalent to a theoretical plate**, h , is the distance the mobile phase must

move, at the rate of normal flow, past the stationary phase until the concentration of solute in the fluid emerging from the theoretical plate is equal to the concentration the solute would have had if the fluid entering the theoretical plate had come into equilibrium with the stationary phase that fills the theoretical plate. For example, if the fluid entering the upstream boundary of the theoretical plate had a concentration of solute A equal to $[A]_{M,ent}'$ and the stationary phase already had solute A immobilized within it at a concentration of $[A]_{S,im}'$, the formal downstream boundary of the theoretical plate would occur at the point where the concentration of the solute in the mobile phase, $[A]_{M,lv}'$, had reached a value

$$[A]_{M,lv}' = \frac{[A]_{M,ent}' + [A]_{S,im}'}{1 + \alpha'_A} \quad (1-2)$$

where all concentrations are expressed in moles (liter of bed)⁻¹.

With this definition, the continuous process of zonal chromatography is equivalent to the following discontinuous sequence of events. A number of moles of solute A equal to $m_{TOT,A}$ is added to the first theoretical plate and allowed to come to equilibrium between the stationary and mobile phases. The entire mobile phase of each plate in the system is then moved to its neighbor downstream, and mobile phase containing no solute is added to the first plate. After the new situation is allowed to come to equilibrium, the same transfers of mobile phases are made. The cycle of equilibrium and transfer is repeated n times. A machine⁸ that performs **chromatography by countercurrent distribution** mechanically proceeds through this exact sequence of transfers. The theoretical plates in the countercurrent machine are individual glass vials, the mobile phases are equal volumes of an aqueous solution, the stationary phases are equal volumes of an immiscible organic solvent, and the steps of equilibration and transfer are discrete. Normally, however, this sequence of events is a theoretical simplification only formally equivalent to what actually happens.

At the conclusion of n steps, the downstream boundary of the mobile phase will have moved through the chromatographic system a distance d_M where

$$d_M = nh \quad (1-3)$$

the number of steps times the height of a theoretical plate. It can be shown^{7,9} that after n steps the mean of the distribution of solute A, and hence the peak of its distribution, will have moved a distance d_A where

$$d_A = \frac{nh}{1 + \alpha'_A} \quad (1-4)$$

while the width of its distribution at half height will be

$$w_{1/2,A} = \frac{h\sqrt{n}[\sqrt{8(\ln 2)\alpha'_A}]}{1 + \alpha'_A} \quad (1-5)$$

In thin-layer chromatography and paper chromatography, the flow of mobile phase up the thin layer or down the paper is stopped before the downstream boundary of the mobile phase reaches the end of the stationary phase. The number of steps n that have occurred is defined by the fact that the boundary has moved a distance nh from the origin at which the solutes were applied. If the relative mobility of solute A is defined as the distance solute A has moved, d_A , divided by the distance the boundary has moved, d_M , then

$$R_{f,A} = \frac{nh}{nh(1 + \alpha'_A)} = \frac{1}{1 + \alpha'_A} \quad (1-6)$$

By combining Equations 1-3, 1-5, and 1-6

$$w_{1/2,A} = \sqrt{d_M} h \left[\sqrt{8(\ln 2) R_{f,A} (1 - R_{f,A})} \right] \quad (1-7)$$

The distance that solute A has moved through a given chromatographic system, d_A , is directly proportional to the number of theoretical plates through which the mobile phase has moved (Equation 1-4), but the width of its distribution, $w_{1/2,A}$, is proportional to the square root of the number of theoretical plates through which the mobile phase has moved (Equation 1-5). As a result, as the chromatography progresses, solutes separate from each other more rapidly than they spread, and it is this property that permits chromatography to perform separations. This property can be quantified as the resolution between any two solutes.

If the resolution, R_{AB} , between the distribution of solute A and the distribution of solute B is defined as

$$R_{AB} = \frac{2|d_A - d_B|}{w_{1/2,A} + w_{1/2,B}} \quad (1-8)$$

then by assuming that h is the same for both solutes and combining Equations 1-3 through 1-5

$$R_{AB} = \sqrt{\frac{d_M}{h}} \left\{ \frac{2|\alpha'_A - \alpha'_B|}{\sqrt{8(\ln 2)} [\sqrt{\alpha'_A(1 + \alpha'_B)} + \sqrt{\alpha'_B(1 + \alpha'_A)}]} \right\} \quad (1-9)$$

Because α'_A and α'_B are fixed properties of the stationary phase, the solvent, and the solutes, this equation demon-

strates that resolution is increased either by decreasing the height of a theoretical plate or by running the chromatography over a greater distance.

In column chromatography the effluent emerging from the end of the column is collected and the concentration of solute A in this effluent is monitored as a function of the total volume that has emerged since the chromatogram was begun. If the column of stationary phase contains p theoretical plates, the effluent collected and monitored is, by definition, the mobile phase entering plate $p + 1$. As mobile phase emerges from the end of the system, the concentration of solute A that it contains increases, reaches a maximum, and then declines. This results from the approach of the peak of the distribution of solute A to plate $p + 1$, its arrival at plate $p + 1$, and its passage beyond plate $p + 1$. The volume at which the maximum passes through plate $p + 1$ is the elution volume of solute A, $V_{e,A}$. It corresponds to the volume of mobile phase that must pass through the system to bring the maximum of the distribution of solute A into plate $p + 1$. Because it takes p steps for a volume equal to the void volume V_0 to emerge from the column but the peak of the distribution of solute A will have entered only theoretical plate $p(1 + \alpha'_A)^{-1}$ after p steps, the peak of the distribution of solute A will enter plate $p + 1$ only after a volume equal to $V_0(1 + \alpha'_A)$ has passed through the system. It follows that

$$V_{e,A} = V_0(1 + \alpha'_A) \quad (1-10)$$

and

$$R_{f,A} \equiv \frac{V_0}{V_{e,A}} = \frac{1}{1 + \alpha'_A} \quad (1-11)$$

This is the fundamental equation governing column chromatography. It connects the volume at which the solute A emerges from the end of the chromatographic column with its bulk partition coefficient for the material composing the stationary phase. The relationship between the relative mobility $R_{f,A}$ and the partition coefficient α'_A is identical to that governing thin-layer chromatography and paper chromatography (Equation 1-6). This is reassuring because it is reasonable that the same process occurs in all types of chromatography. The validity of this equation was verified experimentally by Martin and Synge.⁷

The width of the peak of concentration at half height, in units of eluted volume, can be shown to be a function of the number of theoretical plates:^{4,7,10,11}

$$w_{1/2,A} = \frac{\sqrt{8(\ln 2)}}{\sqrt{p}} V_{e,A} \quad (1-12)$$

If the resolution between the distribution of solute A and the distribution of solute B is defined as

$$R_{AB} \equiv \frac{2|V_{e,A} - V_{e,B}|}{w_{1/2,A} + w_{1/2,B}} \quad (1-13)$$

then

$$R_{AB} = \frac{\sqrt{8(\ln 2)}}{\sqrt{p}} \left(\frac{2|\alpha'_A - \alpha'_B|}{2 + \alpha'_A + \alpha'_B} \right) \quad (1-14)$$

If the solvent, ionic strength, temperature, and pH of the mobile phase and the volume and chemical structure of the stationary phase remain the same so that the values of α' are unchanged, the resolution of the separation can be improved by increasing the number of theoretical plates, p , that the column contains. The most obvious way to accomplish this is to increase the length of the chromatographic column, but this can become both cumbersome and expensive.

Because the height of a theoretical plate, h , is defined as the distance of passage required for equilibrium to be reached, h decreases and p increases as the **flow rate** of the chromatographic column is decreased, at least until diffusion between the plates becomes a significant factor. In most cases, however, diffusion is severely hindered by the structure of the stationary phase itself and almost never becomes important, and the slower the flow, the better the resolution. This is particularly important in the chromatography of proteins, especially when they are unfolded, because their slow rates of diffusion significantly decrease rates of equilibration with the stationary phase.

The height of the theoretical plate decreases as the **diameter of the particles** in a solid stationary phase decreases,⁷ and it is advantageous to use particles of solid phase that are as small as possible. The small size of the particles increases the surface area available for equilibration and decreases the distances over which the solute molecules must diffuse. The realization of this fact has led to the recent development of the **high-pressure liquid chromatography** foreseen by Martin and Synge.⁷ In such systems, the high pressure is inconsequential to the process of separation but is required to force the liquid mobile phase through the small, finely divided solid particles of the stationary phase at a realistic rate. The particles themselves are spherical in shape and of uniform diameter to promote uniform flow of as rapid a rate as possible over the bed. Because the smaller particles of the solid phase decrease the height of a theoretical plate, more theoretical plates can exist in a given length of bed. This advantage can be exploited either to increase the resolution or to decrease the length of the chromatographic column or both.

Because high pressures are used to increase the rate of flow through a shorter column, the major advantage of high-pressure liquid chromatography is the **speed** with which the chromatograms can be run. For example, when peptides are separated on chromatography by cation exchange with sulfonated polystyrene,¹² at low pressure (<500 psi), the chromatography takes about 25 h; when peptides are separated by reverse-phase adsorption chromatography,¹³ at high pressure (>1000 psi), the chromatography takes only 1 h, even though the resolution in each case is about the same. With reverse-phase adsorption chromatography, the solvents used are also more transparent to ultraviolet light, so peptides can be followed simply by their absorbance with a continuous-flow spectrophotometer.

Improvements in the size, uniformity, and rigidity of the particles of the stationary phase have permitted similar increases in the rate at which chromatography of proteins can be performed. These developments are referred to commercially as fast protein liquid chromatography. In both high-pressure liquid chromatography and fast protein liquid chromatography, the principles remain the same as before, often the solid phases remain the same as before, and the technological improvements of the original techniques are based on previously noted predictions of the original theory.

The discontinuous model presented here for chromatography has been developed for regions of the partition curves (Figure 1-1) where solute A distributes with a constant partition coefficient, α'_A . It turns out that the most usual deviation from such ideal behavior is for the stationary phase to display saturation (curves B and C, Figure 1-1). The more prominent this behavior becomes, the poorer the resolution of the chromatogram becomes.⁴ As a rule, uniform stationary phases of high capacity, by promoting the linearity of the partition function, provide the highest resolution.

The fact that, unless the number of theoretical plates is increased, peak height decreases in almost inverse proportion to α'_A (Equations 1-11 and 1-12) precludes the use of conditions where the solute has a high **affinity for the stationary phase**. Usually, conditions such as solvent, temperature, ionic strength, and pH of the mobile phase and the chemical structure of the stationary phase are manipulated to bring the values of α'_A for the solutes to be separated into a useful range, usually between 1 and 10. A variation in one of these properties of the mobile phase, however, can also be incorporated into the chromatography itself.

To this point, only isocratic zonal chromatography has been described. **Isocratic zonal chromatography** is chromatography in which the mobile phase introduced continuously into the chromatographic system remains of constant composition. It is possible, however, to vary continuously and monotonically the composition of the mobile phase entering a column. This systematic variation produces a **gradient** of one or more properties of the

mobile phase. For example, the ionic strength of the entering mobile phase can be increased continuously over time so that it is a linear function of the volume introduced into the system. Mechanical devices are available to produce linear gradients or gradients that are exponential or logarithmic or some other function of the volume by mixing two or more solutions that differ in the property to be varied. When a gradient of pH is required, the situation becomes somewhat more complicated because the pH of a solution is usually controlled with a buffer. Not only is the pH a logarithmic function of the concentrations of the conjugate acid and base of the buffer, but changing the concentrations of conjugate acid and base often affects the ionic strength. There is no requirement, however, that the gradient be some particular function of a particular property; the only requirement is that the property be varied continuously and monotonically.

The method of **gradient chromatography** is an important tool because it permits the partition coefficient of solute A, α'_A , to be decreased during the chromatographic run. This often is essential because if the partition coefficient for a particular solute is too large, it emerges from the system with such a large elution volume, $V_{e,A}$, that the width of its band is unacceptably large. To produce satisfactory chromatography, the partition coefficient must be less than 10 in most situations, but frequently the values of the partition coefficient of solutes in a complicated mixture can spread over a large range for one particular mobile phase of constant composition. By using a gradient formulated so that all of the partition coefficients for the solutes decrease continuously, even those solutes with the highest affinity for the stationary phase eventually have low enough partition coefficients to emerge from the system within a reasonable time. Usually, a gradient of ionic strength, cosolvent, or pH is employed. It is constructed in such a way that the chosen property continuously changes in a direction that will cause the solutes to have smaller and smaller affinities for the stationary phase and elute earlier than they would under isocratic conditions. For example, if a solute is being adsorbed to a nonpolar stationary phase, a gradient that increases in the concentration of a miscible nonpolar solvent in water is used to decrease gradually the affinity of the solute for the stationary phase.

The stationary phase in a chromatographic system is the **chromatographic medium**. The solid matrix composing a chromatographic medium is almost always a polymer. Both natural polymers, for example, cellulose, and unnatural polymers, for example, polymers of polystyrene cross-linked with divinylbenzene, are used. The basic polymer is often cross-linked appropriately to increase its rigidity and manufactured in the form of small spherical beads of uniform size to improve flow rates. For chromatography of small molecules such as metabolites or peptides, beads of polystyrene or silica gel are used; for chromatography of proteins, cellulose or

beads of dextran, allyldextran, polyethers, agarose, or polymethacrylate are used.* Each of these intentionally inert polymeric matrices is then modified chemically. The type of modification performed determines the molecular property used by the chromatographic system to separate the solutes.

Media for **chromatography by adsorption** are solid phases with which the solutes physically associate by noncovalent forces. Certain amorphous or heterogeneous solids such as hydroxylapatite and silica gel have long been used as chromatographic media for chromatography by adsorption. Amorphous hydroxylapatite has been used extensively in protein purification. Unfortunately, it is prone to significant irreversible adsorption,² and it is heterogeneous and saturates readily, which causes it to have nonlinear distribution behavior. All of these properties limit its resolution. Although it separates nonpolar solutes successfully, silica gel has the unfortunate property of strongly adsorbing hydrogen-bonding solutes, which precludes its use with most biological substances. **Reverse-phase chromatographic media**, however, have found wide use in protein chemistry in the separation of small molecules such as peptides and metabolites. Such media are composed of spherical beads of silica gel that have been heavily alkylated with hydrocarbons of uniform length, for example, octadecyl or octyl groups. This blocks the sites of hydrogen bonding and creates an apolar surface on the beads that adsorbs apolar functional groups on the otherwise polar solutes. Such a chromatographic medium, however, shows little affinity for completely polar solutes unless a significant portion of the silica gel has lost its apolar coating.

Beaded, cross-linked dextran, agarose, or polymethacrylate are covalently modified to produce chromatographic media for the chromatography of proteins by adsorption. The functional groups that are attached to these hydrophilic matrices during the covalent modification are hydrophobic groups such as phenyl, methyl, butyl, propyl, or *tert*-butyl groups. These hydrophobic groups associate directly with hydrophobic groups on the surface of a molecule of protein that are the side chains of the amino acids valine, leucine, isoleucine, and phenylalanine.

In media for chromatography by adsorption, the affinity of the molecules of the solute for the stationary phase arises from their direct physical attachment to the molecular surface of the stationary phase. These transient associations are noncovalent in nature and can be considered as hydrophobic contacts or hydrogen bonding—designations that imply direct molecular contact

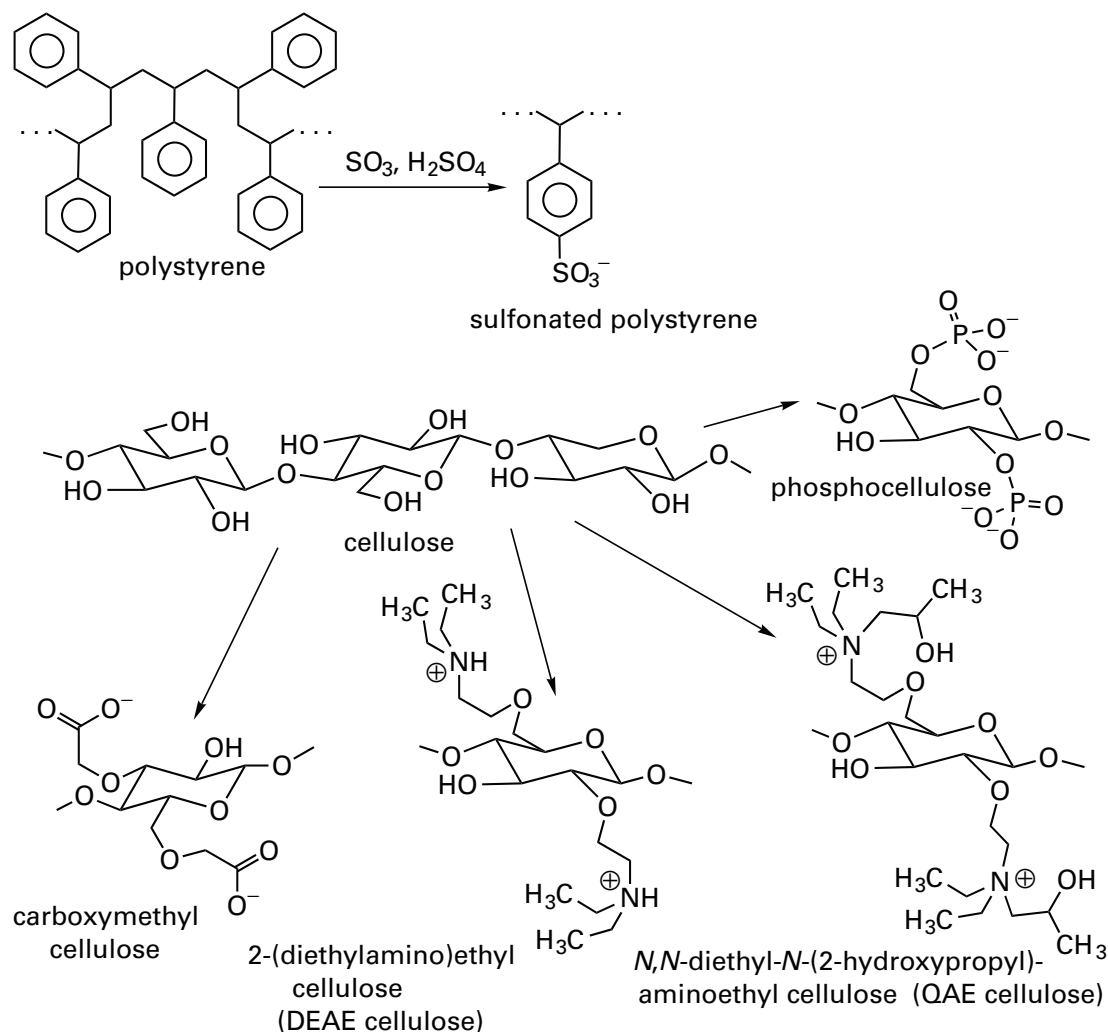
between solid phase and solute. It is this molecular contact that distinguishes chromatography by adsorption from chromatography by ion exchange.

Media for **chromatography by ion exchange** are solids formed from all of the usual neutral polymers to which charged organic functional groups have been covalently attached (Figure 1–2). **Anion-exchange media**, or basic media, are solid phases to which functional groups of positive charge at neutral pH have been covalently attached, and **cation-exchange media**, or acidic media, are solid phases to which functional groups of negative charge at neutral pH have been attached. A distinction can be made between weakly basic or acidic and strongly basic or acidic ion-exchange media based on whether the fixed charges can or cannot be neutralized, respectively, by variation of the pH within the ranges normally employed for chromatography. This is an important distinction because the density of charge, and hence the capacity of the medium, can be changed by changing the pH when weakly basic or weakly acidic ion-exchange media are used but not when strongly basic or strongly acidic ion-exchange media are used. Examples of weakly basic functional groups are tertiary amines such as those on [2-(diethylamino)ethyl]cellulose (DEAE-cellulose); examples of strongly basic functional groups are quarternary ammonium cations such as those on *N,N*-diethyl-*N*-(2-hydroxypropyl)ammonioethyl agarose (QAE agarose) or trimethylammonioethyl polymethacrylate; examples of weakly acidic functional groups are carboxylates, such as those on carboxymethyl cellulose, or phosphates, such as those on phosphocellulose; and examples of strongly acidic functional groups are sulfonates, such as those on sulfonated polystyrene (Figure 1–2) or sulfonated polymethacrylate.

The fixed charges on the stationary phase are responsible for the tendency of ionic solutes of an opposite charge to associate with it. A **counterion** is a mobile ion that is dissolved in the surrounding solution and has a charge opposite in sign to the fixed charges on the stationary phase; a **co-ion** is a mobile ion that is dissolved in the surrounding solution and has a charge of like sign to the fixed charges on the stationary phase. Solutes containing simple univalent ionic functional groups do not form physical contacts with the isolated univalent fixed charges of opposite sign that are attached to the stationary phase when chromatography by ion exchange is performed in aqueous solution. Rather, such charged solutes (for example, nucleotides, amino acids, or proteins) can be considered to be trapped as mobile counterions surrounding the covalently fixed charges in an ionic double layer.¹⁴ The two layers in an **ionic double layer** are a layer of covalently fixed charges on the surface of the polymer forming the stationary phase and a layer of solution, adjacent to that surface, that is enriched in counterions and depleted of co-ions. The molecular surface of the layer of fixed charge is considered to be the boundary between the layers of the double layer.

* The commercial forms of these beaded, cross-linked polymers each have their own uninformative names, but it is possible to learn their compositions if one is perseverant. Although one or the other chromatographic medium may have the same composition, each manufacturer claims unique benefits for his product.

Figure 1-2: Covalent modifications that produce media for ion exchange. Media derived from polystyrene and cellulose are presented.



The enrichment of counterions, which in this case are the solutes being separated, in the layer of solution results from the requirement for maintaining electroneutrality. The layer of solution contains solutes of both net positive and net negative charge but has an excess of solutes of net charge opposite to the charge of the functional groups in the layer of covalently fixed charges and is depleted in solutes of opposite charge. The layer of covalently fixed charges is usually considered to be localized in a geometric surface representing the molecular surface of the polymer, and the layer of solution enriched in the respective counterions is considered to have the properties of a space charge extending into the surrounding solvent.¹⁴

The reason that the diffuse space charge extends a significant distance into the solution beyond this boundary is that the positive and negative charges in the solution are on mobile, dissolved cations and anions, and the enthalpic tendency of the counterions to gather at the charged surface of the boundary and the tendency of the co-ions to avoid the charged surface of the boundary is counterbalanced by the entropic tendency for each of them to diffuse randomly throughout the surrounding

solution. Because the imbalance in charge that defines the ionic double layer falls off exponentially, the layer of solution in which the imbalance in charge occurs theoretically has no outer boundary. It is, however, arbitrarily assigned a thickness that is approximately that distance, from the surface of fixed charges, at which the space charge has decreased by a factor of $\exp(-1)$. Under the normal conditions of chromatography, the thickness of the layer of solution in the double layer would be less than 10 nm.¹⁴ It can be assumed that the boundary that separates the stationary phase from the mobile phase during the chromatography, namely the outside surface of the bead, lies at a much greater distance than this from the molecular surface of the charged strands of polymer within the bead because flow occurs around beads of dimensions at least a thousand times larger. Therefore, the entire ionic double layer must be within the chromatographic stationary phase.

If this assumption is made, the **distribution of counterions** between the stationary phase and the mobile phase becomes formally equivalent to the distribution of permeant counterions across a permeable membrane when a charged, impermeant macromole-

cule is present on only one side of the membrane. In the case of chromatography by ion exchange, the charged polymer of the bead is formally equivalent to the trapped, charged macromolecule. If this is the case, the sum of the fixed charges and the dissolved mobile charges of the same sign in the stationary phase must equal the sum of the dissolved mobile charges of the opposite sign in the stationary phase. It follows that the concentration within the stationary phase of any solute of charge opposite to the fixed charges must always be greater than its concentration within the mobile phase, and it is this bias that can produce significant values of α'_i . This bias can be treated by the Donnan formalism.¹⁵

Consider the situation of an anion-exchange medium of univalent fixed positive charges, N^+ [an example would be *N,N*-diethyl-*N*-(2-hydroxypropyl) aminoethyl] cellulose, Figure 1-2], and a univalent anionic solute, A^- (an example would be AMP^-), in the presence of a dissolved univalent salt, K^+Cl^- , referred to as the electrolyte. Assume that the original stationary phase was the chloride salt of N^+ and that the solute before it was added to the stationary phase was the potassium salt of A^- . All concentrations are expressed in terms of moles (liter of phase)⁻¹, hence the unprimed values. From the requirement for electroneutrality

$$[K^+]_S + [N^+]_S = [Cl^-]_S + [A^-]_S \quad (1-15)$$

$$[K^+]_M = [Cl^-]_M + [A^-]_M \quad (1-16)$$

where the subscripts refer to the stationary and mobile phases. Since the electrolytes are at equilibrium within the theoretical plate

$$[K^+]_M [Cl^-]_M = [K^+]_S [Cl^-]_S \quad (1-17)$$

$$[K^+]_M [A^-]_M = [K^+]_S [A^-]_S \quad (1-18)$$

In the particular circumstance where the concentration of solute A^- is significantly less than the concentration of Cl^- so that $[A^-]$ becomes negligible in both Equations 1-15 and 1-16 and the concentration of fixed charges in the stationary phase, $[N^+]_S$, is so large that $[K^+]_S$ in Equation 1-15 becomes negligible, then

$$\alpha_{A^-} \equiv \frac{[A^-]_S}{[A^-]_M} \equiv \frac{[N^+]_S}{[K^+]_M} \quad (1-19)$$

where α_{A^-} is defined somewhat differently from the partition coefficient described so far. Instead of units of concentration in moles (liter of bed)⁻¹, the units of

concentration are moles (liter of stationary phase)⁻¹ and moles (liter of mobile phase)⁻¹.

Equation 1-19 predicts that the partition coefficient, α_{A^-} , for solute A should be inversely proportional to the concentration of K^+ in the mobile phase. Because the internal volumes of the stationary phases in chromatography by ion exchange are fairly small and the capacities of most media are large even in terms of equivalents (liter of bed)⁻¹, the situation in which $[K^+]_S = [N^+]_S$ is probably rarely approached, and Equation 1-19 should govern most concentrations of salt employed. The effect of adding a univalent salt to the mobile phase is to decrease the value of the partition coefficient for the anion A^- between the cationic stationary phase and the aqueous mobile phase. In this way, the value of α_{A^-} can be adjusted by varying the concentration of electrolyte to optimize an isocratic separation, or a gradient of the electrolyte can be used to vary α_{A^-} continuously. If the concentration of electrolyte is low, α_{A^-} will be large and the mobility of A^- will be negligible. Therefore, a charged solute can be gathered tightly at the origin of the chromatographic system from a large volume of a dilute solution at low ionic strength, and chromatography can then be initiated by increasing the concentration of the electrolyte.

In a weakly basic or acidic ion-exchange medium, the **titration of the charges** that occurs upon adding acid or base, respectively, occurs over a broad range of pH because of electrostatic repulsion among the fixed cations or anions. This permits the density of charge on the medium ($[N^+]_S$ or $[O^-]_S$) to be continuously decreased by incorporating a gradient of pH into the entering mobile phase. For example, if the stationary phase has fixed, protonated tertiary ammonium cations, a gradient of increasing pH would decrease $[R_3NH^+]_S$ as it progresses. The decrease in the density of charge ($[N^+]_S$ or $[O^-]_S$) produces a decrease in α_{A^-} or α_{A^+} (Equation 1-19), causing the solutes to emerge sooner than they would under isocratic conditions. When the solutes themselves are weak acids or bases, however, their ionization may also vary as the gradient of pH progresses, but in the opposite sense to the stationary phase; their effective charge will be increasing as the gradient progresses.

There is no question that Equation 1-19, although intuitively informative, does not describe real ion-exchange processes. At face value it predicts that α_{A^-} should be a function only of the charge density on the stationary phase and the concentration of electrolyte, and this is often not the case. Even simple solutes upon ion exchange display affinities for the supporting polymeric matrix or the functional groups on the fixed charges that sometimes differ greatly from this expectation. The reason for these deviations is almost certainly due to the fact that solutes, brought to high concentration within the double layer by ion exchange, adsorb physically to these constituents, and as a result chromatography by adsorption is superimposed upon the basic process of chromatography by ion exchange. The

clearest example of this is found in the separation of amino acids on sulfonated polystyrene (Figure 1-3).^{16,17} Even though the solutes in the series alanine, valine, leucine, and phenylalanine have almost identical acid dissociation constants, and hence ionic charge, they are cleanly separated. There is little doubt that the separation observed in this series is due to chromatography by adsorption performed by the styrene-divinylbenzene copolymer of the matrix.¹⁶ An ion-exchange medium can also participate in adsorbing simple cations or anions by chelation, such as occurs in the binding of alkali metal cations to polygalacturonic acid.¹⁸

A molecule of protein is a macromolecular poly-electrolyte, the effective charge of which is a function of

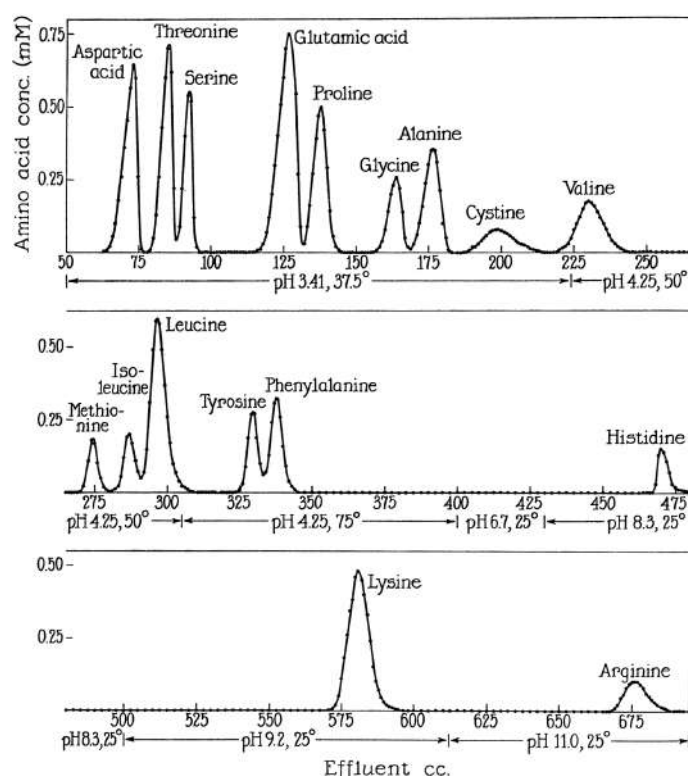


Figure 1-3: Separation of amino acids on chromatography by cation exchange.¹⁶ A mixture of amino acids in the ratios typical of those found in a protein was submitted to chromatography on a column (0.90 cm × 100 cm) of sulfonated polystyrene (Figure 1-2) in the sodium form. The values of the pH and temperatures of the buffered mobile phases are noted below the horizontal axes, which register the volume of the mobile phase that has passed through the column (in centimeters³) since initiation of the chromatography. Changes from one mobile phase to the next were made discontinuously at the times noted. Individual fractions of the effluent emerging from the bottom of the column were collected and assayed for their concentration of amino acid (millimolar). The relative mobility, R_f , of each amino acid in the initial isocratic separation at pH 3.41 would be the void volume of the column divided by the volume at which its peak of concentration emerged from the column. The width at half height, $w_{1/2}$, of each peak is its width in milliliters at a level of concentration half that of the concentration at its peak. Reprinted with permission from ref 16. Copyright 1951 *Journal of Biological Chemistry*.

pH and varies over a wide range. If the pH is changed, the partition coefficients of proteins upon ion exchange vary, and gradients of pH as well as gradients of ionic strength are used in their chromatography. In the case of poly-electrolytes of this type, interactions between the solute and the stationary phase may also lead to direct adsorption. Although simple univalent ions when they are at normal concentrations almost certainly do not physically associate with each other in aqueous solution, poly-electrolytes of opposite charge, such as proteins and ion-exchange media, sometimes do. This results from a **cooperative association** of the opposite charges on the two polymers that arises from the fact that the charges on the ion-exchange medium are covalently fixed and those of the opposite sign on the protein are also covalently fixed. It is always possible that there is a population of sites on the ion-exchange medium where the distribution of charge complements the distribution of charge on the protein, a possibility that will produce physical adsorption. This, however, is probably a rare phenomenon; most of the time the molecules of protein are simply trapped inside the ion-exchange medium as mobile counterions in the ionic double layer.

Media for **chromatography by molecular exclusion*** separate molecules on the basis of differences in their size and shape. The beaded solids used as stationary phases are tangled webs of hydrophilic, linear polymers—dextran, agarose, polyacrylamide, polyether, or polymethacrylate—cross-linked among themselves randomly along their length. These matrices can be produced in two ways. First, polysaccharides such as agarose and dextran spontaneously imbibe water and swell when the dry solid is exposed to an aqueous solution. The degree to which the linear polymers are cross-linked among themselves determines how much water they will imbibe at saturation. This is designated as their water regain, W_r , in milliliters (gram of polysaccharide)⁻¹. This in turn determines the fraction of the volume of the stationary phase occupied by solid polymer, f_{poly} :

$$f_{\text{poly}} = \frac{V_{\text{poly}}}{V_{\text{H}_2\text{O}} + V_{\text{poly}}} = \frac{\bar{v}_{\text{poly}}}{W_r + \bar{v}_{\text{poly}}} \quad (1-20)$$

where V_{poly} is the volume occupied by polysaccharide, $V_{\text{H}_2\text{O}}$ is the volume occupied by water, and \bar{v}_{poly} is the partial specific volume of the polysaccharide in milliliters gram⁻¹. Second, polyacrylamide does not swell readily but can be polymerized from acrylamide monomers and a small amount of the cross-linker N,N' -methylenebis(acrylamide), both dissolved at a certain concentration in an aqueous solution. This produces a rigid gel that can be fragmented. The majority of

* This method is also called size exclusion, gel filtration, and gel permeation.

the volume inside the beads of any of these stationary phases for chromatography by molecular exclusion is occupied by water. When water is within the tangled web of the bead, however, it is no longer mobile but stationary. The mobile phase percolates around the beads and flow occurs only in the interstices among the beads. The void volume, V_0 , is the volume of this space outside of the beads.

The larger the molecule of solute, the less of the open space inside the beads of the stationary phase is available to it. If solute A is too large, it cannot enter the beads at all, and its peak emerges from the system at the void volume, V_0 . Therefore, the elution position on the chromatogram of the completely excluded molecules marks the position of V_0 . A small molecule (in theory, water itself or something equivalent to it) can enter the entire open space in each bead, and its elution position marks V_i , the **included volume**. Unlike with most other chromatographic separations, there is an end to a molecular exclusion chromatogram because no solute can see a larger volume than V_i . The only useful separation that occurs in such a system is of those solutes that emerge between V_0 and V_i , because all solutes larger than a certain size travel together at V_0 and all solutes smaller than a certain size travel together at V_i (Figure 1–4). Between V_0 and V_i on the chromatogram, the larger solutes are the first to emerge.

Because the fluid contained within the beads is identical in composition to the mobile phase percolating around the beads and because a polymer that theoretically has no affinity for the solutes being separated has been chosen, the partition coefficient for solute A, α'_A , between stationary and mobile phases is the ratio of the volume within the stationary phase that solute A can

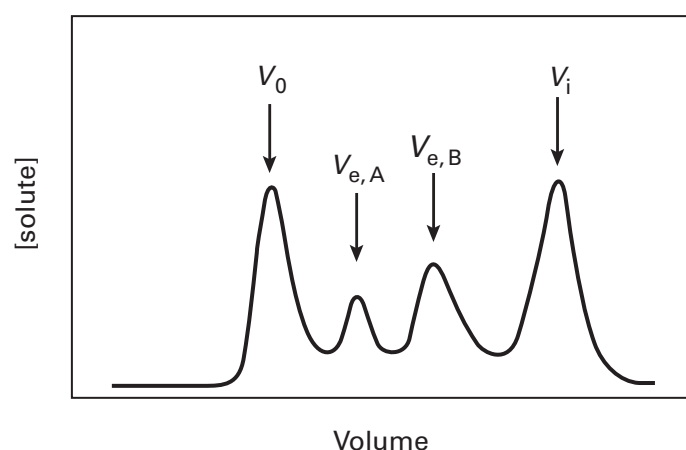


Figure 1–4: Chromatography by molecular exclusion. As in Figure 1–3, the concentration of solute in the fluid emerging from the chromatographic system is plotted as a function of the volume that has emerged. All molecules larger than a certain size move at the void volume, V_0 ; all molecules below a certain size travel at the included volume, V_i ; and solutes A and B travel at their elution volumes, $V_{e,A}$ and $V_{e,B}$, respectively, and are separated from each other because a molecule of solute A is larger than a molecule of solute B.

enter, which is its elution volume minus the void volume, divided by the volume of the mobile phase within the bed, which is, by definition, the void volume V_0 . Parameters other than the partition coefficient, however, are usually used to define the behavior of a solute on chromatography by molecular exclusion. If V_T is the total volume of the bed of the chromatographic system, then the volume of the bed occupied by the stationary phase is $V_0 - V_T$. The fraction of the volume of the stationary phase that is available to solute A is designated $K_{av,A}$:

$$K_{av,A} \equiv \frac{V_{e,A} - V_0}{V_T - V_0} = \frac{\alpha'_A}{\frac{V_T}{V_0} - 1} \quad (1-21)$$

Another parameter is often used to describe the elution during chromatography by molecular exclusion. This is the fraction of the volume within the stationary phase available to a small reference solute, solute R, that is also available to solute A, and it is designated $K_{D,A}$ so that

$$K_{D,A} \equiv \frac{V_{e,A} - V_0}{V_{e,R} - V_0} \quad (1-22)$$

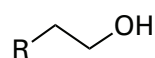
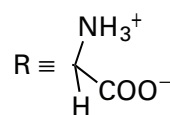
where $V_{e,R}$ is the volume at which solute R elutes. If the reference solute were able to enter the entire aqueous phase within the stationary phase, V_i , then $V_{e,R}$ would be equal to V_i and $K_{D,A}$ would equal $K_{av}(1 - f_{poly})^{-1}$. The difficulty with this definition is that it depends on the identity of the reference solute.

Suggested Reading

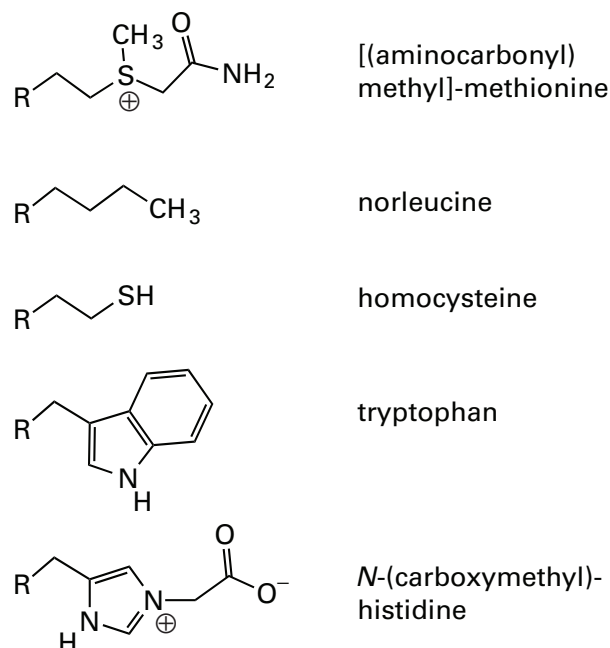
Moore, S., & Stein, W.H. (1951) Chromatography of amino acids on sulfonated polystyrene resins, *J. Biol. Chem.* 192, 663–681.

Problem 1–1: Assume that the total volume of mobile phase in the column described in Figure 1–3 is 45 cm³. Calculate α'_i for aspartic acid, threonine, glutamic acid, proline, glycine, alanine, and valine. Calculate the number of theoretical plates in the column used for the separation shown in Figure 1–3 from the peaks for threonine, serine, proline, glycine, alanine, and valine.

Problem 1–2: List the following amino acids in order of their elution from a column of sulfonated polystyrene.



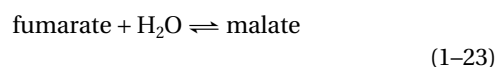
homoserine



Assay

Homogenization of a biological specimen produces a complex mixture of proteins. Before any one of these proteins can be purified, there must be a way to identify it; an assay serves this purpose. An **assay** is any connection between a specific biological phenomenon and a solution containing the protein responsible for this phenomenon. During a purification, separations of high resolution are performed that produce large numbers of separate samples, and the need to locate the protein of interest within these separated fractions requires that they be individually assayed (see points in Figure 1–3). This fact puts a premium on the speed and efficiency of the assay used.

One of the most common types of assay is one that monitors a **chemical reaction catalyzed by an enzyme**. One of the phenomena that occurs in living organisms is the conversion of fumarate to malate



catalyzed by the enzyme fumarate hydratase.*

When the homogenate from a porcine heart, which contains fumarate hydratase, is added to a solution of fumarate, which is otherwise quite stable, the fumarate begins to disappear and malate appears.³ Because fumarate has significant absorbance at 300 nm (A_{300}), the

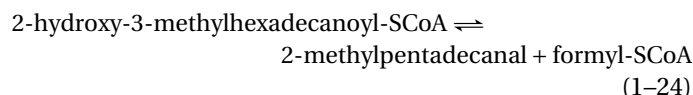
decrease in A_{300} with time can be followed to assay the enzyme.

The goal of any assay is to produce a signal that is directly **proportional to the molar concentration** of the particular protein being assayed. A necessary and sufficient condition for this to be the case is that the quantity measured, for example, the rate of decrease in A_{300} in the first few minutes, be directly proportional to the amount of sample added. This proportionality must be demonstrated directly by examining the magnitude of the signal used in the assay as a function of the amount of sample added.^{19–21} The range over which this direct proportionality between the measurement and the added sample occurs must be known. Because the assay should always be performed within this range, it must be fairly broad to avoid the problem of having to assay every sample at a series of different dilutions to find the range. It is also helpful if the quantity measured, such as A_{300} in the case of fumarate hydratase, changes as a linear function of time within the interval chosen to monitor the reaction.

The rate at which an enzyme converts a reactant into a product usually changes as the **pH** of the solution changes. It is always a good idea to measure the enzymatic activity as a function of the pH of the solution to find the pH at which the rate of the reaction is at its maximum and then use that pH in the routine assay.²²

A **coenzyme** is a molecule that is not a protein but nevertheless must be added to the assay of an enzyme for the reaction it catalyzes to occur. Because the coenzyme is not converted during the enzymatic reaction into a product, it is not a reactant. Coenzymes are used by proteins to provide chemical capabilities that cannot be provided by the side chains of its amino acids alone. Examples of coenzymes are pyridoxal phosphate, thiamin pyrophosphate, flavin adenine dinucleotide, biotin, lipoic acid, heme, chlorophyll, and ubiquinone. If an enzyme catalyzes a reaction requiring a coenzyme, that coenzyme usually must be added to the assay. Often the nature of the reaction is such that the coenzyme required is obvious. For example, most enzymes that catalyze transaminations require pyridoxal phosphate. Often, however, the requirement for a coenzyme is not obvious.

2-Hydroxyphytanoyl-CoA lyase catalyzes the reaction



The enzymatic activity could be readily assayed in the homogenate from a rat liver but disappeared as the purification proceeded. It was found that if thiamin pyrophosphate was added to the assays, however, the activity did not disappear.²³ This result demonstrates that thiamin pyrophosphate is a coenzyme for 2-hydroxyphytanoyl-CoA lyase. There was enough of it in the initial homogenate to satisfy the enzyme, but it was lost as the purification proceeded. It might have been argued that

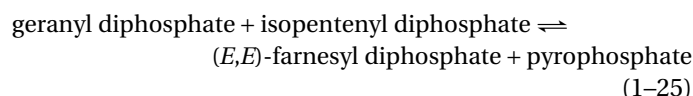
* In this section and in the rest of the book, enzymes are named according to the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (us.expasy.org/enzyme/).

this coenzymatic requirement should have been expected because the enzyme catalyzes a cleavage immediately adjacent to an acyl carbon, but such arguments are usually after the fact. Often the requirement for a coenzyme is not obvious and is both difficult and frustrating to discover.

It also often happens that, as with a coenzyme, a **metallic cation**, such as Mg^{2+} , Ca^{2+} , Zn^{2+} , Cu^{2+} , Fe^{2+} , or K^+ , is required by a protein to perform its function and must be added to the assay. Although there are often obvious choices, such as Mg^{2+} for enzymes having phosphoesters as reactants, the requirement for a particular metal is often unpredicted.

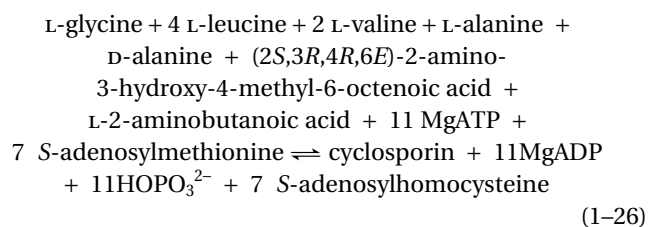
The most unambiguous assay of the reaction catalyzed by an enzyme is one in which the reactants and products are **chromatographically separated** after the reaction and the quantities of each are determined. The introduction of rapid, automated, high-pressure liquid chromatographic systems with associated monitoring systems of high sensitivity has made this approach convenient and efficient. If **radioactive reactants** are available that can be turned into radioactive products, reactants and products from a large number of assays can be separated in arrays of simple, inexpensive chromatographic systems and their respective quantities can be determined by scintillation counting.

Examples of assays in which reactants and products are chromatographically separated have been used for the purifications of the proteins geranyltransferase, cyclosporin synthase, methylamine-glutamate *N*-methyltransferase, and lysine *N*-methyltransferase. Geranyltransferase catalyzes the reaction



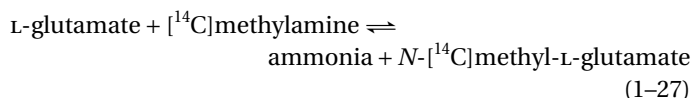
A sample of protein to be assayed for this enzymatic activity can be mixed with geranyl diphosphate and $[1\text{-}^{14}\text{C}]\text{isopentenyl diphosphate}$ and incubated for a set time. The reaction can then be terminated by adding alkaline phosphatase to hydrolyze rapidly the various diphosphates. After extraction, the resulting $[1\text{-}^{14}\text{C}]\text{farnesol}$ and $[1\text{-}^{14}\text{C}]\text{isopentenol}$ in each sample can be separated on small plates by thin-layer chromatography and separately quantified.²⁴ That the product was entirely the expected (*E,E*) isomer of $[1\text{-}^{14}\text{C}]\text{farnesol}$ was demonstrated by gas-liquid chromatography.

Cyclosporin synthase catalyzes the reaction



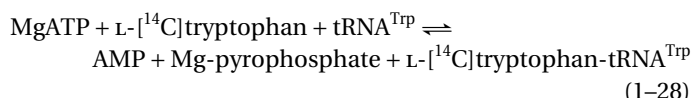
If the reaction is run with *S*-adenosyl[*methyl*- ^{14}C]methionine, the $[^{14}\text{C}]\text{cyclosporin}$ produced can be isolated, after extraction, by thin-layer chromatography.²⁵

Methylamine-glutamate *N*-methyltransferase catalyzes the reaction



The $[^{14}\text{C}]\text{methylammonium cation}$ and the $[^{14}\text{C}]\text{methyl-L-glutamate}$ can be separated by isocratic chromatography by cation exchange.²⁶ L-Lysine, *N*^ε-methyl-L-lysine, and *N*^ε,*N*^ε-dimethyl-L-lysine are converted in the presence of *S*-adenosyl[*methyl*- ^3H]methionine into mixtures of *N*^ε-[^3H]methyl-L-lysine, *N*^ε,*N*^ε-[^3H]dimethyl-L-lysine, and *N*^ε,*N*^ε,*N*^ε-[^3H]trimethyl-L-lysine by lysine *N*-methyltransferase. After removal of unreacted *S*-adenosyl[*methyl*- ^3H]methionine with activated charcoal, the three radioactive products can be separated by thin-layer chromatography and quantified individually.²⁷

As in the previous example, where unreacted *S*-adenosyl[*methyl*- ^3H]methionine was removed by adsorption to activated charcoal, the chemical transformation performed by an enzyme often produces a product that can be exclusively **transferred to a separable phase**. For example, tryptophan-tRNA ligase catalyzes the reaction



The L- $[^{14}\text{C}]\text{tryptophan-tRNA}^{\text{Trp}}$ can be isolated from the assay solution as a precipitate, free of L- $[^{14}\text{C}]\text{tryptophan}$, by treatment with acid and filtration through filters of glass fiber.²⁸ The $[^{14}\text{C}]\text{CO}_2$ released from L- $[1\text{-}^{14}\text{C}]\text{glutamate}$ by glutamate decarboxylase²⁹ or from 4-hydroxyphenyl[$1\text{-}^{14}\text{C}]\text{pyruvate}$ by 4-hydroxyphenylpyruvate dioxygenase³⁰ can be released as a gas from the assay solutions by treatment with acid and collected in a separate well containing a strong base. The enzyme encoded by the *murG* gene of *Escherichia coli* catalyzes the addition of the *N*-acetylglucosamine from UDP-*N*-acetylglucosamine to the 4' position of the muramoyl group in 1'-*O*-β-[3(*R*)-3,7-dimethylhept-6-enyl]-1'-diphospho-2'-*N*-acetylmuramoyl-L-alanyl-D-γ-glutamyl-6-carboxy-L-lysyl-D-alanyl-D-alanine. A derivative of the heptenyldiphospho-*N*-acetylmuramoyl pentapeptide to which a molecule of biotin has been covalently attached can be used in an assay for this enzyme³¹ along with UDP-*N*- $[^{14}\text{C}]\text{acetylglucosamine}$. The resulting biotinylated β(1,4)-*N*- $[^{14}\text{C}]\text{acetylglucosaminylheptenyldiphospho-}N\text{-acetylmuramoyl pentapeptide}$ can be separated cleanly and quantitatively from the remaining UDP-*N*- $[^{14}\text{C}]\text{acetylglucosamine}$ by adsorbing it to a solid phase on which has been attached covalently the protein

avidin, which binds the biotin in the product with high affinity.

A special case of assays that depend on transferring a product or a reactant to a separate phase are those used to monitor the binding of a small molecule to a protein. Certain proteins, known loosely as **receptors**, often do not catalyze a chemical reaction but respond to specific small molecules, referred to as **agonists**, by binding them and then undergoing a change in structure. Receptors are assayed by their ability to bind either these agonists or similar molecules that also bind but do not elicit the response, referred to as **antagonists**. In such **binding assays**, the receptor and a suitable radioactive agonist or antagonist are mixed together, the binding is allowed to come to equilibrium, and the receptor–agonist or receptor–antagonist complex is separated from unbound agonist or antagonist, respectively. Because receptors are usually proteins dissolved in membranes, the separation of bound from unbound ligand often takes advantage of the large size of the fragments of membrane produced by homogenization, which can be separated from the rest of the solution by filtration or centrifugation. After the separation, the amount of bound radioactivity is then determined by scintillation counting.

Chemically stable agonists or antagonists of high affinity for a receptor are required to ensure that the binding is at saturation so that all receptors are counted and to prevent dissociation of receptor and agonist or receptor and antagonist during the separation of bound and free radioactivity. These reagents are often produced by the synthesis of analogues of the natural compounds. For example, [^3H]dihydroalprenolol is a radioactive synthetic compound that binds tightly (dissociation constant = 2 nM)³² to the β -adrenergic receptor, which physiologically responds to epinephrine. Its binding has been used as an assay during the purification of this receptor.³³ Often a synthetic compound the binding of which to a receptor is strong has been obtained during a search for pharmaceutically useful agents. An example of this kind of product is prazosin, which was developed as a drug specific for α_1 -adrenergic receptors and the binding of which (dissociation constant = 1 nM) could be used as an assay during the purification of the α_1 -adrenergic receptor.³⁴ Often the naturally occurring agonist has an affinity great enough that it can be used in an assay during the purification of the receptor. For this purpose, it is synthesized in a radioactive form. Examples would be the use of the binding of ^{125}I -epidermal growth factor³⁵ (dissociation constant = 20 nM) and the binding of [$1,2\text{-}^3\text{H}_2$]progesterone³⁶ (dissociation constant = 1 nM) as assays for their receptors.

In all binding assays for receptors, the difficulty is to separate the complex between the receptor and the agonist from the unbound agonist without losing the bound agonist through the dissociation of the complex. It is often possible to sediment the complex in a preparative ultracentrifuge.³⁷ This strategy is particularly useful for

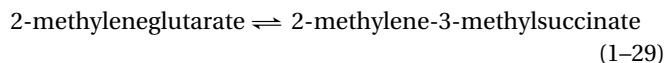
weakly bound agonists that dissociate rapidly after the unbound agonist is removed, because during sedimentation the concentration of unbound agonist does not change so the amount of bound agonist does not either. The small amount of unbound agonist in the pellet can be estimated and a correction made to obtain an accurate measurement of the bound agonist. With agonists and antagonists that bind tightly, the complex can be separated rapidly with little loss of bound radioactivity on rapid chromatography by molecular exclusion on small, disposable columns.³⁸

Binding assays have also been developed for proteins that associate with specific nucleotide sequences in DNA,³⁹ such as promoters or other regulatory elements. A short fragment of DNA labeled with [^{32}P]phosphate at one end and containing the sequence of interest is used as a reagent. When such a fragment is digested with deoxyribonuclease I and the products are then separated by electrophoresis, a characteristic pattern of shorter segments of DNA of various lengths is obtained as a result of random cleavage by the nuclease of the phosphodiester bonds along the double-stranded DNA. The presence of a protein that binds specifically to a particular nucleotide sequence in a short fragment of end-labeled DNA results in prevention of cleavage of the DNA by the nuclease at that site. The fragments resulting from cleavages in this region disappear from the display, and this **footprint** demonstrates that the **DNA-binding protein** is present. Such an assay can be used to determine the relative concentration of the DNA-binding protein by examining the patterns produced as a series of dilutions is performed in the solution of the protein added to the end-labeled DNA.

An enzyme that catalyzes a physical or chemical transformation of DNA can often be assayed by separating the product of the transformation from the reactant by electrophoresis. Deoxyribonucleic acid primase/helicase from T7 bacteriophage catalyzes the unwinding of double-stranded DNA. Double-stranded DNA, one of the strands of which has been labeled with [^{32}P]phosphate at its 5' end, is mixed with a sample of protein to be assayed for this activity, and after a few seconds the reaction is quenched with dodecyl sulfate. The ^{32}P -labeled single-stranded DNA produced by the unwinding can be separated from the ^{32}P -labeled double-stranded DNA by electrophoresis.⁴⁰

Up to this point, with the exception of that for fumarate hydratase, the assays described have been discontinuous ones. The reaction is allowed to proceed for a certain interval, it is quenched in some way, and the amount of product formed is then measured, usually by dissecting the final, quenched solution. Because less manipulation is required and because the result is immediate, **continuous assays** in which the product of the live enzyme is monitored as it is formed are more convenient. As in the assay for fumarate hydratase, the continuous **change in absorbance** of a reactant or

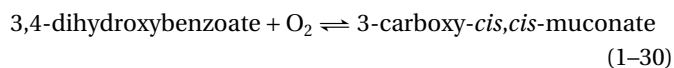
product is often followed. The reaction catalyzed by 2-methyleneglutarate mutase



can be assayed⁴¹ by monitoring the change in absorbance of the solution if the 2-methylene-3-methylsuccinate is converted immediately as it is produced into 2,3-dimethylmaleate with the enzyme methylitaconate Δ -isomerase. 2,3-Dimethylmaleate absorbs most strongly at 230 nm, but in the assay for the enzyme, absorbance changes between 240 and 256 nm were monitored in order to avoid interference from the absorbance of nucleic acids and protein in the samples.

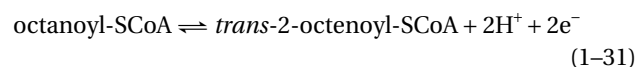
Because the enzymatic activity in the initial homogenate and at early steps in the purification of a protein is usually quite low, large amounts of these heterogeneous mixtures of protein and nucleic acid often must be added to the assay and their intrinsic absorbance can be appreciable. This problem precludes the use of absorbance changes at wavelengths below 240 nm for continuous assays based on absorbance because all proteins absorb too strongly in this range. A crude mixture of protein and nucleic acid also has a significant and rather uniform absorbance between 240 and 290 nm so the change in absorbance being monitored in the assay must be great enough to overcome this **interference**, as it is in the assay for 2-methyleneglutarate mutase (between $\Delta\epsilon_{240} = 3700 \text{ M}^{-1} \text{ cm}^{-1}$ and $\Delta\epsilon_{256} = 660 \text{ M}^{-1} \text{ cm}^{-1}$ for the production of 2,3-dimethylmaleate).⁴¹

In any type of enzymatic assay, the proteins and nucleic acids in the early, crude mixtures can interfere in other, unexpected ways with many otherwise useful assays. In particular, contaminating proteins may catalyze the transformation of either the reactants or the products of the enzyme being assayed. The reaction catalyzed by protocatechuate 3,4-dioxygenase



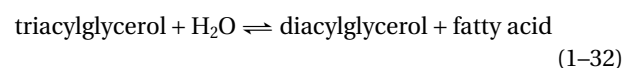
could be followed by the increase in A_{290} as the reaction proceeded.⁴² In crude homogenates, however, the 3-carboxy-*cis,cis*-muconate was converted by a contaminating enzyme into 3-carboxy-*cis,cis*-muconolactone, and absorbances had to be corrected for this further transformation until the contaminating enzyme had been lost at an intermediate step in the purification.

Even though an enzyme normally produces a product the absorbance of which is no different from that of the reactant, it is sometimes possible to design a **synthetic reactant** in such a way that its absorbance does change upon its conversion in a continuous assay. Medium-chain acyl-CoA dehydrogenase catalyzes the reaction



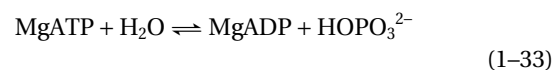
involving no change in absorbance. If synthetic 4-thiaoctanoyl-SCoA is used as a reactant instead of octanoyl-SCoA, the 4-thia-*trans*-2-octenoyl-SCoA produced, because it is a vinylthioether, absorbs strongly at 312 nm ($\epsilon_{312} = 22,000 \text{ M}^{-1} \text{ cm}^{-1}$).⁴³

Enzymes that operate at **lipid-water interfaces** are often difficult to monitor with a continuous assay because of the requirement that lipid and water be present together, which can lead to cloudy suspensions that interfere with spectrometry. Triacylglycerol lipase catalyzes the reaction



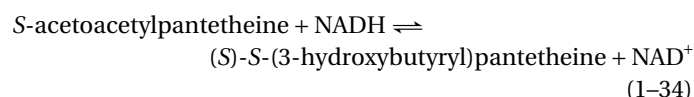
Both the triacylglycerol and the diacylglycerol are fats or oils. If the triacylglycerol is presented to the enzyme on the surface of a droplet of oil in an oil drop tensiometer, the change in surface tension of the droplet resulting from the hydrolysis of the triacylglycerol can be monitored continuously.⁴⁴ It is also possible, however, to produce a clear emulsion of water-filled micelles suspended in an organic solvent. If the lipase is incorporated into the aqueous phase of these "reverse" micelles and the triacylglycerol is dissolved in the organic solvent, the shift in the wavelength of the absorbance of the C=O stretch from 1751 to 1715 cm^{-1} occurring upon hydrolysis of the triacylglycerol at the interface between water and oil can be monitored in an infrared spectrophotometer.⁴⁵ Infrared spectrometry is useful in this instance because it is much less affected by light scattering than is visible or ultraviolet spectrometry.

In the assay for 2-methyleneglutarate mutase (Equation 1-29), the 2-methylene-3-methylsuccinate is converted immediately and continuously upon its production into 2,3-dimethylmaleate by another enzyme, methylitaconate Δ -isomerase, that has been added intentionally to the solution. This is an example of a coupled continuous assay. A **coupled assay** is an assay in which one or more purified enzymes, usually commercially available, and any reactants required by those additional enzymes are added to transform the immediate product of the enzyme of interest by a subsequent enzymatic reaction or reactions that produce a change in absorbance or fluorescence, the rate of which is directly proportional to the rate at which the product is produced. The other enzymes and their reactants must be added at high enough concentrations that the complete transformation of the initial product is effectively immediate, and the initial product is converted continuously as it is formed. Another example of a coupled assay is that used to follow the ATPase reaction



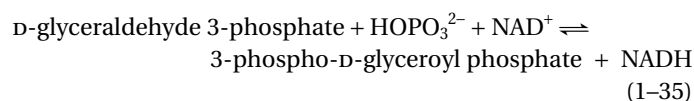
catalyzed by myosin subfragment 1.⁴⁶ Both the reactant 2-amino-6-mercapto-7-methylpurine ribonucleoside and the enzyme purine-nucleoside phosphorylase are added to the solution in addition to MgATP and the ATPase. The inorganic phosphate produced is immediately and continuously used by the phosphorylase to cleave the purine ribonucleoside to ribose-1-phosphate and 2-amino-6-mercapto-7-methylpurine that, unlike the ribonucleoside, absorbs strongly at 360 nm ($\Delta\epsilon_{360} = 11,000 \text{ M}^{-1} \text{ cm}^{-1}$). This coupled continuous assay is useful for monitoring any one of the many enzymes that have **inorganic phosphate** as one of their products.

Of all of the changes of absorbance that are employed in continuous enzymatic assays, none is more heavily used than the decrease in A_{340} of **dihydronicotinamide adenine dinucleotide** (NADH; $\epsilon_{340} = 6220 \text{ M}^{-1} \text{ cm}^{-1}$)⁴⁷ or its phosphate (NADPH; $\epsilon_{340} = 6100 \text{ M}^{-1} \text{ cm}^{-1}$), when it is oxidized to nicotinamide adenine dinucleotide (NAD^+) or to its phosphate (NADP^+), respectively, or the increase in A_{340} that occurs in the reverse reaction. There is a large class of enzymes, known as **dehydrogenases**, that use the oxidation–reduction pairs of either NAD^+ and NADH or NADP^+ and NADPH, and they can be assayed directly and continuously. For example, 3-hydroxyacyl-CoA dehydrogenase catalyzes the reaction

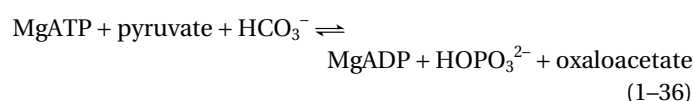


The loss of NADH can be followed at 340 nm.⁴⁸ To insure that only 3-hydroxyacyl-CoA dehydrogenase is being assayed, a control in the absence of S-acetoacetylpanthetheine is run.

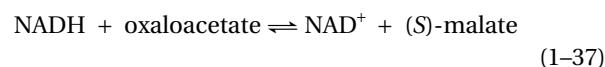
In the opposite sense, the increase in A_{340} can be used to follow the reaction catalyzed by glyceraldehyde-3-phosphate dehydrogenase (phosphorylating):⁴⁹



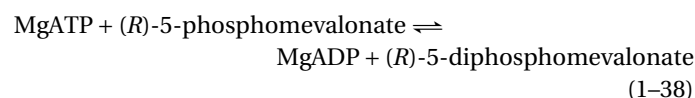
The absorbance change produced by a particular dehydrogenase can also be used in a coupled, continuous assay to monitor enzymatically catalyzed reactions that do not involve NADH. Examples of such coupled assays have been used for the purifications of pyruvate carboxylase, phosphomevalonate kinase, and imidazoleglycerol-phosphate dehydratase. The oxaloacetate produced by pyruvate carboxylase



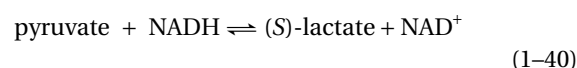
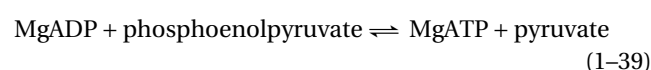
can be monitored by adding NADH and excess malate dehydrogenase:⁵⁰



Phosphomevalonate kinase catalyzes the reaction

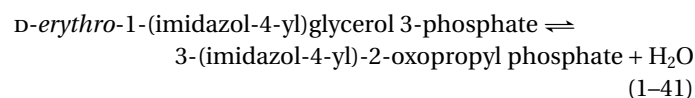


The MgADP can be monitored continuously as it is produced by adding phosphoenolpyruvate, NADH, and an excess of both pyruvate kinase and L-lactate dehydrogenase:⁵¹

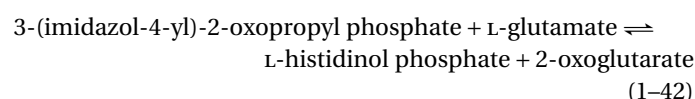


This coupled assay is widely used for enzymes that produce **MgADP**.

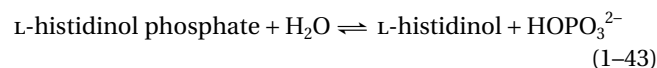
The 3-(imidazol-4-yl)-2-oxopropyl phosphate produced by imidazoleglycerol-phosphate dehydratase



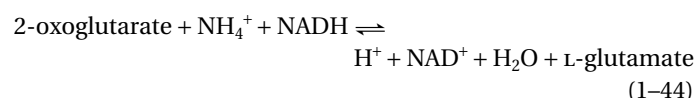
during its assay⁵² is consumed as it is produced by histidinol-phosphate transaminase:



The unfavorable equilibrium of Reaction 1–42 is pulled in the direction written by the exergonic hydrolysis catalyzed by histidinol-phosphatase



The 2-oxoglutarate produced by the transaminase (Equation 1–42) is converted back to L-glutamate by glutamate dehydrogenase:

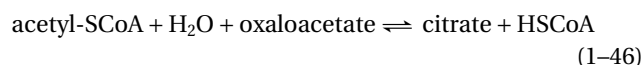


to produce the decrease in A_{340} . Similar coupled assays are used to monitor other transaminases.

One of the more subtle uses of a coupled assay based on the A_{340} of NADH is the one devised⁵³ for hydroxymethylglutaryl-CoA lyase:

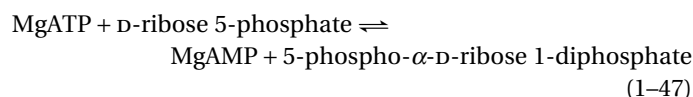


This coupled assay takes advantage of the fact that the equilibrium of the malate dehydrogenase reaction (Equation 1-37) lies in the direction of NAD^+ and (S)-malate so that if NAD^+ , malate, and malate dehydrogenase are mixed together, little oxaloacetate and NADH are formed. With this in mind, it can be seen that if (S)-malate, NAD^+ , and excesses of citrate (*si*) synthase and malate dehydrogenase are present during the progress of the reaction catalyzed by hydroxymethylglutaryl-CoA lyase, the conversion of the acetyl-CoA into citrate by citrate (*si*) synthase

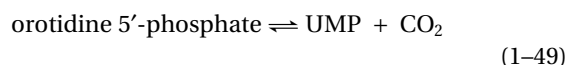
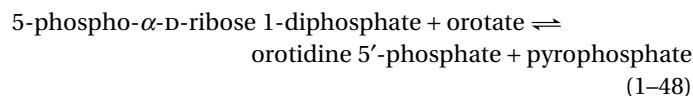


consumes oxaloacetate and pulls the unfavorable equilibrium of the malate dehydrogenase reaction in the direction of NADH production, and hence an increase in the A_{340} of the solution is observed.

The two or more enzymatic steps in a coupled assay are sometimes disconnected rather than allowed to proceed simultaneously. An example would be an assay⁵⁴ for ribose-phosphate diphosphokinase:

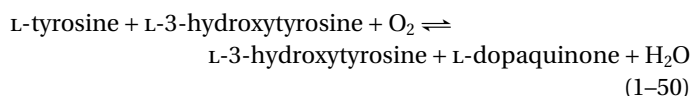


The reaction is quenched by boiling, and the amount of 5-phospho- α -D-ribose 1-diphosphate that has accumulated is determined by adding orotate, orotate phosphoribosyltransferase, and orotidine-5'-phosphate decarboxylase:

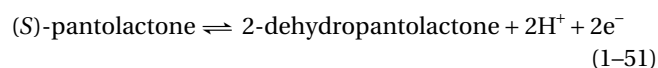


The decrease in A_{295} due to the loss of orotate is proportional to the 5-phospho- α -D-ribose 1-diphosphate originally present in the quenched samples. The decarboxylation has been incorporated in the assay to draw the reactions to completion.

Colorimetric assays are assays in which a reagent is added that reacts chemically rather than enzymatically with a product of the enzymatic reaction being monitored to produce a change in absorbance, often observed visually as a dramatic change in the color of the solution. Monophenol monooxygenase catalyzes the reaction

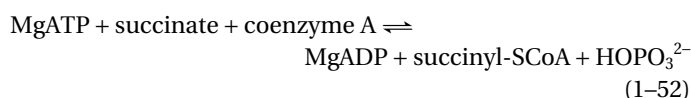


When 3-methyl-2-benzothiazolinonehydrazone has been added to the solution of the assay,²¹ the L-dopaquinone reacts rapidly and quantitatively with it to produce a dark pink color ($\epsilon = 29,000 \text{ M}^{-1} \text{ cm}^{-1}$), the appearance of which can be monitored continuously. As is medium-chain acyl-CoA dehydrogenase (Equation 1-31), (S)-pantolactone dehydrogenase, which catalyzes the oxidation



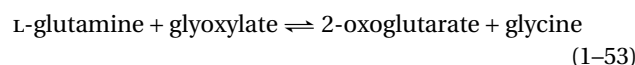
is a member of a large class of enzymes that catalyze oxidation-reduction reactions and then transfer the electrons involved either to or from small proteins or natural compounds the role of which is to receive or provide electrons. These natural donors or acceptors can often be replaced by synthetic donors or acceptors. (S)-Pantolactone dehydrogenase accepts phenazine methosulfate as an oxidant in place of the acceptor it uses naturally, and reduced phenazine methosulfate readily oxidizes nitrotetrazolium blue. When both of these compounds are present in the assay, the appearance of diformazan, which is the product of the oxidation of nitrotetrazolium blue, can be followed by its strong absorbance ($\epsilon_{570} = 40,200 \text{ M}^{-1} \text{ cm}^{-1}$).⁵⁵ It is possible to monitor the production of coenzyme A by citrate (*si*) synthase (Equation 1-46) continuously⁵⁶ by the addition of 5,5'-dithiobis(2-nitrobenzoate). This reagent reacts with the thiol of the coenzyme A as it is formed to release the bright yellow 2-nitro-5-thiolatobenzoate dianion. This assay is useful for monitoring any enzyme that produces **coenzyme A**.

The colorimetric assays described so far are continuous assays in which the chemistry of the colorimetric reagent is compatible with the aqueous solution and neutral pH required to avoid denaturation and inactivation of the protein being assayed. This is usually not the case with colorimetric reagents. In the instances in which it is not, the assay must be **quenched** after a convenient interval before the colorimetry is performed. 2-Hydroxy-6-ketono-2,4-diene-1,9-dioic acid 5,6-hydrolase produces succinate as one of its two products. The production of succinate is coupled in the assay²² to the reaction catalyzed by succinate-CoA ligase (ADP-forming):

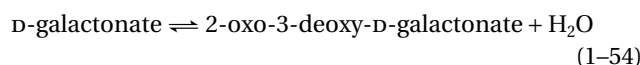


After 15 min, the solution is heated to 100 °C to quench the enzymatic reaction and the inorganic phosphate is

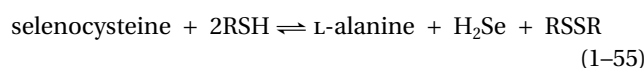
assayed by its reaction with Malachite green in the presence of citrate, which produces strong absorbance at 600 nm. Phosphate produced during an enzymatic reaction can also be determined colorimetrically by the addition of ammonium molybdate in dilute sulfuric acid and a strong reductant, which together produce a blue color proportional in magnitude to the phosphate present.⁵⁷ Glutamine–pyruvate transaminase will also catalyze the reaction



The glycine produced and the L-glutamine remaining will react with *o*-phthalaldehyde and a thiol, after the enzymatic conversion has been terminated, to produce complexes that absorb in the near ultraviolet.⁵⁸ The glycine complex, however, absorbs at a higher wavelength ($\lambda_{\text{max}} = 330 \text{ nm}$). Galactonate dehydratase catalyzes the reaction



After the reaction is quenched, the ketonic product is reacted with semicarbazide⁵⁹ to produce a semicarbazone that absorbs at 250 nm.⁵⁹ Selenocysteine lyase catalyzes the reaction



where RSH is a mercaptan such as 2-mercaptoethanol. After the enzymatic reaction is stopped, the H_2Se can be assayed colorimetrically by its reaction with lead acetate, a reaction that yields a yellow color.⁶⁰

Biological assays are assays in which the ability to evoke a complex biological response by samples added to cells or whole organisms is determined. For example, the assay for a protein referred to as the Hurler corrective factor measures the ability of this protein to prevent the accumulation of sulfated mucopolysaccharide in lysosomes of intact cells. It is this accumulation that causes Hurler's syndrome. Samples are added to a series of petri dishes on which fibroblasts from a patient with Hurler's syndrome have been grown and $^{35}\text{S}]\text{SO}_4$ is added. After several days, the accumulation of ^{35}S -sulfated mucopolysaccharide is assessed by washing the cells and submitting them to scintillation counting.⁶¹ In this particular assay, the decrease in accumulation of radioactivity was not directly proportional to the amount of sample added, and this problem was overcome by constructing a dose–response curve.

A biological assay was also used for the maturation-promoting factor, which is a protein involved in controlling the cell cycle.⁶² Samples containing this protein could be assayed for its activity by injecting sequentially

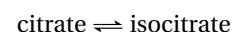
diluted aliquots into individual oocytes from the frog *Xenopus laevis* and scoring the cells for the disappearance of geminal vesicles.⁶³ With the use of this assay, the protein could be followed during a purification procedure⁶³ and the remarkable fluctuation of its concentration during the cell cycle could be documented.⁶²

The success of a particular assay is usually judged on the bases of its accuracy, sensitivity, and selectivity. For following the distribution of a protein during its purification, the accuracy of an assay is not critical—all that is needed is a way to decide whether or not it is present in a particular fraction—but for kinetic studies of the reaction catalyzed by an enzyme, accuracy is often critical.⁶⁴ If only small amounts of a protein are present, the sensitivity of an assay is also often critical.⁶⁵ It is usually to increase the sensitivity of an assay that radioactive reactants are used so that the small amounts of product produced or ligand bound can be identified. Fluorescence is often used for the same purpose. For example, continuous assays monitoring the absorbance of NADH can detect its production at $10 \text{ nmol min}^{-1} \text{ mL}^{-1}$ but those monitoring its fluorescence can detect its production at $0.1 \text{ nmol min}^{-1} \text{ mL}^{-1}$. When following the increase in a particular product produced from a particular reactant or the binding of a particular ligand, an assay is usually selective for a particular protein, but sufficient selectivity is often difficult to achieve. It was only when agonists and antagonists of high affinity and high selectivity were synthesized that the various receptors for epinephrine could be separately identified and purified. A suspension of cellular membranes displays a rather high level of adenosine triphosphatase activity arising from a number of different proteins. It was only when that portion of this activity for which sodium/potassium-exchanging ATPase was responsible could be clearly distinguished⁶⁶ that it became possible to purify the enzyme.⁶⁷

Suggested Reading

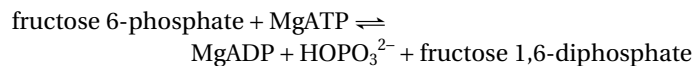
- Winder, A.I. & Harris, H. (1991) New assays for the tyrosine hydroxylase and Dopa oxidase activities of tyrosinase, *Eur J. Biochem.* 198, 317–326.
- McClure, W.R. (1969) A kinetic analysis of coupled enzyme assays, *Biochemistry* 8, 2782–2786.

Problem 1–3: Design a coupled assay, based on the release of $^{14}\text{C}]\text{CO}_2$, for the enzyme *cis*-aconitase, which catalyzes the reaction

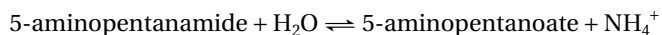


Problem 1–4: Design a coupled assay based on the reduction of NAD^+ for the enzyme fumarate hydratase.

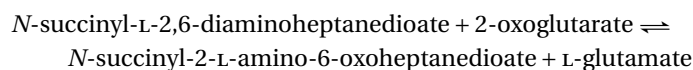
Problem 1–5: Design a coupled assay for phosphofructokinase, the enzyme that catalyzes the reaction



Problem 1-6: Design a coupled assay for 5-aminopentanamide:



Problem 1-7: Design a coupled assay for succinyl-diaminopimelate transaminase:



Purification of a Protein

The **homogenization** of a biological specimen produces a clarified solution of protein and nucleic acid. Animal tissues can be diced, blended, and then processed with a homogenizer to produce a turbid suspension that is then clarified by centrifugation. Plant tissues, because their cells are surrounded by strong cell walls, must be homogenized more forcefully.⁶⁸ If the protein of interest is located in one of the organelles within a plant or animal cell, that organelle is often isolated from the homogenate and then separately fragmented. For example, a protein required for the activation of transcription was purified from an extract of nuclei that had been isolated from a homogenate of HeLa cells,⁶⁹ and 2-hydroxyphytanoyl-CoA lyase was purified from sonicated peroxisomes that had been isolated from a homogenate of rat liver.²³ Bacteria, because they are small, single cells surrounded by a tough integument, are particularly difficult to homogenize. Sonication or passage through a French pressure cell is usually required.

If the source that has been chosen is a plant or an animal, different organs and different species are scanned with the assay to find a source in which the particular protein is present at the highest relative concentration. If the source is a bacterium, various strategies are employed to increase the concentration of the protein of interest. For example, the *Xanthobacter* from which cyclohexane monooxygenase was purified were grown on cyclohexane as the sole carbon source because this enzyme is one of those in the pathway that catabolizes cyclohexane.⁷⁰

The goal of the **purification** of a protein from the clear solution produced by centrifugation of a homogenate is to isolate that protein, whose presence and relative molar concentration can be followed by a specific assay, from all of the other proteins present. To do this, advantage is taken of the properties that distinguish a molecule of one protein from a molecule of another. Proteins are macromolecules of molar mass 10,000–10,000,000 g mol⁻¹. Unless the molecules of a protein have been posttranslationally modified hetero-

geneously by processes such as glycosylation, phosphorylation, endopeptidolytic digestion, or acetylation, each molecule of a given protein has the same covalent structure, the same distribution of polar and nonpolar functional groups over its surface, and the same shape as every other molecule of the same protein. Different proteins are distinguished from each other by differences in these properties.

A molecule of protein can be a **globular macromolecule**, the shape of which resembles a hollow metal sphere that has been dented at random, or it can be a **fibrous macromolecule**, the shape of which is elongated, often dramatically, in one dimension, irregular, and either rigid or flexible. The diameters of globular proteins vary from 2 to 10 nm; the lengths of fibrous proteins can be as great as 300 nm. Positively and negatively charged functional groups are distributed in a characteristic array over the surface of each molecule of a particular protein. In addition to guanidinium ions from the arginines, these charged functional groups are carboxylate ions from the glutamates and aspartates, ammonium ions from the lysines, and imidazolium cations from the histidines, each of which can be neutralized by lowering or raising the pH. As a result, the net charge on a molecule of protein varies with the pH and can be negative or positive within normal physiological ranges. Patches of nonpolar functional groups are distributed in a characteristic array over the surface of each molecule of protein. The affinity of these patches for nonpolar solid phases can be exploited to separate molecules of one protein from those of another. Chromatography is used to separate molecules of protein by differences in their size, their shape, their charge as a function of pH, and the unique distribution of polar and nonpolar groups on their surfaces.

The strategy for the purification of a protein is tailored to the particular problems faced in each instance. Usually it includes a series of steps, each involving a fractionation of the solution by chromatography or adsorption. Each step produces a series of fractions, from two to several hundred, each contained in a volume of aqueous solution. Those fractions containing the protein of interest are identified by the assay, they are pooled together, and the protein in the pool is submitted to the next step of fractionation. The three requirements for the successful purification of a protein are an assay for the protein, the ability to minimize or prevent the loss of the protein through endopeptidolytic degradation or denaturation, and a source of the protein of sufficient abundance.

The progress of the purification of a protein is usually evaluated by examination of both the total activity recovered, which is a measure of the yield of the particular protein being purified at each step, and the specific activity, which is a measure of the enrichment of the protein of interest relative to the other proteins present (Table 1-1). The assay provides a numerical value for the amount of biological activity in a milliliter of the solution

Table 1–1: Purification of Aryl-acylamidase from *Nocardia globerule*⁷¹

purification step	total protein (mg)	total activity ($\mu\text{mol min}^{-1}$)	specific activity ($\mu\text{mol min}^{-1} \text{mg}^{-1}$)	yield of activity (%)	enrichment (x -fold)
cell-free extract	4400	560	0.13	100	1
ammonium sulfate	4100	540	0.13	97	1
phenyl-agarose ^a	460	450	0.97	80	7
DEAE-Sephacel ^b	38	210	5.4	37	41
Sephadex G-150 ^c	18	140	8.1	26	60
anion exchange ^d	4.7	90	19	16	150
ammonium sulfate	2.4	55	23	10	180
Superose ^e	1.2	45	37	8	290

^aBeaded, cross-linked agarose (Figure 1–7) to which phenyl groups are attached in ether linkage. ^bFigure 1–3. ^cBeaded, cross-linked dextran for chromatography by molecular exclusion. ^d(CH₃)₃N⁺CH₂ – groups covalently linked to a beaded hydrophilic polyether. ^eBeaded, cross-linked agarose for chromatography by molecular exclusion.

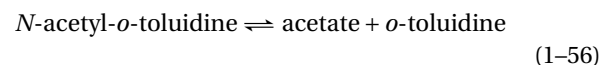
being assayed. For an enzymatic assay, this value is the number of micromoles of reactant that would be converted to product every minute if one milliliter of the solution had been added to the assay. The **total activity** present after any step in the purification is the activity milliliter^{–1} multiplied by the total number of milliliters in the pool of fractions. The **yield of activity** is the percentage of the initial total activity remaining after each step. Although the yield of activity usually decreases as the purification proceeds, sometimes it increases, for example if an inhibitor of the activity is removed during a step.⁷²

The **concentration of protein**, in units of milligrams milliliter^{–1}, in the pool of fractions is also assayed. The most accurate method for making this determination⁷³ is quantitative amino acid analysis (Figure 1–3), but this procedure is too tedious and time-consuming for routine assays. The Biuret colorimetric assay⁷⁴ is the most accurate rapid method, but its low sensitivity often requires that an unreasonable portion of a precious sample be sacrificed. The Lowry⁷⁵ colorimetric method, because of its sensitivity, is the most widely used method for determining the concentration of protein in a sample, but it suffers from the drawbacks that many solutes other than protein also produce color and that different proteins give different yields of color. For example, it was shown that the concentration of protein in samples of purified hydrogenase I from *Clostridium pasteurianum*, which had been accurately quantified by quantitative amino acid analysis, was overestimated by the Lowry procedure by a factor of 1.37 ± 0.03 .⁷⁶ The least quantitative but most convenient and rapid methods for assessing the concentration of protein are the colorimetric method of Bradford⁷⁷ and the absorbance of the solution at 280 nm. The **specific activity** of a pool of fractions from a step in the procedure for purifying the protein is the amount of biological activity displayed by a milligram of the proteins in that solution—the activity milliliter^{–1} divided by the amount of protein milliliter^{–1}. For an enzyme, the units of specific activity are (micromoles

of reactant converted) minute^{–1} (milligram of protein)^{–1}. The **enrichment** in the protein of interest during a particular series of steps is the increase in its specific activity relative to its initial specific activity in the homogenate.

There is a **conventional order** in which the various steps of the purification are carried out. This order is usually determined by the amount of material a certain procedure can accommodate, because the amounts that must be processed, if the samples have been concentrated after each step, always decrease as the purification proceeds because of the decrease in the total amount of protein. Precipitations can be carried out on large volumes and are usually the first step in a purification. If appropriate, selective adsorption is used in the next step because it is an efficient method for handling large samples and the media are usually inexpensive. Chromatography by ion exchange is usually used before chromatography by adsorption because the media used for the former are usually less expensive and have higher capacity. Chromatography by molecular exclusion is usually used as a late step because it is most successful when the samples, and hence the amount of protein, are as small as possible.

The purification of aryl-acylamidase



from *Nocardia globerule* (Table 1–1) illustrates this systematic strategy. In each step of the purification the specific enzymatic activity increases as extraneous proteins are separated from the desired protein, and the yield of enzymatic activity after each step is high. Nevertheless, because there are so many steps, the overall yield is only 8%, but an 8% yield is high for the purification of a protein. In this example, chromatography by molecular exclusion on Superose is used in the last step when total amounts of protein are small so that the samples can be concentrated to the small volumes required by this pro-

cedure. Chromatography by anion exchange (DEAE-Sephacel), however, can be used early in the purification because large volumes at low concentration of electrolyte can be passed through the ion-exchange medium to concentrate the protein on the top of the column. The chromatography itself is then initiated by increasing the concentration of electrolyte.

The **precipitation of proteins** from an aqueous solution that is effected by the addition of a high concentration of another solute has a long history. Originally, such precipitations were observed upon the addition of certain salts to solutions of proteins. This observation led to the terms **salting out**, to describe a precipitation caused by a salt, and **salting in**, to describe the dissolution of a precipitate caused by a salt. For example, sulfate ion salts out, and thiocyanate ion and guanidinium ion salt in. A systematic study of the effect of salts on the solubility of proteins led to the **Hofmeister series**,⁷⁸⁻⁸⁰ an ordering of various ions on the basis of their ability to salt out or salt in.* Similar effects, however, are observed with nonionic solutes as well, somewhat confounding the words chosen. Urea salts in, and poly(ethylene glycol) salts out.

It has been shown that these capacities of solutes, both ionic and nonionic, to affect the solubility of a protein can be ascribed to differences in **preferential solvation**.⁸³⁻⁸⁵ The preferential solvation of a particular protein by a particular solute can be defined by the equation

$$\text{preferential solvation} \equiv \frac{\left(\frac{\partial m_s}{\partial m_p} \right)_{T, \mu_{H_2O}, \mu_s}}{\gamma_s} \quad (1-57)$$

where m_s is the grams of that solute in the solution for every gram of water, m_p is grams of that protein in the solution for every gram of water, γ_s is the concentration of the solute in the solution in grams milliliter⁻¹, T is the temperature, and μ_{H_2O} and μ_s indicate that both the chemical potential of the water and the chemical potential of the solute must remain constant as the grams of protein, ∂m_p , change. Solutes that display negative values of preferential solvation salt out and solutes that display positive values of preferential solvation salt in, and the magnitude of their values of preferential solvation correlates with the potency of their ability to salt out or salt in.

A negative value for preferential solvation, indicating salting out, states that grams of solute, ∂m_s , must be

removed from the solution whenever grams of anhydrous protein, ∂m_p , are added to maintain constant chemical potential. The usual reason⁸³ given for the observation of negative preferential solvation is that, in an aqueous solution, the layer of water surrounding the protein has properties distinct from those of the rest of the water in the solution and a salting-out solute is preferentially excluded from that layer of solvation. The reason grams of solute must be removed to maintain a constant chemical potential is that water is removed from the bulk solution to form this layer of hydration and solute must be removed from the overall solution to keep its concentration the same in the bulk solution surrounding the hydrated protein.

A positive value for the preferential solvation of a particular solute states that the grams of that solute in the solution must be increased when the grams of protein are increased in order to maintain constant chemical potential. Therefore, the solute prefers to interact with the protein rather than with water; for example, it has a higher solubility in the layer of water around the protein or it simply binds to the protein. Positive preferential solvations mean that the protein becomes more soluble as the solute is added to the solution. Such salting-in is displayed by urea, potassium thiocyanate, and guanidinium chloride. At concentrations of 1 M, the value of the preferential solvation of bovine serum albumin by potassium thiocyanate⁸³ is +0.07 mL g⁻¹ and that of bovine serum albumin by guanidinium chloride⁸⁴ is +0.26 mL g⁻¹. The ability of urea to increase the solubility of proteins is frequently used during their purification. For example, the proteins that form intermediate filaments, which are naturally occurring, insoluble polymeric aggregates of protein, are purified from the solution that is obtained by dissolving the filaments in 7 M urea.⁸⁶ The advantage of using urea is that because it is a neutral molecule, it has no effect on chromatography by ion exchange.

It is for precipitation, however, that preferential solvation is usually exploited during purification of a protein. Assume that a solution is at saturation in the concentration of a particular protein; in other words, the chemical potential of that protein in the saturated solution is equal to the chemical potential of that protein in its precipitate. If a solute with negative value of preferential solvation is added to the saturated solution of protein, some of the protein must precipitate to maintain a constant chemical potential. In reality, what happens is that as more and more of the solute is added, the chemical potential of the protein decreases until it equals that of its precipitate and then it begins to precipitate. The more negative the value of preferential solvation for the solute being added, the more rapidly does the concentration of protein reach and then surpass saturation. At 1 M concentration, the value for the preferential solvation of bovine serum albumin by sodium sulfate⁸³ is -0.52 mL g⁻¹. As a comparison, the preferential solvation

* The effects of salts on many properties of proteins, such as their enzymatic activity⁸¹ and their specific associations with each other,⁸² are often governed by the Hofmeister series.⁸⁰

of bovine serum albumin by NaCl, a salt that shows weak salting-out, is -0.26 mL g^{-1} at a concentration of 1 M. Although sodium sulfate has been used to precipitate proteins during purifications, ammonium sulfate is preferred because it is more soluble than sodium sulfate and it is also lethal to fungi or bacteria that would otherwise be happy to use the precipitated protein as a source of food. A protein as a precipitate in a concentrated solution of ammonium sulfate at 4°C is usually stable for decades. Traditionally, the concentration of ammonium sulfate used to precipitate a protein is expressed as the percentage that the final concentration in the solution is of the concentration of ammonium sulfate at saturation (0.52 g mL^{-1} at 4°C).

Ammonium sulfate at high concentrations causes most proteins to precipitate from solution. In the example of aryl-acylamidase (Table 1–1), the enzyme was precipitated between 25% and 60% ammonium sulfate. No purification was observed in this instance; the step was used to concentrate the protein and rapidly remove it from all of the other metabolites in the clarified homogenate. Usually, however, an attempt is made to obtain some purification. Each protein precipitates in a given range of ammonium sulfate concentration. Extraneous proteins that precipitate at lower concentrations can be removed first, and then the protein being purified can be precipitated by raising the concentration of ammonium sulfate and thus be separated from proteins that remain soluble at the higher concentration. For example, formate–tetrahydrofolate ligase was purified 10-fold by bringing the solution of ammonium sulfate to 50% of saturation to precipitate other proteins, then increasing the concentration of ammonium sulfate to 70% of saturation to precipitate the synthase while leaving yet other proteins in the supernatant.⁶⁸ Purification by ammonium sulfate precipitation is usually not so large as in this example, but the procedure is a mild one, usually of high yield. Precipitation with ammonium sulfate can be used to concentrate rapidly and gently a solution of protein between later steps in a purification (Table 1–1).

Poly(ethylene glycol) has also been used to precipitate proteins selectively and reversibly. It is easy to imagine why a large hydrophilic polymer such as poly(ethylene glycol) would be excluded from the layer of water surrounding a protein and thus have a negative value for preferential solvation. Tryptophan 5-mono-oxygenase can be purified 5-fold after precipitation with poly(ethylene glycol) and redissolution in aqueous buffer.⁸⁷ **Trimethylamine oxide**, a naturally occurring solute in the serum of fish,⁸⁸ is also able to precipitate proteins.⁸⁹

Several other types of precipitation are used during the purification of a protein. At the pH at which a given protein bears no net charge, known as its isoelectric pH, it is least soluble in water. If the pH is adjusted to this value and the salts in the solution are removed by

dialysis, the protein will often precipitate, while other proteins, which have different isoelectric points, do not. Such an **isoelectric precipitation** has been used in the purification of aspartate carbamoyltransferase⁹⁰ and in the purification of fibrinogen.⁹¹ One traditional method of concentrating protein and removing it from other molecules in a homogenate is to precipitate it by adding acetone.⁹² The resulting dry **acetone powder** can be extracted with a buffered aqueous solution, and if one is lucky, the protein of interest will dissolve. Because their DNA is not contained in nuclei, when bacterial cells are fragmented by homogenization, the DNA is released as an intractable, gelatinous mass. Before the solution can be processed further, the DNA must be precipitated with **streptomycin sulfate**⁹³ or the DNA must be hydrolyzed to small fragments that are not gelatinous by adding nucleases to the solution.

Isoelectric precipitation and precipitations with poly(ethylene glycol) and ammonium sulfate are reversible, and the protein is readily redissolved by decreasing the concentration of precipitant or changing the pH. In contrast, precipitation by acid or heat is usually not reversible. In these situations advantage is taken of the ability of the protein of interest to remain in solution while other proteins precipitate irreversibly. An example of the use of **precipitation with heat** occurs in the purification of 6-phosphofructokinase, and during this step a 2.5-fold increase in specific activity was recorded.⁹⁴ These techniques are quite harsh and can lead to degradation of the protein being purified by endopeptidases or to chemical alterations such as deamidation of glutamine and asparagine side chains⁹⁵ even though little loss of enzymatic activity is recorded.

Proteins are separated chromatographically by exploiting differences among them in particular properties. Different proteins have different sizes and shapes and can be separated on chromatography by molecular exclusion. Different proteins also have different charges at a given pH and can be separated on **chromatography by ion exchange**. In the case of chromatography by ion exchange, a pH is chosen at which the protein to be purified has a net charge opposite to the fixed charge on the chromatographic medium so that it will participate in ion exchange with the stationary phase as the chromatography progresses. The elution of bound protein is usually performed with a gradient of increasing concentration of a simple monovalent salt such as KCl. If a gradient of pH is used, the change in pH is usually in the direction that would decrease the magnitude of the net charge on the protein. Because the value of α' is changing continuously, the use of a gradient always produces chromatographic separations of much lower resolution than those performed isocratically without a gradient. The advantage of a gradient, however, is that it bypasses the problem of finding conditions of pH and ionic strength at which the value of α' for the protein being purified is in a usable range. Because molecules of pro-

tein are multivalent ions, their values of α' change rapidly as the ionic strength is varied, and such a search is often tedious and fruitless.

Glyceraldehyde-3-phosphate dehydrogenase (GDH), phosphoglycerate mutase (PGM), and phosphoglycerate kinase (PGK) in the ammonium sulfate precipitate from a clarified homogenate could be separated on **chromatography by molecular exclusion** (Figure 1–5A).⁹⁶

Each of the three enzymes migrates with a characteristic elution volume, V_e , and the glyceraldehyde-3-phosphate dehydrogenase is cleanly separated from the other two enzymes by molecular exclusion chromatography on a column of Sephadex G-150. The fractions containing the activities of phosphoglycerate mutase and phosphoglycerate kinase were combined and submitted directly to ion-exchange chromatography on DEAE-cellulose devel-

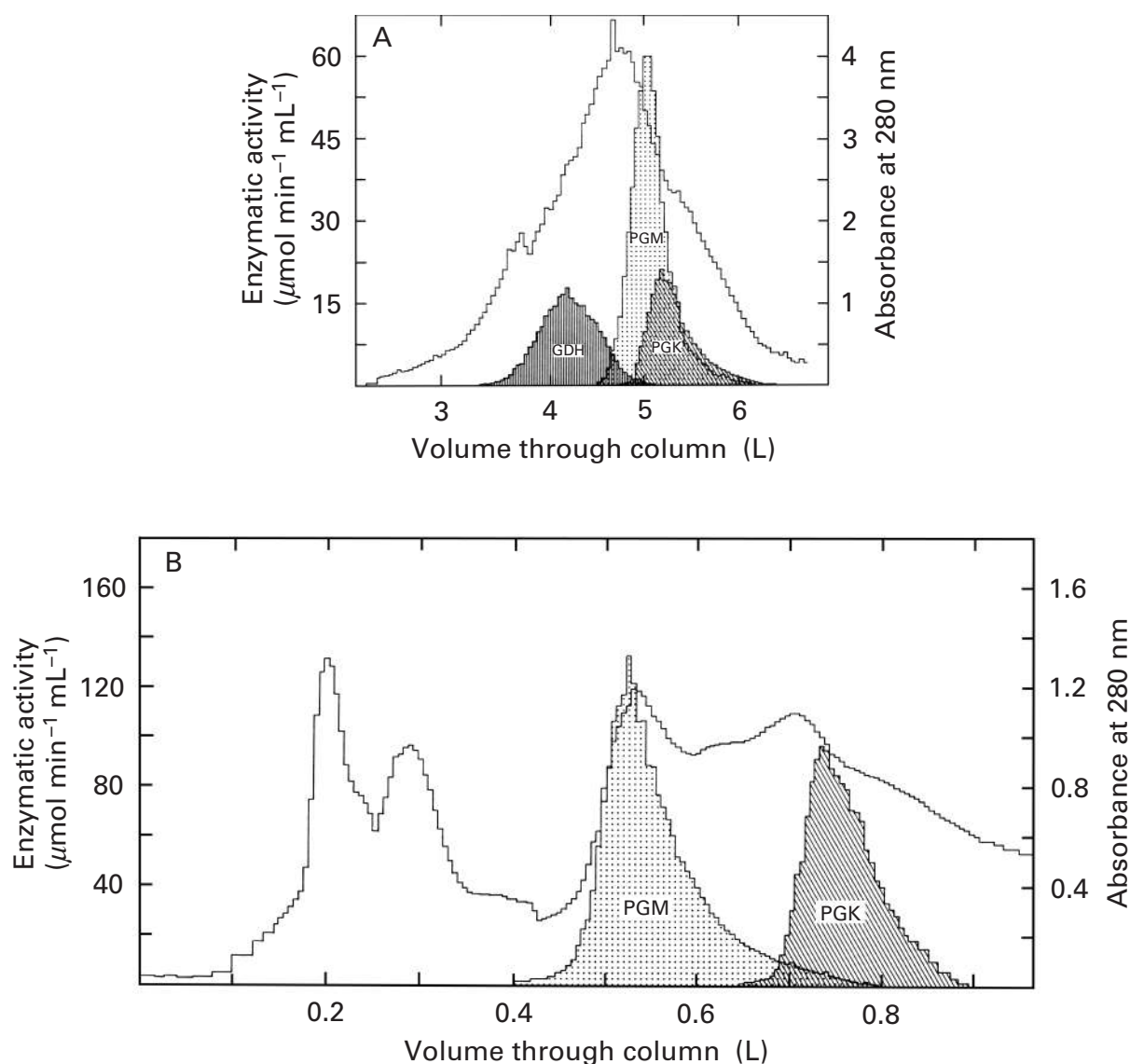


Figure 1–5: Chromatography by molecular exclusion (A) and chromatography by anion exchange (B) of proteins in a homogenate from the bacterium *E. coli*.⁹⁶ The clarified homogenate was submitted to precipitation with ammonium sulfate (30–45%). The precipitate (7.2 g of protein) was redissolved in a minimum volume (120 mL) of aqueous buffer and submitted to zonal chromatography on a column (10 cm \times 120 cm) of cross-linked dextran (Sephadex G-150). (A) Fractions were assayed for protein (absorbance at 280 nm) and enzymatic activity (micromoles minute⁻¹ milliliter⁻¹) of glyceraldehyde-3-phosphate dehydrogenase (GDH), phosphoglycerate mutase (PGM), and phosphoglycerate kinase (PGK), respectively. The proteins contained in the fractions from 4.9 to 5.8 L in the chromatogram in panel A were combined and submitted to chromatography by anion exchange. (B) The ionic strength of the buffer used for the chromatography by molecular exclusion was low enough that the sample (900 mL) could be passed directly through the column (2.2 cm \times 25 cm) of diethylaminoethyl- (DEAE-) cellulose while the proteins gathered at the top of the medium for ion exchange. Chromatography was then initiated with a gradient of NaCl (0–0.15 M in the same buffer at pH 8). Fractions were again assayed for protein and enzymatic activity. Reprinted with permission from ref 96. Copyright 1971 *Journal of Biological Chemistry*.

oped with a gradient of sodium chloride (Figure 1-5B). In this step the phosphoglycerate mutase was cleanly separated from the phosphoglycerate kinase. These examples illustrate the use of column chromatography, monitored by enzymatic assay, to separate proteins.

An example of the use of a sequence of steps of column chromatography to purify a particular protein is found in the purification of α -ketoisocaproate oxygenase from rat liver (Figure 1-6).⁹⁷ Aside from an initial ammonium sulfate precipitation, only three consecutive steps, chromatography by ion exchange (Figure 1-6A), chromatography by adsorption (Figure 1-6B), and chromatography by molecular exclusion (Figure 1-6C), were necessary to purify the enzyme to homogeneity.

Because the resolution of chromatography by ion exchange run with a gradient and the resolution of chromatography by molecular exclusion are not great (Figures 1-5 and 1-6), the increase in specific activity seen in each of the chromatographic steps is usually around 5-fold. Extreme examples of purification, such as the 100-fold purification of 3-deoxy-7-phosphoheptulonate synthase on phosphocellulose⁹⁸ or the 100-fold purification of methylcrotonyl-CoA carboxylase on DEAE-cellulose,⁹⁹ are rare. For reasons that are not obvious, however, it has recently been discovered that the magnitude of the purification on chromatography by adsorption is often significantly greater than that

observed on chromatography by ion exchange or molecular exclusion.

Traditionally, hydroxylapatite, because of its physical properties, has been used mainly for selective adsorption of proteins, but recently much more effective, beaded forms of hydroxylapatite that can be used for chromatography by adsorption have become available. Nitric-oxide reductase could be purified 100-fold by chromatography on one of these media,¹⁰⁰ and acetyl-CoA hydrolase, 60-fold.¹⁰¹ The media most widely used, however, for **chromatography by adsorption** (Table 1-2) are produced by synthetically coupling defined organic functional groups or molecules or chelated metal ions¹⁰⁵ to beaded hydrophilic matrices, usually cross-linked agarose or polymethacrylate. Although the intention in the syntheses in which organic molecules are covalently attached to the polymer has often been to produce a chromatographic medium with a specific affinity for one particular protein or class of proteins, most of these products have turned out to be simple adsorption media with useful and unexpected affinities for proteins in general.¹⁰⁶ Ironically, this makes them more valuable than they were originally intended to be.

Successful purification of a minor component from a complex mixture requires that the set of distribution coefficients, α'_i , for the components present assume a new and randomly permuted sequence of magnitudes as each new chromatographic medium is used. If it were possible to do so, a series of chromatographic steps would be designed so that all of the components that had similar values of α'_i in the preceding step, and that were not separated, have different values of α'_i in the next step and are separated. The availability of a collection of microscopically uniform adsorption media with peculiar and unexpected affinities for proteins in general assists in this strategy. The purification achieved on these media is often dramatic (Table 1-2). There are two methods, how-

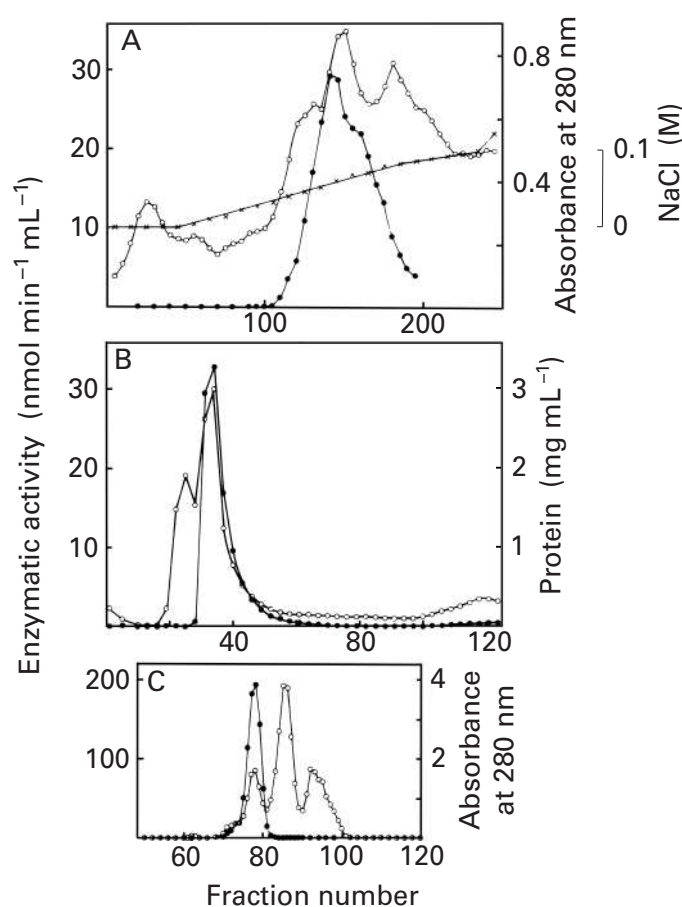


Figure 1-6: Column chromatography of α -ketoisocaproate oxygenase from rat liver.⁹⁷ An ammonium sulfate (45–75%) precipitate (35 g of protein) of the clarified homogenate was redissolved, dialyzed to remove salt, and applied to a column (5 cm \times 80 cm) of DEAE-cellulose (A). The chromatogram was developed with a gradient of NaCl (\times) from 0 to 0.1 M. Fractions containing enzymatic activity (4 g of protein) were pooled, concentrated, brought to 2.5 M NaCl, and applied to a column (4 cm \times 40 cm) of cross-linked agarose to which phenyl groups had been covalently attached. The proteins were eluted with a gradient between 2.5 M NaCl and buffer without added NaCl (B). The fractions containing enzymatic activity (500 mg of protein) were pooled, concentrated, and applied to a column (5 cm \times 80 cm) of allyldextran cross-linked with *N,N'*-methylenebis(acrylamide) for chromatography by molecular exclusion (C). In each panel, α -ketoisocaproate oxygenase activity (nanomoles of CO_2 released minute⁻¹ milliliter⁻¹; \times) is presented as a function of fraction number. The total protein in each fraction (\circ) was also monitored by absorbance at 280 nm. The final yield was 70 mg of protein in the peak of enzymatic activity. Reprinted with permission from ref 97. Copyright 1982 *Journal of Biological Chemistry*.

Table 1–2: Purification of Proteins on Chromatography by Adsorption

protein purified	molecule attached covalently to agarose ^a	property of solution varied for elution	enrichment ^e (x-fold)
coproporphyrinogen oxidase ¹⁰²			
step 1	Cibacron blue	increasing [sodium cholate] ^c	80
step 2	phenyl group ^b	increasing [Tween 80] ^c	2.5
isocitrate dehydrogenase (NADP ⁺) ¹⁹			
step 1	reactive red	increasing [NaCl]	20
step 2	reactive red	increasing [NADP] ^d	15
step 3	phenyl group ^b	decreasing [(NH ₄) ₂ SO ₄]	2
formate–tetrahydrofolate ligase ⁶⁸	Matrex green	increasing [KCl]	5
glutamyl-tRNA reductase ¹⁰³	phenyl group ^b	decreasing [KCl]	9
aminodeoxychorismate lyase ¹⁰⁴	reactive yellow	increasing pH	260

^aIn all cases cited, cross-linked agarose (Figure 1–7) was used as the polymeric support to which the organic molecules were covalently attached. ^bIn ether linkage to agarose. ^cThese solutes are detergents. ^dAffinity elution. ^eEnrichment during each step.

ever, that do not rely on chromatography and that often produce even greater degrees of purification. They are based on the selective elution from or selective adsorption to a stationary phase and can be referred to as affinity elution or affinity adsorption, respectively.

When a protein is purified by **affinity elution**, it is first adsorbed to a stationary phase, such as a chromatographic medium; and after all unabsorbed proteins have been washed away, a compound that binds with high specificity to the protein of interest and leads to its elution is added (as for example in the second step of the purification of isocitrate dehydrogenase, Table 1–2). The presence of this compound can sometimes cause only that protein to which it binds to elute from the stationary phase. For example, when (carboxymethyl)cellulose is added to a crude, clarified homogenate from liver at pH 6, all of the fructose 1,6-bisphosphatase is adsorbed along with many other proteins. When the (carboxymethyl)cellulose is collected, washed well with 5 mM sodium malonate, pH 6, and then rinsed with 0.06 mM fructose 1,6-bisphosphate in 5 mM sodium malonate, pH 6, only the fructose 1,6-bisphosphatase elutes in the rinse. In one step the enzyme can be purified 400-fold, to homogeneity.¹⁰⁷ Transketolase, after initial purification by DEAE-cellulose from homogenates of human leukocytes, will adsorb tightly to the top of a small column (16 mL) of (carboxymethyl)cellulose when a dilute solution (90 mL) of the protein dissolved at low ionic strength is passed over the column. After the column has been washed extensively, the transketolase is eluted with buffer to which xylulose 5-phosphate (0.2 mM) and ribose 5-phosphate (0.3 mM) have been added. The transketolase is purified 40-fold to homogeneity.¹⁰⁸ Protein kinase N, bound to a methylenesulfonate cation-exchange medium, can be eluted specifically with ATP (0.1 mM) for a purification of 2500-fold.¹⁰⁹

Although it requires much more effort, affinity adsorption is more widely used than affinity elution and has been successful in a number of instances. The basic

idea in **affinity adsorption** is to synthesize a stationary phase to which has been covalently attached a chemical compound that binds specifically and with high affinity to the protein being purified. The compound synthetically attached to the stationary phase is usually an analog or a derivative of a reactant or product in the reaction catalyzed by an enzyme, an inhibitor of the enzyme, an allosteric activator of the enzyme, or an agonist or antagonist of a receptor. This compound, when covalently attached to the stationary phase, is referred to as an immobilized ligand for the protein. Cross-linked agarose^{110,111} is the stationary phase to which the immobilized ligand is usually attached.

One of the original examples of this technique¹¹² can serve to illustrate the strategy. Micrococcal nuclease is an enzyme from *Staphylococcus aureus* that can hydrolyze the phosphodiester of either single-stranded RNA or double-stranded DNA to produce as its final products 3'-phosphomononucleotides or dinucleotides. Thymidine 3',5'-bisphosphate is a specific inhibitor of the nuclease that binds to it tightly. A *p*-aminophenyl derivative of this inhibitor was synthesized and attached covalently to agarose through its aniline nitrogen to produce a stationary phase displaying the thymidine 3',5'-bisphosphate (Figure 1–7).¹¹² When a crude supernatant containing micrococcal nuclease was passed over this affinity medium, none of the nuclease emerged but almost all of the protein did. The nuclease could then be eluted nonspecifically with dilute acetic acid in greater than 90% yield. It was completely purified in this one step.

Since this early report, the technical aspects of affinity adsorption have been exhaustively explored. The main difficulty to which many of these investigations have been directed is positioning the ligand far enough from the polymeric matrix of the agarose to minimize steric hindrance and thus interact effectively with the protein.^{113,114} This problem may explain many of the failed attempts to use the technique of affinity adsorp-

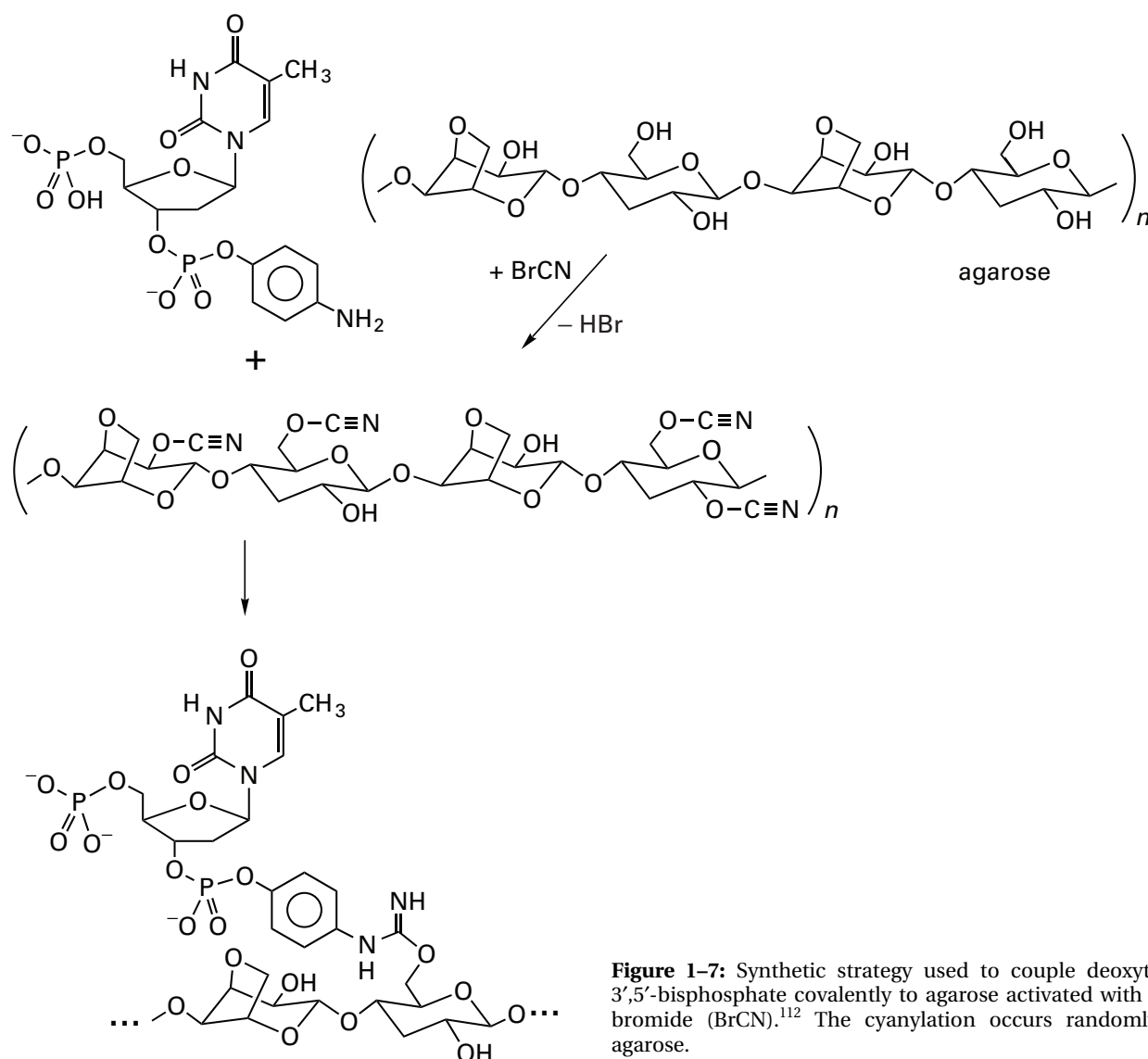


Figure 1-7: Synthetic strategy used to couple deoxythymidine 3',5'-bisphosphate covalently to agarose activated with cyanogen bromide (BrCN).¹¹² The cyanylation occurs randomly on the agarose.

tion. Several long, hydrophilic connecting links, usually referred to as **spacers**, that serve the purpose of the *p*-aminophenyl in the original example (Figure 1-7) have been developed to solve this problem. Often a long hydrophilic spacer is created during the set of reactions used to attach the ligand to the solid phase (Figure 1-8).³⁴ Many different strategies for attaching ligands of various structures to the stationary phase have been developed.

The cases in which affinity adsorption has been successful in the purification of proteins provide a provocative collection of examples (Table 1-3). Because purifications of 100-fold in one step are not unusual, this approach has obvious advantages over the traditional strategy that combines chromatography by ion exchange, chromatography by molecular exclusion (Table 1-1), and chromatography by adsorption (Table 1-2), where several steps are required to achieve the same degree of purification. Affinity adsorption, how-

ever, often requires a greater investment than assembling a sequence of simple chromatographic steps and has a higher risk of failure. Often the affinity adsorbent produces only a modest purification of 10-fold or less under conditions that suggest that the process occurring is either nonspecific ion exchange¹³⁹ or simple adsorption¹⁴⁰ or affinity elution from a nonspecific stationary phase.¹⁴¹ Often the desired protein adsorbs so tightly to the affinity medium that it can be eluted only in low yield.¹⁴²

The central, defining feature of affinity adsorption is the design of the stationary phase, but the conditions used for **elution** of the bound protein are also characteristic. Often they are merely the application of a mobile phase of extreme pH or ionic strength such as in the original example of micrococcal nuclease. The ideal approach, however, is to combine affinity adsorption with affinity elution to gain an advantage in each of the

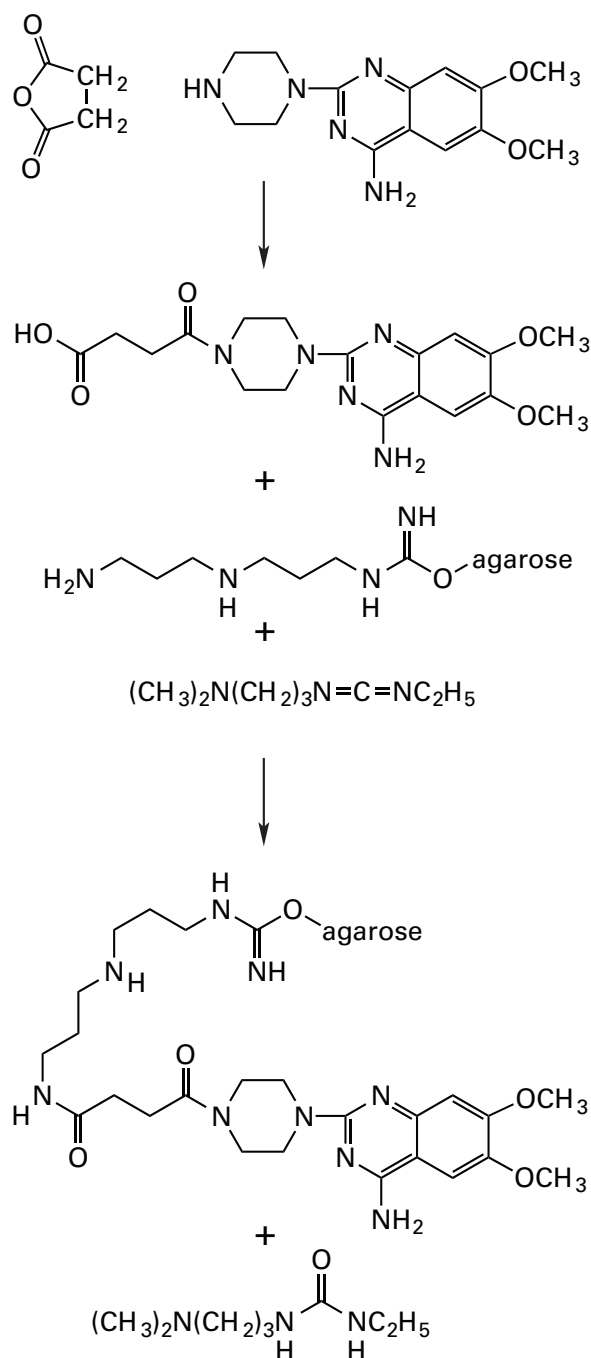


Figure 1-8: Use of a hydrophilic spacer to connect a specific ligand to a polymeric support.³⁴ N,N -Di-(3-aminopropyl)amine was attached to agarose by activating the polysaccharide with cyanogen bromide (Figure 1-7). 1-(4-Amino-6,7-dimethoxy-2-quinazolinyl)piperazine, which is a portion of prazosin, a specific antagonist for α_1 -adrenergic receptors, was succinylated and then attached to the aliphatic amine by activation of the resulting carboxylic acid with 1-[(N,N -dimethylamino)propyl]-3-ethylcarbodiimide. This produced a spacer of 14 atoms connecting an oxygen of the polysaccharide with the nitrogen of the ligand. The spacer is hydrophilic by virtue of the O -alkyl- N -alkyl urea, the amine, and the two N -alkyl amides. This affinity medium was used to purify α_1 -adrenergic receptor.

two steps, and the protein is often eluted with a solution of the soluble ligand from which the immobilized ligand was derived (Table 1-3, Figure 1-9).¹²⁷

Affinity adsorption has also been used to purify proteins that bind to particular nucleotide sequences in DNA.¹⁴³ The spacer holding the DNA recognized by the protein away from the surface of the agarose can be produced by polymerizing short fragments of DNA containing the target sequence to produce a long repeating double strand of DNA and then attaching this long repeating polymer to the agarose through one of its ends. The DNA closest to the surface of the agarose acts as a spacer for the more peripheral segments. Such an affinity adsorbent was used to purify the promoter-specific transcription factor Sp1.³⁹ This protein binds to the nucleotide sequence GGGGCGGGGC in double-stranded DNA, and its concentration in a particular solution can be assayed by observing its footprint on DNA

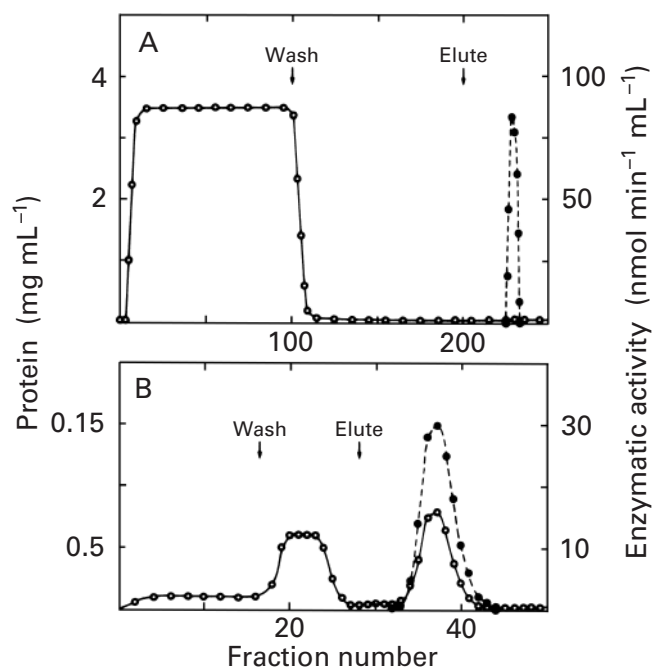


Figure 1-9: Affinity adsorption and affinity elution used in combination to purify 5-formyltetrahydrofolate cyclo-ligase.¹²⁷ (A) A crude extract (7.3 g of protein in 2 L) from the bacterium *Lactobacillus casei* was passed over a column (4 cm \times 18 cm) of agarose to which 5-formyltetrahydropteroylglutamate had been attached. After the affinity adsorbent had been washed with 2 L of buffer until no more protein emerged, the bound enzyme was eluted with a solution of 5-formyltetrahydrofolate, a reactant in the enzymatic reaction. (B) A purified fraction (0.7 mg of protein in 40 mL from a later step in the procedure) was passed over a column (2 cm \times 13 cm) of agarose to which ATP had been covalently attached. After the affinity adsorbent had been washed with 100 mL of buffer, the bound enzyme was eluted with a solution of ATP, another reactant in the enzymatic reaction. Protein concentration (milligrams milliliter⁻¹; \circ) and enzymatic activity (nanomoles minute⁻¹ milliliter⁻¹; \bullet) were measured for each fraction collected from each column. Reprinted with permission from ref 127. Copyright 1984 *Journal of Biological Chemistry*.

Table 1–3: Examples of the Use of Affinity Adsorption in the Purification of Proteins

protein	ligand	point of connection to ligand	elution conditions	enrichment (x-fold)
3-deoxy-7-phosphoheptulonate synthase ¹¹⁵	tyrosine (allosteric inhibitor)	amino group		100
procollagen-proline dioxygenase ¹¹⁶	(Pro-Gly-Pro) _n (<i>n</i> ≅ 10)	amino terminus	solution of (Pro-Gly-Pro) _n	1500
UDP-glucose 4-epimerase ¹¹⁷	UDP	β-phosphate	UMP	100
L-lactate dehydrogenase ¹¹⁸	NAD ⁺	adenosine N6	phosphate	4
	NAD ⁺	adenosine C8	pyruvyl NAD ⁺	40
	AMP	adenosine C8	pyruvyl NAD ⁺	40
isocitrate dehydrogenase (NAD ⁺) ¹¹⁹	AMP	ribose	NAD ⁺	40
choline <i>O</i> -acetyltransferase ¹²⁰	coenzyme A		gradient of NaCl	100
cathepsin D ¹²¹	pepstatin	carboxy group	pH 8.5	100
<i>N</i> -acetylglucosamine kinase ¹²²	glucosamine	N2	glucose	20
hexokinase ¹²³	glucosamine	N2	glucose	40
dihydrofolate reductase ¹²⁴	methotrexate	carboxy group	dihydrofolate	<200
<i>N</i> -acetylgalactosaminide-mucin β-1,3-galactosyltransferase ¹²⁵	asialomucin	bound to DEAE-cellulose	EDTA	20
ornithine decarboxylase ¹²⁶	pyridoxamine phosphate	amino group	pyridoxal phosphate	1000
5-formyltetrahydrofolate cyclo-ligase ¹²⁷	5-formyltetrahydropteroylglutamate	carboxy group	5-formyltetrahydrofolate	4000
β-adrenergic receptor ¹²⁸	alprenolol	olefin addition	isoproterenol	100
α ₁ -adrenergic receptor ³⁴	analogue of prazosin	carboxy group	prazosin	200
plasminogen ^{129,130}	L-lysine	amino group	ε-aminocaproic acid	200
choline <i>O</i> -acetyltransferase ¹³¹	3-[<i>O</i> -(2''-aminoethyl)-3'-hydroxyphenyl]- 3-oxopropyltrimethylammonium	ethylamine	3-(3'-hydroxyphenyl)- 3-oxopropyltrimethylammonium	70
protein geranylgeranyltransferase ¹³²	undecapeptide with sequence YREKKFFCAIL	lysylamines	pH 5.0	130
adenylate cyclase ¹³³	succinylated deacetylforskolin	succinyl carboxylate	forskolin	2000
α subunit of GTP-binding regulatory protein ¹³⁴	βγ subunits of the complete protein	thiols of cysteines	AlF ₄ ⁻	
myristoylated alanine-rich c-kinase substrate ¹³⁵	calmodulin	lysylamines	NaCl, EGTA	100
[heparan sulfate]-glucosamine <i>N</i> -sulfotransferase ¹³⁶	adenosine 3',5'-bisphosphate	adenosine N6	adenosine 3',5'-bisphosphate	40
malate dehydrogenase (oxaloacetate-decarboxylating) (NADP ⁺) ¹³⁷	adenosine 2',5'-bisphosphate	adenosine N6	NADP ⁺	50
binding protein for complement component C3 ¹³⁸	complement component C3	thiol of a cysteine	20% ethanol	

containing this specific sequence. An extract of nuclei from HeLa cells was purified on chromatography by molecular exclusion, chromatography by adsorption on heparin bound to agarose, chromatography by cation exchange on sulfated dextran, and affinity adsorption on agarose to which the specific DNA was attached. In the last step, the protein was eluted with a high concentration (0.5 M) of KCl. The first three steps produced 100-fold purification with a 20% yield, and the last step alone produced a further 100-fold purification with a 50% yield.

The inhibition of the DNA polymerase from herpes simplex virus by the antiviral agent 9-[O-(2-hydroxyethyl) hydroxymethyl]guanosine (acyclovir) results from the formation of a tight complex between the polymerase, a duplex of DNA containing a template and a primer into which 9-[O-(2-hydroxyethyl)hydroxymethyl]guanosine has been incorporated at the 3' end of the primer of DNA as it is being elongated, and the triphosphate of the next nucleotide encoded by the template. When a duplex of template and primer into which 9-[O-(2-hydroxyethyl)hydroxymethyl]guanosine had been incorporated was covalently attached to cross-linked agarose, the resulting affinity adsorbent bound the polymerase strongly, but only when GTP, the next nucleotide encoded by the template strand, was present in the solution. A homogenate was passed over the adsorbent in the absence of GTP to adsorb proteins binding nonspecifically to DNA. The unbound proteins from this first step were then added to the adsorbent in the presence of GTP, and the DNA polymerase was bound strongly. The column was rinsed with high salt and brought to low ionic strength, the GTP was removed to eliminate the binding with high affinity, and the DNA polymerase was then eluted with a gradient of NaCl.¹⁴⁴

It is sometimes the case that during a step of the purification the activity of the enzyme of interest disappears, which can be discouraging. A common reason for such disappearance is digestion of the protein of interest by endopeptidases in the homogenate.^{145,146} A more interesting reason, however, for loss of enzymatic activity is that the function being assayed requires more than one protein and that these proteins are separated from each other during a step in the purification. The removal of mismatched uracil bases from double-stranded DNA is conveniently assayed by following the replacement by [α -³²P]dCTP of a mismatched uracil in the center of a short segment of double-stranded DNA.¹⁴⁷ Although the homogenate from HeLa cells displayed significant activity in this assay, that activity disappeared upon fractionation by sulfopropyl-agarose. It could be regained by combining two of the fractions produced by this step.¹⁴⁷ The active protein in one of these two fractions was then purified to homogeneity by use of the assay supplemented with the other fraction. The ability of this other fraction to support the enzymatic activity in the presence of the purified protein was lost upon its fractionation by phenyl-agarose. When two of the fractions from this step

were combined, activity returned. The single proteins in each of these two fractions were then purified separately, in each case by use of assays supplemented with the other two necessary components. In the end, the three distinct proteins that together perform the reaction were each purified to homogeneity.¹⁴⁸ Only when all three are mixed together is enzymatic activity observed.

The goal of purification is to obtain the protein of interest isolated from all of the other proteins that were originally in the homogenate derived from the biological specimen. That this has been achieved is often suggested by the **coelution** of the protein present and the biological or enzymatic activity in the last chromatographic step of the purification (Figures 1–6 and 1–10).¹⁴⁹ This is only an indication of purity, and the absolute purity of the final preparation must always be demonstrated independently by electrophoresis.

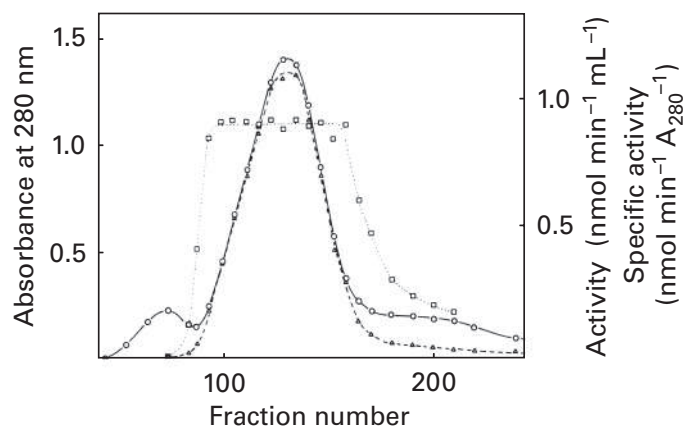


Figure 1–10: Chromatography by molecular exclusion of malate synthase.¹⁴⁹ A solution (280 mg of protein in 7 mL) of malate synthase, from the penultimate step in the purification procedure from *Saccharomyces cerevisiae*, was loaded onto a column (1.8 L) of cross-linked dextran. The fractions (7 mL) collected from the bottom of the column were assayed for absorbance at 280 nm (○) and enzymatic activity (nanomoles minute⁻¹ milliliter⁻¹; △) and specific activity (□) was calculated by dividing enzymatic activity (milliliter)⁻¹ by the absorbance at 280 nm. Reprinted with permission from ref 149. Copyright 1981 Springer-Verlag.

Suggested Reading

- Grimshaw, C.E., Henderson, G.B., Soppe, G.G., Hansen, G., Mathur, E.J., & Huennekens, F.M. (1984) Purification and properties of 5,10-methenyltetrahydrofolate synthetase from *Lactobacillus casei*, *J. Biol. Chem.* 259, 2728–2733.
- Nicholl, I.D., Nealon, K., & Kenny, M.K. (1997) Reconstitution of human base excision repair with purified proteins, *Biochemistry* 36, 7557–7566.

Problem 1–8: Sugars such as lactose have negative values of preferential solvation. Unlike the preferential solvation of most salts, which are affected by activity coefficients, the preferential solvation of lactose is invari-

ant with its concentration. The preferential solvation of bovine serum albumin by lactose⁸⁵ is -0.35 mL g^{-1} . What molar concentration of lactose should have an effect on the solubility of bovine serum albumin equal to a 1 M concentration of Na_2SO_4 ? What is the percent saturation of a 1 M solution of $(\text{NH}_4)_2\text{SO}_4$? Why isn't lactose used to precipitate protein?

Problem 1-9: Calculate the number of theoretical plates in the column used for the separation displayed in Figure 1-5A from the width of the peak of phosphoglycerate mutase. Use the number of theoretical plates to calculate the width the peak of glyceraldehyde-3-phosphate dehydrogenase should have. Why might its peak be wider than the width calculated?

Problem 1-10: Calculate the number of theoretical plates in the column used in Figure 1-6C.

Problem 1-11: The table below describes the purification of glutamyl-tRNA reductase. Calculate, in the proper units, the total enzymatic activity, the yield, the total protein, the specific activity, and the cumulative enrichment at each step.

Problem 1-12: Alprenolol (Al) binds tightly and specifically to β -adrenergic receptor (βAR), which is a protein in the plasma membranes of certain animal cells. The dissociation constant for this binding is the equilibrium constant defined by the equation

$$K_d = \frac{[\text{Al}][\beta\text{AR}]}{[\text{Al} \cdot \beta\text{AR}]}$$

where all concentrations are in moles (liter)⁻¹. They are the concentration of free alprenolol, [Al], the concentration of uncomplexed β -adrenergic receptor, [βAR], and the concentration of the complex between the alprenolol and β -adrenergic receptor, [Al \cdot βAR]. The value for K_d is 8 nM.

Alprenolol was covalently attached to agarose to produce an affinity adsorbent for the purification of β -adrenergic receptor. The final concentration of the alprenolol covalently bound to the solid phase, $[\text{Al}_B]_{\text{TOT}}'$, was 2 mM in units of millimoles (liter of bed)⁻¹. All molar concentrations designated with primes are in moles (liter of bed)⁻¹. Assume that the dissociation constant between covalently bound alprenolol and β -adrenergic receptor is the same as that for unbound alprenolol (8 nM).

Consider what happens when a solution containing β -adrenergic receptor is added to a chromatographic column containing the affinity adsorbent. If, as is reasonable, $[\beta\text{AR}]' \ll [\text{Al}_B]_{\text{TOT}}'$, where $[\text{Al}_B]_{\text{TOT}}'$ is the molar concentration of covalently bound alprenolol (2 mM), then $[\text{Al}_B]_{\text{TOT}}' = [\text{Al}_B]'$, the molar concentration of covalently attached alprenolol to which β -adrenergic receptor is not bound; and from the equation for K_d

$$\frac{[\text{Al}_B]_{\text{TOT}}'}{K_d} = \frac{[\text{Al}_B \cdot \beta\text{AR}]'}{[\beta\text{AR}]'} \cong \alpha'_{\beta\text{AR}}$$

where α' is the partition coefficient for β -adrenergic receptor between the mobile phase, βAR , and its complex with alprenolol covalently bound to the stationary phase, $\text{Al}_B \cdot \beta\text{AR}$.

- (A) If the chromatographic column has a volume of mobile phase, V_0 , of 2.0 mL, calculate the elution volume, $V_{e,\beta\text{AR}}$, of β -adrenergic receptor.

One way to decrease the elution volume of β -adrenergic receptor would be to add free alprenolol to the mobile phase at a particular molar concentration $[\text{Al}_M]'$ in moles (liter of bed)⁻¹. Again, if $[\beta\text{AR}]' \ll [\text{Al}_B]_{\text{TOT}}'$, then

$$\alpha'_{\beta\text{AR}} \cong \frac{[\text{Al}_B \cdot \beta\text{AR}]'}{[\beta\text{AR}]' + [\text{Al}_M \cdot \beta\text{AR}]'}$$

- (B) Derive an equation for $\alpha'_{\beta\text{AR}}$ in terms of $[\text{Al}_M]_{\text{TOT}}'$, $[\text{Al}_B]_{\text{TOT}}'$, and K_d , if

Table for Problem 1-11¹⁰³

purification step	volume of final pool (mL)	enzymatic activity ($\mu\text{mol min}^{-1} \text{ mL}^{-1}$)	protein concentration (mg mL^{-1})
clarified homogenate	2,270	ND ^e	4.6
DEAE-cellulose ^a	720	2.1	4.0
phosphocellulose ^a	300	1.7	1.0
phenyl-agarose ^b	110	2.9	0.20
blue-agarose ^c	7	19	0.45
methylsulfonated polyether ^d	1	25	0.10
Superose molecular exclusion	2	5	0.005

^aFigure 1-3. ^bAgarose (Figure 1-7) to which phenyl groups are attached through ether linkage. ^cAgarose to which Cibacron blue has been covalently attached. ^dBeaded hydrophilic polyether resin to which methylsulfonate groups are covalently attached. ^eInterfering enzymatic activities prohibited assay.

$$K_d = \frac{[Al_M]' [\beta AR]'}{[Al_M \cdot \beta AR]'}$$

- (C) Calculate the elution volume of β -adrenergic receptor from the same chromatographic column ($V_0 = 2$ mL) if the concentration of alprenolol in the mobile phase, $[Al_M]_{TOT}'$, is 0.10 mM.

Molecular Charge

Before electrophoresis can be understood, the property of a molecule of protein that permits its electrophoresis to occur, namely, its molecular charge, must be understood. The **mean net molecular charge number*** of a molecule of protein i , \bar{Z}_i [mean number of elementary charges (molecule of protein)⁻¹], is the difference between the number of its many positive elementary charges and the number of its many negative elementary charges averaged over time. These individual elementary charges are those of adsorbed ions from the solution; those of any tightly bound coenzymes, inorganic anions, and metallic cations; and those of the covalent post-translational modifications of the protein as well as the more obvious positive elementary charges of the guanidinium, ammonium, and imidazolium cations and the negative elementary charges of the carboxylates, thiolates, and phenolates that are the side chains of the amino acids incorporated into the covalent molecular structure of the protein itself.

Each of the charged side chains of the amino acids is the conjugate acid or conjugate base of a weak neutral base or weak neutral acid, respectively (Table 2-2), and the degree to which each is ionized is a function of the pH of the solution. The **mean net proton charge number** of protein i , $\bar{Z}_{H,i}$, is the difference, averaged over time at a particular pH, between the number of all the positive elementary charges and the number of all the negative elementary charges on a molecule of that protein that arise from ions or functional groups that remain affixed to the protein, covalently or noncovalently, in pure water in the

absence of any other dissolved electrolytes in the solution. Because the rates of protonic equilibria are extremely rapid, all molecules of a protein i , even though each has a different number of elementary charges at any instant, will have the same mean net proton charge number, $\bar{Z}_{H,i}$, if the fluctuations of charge are averaged over a time as long as a second.

The change in the mean net proton charge number on a protein as a function of pH is measured by performing a simple **acid-base titration** (Figure 1-11).^{150,151} The number of moles of protons or hydroxide ions necessary to adjust an unbuffered solution containing a known molar concentration of protein i to a given final pH from a given initial pH is measured. The number of moles of protons or hydroxide ions necessary to adjust an identical unbuffered solution, lacking the protein, to the same final pH from the same initial pH is then measured. The solution containing the protein will always consume more moles of protons or hydroxide ions than the control, and this additional amount can be converted into the equivalents of positive charge gained by the protein upon association of the excess protons or the equivalents of positive charge lost upon dissociation of the protons and their combination with the excess hydroxide ions, respectively, as the pH of the solution is changed from the initial value to the final value. In order to anchor this titration curve at some absolute number of net elementary charges on the protein rather than equivalents of elementary charge relative to those on the protein at some arbitrary initial pH, three distinct properties of a protein, which are often confused with each other, must be understood and clearly distinguished. These properties are the isoionic point, the point of zero net proton charge, and the isoelectric point.

The **isoionic point**, $pH_{isoionic,i}$ of protein i is the pH of a solution containing only water and protein i including all of its tightly bound ions, coenzymes, and post-translational modifications.¹⁵⁰ A solution containing only water and protein i , an isoionic solution, is usually obtained by passing a solution of protein i over a mixed-bed ion-exchange medium to remove all salts, and the pH of the resulting solution is then measured.¹⁵¹ Since the only cations and anions in such a solution, other than the ions bound tightly or covalently to the protein, are protons and hydroxide ions

$$(\bar{Z}_{H,isoionic,i}) [\text{protein } i] + [H^+]_{isoionic} = [OH^-]_{isoionic} \quad (1-58)$$

where $[\text{protein } i]$ is the molar concentration of the protein, $\bar{Z}_{H,isoionic,i}$ is the mean net proton charge number on the protein at its isoionic point, and $[H^+]_{isoionic}$ and $[OH^-]_{isoionic}$ are the molar concentrations of protons and hydroxide ions in this isoionic solution. This equation can be combined with the expression for the ionization

* The charge number of an ion, a molecule, or a functional group should be distinguished from its charge. The charge number is the number of elementary charges borne by the ion, the molecule, or the functional group. The charge is the number of coulombs borne by the ion, the molecule, or the functional group. The charge on an ion, a molecule, or a functional group is the elementary charge (1.602×10^{-19} C) multiplied by its charge number. Chemists are accustomed to refer to the number of elementary charges borne by an ion, a molecule, or a functional group, its charge number, as its "charge". This habit involves no misunderstanding so long as the actual charge on the ion, the molecule, or the functional group is never involved in the discussion. Unfortunately, the property that determines its electrophoretic mobility is the charge on the molecule, not its charge number, so in this instance the distinction between charge number and charge must be clear.

of water ($K_w = [H^+][OH^-]$) and the definition of pH ($[H^+] = 10^{-pH}$) to give

$$\bar{Z}_{H, \text{isoionic}, i} = \frac{K_w - 10^{-2 \text{pH}_{\text{isoionic}, i}}}{[\text{protein } i] 10^{-\text{pH}_{\text{isoionic}, i}}} \quad (1-59)$$

Equation 1-59 can be used to calculate the mean net proton charge number on the protein i at its isoionic point, and this provides a measurement of the absolute mean net proton charge number on the protein i at one pH in the absence of electrolytes. It is this direct measurement of the mean number of charges on the protein at a given pH in the absence of electrolyte that is usually used to anchor the titration curve of a protein (Figure 1-11).¹⁵¹

The **point of zero net proton charge** is the pH at which the mean net proton charge number on protein i is zero. The isoionic point, $\text{pH}_{\text{isoionic}, i}$ is formally distinguished from the point of zero net proton charge because at the isoionic point the protein does bear a mean net proton charge. It is clear from Equation 1-58, however, that if $[\text{protein } i]$ is significant and $\text{pH}_{\text{isoionic}, i}$ is between pH 5 and 9, there is little difference between the isoionic point and the point of zero net proton charge. This is not the case, however, for acidic or basic proteins.

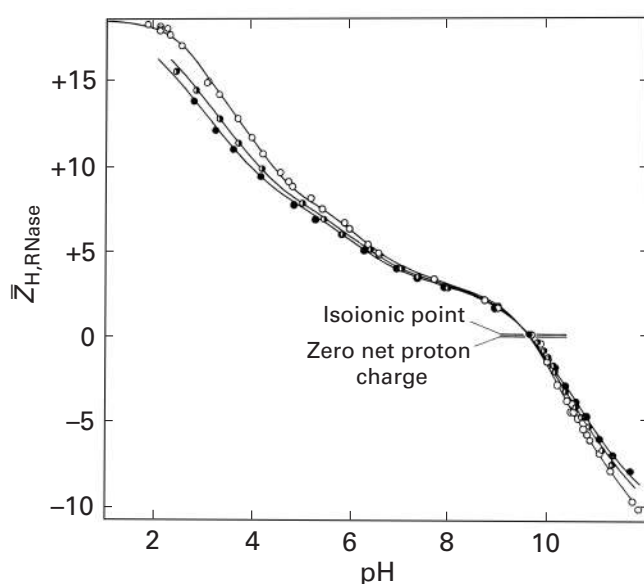


Figure 1-11: Net mean proton charge number on ribonuclease as a function of pH. Solutions of ribonuclease at ionic strengths 0.01 M (●), 0.03 M (●), and 0.15 M (○), produced with KCl, were titrated with either KOH or HCl.¹⁵¹ The changes in pH as a function of the equivalents of acid or base added (mole of protein)⁻¹ were recorded. The isoionic point was determined by passing a solution of the protein over a mixed-bed medium for ion exchange to remove all electrolytes except the protein, H^+ , and OH^- . The point of zero net proton charge was then calculated with Equation 1-59. The absolute mean net proton charge number, $\bar{Z}_{H, \text{RNase}}$, is presented as a function of pH. Reprinted with permission from ref 151. Copyright 1956 American Chemical Society.

It is its point of zero net proton charge that is routinely **estimated from the sequence** of a protein. If it is assumed that the protein bears no unknown tightly bound ions or coenzymes and has no unknown post-translational modifications and if it is assumed that each side chain of each type of amino acid has its ideal, unperturbed value of pK_a (Table 2-2), then it is possible to estimate the point of zero net proton charge of the protein from its composition of amino acids and any known tightly bound ions, coenzymes, and posttranslational modifications (Problem 1-15). Such estimates of points of zero net proton charge are commonly performed by simple algorithms available at data banks on the internet. Such calculations are usually rather inaccurate estimates of the actual points of zero net proton charge because the values for the pK_a of the amino acids are seldom the same in the native protein as their ideal, unperturbed values, which are accurate estimates only when the amino acid is in an unfolded polypeptide and does not have an immediate neighbor with an ionized side chain. For example, Glutamate 89 of β -lactoglobulin is buried within the protein at low pH and does not titrate with the rest of the glutamates but becomes exposed during a change that occurs in the structure of the protein above pH 7 and titrates as the structural change progresses.¹⁵² In addition, there often are unknown tightly bound ions or post-translational modifications. Finally, the point of zero net proton charge is often between pH 6 and 8, where small shifts in the titration curve lead to large changes in the point of zero net proton charge (Figure 1-11). The result of one of these algorithmic estimates of the point of zero net proton charge is usually referred to, erroneously, as the isoelectric point of the protein.

The **isoelectric point** of protein i , pI_i , is the pH at which, under a given set of conditions, the mean net molecular charge number of protein i , \bar{Z}_i , is zero.⁷⁸ The mean net proton charge number on protein i , $\bar{Z}_{H, i}$, differs from the mean net molecular charge number on protein i , \bar{Z}_i , because proteins have a tendency to bind weakly the ions of electrolytes in the solution, even ones as simple as halides¹⁵³ and alkali metal ions.¹⁵⁴ This binding occurs even at the point of zero net proton charge and is reflected as a decrease or increase in $\text{pH}_{\text{isoionic}, i}$ as a neutral salt is added to an isoionic solution.¹⁵⁰ For example, if protein i in an isoionic solution binds more of the anions than the cations of a neutral salt that has been added, the increase in its negative charge will indirectly cause it to take up more protons, increasing $\text{pH}_{\text{isoionic}, i}$. The reverse effect on the isoionic point is observed when the cations are preferentially bound.

This binding of small simple ions, such as halides and alkali metal cations, to proteins results from **chelation**. Two or more fixed charges or dipoles on the protein, of opposite sign to the bound ion, have to be properly oriented to perform such chelation. Consequently, the number of each type of ion bound at the isoionic point is a unique and unpredictable property of each protein. In

deoxyhemoglobin, a site at which chloride binds to the protein has been identified, and it sits between two functional groups, an ammonium cation of the amino terminus and a guanidinium cation of an arginine, that both bear a positive charge and chelate the chloride.¹⁵⁵ In tryptophanase, a site at which potassium ion binds to the protein is formed from the oxygen of a carboxylate and three acyl oxygens from the backbone of the polypeptide that together chelate the ion.¹⁵⁶ In plasminogen activator inhibitor 1, a site at which a chloride ion binds is surrounded by two ammonium cations of two lysines and two NH groups of two amides from the backbone of the polypeptide that all chelate the ion.¹⁵⁷ In exotoxin A from *Pseudomonas aeruginosa*, a site at which a chloride ion binds is formed from two guanidinium cations of two arginines, and a site at which a sodium ion binds is formed by two acyl oxygens from the polypeptide backbone.¹⁵⁸

Although there is no relation between the number of ions bound and the charge on the protein at a particular pH, proteins with high densities of negative charge seem to bind cations more readily than those with low densities of negative charge.¹⁵⁴ This tendency presumably results from the increase in the probability of proper juxtaposition for chelation with the increase in the density of negative charge. As the pH is lowered from the point of zero net proton charge, the density of positive charge on a protein increases only marginally; rather, the density of negative charge decreases as carboxylates are neutralized. It has been observed that the number of bound anions increases as the pH is lowered,¹⁵³ which results from the decrease in electrostatic repulsion, due to these carboxylates, that at neutral pH inhibits the chelation of dissolved anions by the fixed positive charges on the protein. For reasons that are not well understood but may include the differences in ionic radii, proteins seem to bind halides more readily than they do alkali metal ions.

The mean net molecular charge number, \bar{Z}_i , on protein i in a solution containing simple neutral salts such as $(\text{NH}_4)_2\text{SO}_4$, NaCl, or KCl is the sum of the mean net proton charge number and the net charge number contributed by these loosely bound ions:

$$\bar{Z}_i = \bar{Z}_{\text{H},i} + \sum_{j=1}^m \bar{v}_j z_j \quad (1-60)$$

where \bar{v}_j is the mean number of ions of species j and charge number z_j bound by the protein. It is this net charge on protein i that determines its behavior on chromatography by ion exchange or electrophoresis. In turn, electrophoresis is the usual method for determining the isoelectric point of a protein.

Suggested Reading

Tanford, C., & Wagner, M.L. (1954) Hydrogen ion equilibria of lysozyme, *J. Am. Chem. Soc.* 76, 3331–3336.

Problem 1–13: The commercially available anion-exchange medium DEAE-Bio-Gel A is a beaded polymer formed from the naturally occurring polysaccharide agarose (Figure 1–7) to which are attached 2-(diethylamino)ethyl groups (Figure 1–2). In a column poured with DEAE-Bio-Gel A, the concentration of covalently attached tertiary ammonium cations is 20 mmol (L of bed)^{–1}. The bed of such an ion-exchange resin can be divided theoretically into two compartments that can be referred to as the stationary compartment and the mobile compartment. The stationary compartment, which is the volume within the beads, surrounds the covalently attached tertiary ammonium cations and includes enough of the surrounding volume that the compartment is electroneutral. The mobile compartment, which is the volume surrounding the beads, is the remainder of the volume that is accessible to the protein being submitted to the chromatography.

An isoionic solution of bovine serum albumin at 50 mg mL^{–1} has a pH of 5.48. This solution is adjusted to the desired pH with KOH to produce a potassium salt of bovine serum albumin, K_nBSA. Samples of this polyanionic form of bovine serum albumin are submitted to chromatography by anion exchange on a column 4.5 cm in diameter and 40 cm in length of DEAE-Bio-Gel A in the chloride form. The solution within the DEAE-Bio-Gel A itself has been adjusted with HCl to the same pH as that of the solution of protein and equilibrated with an unbuffered solution of KCl. No buffer has to be used because the bovine serum albumin and the diethyl aminoethyl groups on the agarose provide adequate buffering.

The movement of the bovine serum albumin through the chromatographic system will be determined by its partition coefficient between the stationary compartment and the mobile compartment

$$\alpha'_{\text{BSA}} = \frac{[\text{BSA}^{n-}]'_s}{[\text{BSA}^{n-}]'_m}$$

where the superscript $n-$ refers to the mean net molecular charge number on the bovine serum albumin at the chosen pH, and as in Equation 1–1, the primes on the concentrations indicate that they are in units of moles (liter of bed)^{–1}. The free energies of transfer of the ions between the stationary compartment and the mobile compartment, however, are governed by the actual molar concentrations of the bovine serum albumin in the two compartments (indicated by the unprimed brackets as usual) according to the partition coefficient

$$\alpha_{\text{BSA}} = \frac{[\text{BSA}^{n-}]_s}{[\text{BSA}^{n-}]_m}$$

(A) Show that

$$\alpha_{\text{BSA}} = \alpha'_{\text{BSA}} \left(\frac{1-f_s}{f_s} \right)$$

where f_s is the fraction of the total accessible volume of the bed, V_T , that is the volume of the stationary compartment, V_s . Note that by definition the sum of V_s and the volume of the mobile compartment, V_M , is V_T .

The ideal distribution of bovine serum between the mobile and stationary compartments in the DEAE-Bio-Gel A is governed by equations equivalent to Equations 1–15 to 1–18 that describe the conservation of charge in the two compartments and the equivalence of the ideal activities of the various dissolved salts in the system.

(B) Write four equations equivalent to Equations 1–15 to 1–18 for the special case of bovine serum albumin on DEAE-Bio-Gel A. Use the explicit abbreviations K^+ , Cl^- , BSA^{n-} , and $DEAE^+$. Remember that for the potassium salt of a multivalent anion, K_nA , where the charge number on anion A is $n-$, the ideal activity of the salt in a solution is

$$a_{K_nA} = [K^+]^n [A^{n-}]$$

In this equation, n does not have to be an integer.

(C) Unlike the derivation in the book, assume that only the concentration of bovine serum albumin, not $[K^+]_s$, is negligible and show that

$$[K^+]_s = \frac{\sqrt{[DEAE^+]_s^2 + 4[K^+]_M^2} - [DEAE^+]_s}{2}$$

that

$$\alpha_{\text{BSA}} = \left(\frac{\sqrt{[DEAE^+]_s^2 + 4[K^+]_M^2} + [DEAE^+]_s}{2[K^+]_M} \right)^n$$

that

$$\alpha'_{\text{BSA}} = \left(\frac{f_s}{1-f_s} \right) \left(\frac{\sqrt{[DEAE^+]_s'^2 + 4f_s^2[K^+]_M^2} + [DEAE^+]_s'}{2f_s[K^+]_M} \right)^n$$

that

$$\frac{(\alpha_{\text{BSA}})^{2/n} - 1}{(\alpha_{\text{BSA}})^{1/n}} = \frac{[DEAE^+]_s}{[K^+]_M}$$

and that

$$\frac{\left[\alpha'_{\text{BSA}} \left(\frac{1-f_s}{f_s} \right) \right]^{2/n} - 1}{\left[\alpha'_{\text{BSA}} \left(\frac{1-f_s}{f_s} \right) \right]^{1/n}} = \frac{[DEAE^+]_s'}{f_s[K^+]_M}$$

where $[DEAE]'$ is the concentration of covalently attached tertiary ammonium cations: 20 mmol (L of bed)⁻¹.

(D) The titration curve of bovine serum albumin¹⁵⁹ is such that the value of the partial derivative $(\partial \bar{Z}_{H,BSA} / \partial pH)_{T,Ic}$ has a constant value over the region from pH 5.5 to 7.0 of -5.9 .

Assume that, under the conditions of the experiment, bovine serum albumin does not bind either K^+ or Cl^- . What is the mean net molecular charge number on the bovine serum albumin at pH 6.00, and what is the mean net molecular charge number on the bovine serum albumin at pH 7.00?

(E) Before solutions containing the potassium salts of bovine serum albumin are run, a sample of the isoionic solution of bovine serum albumin at 50 mg mL⁻¹ is adjusted to pH 5.0 with HCl and run on the column of DEAE-Bio-Gel A equilibrated at pH 5.0 and eluted with 0.04 M KCl. The elution volume of the bovine serum albumin in this run is 537 mL. What parameter of the ion-exchange column is measured by this experiment?

(F) A sample of the isoionic solution of bovine serum albumin at 50 mg mL⁻¹ is adjusted to pH 6.00 with KOH and run on the column of DEAE-Bio-Gel A equilibrated at pH 6.00 and eluted with 85 mM KCl. The elution volume of the bovine serum albumin on this run is 3.38 L. What is the value of α_{BSA} under these conditions?

(G) Show that the value of f_s for the DEAE-Bio-Gel A in the column is 0.060.

(H) A sample of the isoionic solution of bovine serum albumin at 50 mg mL⁻¹ is adjusted to pH 7.00 with KOH and run on the column of DEAE-Bio-Gel A equilibrated at pH 7.00. What concentration of KCl must be used to have the elution volume of the bovine serum albumin be 3.00 L?

(I) Explain which of the assumptions, either implicit or explicit, relied upon in the preceding development are most certainly oversimplifications and explain why each of them is an oversimplification.

Problem 1–14: At a protein concentration of 3×10^{-4} M, the isoionic pH of ribonuclease¹⁵¹ is 9.60. Calculate $\bar{Z}_{H,\text{isoionic,RNase}}$.

Problem 1–15: Assume that the side chains of the acidic and basic amino acids in a native properly folded protein all have the same values for their acid dissociation constants that they do in the unfolded polypeptide (Table 2–2). Let f be the fraction of a particular acidic or basic amino acid that is ionized at a given pH.

- (A) Show that for a particular type of amino acid the conjugate base of which is anionic, such as aspartate, glutamate, cysteine, tyrosine, or a carboxy terminus,

$$f_{\text{anionic}} = \frac{1}{1 + 10^{(\text{p}K_a - \text{pH})}}$$

where the $\text{p}K_a$ is the one found in Table 2–2. Show that for a particular type of amino acid the conjugate acid of which is cationic, such as histidine, lysine, arginine, or an amino terminus,

$$f_{\text{cationic}} = \frac{1}{1 + 10^{(\text{pH} - \text{p}K_a)}}$$

where the $\text{p}K_a$ is the one found in Table 2–2 for that amino acid.

In a molecule of fructose-bisphosphate aldolase from rabbit skeletal muscle, there are four identical polypeptides, each containing one amino terminus, 14 aspartates, 24 glutamates, 11 histidines, eight cysteines, 12 tyrosines, 26 lysines, 15 arginines, and one carboxy terminus. There are no bound coenzymes or posttranslational modifications. The $\text{p}K_a$ of a carboxy terminus is 3.3, and that of an amino terminus is 8.0.

- (B) Calculate the mean net proton charge number on a molecule of fructose-bisphosphate aldolase at pH 8 and at pH 9. (If you are adept at using a computer, go to part E first).
- (C) Estimate the point of zero net proton charge for fructose-bisphosphate aldolase.
- (D) What is the value of the point of zero net proton charge for fructose-bisphosphate aldolase according to the experiments in Figure 1–16?
- (E) Write a program or program a spreadsheet to calculate the mean net proton charge number on fructose-bisphosphate aldolase at any pH.
- (F) Use the program to draw a titration curve of fructose-bisphosphate aldolase.

Problem 1–16: A particular protein is modified within the cells where it is normally located by the covalent attachment of inorganic phosphate in the form of phosphate esters. Anywhere between zero and seven phosphates can be attached to the protein under normal circumstances. The isoelectric points of these eight dif-

ferent forms of the protein are 4.32, 4.29, 4.26, 4.23, 4.20, 4.17, 4.14, and 4.11, respectively. At these values of pH, each phosphate ester would have a charge number of -1.00 so an additional equivalent of negative charge is added to the protein when an additional phosphate is added.

- (A) Explain why the isoelectric point of the protein decreases as each phosphate is added.
- (B) What amino acid side chains are titrating in this range of pH? (See Table 2–2).
- (C) Assume that the aspartates and glutamates of the protein have the same $\text{p}K_a$ (4.2). The decrease in the mean net proton charge number on a protein as the pH is lowered, if only the glutamates and aspartates are titrating, should be

$$\Delta \bar{Z}_H = n_{\text{E+D}} \left[\frac{1}{1 + 10^{(4.2 - \text{pH}_f)}} - \frac{1}{1 + 10^{(4.2 - \text{pH}_i)}} \right]$$

where $n_{\text{E+D}}$ is the total number of glutamates plus aspartates in the protein, pH_f is the final pH, and pH_i is the initial pH. What is the total number of glutamates plus aspartates in the protein?

Electrophoresis

When a molecule of protein i at a given pH in an aqueous solution of electrolytes is placed in an electric field, it will experience a force, F_{el} , in the direction x such that

$$F_{\text{el}} = Q_i E_x = e_a \bar{Z}_i E_x \quad (1-61)$$

where Q_i is the mean charge on protein i (coulombs), e_a is the elementary charge (1.602×10^{-19} C), \bar{Z}_i is the mean net molecular charge number of the molecule of protein i under these circumstances, and E_x is the electrical field (volts centimeter $^{-1}$) or gradient of the electrical potential ($\partial V / \partial x$) in the x direction. The units of force (grams centimeter second $^{-2}$) follow from the fact that one volt is one joule coulomb $^{-1}$ (10^7 gram centimeter 2 second $^{-2}$ coulomb $^{-1}$). Electrophoresis is usually run in an apparatus designed so that $(\partial V / \partial y)$ and $(\partial V / \partial z)$ are zero, and the force F_{el} will cause the molecule of protein i to move only in the x direction.

For the moment, it will be assumed that only the molecule of protein i and its physically bound ions move. As the molecule of protein i moves, a frictional force, F_{fric} , exerted by the surrounding stationary liquid is experienced by the molecule. The frictional force is proportional to the velocity of movement of the molecule

$$F_{\text{fric}} = -f_i \left(\frac{\partial x_i}{\partial t} \right)_E \quad (1-62)$$

where the constant of proportionality, f_i , is the **frictional coefficient** (grams second⁻¹) of the molecule of protein, one of its physical properties.

At this point a digression is necessary to explain the frictional coefficient before continuing with a discussion of electrophoresis. The most direct way to determine the frictional coefficient of a molecule of protein is from its diffusion coefficient, D . The **diffusion coefficient** is a measure of the net tendency of any population of identical molecules to spread from a region of high concentration to a region of low concentration; the driving force behind this movement is not a function of any intrinsic feature of the individual molecules such as their charge number or their mass. The diffusion coefficient D_i (centimeters² second⁻¹) of any substance i in solution is defined by Fick's law

$$J_{x,i} = -D_i \left(\frac{\partial c_i}{\partial x} \right)_t \quad (1-63)$$

where $J_{x,i}$ is the flux (moles centimeter⁻² second⁻¹) of substance i through a planar surface of unit area, c_i is the concentration (moles centimeter⁻³) of the substance i at any point, and x is the distance (centimeters) along an axis normal to the planar surface. The greater $(\partial c_i / \partial x)_t$, the greater the diffusive force, and the greater the net flux. The diffusion coefficient of substance i , D_i , is simply the constant of this proportionality. It can be shown that

$$f_i = \frac{k_B T}{D_i} \quad (1-64)$$

where k_B is Boltzmann's constant (1.38×10^{-16} g cm² s⁻² K⁻¹) and T is the temperature (kelvins).

The diffusion coefficient of a protein is most unambiguously measured by creating a sharp boundary between two solutions, one of which contains the protein at a given initial concentration and the other of which is otherwise identical to the first but does not contain the protein (Figure 1-12).¹⁶⁰ At any time after initiating the experiment, $(\partial c_i / \partial x)_t$, where x is normal to the original boundary, will be a Gaussian function. The width of this function will increase with time as diffusion spreads the boundary, and

$$D = \frac{1}{4\pi t} \left(\frac{A}{H} \right)^2 \quad (1-65)$$

where A is the area (concentration) of the curve of $(\partial c_i / \partial x)_t$ against x and H is its maximum height (concentration centimeter⁻¹). At the present time, however, the diffusion coefficients of proteins are usually measured by dynamic light scattering^{161,162} or by pulsed field gradient nuclear magnetic resonance.^{163,164}

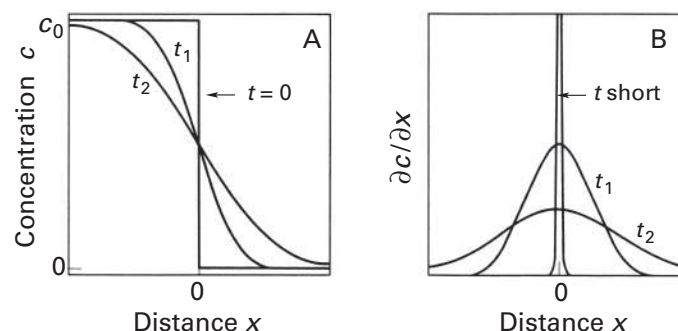


Figure 1-12: Measurement of a diffusion coefficient.¹⁶⁰ (A) Spreading of a boundary of concentration at the interface formed between two solutions, one containing the solute and the other not containing the solute. A solution containing the solute is brought in contact with a solution otherwise identical, but lacking the solute, to form an interface at the origin of the horizontal axis. At the initial time the function of the concentration (c) is discontinuous at the interface at the origin of the horizontal axis, but as time progresses (t_1 and t_2) the solute diffuses in the direction x normal to the interface into the vacant solution and a gradient of concentration develops. (B) The first derivative of the function of concentration with respect to distance in the direction x $[(\partial c / \partial x)_t]$ at any instant is a Gaussian function (curves labelled t_1 and t_2), the width of which increases and the height of which decreases with time, t . Reprinted with permission from ref 160. Copyright 1961 John Wiley.

The frictional coefficients of spheres or ellipsoids of revolution can be calculated. For a sphere

$$f = 6\pi\eta r \quad (1-66)$$

where η is the viscosity (pascal seconds, where a pascal is a kilogram second⁻² meter⁻²) of the solution and r is the radius (centimeters) of the sphere. This equation has led to the formalism of the **effective sphere** or Stokes' sphere representing protein i , the radius of which, a_i , is defined as

$$a_i \equiv \frac{k_B T}{6\pi\eta D_i} \quad (1-67)$$

This radius, a_i , is simply the radius of a sphere the diffusion coefficient of which is the same as the diffusion coefficient of protein i . It is usually referred to as the **Stokes' radius** of the protein.

It is now possible to return from the digression defining the frictional coefficient to the molecule of protein in the electric field. When the electric field is turned on, a steady state⁴ is rapidly reached in which $F_{el} = -F_{fric}$ and which is characterized by a constant terminal velocity $(\partial x_i / \partial t)_E$ of the molecules of protein i in the direction of the electric field. At steady state, because $F_{el} = -F_{fric}$

$$\left(\frac{\partial x_i}{\partial t} \right)_E = \frac{e_a \bar{Z}_i E_x}{f_i} \quad (1-68)$$

Although electrophoresis is usually carried out in an apparatus in which the current passes through a complicated path, the region of the apparatus over which the proteins are actually separated, the **electrophoretic field**, is uniform in its dimensions and in its specific conductance so that E_x is constant over its length. The **free electrophoretic mobility**, u_i° (centimeters² volt⁻¹ second⁻¹) of protein i is defined as

$$u_i^\circ \equiv \frac{(\partial x_i / \partial t)_E}{(\partial V / \partial x)_t} = \frac{d_i l}{Vt} \quad (1-69)$$

where d_i is the distance (centimeters) moved by protein i in time t (seconds) when a particular voltage V is applied to an electrophoretic field of length l . This definition causes the electrophoretic mobility to be only a function of the molecule of protein and the medium through which it is moving.

It follows that, if the assumptions that have been made were correct, the relationship governing electrophoresis would be

$$u_i^\circ = \frac{e_a \bar{Z}_i}{f_i} \quad (1-70)$$

This relationship, however, is an incomplete description of electrophoresis and fails to explain actual behavior.¹⁶⁵ Equation 1-70 states that electrophoretic mobility will be affected by ionic strength only insofar as \bar{Z}_i is affected by ionic strength. In general, \bar{Z}_i increases gradually but not impressively with ionic strength, as ionic shielding permits the molecule of protein i to bear a greater net charge (Figure 1-11), yet it is observed that electrophoretic mobility declines precipitously as ionic strength is increased (Figure 1-13).¹⁶⁵

The inadequacy of Equation 1-70 is due to the erroneous assumption that the only participant responding to the electric field is the molecule of protein and its directly bound ions. This would be true if the molecule of protein were dissolved in pure water with no added electrolyte. In fact, the value for the extrapolation of the experimental curve in Figure 1-13 to zero ionic strength does seem to agree with that calculated from Equation 1-70 (the upper line in the graph). An actual solution of protein, however, must at the very least contain counterions to balance its charge, and in order to perform the electrophoresis, additional electrolyte must be added to the solution as well.

When a charged particle such as a molecule of protein is dissolved in a solution of water containing electrolyte, the solution surrounding the molecule of protein has a net charge of opposite sign due to the existence of an **ionic double layer**.¹⁴ In the present

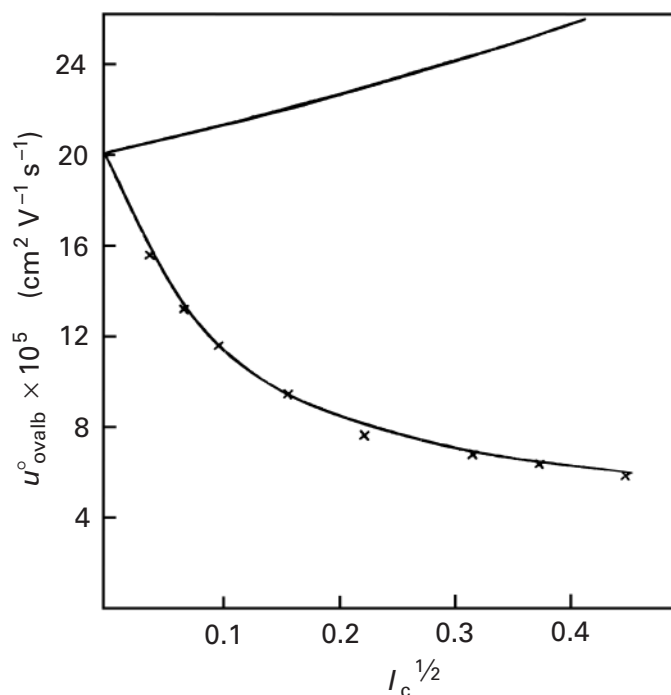


Figure 1-13: Free electrophoretic mobility (u_{ovalb}°) of the protein ovalbumin at pH 7.1 as a function of the square root of the ionic strength of the solution, $I_c^{1/2}$.¹⁶⁵ The upper curve is the behavior of the ideal electrophoretic mobility calculated with Equation 1-70; the points (x) are the observed mobilities. The line through the points is the behavior of the electrophoretic mobility calculated with Equation 1-79. Reprinted with permission from ref 165. Copyright 1940 Royal Society of Chemistry.

case, one layer of the ionic double layer is the layer of fixed charges on or near the surface of a molecule of the protein that produces its mean net charge, and the other layer is the layer of solution surrounding that molecule of the protein. This outer layer of the double layer is enriched in counterions opposite in sign to the net charge on the protein and depleted in co-ions of like sign. A region of solution large enough to contain the entire ionic double layer must be electrically neutral, and consequently, the outer layer of the ionic double layer must have a net charge, which is the difference between the charge carried by the counterions and the charge carried by the co-ions, equal in magnitude but opposite in sign to that of the protein. The distribution of that charge within the outer layer of the ionic double layer is a function of the ionic strength of the solution.

If a molecule of protein i were the sphere of its Stokes radius a_i (centimeters), if that sphere had a uniform density of elementary charges on its surface producing a mean net molecular charge number, \bar{Z}_i , and if that sphere were dissolved in a solution of monovalent electrolyte the positive and negative ions of which were both spherical and both had the same radius a_j , then the

radial distribution of electrostatic potential, $\phi(r)$, in volts, through the outer layer of the ionic double layer would be approximated by

$$[\phi(r)]_{r > (a_i + a_j)} = -\frac{e_a}{\epsilon_r r} \left\{ \frac{\bar{Z}_i \exp [\kappa(a_i + a_j - r)]}{1 + \kappa(a_i + a_j)} \right\} \quad (1-71)$$

where r is the distance (centimeters) from the center of the sphere, ϵ_r is the relative permittivity* of the solvent (dimensionless), and e_a is the elementary charge. The term $(a_i + a_j)$ takes account of the fact that an ion cannot approach the sphere of charge closer than its finite radius a_j permits.

By Coulomb's law

$$1 \text{ coulomb}^2 = 10^{-2} c^2 \text{ gram centimeter} = 8.928 \times 10^{18} \text{ gram centimeter}^3 \text{ second}^{-2} \quad (1-72)$$

where c is the speed of light in a vacuum. Consequently,

$$1 \frac{\text{coulomb}}{\text{centimeter}} = 8.298 \times 10^{11} \text{ V} \quad (1-73)$$

The parameter κ (centimeters⁻¹) in Equation 1-71 is defined by the relationship

$$\kappa^2 \equiv \frac{4\pi e_a^2}{\epsilon_r k_B T} \sum_{j=1}^m n_j z_j^2 \quad (1-74)$$

where z_j is the charge number of the ion of species j composing the electrolyte and n_j is the number of ions j for each cubic centimeter of the solution. The units on e_a^2 can be directly converted (Equation 1-72) from coulombs² to gram centimeter³ second⁻².

The term $\sum n_j z_j^2$ is related to the **ionic strength**, I_c (moles liter⁻¹), defined as

$$I_c \equiv \frac{1}{2} \sum_{j=1}^m [J] z_j^2 \quad (1-75)$$

where $[J]$ is the molar concentration (moles liter⁻¹) of the ion of species j , by the relationship

$$\sum_{j=1}^m n_j z_j^2 = 2I_c N_A \quad (1-76)$$

where N_A is Avogadro's number, and

$$\kappa^2 = \frac{8\pi e_a^2 N_A}{\epsilon_r k_B T} I_c \quad (1-77)$$

The term within the brackets on the right side of Equation 1-71 can be considered to be the **effective charge number** of the sphere, and this effective charge number is a function of the distance r from its center. If there were no electrolytes in the solution so that $\kappa = 0$, the potential would decrease radially only as the inverse of the distance, r , as expected for a sphere of charge in a medium of uniform relative permittivity, ϵ_r , and the full charge number, \bar{Z}_i , would contribute to the potential at all values of r . If κ does not equal zero, however, the effective charge number decreases as r increases due to the presence of the outer layer of the ionic double layer. Because the term κ has the dimensions of centimeters⁻¹, its inverse, κ^{-1} , is used as a measure of the **thickness of the double layer**.¹⁴ Equations 1-71 and 1-77 define the dimensions of the double layer and state that the thickness of the double layer will decrease as ionic strength is increased. As the thickness of the ionic double layer decreases, the layer of counterions tightens around the molecule of protein.

Consequently, there are two distributions of charge that respond to the electric field, the one on the molecule of protein and its bound ions with a total charge number of \bar{Z}_i and the one within the outer layer of the ionic double layer the distribution of whose charge is defined by Equation 1-71 and the total charge number of which is $-\bar{Z}_i$. The protein is drawn in one direction by the electric field; the outer layer of double layer, in the other. The effect of this electrostatic force on the outer layer of solution surrounding the protein applied in a direction opposite to that on the molecule of protein (Equation 1-70) causes the outer layer to move in a direction opposite to that in which the protein is caused to move. The consequence of this retrograde movement is to increase the drag of the solution on each molecule of the protein, and this effect can be described in terms of an increase in the effective frictional coefficient of each molecule.¹⁶⁶ As the outer layer moves one way and the molecule of protein, the reason for the existence of the double layer in the first place, moves in the other, the outer layer continuously dissolves behind the molecule of protein and re-forms around it in front so that the movement of the protein is continuously impeded. Because the thickness of the outer layer of the double layer decreases as the ionic strength of the solution increases, its velocity in the direction opposite to that of the protein increases as the ionic strength increases. The result is an increase in its drag on the molecule of protein, and thereby a decrease

* The relative permittivity or **dielectric constant** of a substance is its permittivity relative to the permittivity of the vacuum.

in the electrophoretic mobility of the protein as the ionic strength increases.*

On the basis of these assumptions, an equation has been derived^{165,167-169} to describe the electrophoretic mobility of protein i if its shape is approximately that of a sphere:

$$u_i^\circ = \frac{e_a \bar{Z}_i}{f_i} \left(\frac{1 + \kappa a_j}{1 + \kappa a_j + \kappa a_i} \right) f(\kappa a_i) \quad (1-78)$$

where $f(\kappa a_i)$, Henry's function, is a function in κa_i for which there is no exact expression¹⁶⁸ but which can be expressed graphically (Figure 1-14).¹⁶⁰ The value of this function varies between 1.0 and 1.5. It can be seen that when $\kappa a_j < 1$, as is usually the case for a solution of protein,

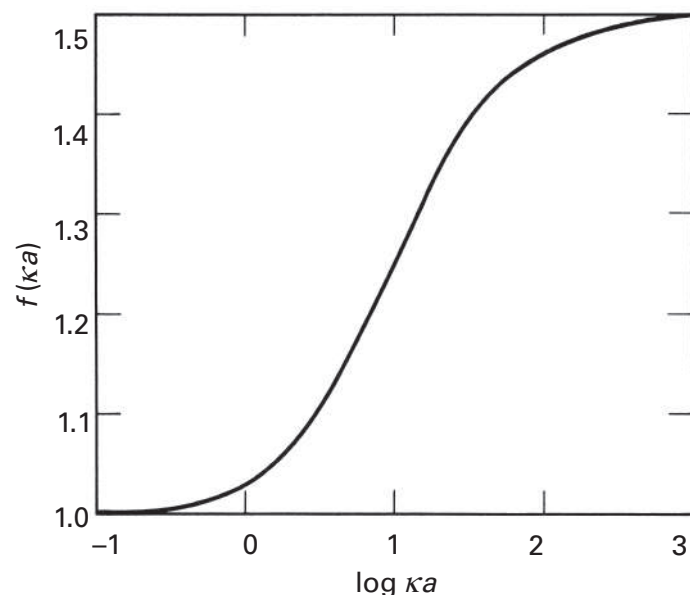


Figure 1-14: Graphic presentation of Henry's function.¹⁶⁰ Reprinted with permission from ref 160. Copyright 1961 John Wiley.

* There is another way to describe the effect of the double layer on electrophoresis. An assumption is made that there exists a discrete surface of shear that defines a boundary between the stationary solution and both the moving molecule of protein and the solution moving with it. The charge within the surface of shear, Q , which includes both the charge of protein i , $e_a \bar{Z}_i$, and the sum of the charges of the enclosed mobile ions, $e_a \bar{Z}_j$, creates a potential at that surface. The potential at the surface of shear is the zeta potential, ζ . If this assumption were a realistic one, then

$$u_i^\circ = \frac{D_i \zeta}{4\pi\eta}$$

The problem is that the relationship between \bar{Z}_i and ζ is a complex one. It is possible to calculate ζ for a molecule of protein from its electrophoretic mobility, but ζ provides little information about that molecule of protein because it includes the potential resulting from both the mobile ions and the molecule of protein itself.

$$u_i^\circ \cong \frac{e_a \bar{Z}_i}{f_i} \left(\frac{1}{1 + \kappa a_i} \right) f(\kappa a_i) \quad (1-79)$$

This equation predicts that the electrophoretic mobility will decrease as the ionic strength increases (Figure 1-13) because κ increases as the ionic strength increases (Equation 1-77).

The points in Figure 1-13 are the observed electrophoretic mobilities of the protein ovalbumin at various ionic strengths as measured by Tiselius and Svensson.¹⁶⁵ The top line is their calculation of the mobilities with Equation 1-70 by use of independent measurements of $\bar{Z}_{\text{ovalbumin}}$ and $f_{\text{ovalbumin}}$. The lower line is their calculation of the mobilities with Equation 1-79. The agreement between calculated values and observed values is surprisingly satisfactory. As the authors point out, the calculated value from Equation 1-70, in the absence of electrolyte, comes close to the extrapolated value of the actual mobilities.

According to Equation 1-78, at a constant ionic strength, the electrophoretic mobility of protein i should be directly proportional to \bar{Z}_i , and this proportionality is reflected in the direct proportionality that obtains between $\bar{Z}_{H,i}$ and u_i° (Figure 1-15)¹⁷⁰ as $\bar{Z}_{H,i}$ is varied by varying the pH at a constant ionic strength.¹⁷⁰ In fact, it is possible to display the titration curve of a protein visually by submitting a sample to electrophoresis on a two-dimensional electrophoretic field prepared so that there is a linear gradient of pH across the field perpendicular to the direction of electrophoresis.¹⁷¹ The absolute values of the electrophoretic mobilities of several proteins have been calculated from experimental values of their mean net proton charge numbers, $\bar{Z}_{H,i}$, by use of Equation 1-78 with the assumption that $\bar{Z}_{H,i} = \bar{Z}_i$ or a more complicated equation derived from a cylindrical model rather than a spherical one. The agreement between calculated values of u_i and experimental values of u_i was within a factor of 2 or less.¹⁶⁹

The lack of exact agreement between calculated and experimental values was assumed to be due to the difference between $\bar{Z}_{H,i}$ and \bar{Z}_i caused by the binding of inorganic ions to the proteins. In this case, the proportionality between $\bar{Z}_{H,i}$ and u_i observed in Figure 1-15 could still be explained, if the binding of counterions increases proportionately as $\bar{Z}_{H,i}$ increases in magnitude.¹⁵³ It has been demonstrated, however, by systematically varying the charge number on human carbonate dehydratase II that electrophoretic mobility is not directly proportional to charge number when charge number becomes large.¹⁷² This deviation from the behavior predicted by Equation 1-79 at high levels of charge number could be accounted for by using the nonlinear Poisson-Boltzmann equation rather than the linear form used in the derivation of Equation 1-79 and by using a correction for ion relaxation and polarization. The latter correction accounts for the local electric field arising

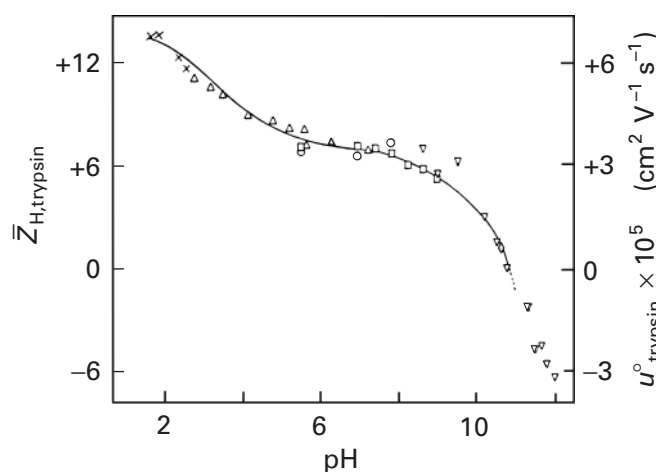


Figure 1-15: Comparison of the electrophoretic mobilities of trypsin ($\text{cm}^2 \text{ volt}^{-1} \text{ second}^{-1}$) at 0°C (u_{trypsin}^o ; points) with the acid-base titration curve of trypsin determined at 20°C ($Z_{\text{H, trypsin}}$, continuous curve).¹⁷⁰ The respective scales on the two vertical axes, those for electrophoretic mobility and mean net proton charge number, respectively, both with respect to pH, were adjusted to produce maximum coincidence. The value for $Z_{\text{H, trypsin}} = 0$ was arbitrarily set to coincide with the isoelectric point. The coincidence displayed is in shape rather than absolute value or excursion. The different symbols denote the different buffers used to maintain the pH during electrophoresis: (\times) Na^+ , H^+ , Cl^- ; (Δ) Na^+ , H^+ , acetate $^-$, Cl^- ; (\square) Ca^{2+} , H^+ , barbiturate $^-$, Cl^- ; (\circ) Mg^{2+} , H^+ , barbiturate $^-$, Cl^- ; (∇) Ca^{2+} , H^+ , glycinate $^-$, Cl^- ; (\diamond) Ca^{2+} , H^+ , NH_3 , Cl^- . The ionic strength was maintained at 0.13 M. Reprinted with permission from ref 170. Copyright 1952 Academic Press.

from the distortion of the outer layer of the double layer caused by its movement in a direction opposite to that of the protein and its inability to dissolve behind it and reform around it instantaneously.

At its isoelectric point, pI_i , the electrophoretic mobility of protein i becomes zero (Equation 1-78), and this fact permits the isoelectric point of a protein to be measured by electrophoresis.¹⁷³ Electrophoretic mobilities are measured at values of pH greater than and less than pI_i , and the pH of zero mobility is determined by interpolation (Figure 1-15).

The effect of ionic strength on the isoelectric point of a protein in the absence of actual binding of the ions in the electrolyte to the protein has been calculated to be smaller than the experimental error in measurement.¹⁵⁰ Nevertheless, significant variations in isoelectric point with ionic strength are generally observed (Figure 1-16),¹⁷⁴ and these depend on the particular neutral salt chosen to adjust the ionic strength. The explanation for this behavior can only be the **preferential binding** of particular ions—in Figure 1-16, always that of the anions—in the chosen electrolyte. The net binding of ions can be calculated from the observed changes in the isoelectric point because, from Equation 1-60, when $Z_i = 0$

$$Z_{\text{H},i} = -\sum_{j=1}^m \bar{v}_j z_j \quad (1-80)$$

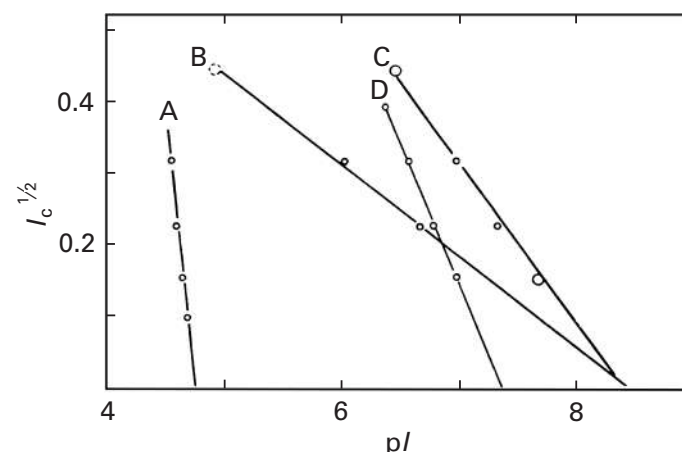


Figure 1-16: Variations in the electrophoretic isoelectric points (pI) of a protein as a function of the square root of the ionic strength ($I_c^{1/2}$).¹⁷⁴ Line A, ovalbumin in acetate; line B, fructose-bisphosphate aldolase in phosphate; line C, fructose-bisphosphate aldolase in acetate; line D, carboxyhemoglobin in phosphate. Reprinted with permission from ref 174. Copyright 1949 American Chemical Society.

and $Z_{\text{H},i}$ is available from titration data (Figure 1-11).

To this point, only the free electrophoretic mobility of a protein, u_i^o , has been discussed. The free electrophoretic mobility is the electrophoretic mobility displayed by a protein in free solution. This property of the protein is measured by moving boundary electrophoresis¹⁷⁵ in an apparatus developed by Tiselius.¹⁷⁶ This technique has been supplanted by electrophoresis in continuous **gels of cross-linked polyacrylamide**. A gel of cross-linked polyacrylamide is a hydrated plastic cast in a mold from a solution of acrylamide and the cross-linker N,N' -methylenebis(acrylamide) along with a buffer and other salts. The total concentration of acrylamide and N,N' -methylenebis(acrylamide) in the final gel can be varied from 3% to 20%.

It has been demonstrated experimentally by Morris¹⁷⁷ that the relative electrophoretic mobilities of proteins in polyacrylamide gels vary regularly with the concentration of acrylamide used to cast the gel (Figure 1-17)¹⁷⁷ and

$$u_i = u_i^o \exp(-K_{r,i} T_a) \quad (1-81)$$

where u_i is the electrophoretic mobility of protein i observed on a gel cast from a solution whose total concentration of acrylamide, in percent, was T_a and $K_{r,i}$ is a **retardation coefficient** unique to protein i . Such behavior was first noted by Ferguson¹⁷⁸ on gels cast from starch in which the same equation applies (Equation 1-81), but the concentration is T_s , the concentration of the starch.¹⁷⁸

According to Equation 1-81, u_i^o should be the free electrophoretic mobility of protein i , and this has been shown to be the case.¹⁷⁷ It follows that

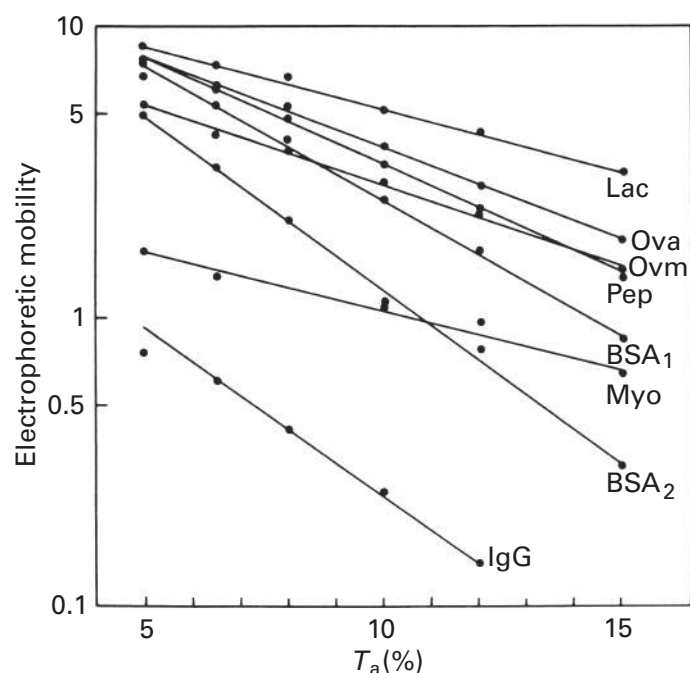


Figure 1-17: Electrophoretic mobility, presented on a logarithmic scale, of various proteins on gels of various concentrations of polyacrylamide (T_a).¹⁷⁷ The gels were cast from solutions of pH 8.88 with total concentrations of acrylamide plus N,N' -methylenebis(acrylamide) (T_a) varying between 5% and 15%. The concentration of N,N' -methylenebis(acrylamide) was always 20-fold less than the concentration of acrylamide. The proteins were β -lactoglobulin (Lac), ovalbumin (Ova), ovomucoid (Ovm), pepsin (Pep), bovine serum albumin monomer (BSA_1) and dimer (BSA_2), myoglobin (Myo), and immunoglobulin G (IgG). Reprinted with permission from ref 177. Copyright 1966 Elsevier.

$$u_i \cong \frac{e_a \bar{Z}_i}{f_i} \left[\frac{f(\kappa a_i)}{1 + \kappa a_i} \right] \exp(-K_{r,i} T_a) \quad (1-82)$$

Examination of this relationship reveals that the electrophoretic mobilities of the proteins in a complex mixture upon a gel of polyacrylamide are directly proportional to their respective charges, which are determined by complex functions of pH (Figure 1-11); are complex functions of their respective frictional coefficients, which are determined by their sizes and shapes; and are exponentially proportional to the product of a constant, which is unique for each, and the concentration of acrylamide. At a given pH, ionic strength, and concentration of polyacrylamide, each of the proteins in this mixture will have a characteristic electrophoretic mobility (Figure 1-17) and they can be separated one from the other. In this way, electrophoresis can provide a catalogue of the number of proteins present in the mixture and the relative amounts of each.

Electrophoresis of native proteins* is the most reliable method available for assessing the homogeneity of a sample of purified protein. A sample of pure protein in its native state should display only one component upon gel electrophoresis. Because the electrophoretic mobilities of two proteins change disproportionately as either the pH (Figure 1-15) or the concentration of acrylamide (Figure 1-17) is changed, the possibility that the single component observed under one set of conditions results from the accidental coelectrophoresis of two or more proteins can be dismissed by running electrophoresis at several values of pH^{179,180} or several concentrations of acrylamide.¹⁷⁷

If they are to be used in the roles of cataloguing mixtures and establishing purity,¹⁸¹ electrophoretic separations of native proteins on polyacrylamide gels must have as high a resolution as possible. This high resolution is achieved by using a discontinuous buffer system and performing what has been referred to as **disc electrophoresis**,^{182,183} the pun apparently intended. This technique relies upon the creation of three **stable moving boundaries** (Figure 1-18).¹⁸² Each of the three is formed between two solutions of different ionic composition. It is the applied electric field that causes the boundaries to move.

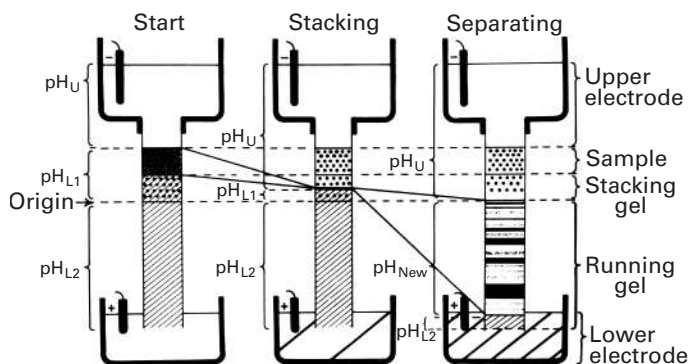


Figure 1-18: Disc electrophoresis.¹⁸² At the start (left) the proteins in the original sample (black rectangle) are in a large volume and at a low pH (pH_{L1}). They are compressed to a small volume, or disc, as they move through the stacking gel by being trapped in the stable boundary between the upper solution (pH_U , buffer) and the solution of the original sample and the spacer (pH_{L1}). (Middle) Upon fusion of this descending boundary and the stable ascending boundary between the solution originally in the running gel (pH_{L2}) and the solution originally in the stacking gel (pH_{L1}), the pH at the boundary increases and the new more rapidly moving boundary outruns the proteins and deposits a newly created solution of higher pH (pH_{new}) behind it as it moves ahead of the separating proteins (right). The proteins also escape the first boundary because, at about the same time as the jump in pH at the fusion of the descending and ascending boundaries, they encounter the running gel, which has a higher percentage of acrylamide and which decreases their mobility. Reprinted with permission from ref 182. Copyright 1964 New York Academy of Sciences.

* The electrophoresis of proteins unfolded in solutions of dodecyl sulfate is quite different and will be discussed later.

The first of these boundaries is used to trap the proteins and sweep them into an extremely narrow band prior to the electrophoretic separation. This process has been called **stacking**. It significantly improves the resolution of the subsequent separation by shrinking the original sample to a hairline so that all of the molecules of protein begin the electrophoretic separation at nearly the same point. The stacking occurs because the proteins are initially placed as a sample that is sandwiched between an upper solution and a lower solution. The upper solution is simply poured on top of the sample, but the lower solution is in a polyacrylamide gel, so that convective turbulence does not disrupt the stable moving boundaries, but it is a gel of a low concentration of polyacrylamide, so that the mobilities of the proteins are as high as possible. This gel of high porosity is the **stacking gel**. The solution in which the protein is dissolved has the same composition as the lower solution.

The upper and lower solutions are prepared so their respective ionic compositions will form a stable moving boundary of a particular type. Although systems for cationic proteins are also available, to describe this boundary, it will be assumed that the direction of electrophoretic movement of both the proteins and this first stable moving boundary is downward and a pH has been chosen such that the proteins are all anionic. In this case, both the upper solution and the lower solution above and below the boundary, respectively, are prepared from salts of the same cationic weak acid [for example, tris(hydroxymethyl)methylammonium ion]. An anion (for example, glycinate ion) the mobility of which is less than the mobilities of all the proteins is used to make the upper solution, and an anion (for example, chloride ion) the mobility of which is greater than the mobilities of all the proteins is used to make the lower solution. The stable descending boundary formed is one between these two anions. If a molecule of one of the proteins finds itself in the lower solution, it is surrounded by anions that are moving faster than it is, and it is overtaken by the boundary. If a molecule of one of the proteins finds itself in the upper solution, it is surrounded by anions that are moving more slowly than it is, and it outstrips them and returns to the boundary. The result of these events is that the proteins all gather within the descending boundary itself, which remains extremely sharp and stable if the upper and lower solutions have the proper ionic compositions.¹⁸⁴

The stacking process is able to compress the proteins to thin lamella, but in order for electrophoretic separation to occur, they must be **released from the boundary** after they have been stacked. This can be done if the upper solution of this initial descending boundary has been made with an anion, α , that is slower than the protein only because it is the conjugate anionic base (for example, glycinate ion) of a weak neutral acid (glycine, $pK_a = 9.6$) and the pH of the upper solution has been chosen to be significantly lower than the pK_a of that weak

acid. Under these conditions, the anion α is slow because only a fraction of the weak acid is anionic at any instant. The acid-base equilibrium has the effect of decreasing the mobility of the anion α from its value in the absence of its conjugate acid to a lower value, and

$$u_{\alpha} = u_{\alpha}^{\circ} f_{\alpha} \quad (1-83)$$

where u_{α} is the mobility of the upper anion at the actual ratio of conjugate base to acid in the solution, u_{α}° is its mobility in the absence of its conjugate acid, and f_{α} is the fraction of the total weak acid that is ionized at the ratio chosen. In such a situation, the proteins can be released from the descending boundary by abruptly increasing the pH, and hence the value of f_{α} , so that u_{α} becomes greater than the mobilities of the proteins, and the new stable, but now rapidly descending, boundary that results drops the proteins behind at the origin of the electrophoretic separation.

The abrupt increase in the pH of the upper solution of the descending boundary can be accomplished by the arrival of a stable ascending boundary between two concentrations of the same cationic buffer. Behind this ascending boundary is a solution of the same cationic weak acid as used for the upper and lower solutions of the initial descending boundary [for example, tris(hydroxymethyl)methylammonium ion] but at a higher concentration and a higher pH than that of the solution behind the initial descending boundary. This ascending boundary between different concentrations of the same cationic buffer has been constructed so that the anion in both its upper and lower solutions (for example, chloride) is the same. Because its upper solution is by definition the lower solution of the initial descending boundary, this anion is already the fast anion of that boundary. The cationic weak acid of the upper solution of the ascending boundary is by definition the cationic weak acid in the two solutions used to make the initial descending boundary. The concentration and pH of the cationic buffer in the lower solution of the ascending boundary, however, is chosen to be high enough to adjust the final pH behind the new descending boundary to a value high enough to release the proteins from the initial descending boundary. If the release is unsuccessful, or only partially successful, the proteins, or some of the proteins, remain trapped in the new descending boundary and are never separated. These trapped unseparated proteins form an extremely sharp but uninformative and deceptive band at the bottom of the final electrophoretogram.¹⁸⁵

The release of the proteins from the initial descending boundary in which they were trapped and stacked can be accomplished even more effectively by using a stable ascending boundary behind which is a solution of a cationic weak acid of a higher pK_a than the cationic weak acid used as the counterion in the initial descending boundary.¹⁸⁶ This ascending boundary has been constructed so that the anion in both its upper and lower solu-

tions is the same. Because its upper solution is the lower solution of the initial descending boundary, this anion is the fast anion of the upper boundary (for example, chloride ion). The cationic weak acid of the upper solution of the ascending boundary must be the cationic weak acid of the two solutions used to make the initial descending boundary (for example, pyridinium ion; $pK_a = 5.14$). The cation of the lower solution of the ascending boundary is chosen to be the cationic conjugate acid [for example, tris(hydroxymethyl)methylammonium ion; $pK_a = 8.10$] of a neutral base strong enough to adjust the final pH behind the new descending boundary to a value higher than the pK_a of the neutral conjugate acid of the anion α (for example, 4-morpholineethanesulfonate ion; $pK_a = 6.15$) and release the proteins from the initial descending boundary. The difficulty with this strategy is that the pH of the upper solution of the initial descending boundary is often so low that the proteins are no longer anionic and move upward instead of downward. But it is effective with complexes of protein and dodecyl sulfate because they are anionic at all reasonable values of pH.

To ensure that as many proteins as possible are released from the initial descending boundary, shortly after the fusion of the ascending boundary and the initial descending boundary, the descending band of proteins in the stacking gel encounters a much higher concentration of polyacrylamide, **the running gel**, which decreases the mobilities of all of the proteins by virtue of the relationship in Equation 1–81. This frictional deceleration of the proteins increases the probability that all of their mobilities will be less than that of the now accelerated anion of the upper solution so that they can escape from the new descending boundary.

The polyacrylamide gel is poured in two stages (Figure 1–18): the running gel, the polyacrylamide concentration of which is high and upon which the separation will occur, and the stacking gel, the polyacrylamide concentration of which is as low as possible to keep the mobilities of the proteins as high as possible and in which the stacking will occur.

Three stable moving boundaries must be constructed (Figure 1–18). At the start of the electrophoresis, the initial descending boundary between the slow anion and the fast anion that will compress the proteins is the boundary between the upper electrode solution (pH_U) and the solution in the sample and the stacking gel (pH_{L1}). At the start of the electrophoresis, the ascending boundary between the two concentrations of the cationic conjugate acid of the weak base or between the cationic conjugate acids of the weaker base and the stronger base that will deliver the pH jump is the boundary between the solutions in the running gel (pH_{L2}) and the stacking gel (pH_{L1}). The third stable moving boundary is the new descending boundary that deposits behind it the solution in which the proteins are actually separated (pH_{new}). It forms upon the fusion of the other two.

As the initial descending boundary moves through

the stacking gel, it must maintain a constant pH and ionic strength behind it (pH_U) to maintain the low and constant mobility of the slow anion in the upper solution. As the new descending boundary moves, it must deposit behind itself a solution of constant pH and ionic composition (pH_{new}) to form a uniform electrophoretic field upon which the proteins can be separated. The pH and ionic composition of the solution that is deposited behind the new descending boundary is different from the pH and ionic composition of any of the solutions initially present, but the cation in this newly created solution is the weak cationic acid of the original lower phase of the ascending boundary and the anion in this solution is the now accelerated anion of the original upper phase of the initial descending boundary. The constant pH deposited behind this new descending boundary is established by the weak cationic acid found on both sides of the boundary and its conjugate base and the now accelerated slow anion found in the upper solution of the boundary, which is a weak anionic base, and its conjugate acid. All four of these species together buffer the deposited solution and determine both the ionic strength and the value of the deposited pH and hence the pH of the actual electrophoresis.

The equations that govern the creation of a stable moving boundary and the ability of that boundary to deposit a solution of uniform pH and ionic composition were derived by Ornstein¹⁸² from the regulating functions described by Kohlrausch.¹⁸⁴ On the basis of these equations, Jovin¹⁸⁷ has developed a more elaborate theoretical description of discontinuous electrophoresis, and he and his colleagues have provided the necessary recipes for a large number of discontinuous systems.¹⁸⁸

Suggested Reading

- Tiselius, A., & Svensson, H. (1940) The influence of electrolyte concentration on the electrophoretic mobility of egg albumin, *Trans. Faraday Soc.* 36, 16–22.
- Carbeck, J.D., & Negin, R.S. (2001) Measuring the size and charge of proteins using protein charge ladders, capillary electrophoresis, and electrokinetic models of colloids, *J. Am. Chem. Soc.* 123, 1252–1253.

Problem 1–17: The uptake of protons by 1 mol of horse carboxyhemoglobin in the range of pH 6–8 is about 9 mol of protons for each drop of 1 unit in pH.¹⁸⁹ Use this value to estimate the moles of phosphate bound by a mole of horse carboxyhemoglobin at its isoelectric point at the phosphate concentration of the last point in curve D of Figure 1–16 ($[phosphate] = 0.12\text{ M}$). Assume that no cations other than protons are binding to the protein under these conditions.

Problem 1–18: Use interpolated values for the free electrophoretic mobility of ovalbumin (Figure 1–13) at ionic strengths of 0.0025, 0.01, and 0.16 M to calculate the charge number on the protein during the elec-

trophoresis. The pH for the measurements was 7.1, and the temperature was 294 K. The viscosity of water at 294 K is 1.0 mPa s. The diffusion coefficient of ovalbumin at 294 K is $4.2 \times 10^{-7} \text{ cm}^2 \text{ s}^{-1}$.

Problem 1-19: The isoelectric point of normal hemoglobin, hemoglobin A, is 6.87, and that of sickle hemoglobin, hemoglobin S, is 7.09 when electrophoresis is carried out under the same conditions.¹⁷³ In the vicinity of the isoelectric point, the charge number on either of these hemoglobins changes by about 13 equiv for every mole of protein for every change of 1 unit in pH. At the same pH, anywhere between their two respective isoelectric points, what is the difference in charge number between hemoglobin A and hemoglobin S?

Problem 1-20: The frictional coefficient of trypsin at 10 °C is $5.5 \times 10^{-8} \text{ g s}^{-1}$. Assume the molecule to be a sphere and calculate its free electrophoretic mobility at 10 °C and at pH 6 and $I_c = 0.13 \text{ M}$ by using the results of the acid-base titration in Figure 1-15, which are for 20 °C, and Equation 1-79. Assume that $\bar{Z}_{\text{trypsin}} = \bar{Z}_{\text{H,trypsin}}$ and that $\bar{Z}_{\text{H,trypsin}}$ at pH 6 is the same at 10 °C as at 20 °C.

Problem 1-21: The frictional coefficient of ribonuclease at 25 °C is $2.6 \times 10^{-8} \text{ g s}^{-1}$. Assume the molecule to be a sphere and calculate its free electrophoretic mobility at pH 6 and $[\text{KCl}] = 0.15 \text{ M}$ by using the results presented in Figure 1-11 and Equation 1-79 with the assumption that $\bar{Z}_{\text{RNase}} = \bar{Z}_{\text{H,RNase}}$. In a field of 20 V cm^{-1} , how far would ribonuclease travel in 3 h if it had this mobility?

Problem 1-22: The mean net proton charge number on bovine serum albumin (BSA) at pH 8.0 and ionic strength 0.15 M is -17 .¹⁵⁹ The diffusion coefficient of bovine serum albumin at 20 °C is $6.0 \times 10^{-7} \text{ cm}^2 \text{ s}^{-1}$. Its retardation coefficient (K_r) on polyacrylamide gels is $0.16 (\%)^{-1}$. The viscosity of water at 20 °C is 1.0002 mPa s.

- Show that the ionic strength of a solution containing 0.10 M sodium phosphate at pH 8.0 is 0.27 M if the three values of the pK_a for phosphate are 2.12, 7.21, and 12.32.
- Estimate the mobility (u_{BSA}) of bovine serum albumin on a 7.5% gel of polyacrylamide run at 20 °C in 0.10 M sodium phosphate at pH 8.0. Assume for now that $\bar{Z}_{\text{H,BSA}}$ equals \bar{Z}_{BSA} at pH 8.0.
- How far would the bovine serum albumin move in 2 h if the field across the gel was 50 V cm^{-1} ?
- Refer to Figure 1-16. On the basis of the behavior displayed in this figure, would you expect the bovine serum albumin to have a larger or a smaller mobility than you estimated? Why?

Problem 1-23: The table below contains information about five imaginary proteins where a is the Stokes'

radius, \bar{Z} is the mean net charge number on the protein at pH 7, $(\partial \bar{Z}_{\text{H},i} / \partial \text{pH})_{I_c}$ is the change in mean net proton charge number with pH, K_r is the retardation coefficient for polyacrylamide, and u° is the free electrophoretic mobility for a temperature of 25 °C, an ionic strength of 0.1 M, and a pH of 7.0.

- Assume that, at constant ionic strength, $(\partial \bar{Z}_{\text{H},i} / \partial \text{pH})_{I_c}$ is equivalent to $(\partial \bar{Z}_i / \partial \text{pH})_{I_c}$ for each of the five proteins, and calculate the electrophoretic mobilities of these five proteins, at 25 °C and an ionic strength of 0.1 M, under each of the following conditions: (1) pH 7.0 on 5% polyacrylamide; (2) pH 7.0 on 10% polyacrylamide; (3) pH 5.0 on 5% polyacrylamide; (4) pH 5.0 on 10% polyacrylamide.
- What is the order of the migration of these five proteins under each of the four conditions?
- What will happen to protein E at pH 7.0 that would not happen at pH 5.0 if a mixture of the proteins is run on vertical polyacrylamide gels with the cathode at the bottom and the anode at the top?
- Assume that \bar{Z}_i does not change as ionic strength changes and calculate the mobilities of the five proteins at an ionic strength of 0.2 M at pH 5 and at 25 °C on 5% polyacrylamide. How does the increase in ionic strength affect the mobilities?

protein	a (nm)	\bar{Z} (pH 7)	$\left(\frac{\partial \bar{Z}_{\text{H}}}{\partial \text{pH}} \right)_{I_c}$	K_r ($\%^{-1}$)	u° $\left(\frac{\text{cm}^2}{\text{Vs}} \right)$
A	2.4	+1.4	-0.2	0.045	1.7×10^{-5}
B	5.3	+9.8	-1.8	0.152	3.2×10^{-5}
C	4.9	+6.2	-2.7	0.146	2.3×10^{-5}
D	2.6	+0.9	-0.5	0.048	1.0×10^{-5}
E	3.4	-3.4	-2.3	0.073	-2.4×10^{-5}

Criteria of Purity

When the purification of a particular protein is monitored analytically by disc electrophoresis (Figure 1-19),¹⁹⁰ the array of other proteins present at the early states of the purification is seen gradually to become less complex in the later stages as one component emerges from the background and becomes more prominent until it alone remains.^{190,191} To be certain that the single component observed at the last step of the purification is the only one present in the purified preparation, electrophoresis should be run at a variety of protein concentrations in addition to a few different acrylamide concentrations and values of pH.¹⁹² At high concentrations of protein, minor impurities are most easily recognized, while at low concentrations, two closely running

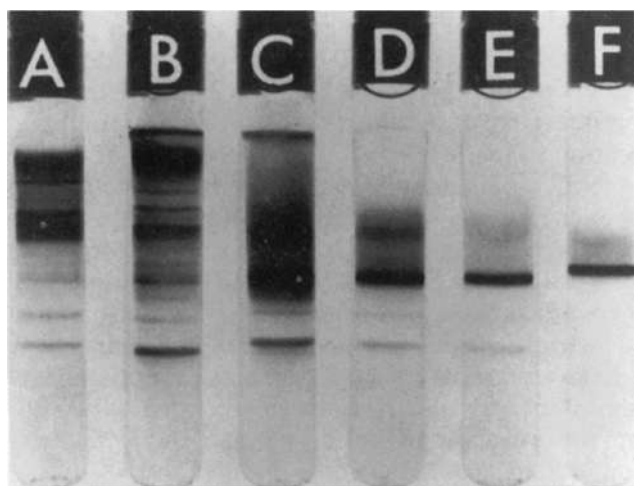


Figure 1-19: Disc electrophoresis on gels of polyacrylamide of native proteins from successive steps in the purification of [acyl-carrier-protein] S-malonyltransferase from *E. coli*.¹⁹⁰ Electrophoresis was performed on polyacrylamide gels cast from 15% solutions of acrylamide in a discontinuous system of tris(hydroxymethyl)methylamine and glycylglycine. The different gels represent samples from successive steps in a complete purification of the enzyme, seen in its final purified state on gel F. The gels were stained for protein with Coomassie brilliant blue. Reprinted with permission from ref 190. Copyright 1973 *Journal of Biological Chemistry*.

components can be resolved. Also, by running polyacrylamide gels loaded with a series of protein concentrations, the number and relative amounts of any minor impurities can be quantified.¹⁹³ The polyacrylamide gels should also be stained with two distinct dyes, for example Coomassie brilliant blue and silver oxide,^{194,195} because some proteins do not stain so strongly as others with a particular dye.

The single component observed upon electrophoresis of a sample from the final step of the purification must be shown to be the protein actually responsible for the biological function being purified. Either the polyacrylamide gel is sliced and the assay is performed on each slice (Figure 1-20),^{51,97,131,196} or the intact polyacrylamide gel is **stained for enzymatic activity** (Figure 1-21).¹⁹ The latter is accomplished by placing the gel in a solution that promotes the incorporation of radioactivity¹⁹⁷ or that gives a fluorescent product or a colored product from the enzymatic reaction. For example, by adding lead acetate, the SeH_2 produced in a polyacrylamide gel from the action of selenocysteine lyase can be made to form a yellow band where the enzyme is located.⁶⁰ The most widely used stain for enzymatic activity is based on the ability of NADH to reduce *p*-nitro tetrazolium blue to give a blue color.^{198,199} It is obvious that through coupled assays this reaction can be used to visualize a large array of different enzymatic activities. At times, the protein being purified is itself col-

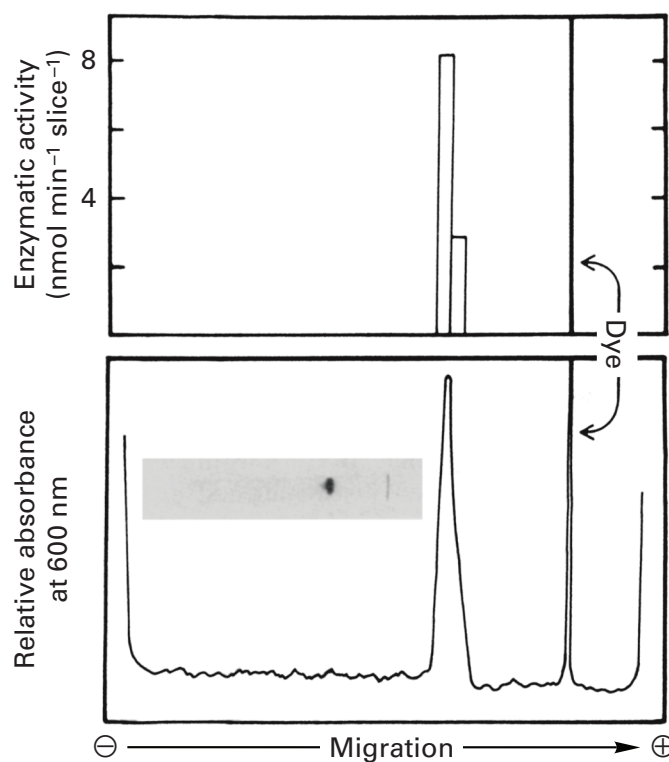


Figure 1-20: Electrophoresis of purified porcine phosphomevalonate kinase (20 μg) on a gel cast from a 10% solution of acrylamide.⁵¹ Following the electrophoresis, the cylindrical gel was divided in half longitudinally. One half was cut into slices laterally, and the slices were assayed individually for enzymatic activity (A). The other half was stained for protein and then scanned for the resulting absorbance (B). The inset in panel B is a photograph of the stained gel. Reprinted with permission from ref 51. Copyright 1980 American Chemical Society.

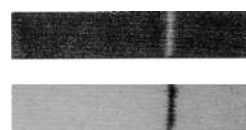


Figure 1-21: Staining a polyacrylamide gel for enzymatic activity.¹⁹ Two samples of purified isocitrate dehydrogenase (NADP^+) from the final step on phenyl agarose (Table 1-2) were submitted to electrophoresis on separate lanes of a thin slab of polyacrylamide. After the electrophoresis, the lanes were cut from the slab. One of the lanes (lower) was stained for protein with Coomassie brilliant blue. The other lane (upper) was placed in a solution of isocitrate and NADP^+ . The intrinsic fluorescence of the NADPH produced by the enzyme was observed by illuminating the gel with ultraviolet light. Reprinted with permission from ref 19. Copyright 1992 Blackwell Publishing.

ored, by virtue of a bound chromophore, such as the coenzyme B_{12} associated with D-lysine 5,6-aminomutase,²⁰⁰ and the coelectrophoresis of the purified protein and that color can be observed directly.

Several artifacts can produce misleading results on electrophoresis. For example, **aggregation** of individual molecules of the same protein can occur^{201,202} during

either the purification or the stacking process, and this produces an array of complexes, each with a different frictional coefficient and retardation coefficient. The neutral amides of glutamines and asparagines on the protein can hydrolyze randomly and in low yield during a harsh purification to produce anionic carboxylates, and this **modification** leads to variations in \bar{Z}_i that produce multiple components from the same protein. Because these or other similar modifications are integral processes, the components that result from them are usually evenly spaced upon the electrophoretogram,¹⁷² and the nature of the artifact can be recognized by this pattern.^{201,203,204} Each component, however, should be biologically active if the protein is pure.^{203,204}

Although the coelectrophoresis of the purified protein and the biological activity is the most convincing criterion of purity, occasionally the electrophoresis itself destroys the activity.²⁰⁵ For this reason, or simply for personal satisfaction, other criteria of purity are often used. Immunoglobulins raised against the purified enzyme should behave on **immunodiffusion** and **immuno-electrophoresis** as expected of immunoglobulins directed against a single antigen. It is also encouraging when these immunoglobulins are able to precipitate all of the protein and all of the biological activity^{19,206} but not essential, because some immunoglobulins are ineffective at immunoprecipitation. Activity and protein should **comigrate** on chromatography (Figures 1–6 and 1–10)²⁰⁷ or **cosediment** upon gradients of sucrose.²⁰⁸ Even more convincing is the observation that the single band of protein observed upon electrophoresis of samples from fractions collected from the final chromatographic step increases in intensity and then decreases in intensity in concert with the increase and decrease of enzymatic activity, respectively, across the peak.^{23,104}

The grams of protein for every mole of binding site is between 15,000 and 100,000 g mol⁻¹ for most proteins. The concentration of protein (milligrams milliliter⁻¹) and the concentration of **binding sites** (moles liter⁻¹) for a ligand, such as an agonist or antagonist, known to be specific for a desired protein, such as the respective receptor, can be determined on samples from the same solution. If the ratio of these two quantities lies within the expected range and if only one protein can be discerned on electrophoresis, these observations are taken to be convincing criteria of purity, especially if the value of grams mole⁻¹ agrees with the measured molar mass of the protomer of the protein that has been purified. For example, purified histidinol-phosphate transaminase binds 1 mol of pyridoxal phosphate for every 37,000 g of protein,¹⁹² purified methylmalonyl-CoA mutase contains 1 mol of adenosylcobalamine for every 73,000 g of protein,²⁰⁹ and purified α_1 -adrenergic receptor binds 1 mol of [³H]prazosin for every 69,000 g of protein.³⁴

Isoelectric focusing is a method for assessing purity that is based on electrophoresis. A gel of polyacrylamide is cast from a solution containing a mixture of

polyelectrolytes known as ampholytes. The isoelectric points of the ampholytes in the mixture vary over a continuous range of pH values. Upon application of an electric field, this mixture forms a stable gradient of pH in the gel. Each protein migrates through this gradient until it reaches a pH equal to its isoelectric point where it can no longer move, and the proteins in a mixture are spread upon the field in order of their respective isoelectric points. It is a technique that is less flexible than disc electrophoresis because it separates molecules on the basis of only one property rather than three. It also seems to be more sensitive to minor heterogeneities of charge than is electrophoresis. Because, however, isoelectric focusing detects heterogeneity of charge more successfully than electrophoresis, it is an even more stringent test of the homogeneity of a protein.²¹⁰ The coisoelectrofocusing of protein and biological activity,^{48,206,211–213} is an additional criterion of purity independent from the observation of coelectrophoresis. Isoelectric focusing has been combined with electrophoresis to resolve complex mixtures of proteins in two dimensions.²¹⁴ When the clarified homogenate produced from the cytoplasm of the bacterium *E. coli* was submitted to such a procedure, more than 1000 different proteins were represented upon the field (Figure 1–22).²¹⁴ This display indicates the complexity of the mixture of proteins in a cell. From such a mixture, a single protein with a single biological activity is purified.

Suggested Reading

Muro-Pastor, M. I., & Florencio, F. J. (1992) Purification and properties of NADP-isocitrate dehydrogenase from the unicellular cyanobacterium *Synechocystis* sp. PCC 6803, *Eur J. Biochem.* 203, 99–105.

Heterogeneity

Often heterogeneity in a preparation of a purified protein, observed as several different proteins capable of being separated, is detected by electrophoresis or isoelectric focusing even though all of the various components are biologically active; often heterogeneity is discovered in later experiments. This heterogeneity may have a biological origin, for example, because of varying levels of glycosylation or phosphorylation, and the various forms of the protein producing this heterogeneity may coexist in the tissue prior to homogenization, but usually the heterogeneity arises during the purification itself. Such heterogeneity is produced by processes that are minimized by avoiding extremes of pH through the use of well-buffered solutions, by working at low temperatures (0–5 °C), and by performing the purification in as short a period of time as possible.

That it is the purification itself producing the heterogeneity often becomes apparent when a new, more rapid, less debilitating method of purification is devised

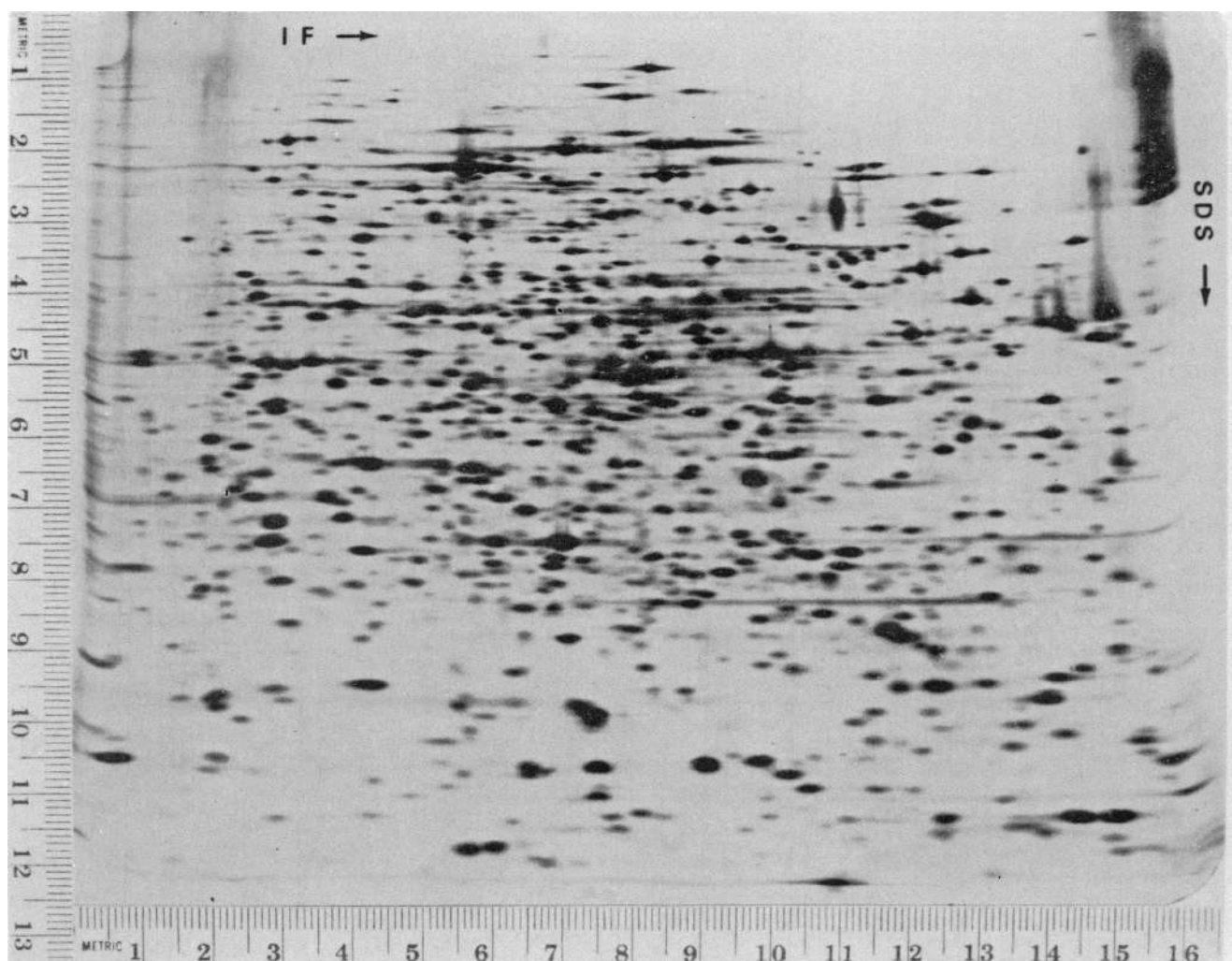


Figure 1-22: Separation of proteins from the cytoplasm of the bacterium *E. coli* by electrophoresis in two dimensions.²¹⁴ A sample (10 μ g of protein) from a homogenate of the bacteria, grown in the presence of [14 C]amino acids, was submitted to isoelectric focusing (pH 3–10), under conditions where the proteins were unfolded (9 M urea), on a cylindrical (0.25 cm \times 13 cm) gel of polyacrylamide. After the unfolded proteins had reached their respective isoelectric points, the gel was removed from its tube, soaked in a solution of sodium dodecyl sulfate (SDS) to coat the unfolded polypeptides with this anionic detergent, and the cylinder was laid across the top of a flat slab (14 cm \times 16 cm \times 0.3 mm). The unfolded polypeptides separated by isoelectric focusing (IF) in the first dimension were then separated by electrophoresis (SDS) in the second dimension. [14 C]Polypeptides were located by placing the slab on photographic film and exposing the film for a long enough time that the radioactive disintegrations in each spot of protein produced the dark spots seen in the figure. Reprinted with permission from ref 214. Copyright 1975 *Journal of Biological Chemistry*.

for a certain protein, and the heterogeneity noted previously, the subject of many publications, simply disappears. When fructose-bisphosphatase was purified by a shorter method,²¹⁵ the previously studied requirement of the enzyme for alkaline conditions was no longer manifest. When aconitate hydratase was purified by a more rapid procedure,²¹⁶ it was isolated with its iron still attached. When glyceraldehyde-3-phosphate dehydrogenase from yeast was purified rapidly by affinity chromatography,²¹⁷ the heterogeneous behavior in its binding of ligands^{218–220} was no longer observed.

One of the most publicized causes of heterogeneity or artifactual alteration of a protein during its purification is **digestion by peptidases**.^{146,221} Proteins the biological role of which is to degrade other proteins are

known as peptidases. With the exception of a few peptidases that are located in the cytoplasm such as the calpains, which can be inactivated by chelating any free calcium, most of the peptidases capable of degrading the normal, native proteins in a cell are present in inactive forms or are segregated from the cytoplasm of the cell in which they are located or in which they were produced. This segregation is accomplished by enclosing the peptidases in tight, membrane-sealed packages, the lysosomes, or excreting them into the extracellular surroundings. Upon homogenization, the natural boundaries between the cytoplasm and the cellular compartments containing these peptidases are destroyed, and artifactual digestion of the proteins being purified can commence.

Peptidases are not always a problem. Most native proteins are remarkably resistant to digestion by peptidases, and in most instances, proteins can be purified without being digested. **Harsh treatments**, however, such as heat, the use of detergents, and extremes of pH encourage digestion by peptidases, and proteins purified by procedures employing these conditions often display evidence of deterioration. Because a protein can be nicked by a peptidase and remain almost unaltered in its functional and physical properties, the cumulative effects of digestion by peptidases become most obvious when proteins are unfolded and assessed by electrophoresis in solutions of sodium dodecyl sulfate,¹⁴⁶ a technique that is used to catalog the number and lengths of the polypeptides present in a given preparation.

There are four major classes of peptidases,^{222–224} and their properties determine the precautions that can be taken to inhibit them during a purification procedure. **Acid peptidases** are active only at acidic ranges of pH, and if the purification is carried out at neutral or slightly alkaline pH, their action can be avoided. **Sulphydryl peptidases** contain a thiol necessary for activity. If there is a suspicion that sulphydryl peptidases are responsible for the heterogeneity or loss of activity that is observed, they can be permanently inactivated by treating the solutions of protein with iodoacetate, iodoacetamide, or *N*-ethylmaleimide before the purification is initiated and at one or two intermediate stages during the purification. **Metallopeptidases** require transition metal cations or alkaline earth cations and can be inactivated by adding chelating agents such as *N,N,N',N'*-tetracarboxymethyl-1,2-diaminoethane or *o*-phenanthroline. **Serine peptidases** are invariably inactivated by diisopropyl fluorophosphate, but this compound is extremely toxic and dangerous to use. They are sometimes inactivated by phenylmethanesulfonyl fluoride or by chloromethyl ketones of various specificities. As with the inhibitors of sulphydryl peptidases, these reagents inactivate serine peptidases permanently so that solutions of protein need only be treated prior to initiating the sequence of steps in the purification and at one or two intermediate stages. Even when such precautions are taken, it is always wise to perform the purification in as short a time as possible.

There is a vast array of natural and synthetic **inhibitors of peptidases**^{222–224} that are more or less specific for one or several members of a particular class, and many of them are appropriate for preventing the action of unwanted peptidases during the purification of a protein.²²⁵ Some have been used successfully as additives during the purification of proteins. For example, acetyl-CoA carboxylase has been purified from chicken liver in the presence of parotid trypsin inhibitor,²²⁶ and phosphoglycerate dehydrogenase, from the same source in the presence of leupeptin.²²⁷ Often, however, inhibitors of the activity of peptidases are used prophylactically in the absence of any evidence that they are effective.

Crystallization

As in the isolation of a natural product in organic chemistry, the production of a crystalline preparation (Figure 1–23)²²⁸ was once considered to be the final step in any isolation of a protein. Although the time-consuming search for the proper conditions necessary to crystallize a given protein has gone out of fashion, the exhilarating gallery of photographs of crystalline enzymes compiled by Dixon and Webb²²⁹ testifies to the pleasure that such a conclusion to a long purification must inspire.

Crystallization as a method of purification is usually less effective than chromatography. Some examples of crystallization as the last step in a purification are the purification of 1.5-fold seen upon recrystallization of phosphoenolpyruvate carboxykinase (ATP),²³⁰ the purification of 1.4-fold with a 40% yield seen upon recrystallization of acylphosphatase,²³¹ and the 2-fold purification with a 90% yield seen upon recrystallization of nicotinate-nucleotide diphosphorylase (carboxylating).²³² Recrystallization has been observed to eliminate some of the heterogeneous behavior displayed by a purified protein,²³³ presumably due to an increase in its homogeneity.

Crystals of a protein, aside from their intrinsic beauty, are the specimens required to determine the molecular structure of the purified protein by **X-ray crystallography**, and there is considerable interest in crystallizing proteins.²³⁴ For crystallographic studies, single, untwinned crystals of 0.1–1 mm in size are required, and to produce suitable crystals is a process involving a good deal of trial and error. Because of the number of attempts required by this trial and error and because suitable crystals only form in concentrated solutions of the protein, about 10 mg of protein is required. Furthermore, crystals

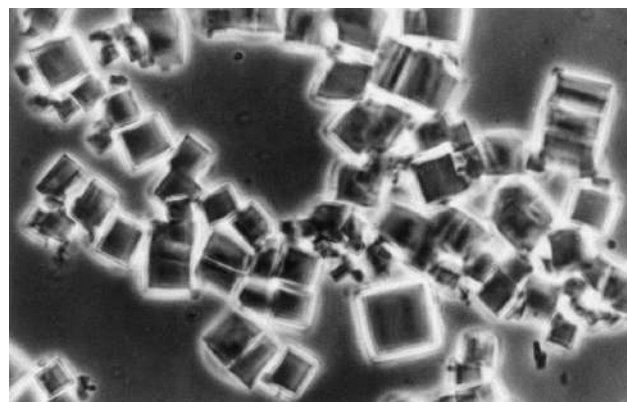


Figure 1–23: Crystals of α -galactosidase isolated from *Mortierella vinacea*.²²⁸ Homogenates of cells of *M. vinacea* were submitted sequentially to ammonium sulfate fractionation, chromatography by anion exchange on (diethylaminoethyl)dextran, and chromatography by molecular exclusion. The final protein was crystallized from a solution of ammonium sulfate. The largest crystal in the field is about 10 μ m across. Reprinted with permission from ref 228. Copyright 1970 *Journal of Biological Chemistry*.

usually will grow most readily from homogeneous solutions of monodisperse protein, so the preparations used must be as pure as possible. Therefore, it is essential to examine the final protein from the purification by electrophoresis in its native state (not after it has been unfolded with dodecyl sulfate), by isoelectric focusing, and by dynamic light scattering before time is wasted trying to crystallize an inhomogeneous sample. The protein also must have suffered as little damage as possible during the procedures used to purify it. For all of these reasons, proteins purified rapidly in one or two steps from overproducing microorganisms, in which more than 10% of the cellular protein can be the protein of interest, are usually the ideal starting material for crystallizations. The construction of such a microorganism is now usually the first step in an attempt to crystallize a protein for crystallographic studies.

Crystals of proteins suitable for crystallography are produced by slowly and continuously increasing the concentration of both the protein and a solute that promotes crystallization. Any of the solutes, such as ammonium sulfate, poly(ethylene glycol), or trimethylamine oxide, that have negative preferential solvations and cause proteins to precipitate from solution can be used to promote crystallization. Choosing conditions of pH and ionic strength within ranges in which the second virial coefficient of the osmotic pressure of a solution of the protein is negative increases the odds of producing crystals.²³⁵ A solution is prepared of the protein and the solute promoting the crystallization, both at concentrations slightly below those at which precipitation would begin. A drop of this solution (1 μ L) is placed upon a glass cover slip. The cover slip is inverted over a well containing a concentrated solution of some salt or other solute in which the activity of water is less than that in the solution of the **hanging drop**. The system is sealed and left in the cold for several weeks. Slowly, water evaporates from the hanging drop and condenses in the well, and if one is lucky, crystals of protein form in the drop. As this is a rare event, hundreds of hanging drops are made, each with a different pH, ionic strength, or concentration of protein over wells with different solutions in them. Small molecules, for example, substrates or ligands, that are known to bind to the protein are also added to some of the drops in the hope that they might encourage crystallization.

Once the crystals are obtained, a sample of them should be dissolved and the protein that they contain submitted to electrophoresis to be certain that it is the one desired rather than a contaminant in that preparation that happened to crystallize while the desired protein did not.²³⁶

Suggested Reading

Noyes, B.E., & Bradshaw, R.A. (1973) Purification and characterization of beef liver dihydrofolate reductase, *J. Biol. Chem.* 248, 3052–3059.

References

1. Cassidy, H.G. (1957) in *Fundamentals of Chromatography; Techniques of Organic Chemistry* (Weissberger, A., Ed.) Vol. 10, pp 31–33, Interscience, New York.
2. Mura-Galelli, M.J., Voegel, J.C., Behr, S., Bres, E.F., & Schaaf, P. (1991) *Proc. Natl. Acad. Sci. U.S.A.* 88, 5557–5561.
3. Massey, V. (1952) *Biochem. J.* 51, 490–494.
4. Morris, C.J.O.R., & Morris, P. (1963) *Separation Methods in Biochemistry*, pp 47–82, Interscience Publishers, New York.
5. Lewis, W.K. (1922) *Ind. Eng. Chem.* 14, 492–497.
6. Peters, W.A. (1922) *Ind. Eng. Chem.* 14, 476–479.
7. Martin, A.J.P., & Synge, R.L.M. (1941) *Biochem. J.* 35, 1358–1368.
8. Craig, L.C., Hausmann, W., Ahrens, E.H., Jr., & Harfenist, E.J. (1951) *Anal. Chem.* 23, 1236–1244.
9. Kyte, J. (1995) *Structure in Protein Chemistry*, pp 4–5, Garland Publishing, New York.
10. Said, A.S. (1956) *AIChE J.* 2, 477–481.
11. Keulemans, A.I.M., & McNair, H.M. (1961) in *Chromatography* (Heftmann, E., Ed.) pp 169–171, Reinhold, New York.
12. Bradshaw, R.A., Garner, W.H., & Gurd, F.R. (1969) *J. Biol. Chem.* 244, 2149–2158.
13. Mahoney, W.C., & Hermodson, M.A. (1980) *J. Biol. Chem.* 255, 11199–11203.
14. Overbeck, J.T.G., & Lijklema, J. (1959) in *Electrophoresis* (Bier, M., Ed.) pp 1–34, Academic Press, New York.
15. Stein, W.D. (1967) *The Movement of Molecules across Cell Membranes*, pp 60–62, Academic Press, New York.
16. Moore, S., & Stein, W.H. (1951) *J. Biol. Chem.* 192, 663–681.
17. Spackman, D.H., Stein, W.H., & Moore, S. (1958) *Anal. Chem.* 30, 1190–1202.
18. Walkinshaw, M.D., & Arnott, S. (1981) *J. Mol. Biol.* 153, 1055–1073.
19. Muro-Pastor, M.I., & Florencio, F.J. (1992) *Eur. J. Biochem.* 203, 99–105.
20. Zeidel, M.L., Nielsen, S., Smith, B.L., Ambudkar, S.V., Maunsbach, A.B., & Agre, P. (1994) *Biochemistry* 33, 1606–1615.
21. Winder, A.J., & Harris, H. (1991) *Eur. J. Biochem.* 198, 317–326.
22. Lam, W.W., & Bugg, T.D. (1997) *Biochemistry* 36, 12242–12251.
23. Foulon, V., Antonenkov, V.D., Croes, K., Waelkens, E., Mannaerts, G.P., Van Veldhoven, P.P., & Casteels, M. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 10039–10044.
24. Stremler, K.E., & Poulter, C.D. (1987) *J. Am. Chem. Soc.* 109, 5542–5544.
25. Lawen, A., & Zocher, R. (1990) *J. Biol. Chem.* 265, 11355–11360.
26. Pollock, R.J., & Hersch, L.B. (1971) *J. Biol. Chem.* 246, 4737–4743.
27. Borum, P.R., & Broquist, H.P. (1977) *J. Biol. Chem.* 252, 5651–5655.
28. Joseph, D.R., & Muench, K.H. (1971) *J. Biol. Chem.* 246, 7602–7609.
29. Wu, J.Y., Matsuda, T., & Roberts, E. (1973) *J. Biol. Chem.* 248, 3029–3034.

30. Roche, P.A., Moorehead, T.J., & Hamilton, G.A. (1982) *Arch. Biochem. Biophys.* 216, 62–73.
31. Ha, S., Chang, E., Lo, M.C., Men, H., Park, P., Ge, M., & Walker, S. (1999) *J. Am. Chem. Soc.* 121, 8415–8426.
32. Caron, M.G., & Lefkowitz, R.J. (1976) *J. Biol. Chem.* 251, 2374–2384.
33. Shorr, R.G., Strohsacker, M.W., Lavin, T.N., Lefkowitz, R.J., & Caron, M.G. (1982) *J. Biol. Chem.* 257, 12341–12350.
34. Graham, R.M., Hess, H.J., & Homcy, C.J. (1982) *J. Biol. Chem.* 257, 15174–15181.
35. Cohen, S., Ushiro, H., Stoscheck, C., & Chinkers, M. (1982) *J. Biol. Chem.* 257, 1523–1531.
36. Kuhn, R.W., Schrader, W.T., Smith, R.G., & O'Malley, B.W. (1975) *J. Biol. Chem.* 250, 4220–4228.
37. Sherrill, J.M., & Kyte, J. (1996) *Biochemistry* 35, 5705–5718.
38. Penefsky, H.S. (1977) *J. Biol. Chem.* 252, 2891–2899.
39. Briggs, M.R., Kadonaga, J.T., Bell, S.P., & Tjian, R. (1986) *Science* 234, 47–52.
40. Hacker, K.J., & Johnson, K.A. (1997) *Biochemistry* 36, 14080–14087.
41. Michel, C., Hartrampf, G., & Buckel, W. (1989) *Eur. J. Biochem.* 184, 103–107.
42. Bull, C., & Ballou, D.P. (1981) *J. Biol. Chem.* 256, 12673–12680.
43. Lau, S.M., Brantley, R.K., & Thorpe, C. (1989) *Biochemistry* 28, 8255–8262.
44. Labourdenne, S., Brass, O., Ivanova, M., Cagna, A., & Verger, R. (1997) *Biochemistry* 36, 3423–3429.
45. Walde, P., & Luisi, P.L. (1989) *Biochemistry* 28, 3353–3360.
46. Webb, M.R. (1992) *Proc. Natl. Acad. Sci. U.S.A.* 89, 4884–4887.
47. Horecker, B.L., & Kornberg, A. (1948) *J. Biol. Chem.* 175, 385–390.
48. Noyes, B.E., & Bradshaw, R.A. (1973) *J. Biol. Chem.* 248, 3052–3059.
49. Duggleby, R.G., & Dennis, D.T. (1974) *J. Biol. Chem.* 249, 162–166.
50. McClure, W.R., Lardy, H.A., & Kneifel, H.P. (1971) *J. Biol. Chem.* 246, 3569–3578.
51. Bazaes, S., Beytia, E., Jabalquinto, A.M., Solis de Ovando, F., Gomez, I., & Eyzaguirre, J. (1980) *Biochemistry* 19, 2300–2304.
52. Parker, A.R., Moore, J.A., Schwab, J.M., & Davisson, V.J. (1995) *J. Am. Chem. Soc.* 117, 10605–10613.
53. Kramer, P.R., & Miziorko, H.M. (1980) *J. Biol. Chem.* 255, 11023–11028.
54. Switzer, R.L. (1969) *J. Biol. Chem.* 244, 2854–2863.
55. Kataoka, M., Shimizu, S., & Yamada, H. (1992) *Eur. J. Biochem.* 204, 799–806.
56. Moriyama, T., & Srere, P.A. (1971) *J. Biol. Chem.* 246, 3217–3223.
57. Leloir, L.F., & Cardini, C.E. (1957) *Methods Enzymol.* 3, 843–844.
58. Cooper, J.L., & Meister, A. (1972) *Biochemistry* 11, 661–671.
59. Donald, A., Sibley, D., Lyons, D.E., & Dahms, A.S. (1979) *J. Biol. Chem.* 254, 2132–2137.
60. Esaki, N., Nakamura, T., Tanaka, H., & Soda, K. (1982) *J. Biol. Chem.* 257, 4386–4391.
61. Barton, R.W., & Neufeld, E.F. (1971) *J. Biol. Chem.* 246, 7773–7779.
62. Gerhart, J., Wu, M., & Kirschner, M. (1984) *J. Cell Biol.* 98, 1247–1255.
63. Wu, M., & Gerhart, J.C. (1980) *Dev. Biol.* 79, 465–477.
64. Canals, F. (1992) *Biochemistry* 31, 4493–4501.
65. Gilbert, W., & Mueller-Hill, B. (1966) *Proc. Natl. Acad. Sci. U.S.A.* 56, 1891–1898.
66. Skou, J.C. (1964) In *Progress in Biophysics and Molecular Biology* (Butler, J.A.V., & Huxley, H.E., Eds.) Vol. 14, pp 131–166, Pergamon, New York.
67. Kyte, J. (1971) *J. Biol. Chem.* 246, 4157–4165.
68. Nour, J.M., & Rabinowitz, J.C. (1991) *J. Biol. Chem.* 266, 18363–18369.
69. Ryu, S., & Tjian, R. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 7137–7142.
70. Trower, M.K., Buckland, R.M., & Griffin, M. (1989) *Eur. J. Biochem.* 181, 199–206.
71. Yoshioka, H., Nagasawa, T., & Yamada, H. (1991) *Eur. J. Biochem.* 199, 17–24.
72. Smigel, M.D. (1986) *J. Biol. Chem.* 261, 1976–1982.
73. Moczydlowski, E.G., & Fortes, P.A. (1981) *J. Biol. Chem.* 256, 2346–2356.
74. Layne, E. (1957) *Methods Enzymol.* 3, 447–454.
75. Lowry, O.H., Rosebrough, N.J., Farr, A.L., & Randall, R.J. (1951) *J. Biol. Chem.* 193, 265–275.
76. Adams, M.W., Eccleston, E., & Howard, J.B. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 4932–4936.
77. Bradford, M.M. (1976) *Anal. Biochem.* 72, 248–254.
78. Edsall, J.T., & Wyman, J. (1958) *Biophysical Chemistry*, Vol. I, pp 263–282, Academic Press, New York.
79. Hofmeister, F. (1888) *Arch. Exp. Pathol. Pharmacol.* 24, 247–260.
80. Cacace, M.G., Landau, E.M., & Ramsden, J.J. (1997). *Q. Rev. Biophys.* 30, 241–277.
81. Huang, X., Knoell, C.T., Frey, G., Hazegh-Azam, M., Tashjian, A.H., Jr., Hedstrom, L., Abeles, R.H., & Timasheff, S.N. (2001) *Biochemistry* 40, 11734–11741.
82. Vogel, R., Fan, G.B., Sheves, M., & Siebert, F. (2001) *Biochemistry* 40, 483–493.
83. Arakawa, T., & Timasheff, S.N. (1982) *Biochemistry* 21, 6545–6552.
84. Arakawa, T., & Timasheff, S.N. (1984) *Biochemistry* 23, 5924–5929.
85. Arakawa, T., & Timasheff, S.N. (1982) *Biochemistry* 21, 6536–6544.
86. Geisler, N., & Weber, K. (1981) *FEBS Lett.* 125, 253–256.
87. Tong, J.H., & Kaufman, S. (1975) *J. Biol. Chem.* 250, 4152–4158.
88. Doolittle, R.F., Thomas, C., & Stone, W., Jr. (1960) *Science* 132, 36–37.
89. Yang, Z., Kollman, J.M., Pandi, L., & Doolittle, R.F. (2001) *Biochemistry* 40, 12515–12523.
90. Gerhart, J.C., & Holoubek, H. (1967) *J. Biol. Chem.* 242, 2886–2892.
91. Fuller, G.M., & Doolittle, R.F. (1971) *Biochemistry* 10, 1305–1311.
92. Kautz, J., & Schnackerz, K.D. (1989) *Eur. J. Biochem.* 181, 431–435.
93. Beebe, J.A., & Frey, P.A. (1998) *Biochemistry* 37, 14989–14997.

94. Uyeda, K., & Kurooka, S. (1970) *J. Biol. Chem.* 245, 3315–3324.
95. Ahern, T.J., & Klibanov, A.M. (1985) *Science* 228, 1280–1284.
96. D'Alessio, G., & Josse, J. (1971) *J. Biol. Chem.* 246, 4319–4325.
97. Sabourin, P.J., & Bieber, L.L. (1982) *J. Biol. Chem.* 257, 7460–7467.
98. Nimmo, G.A., & Coggins, J.R. (1981) *Biochem. J.* 197, 427–436.
99. Lau, E.P., Cochran, B.C., & Fall, R.R. (1980) *Arch. Biochem. Biophys.* 205, 352–359.
100. Sakurai, N., & Sakurai, T. (1997) *Biochemistry* 36, 13809–13815.
101. Lee, F.J., Lin, L.W., & Smith, J.A. (1989) *Eur. J. Biochem.* 184, 21–28.
102. Bogard, M., Camadro, J.M., Nordmann, Y., & Labbe, P. (1989) *Eur. J. Biochem.* 181, 417–421.
103. Chen, M.W., Jahn, D., O'Neill, G.P., & Soll, D. (1990) *J. Biol. Chem.* 265, 4058–4063.
104. Green, J.M., & Nichols, B.P. (1991) *J. Biol. Chem.* 266, 12971–12975.
105. Zachariou, M., & Hearn, M.T. (1996) *Biochemistry* 35, 202–211.
106. Hutchens, T.W., & Porath, J. (1986) *Anal. Biochem.* 159, 217–226.
107. Sarngadharan, M.G., Watanabe, A., & Pogell, B.M. (1970) *J. Biol. Chem.* 245, 1926–1929.
108. Mocali, A., & Paoletti, F. (1989) *Eur. J. Biochem.* 180, 213–219.
109. Volonte, C., & Greene, L.A. (1992) *J. Biol. Chem.* 267, 21663–21670.
110. Araki, C., & Arai, K. (1957) *Bull. Chem. Soc. Jpn.* 30, 287–293.
111. March, S.C., Parikh, I., & Cuatrecasas, P. (1974) *Anal. Biochem.* 60, 149–152.
112. Cuatrecasas, P., Wilchek, M., & Anfinsen, C.B. (1968) *Proc. Natl. Acad. Sci. U.S.A.* 61, 636–643.
113. Cuatrecasas, P. (1970) *J. Biol. Chem.* 245, 3059–3065.
114. Steers, E., Jr., Cuatrecasas, P., & Pollard, H.B. (1971) *J. Biol. Chem.* 246, 196–200.
115. Chan, W.W., & Takahashi, M. (1969) *Biochem. Biophys. Res. Commun.* 37, 272–277.
116. Berg, R.A., & Prockop, D.J. (1973) *J. Biol. Chem.* 248, 1175–1182.
117. Geren, C.R., & Ebner, K.E. (1977) *J. Biol. Chem.* 252, 2082–2088.
118. Lee, C.Y., Lappi, D.A., Wermuth, B., Everse, J., & Kaplan, N.O. (1974) *Arch. Biochem. Biophys.* 163, 561–569.
119. Nealon, D.A., & Cook, R.A. (1979) *Biochemistry* 18, 3616–3622.
120. Ryan, R.L., & McClure, W.O. (1979) *Biochemistry* 18, 5357–5365.
121. Huang, J.S., Huang, S.S., & Tang, J. (1979) *J. Biol. Chem.* 254, 11405–11417.
122. Allen, M.B., & Walker, D.G. (1980) *Biochem. J.* 185, 565–575.
123. Magnani, M., Serafini, G., Stocchi, V., Bossu, M., & Dacha, M. (1982) *Arch. Biochem. Biophys.* 216, 449–454.
124. Kaufman, B.T., & Pierce, J.V. (1971) *Biochem. Biophys. Res. Commun.* 44, 608–613.
125. Mendicino, J., Sivakami, S., Davila, M., & Chandrasekaran, E.V. (1982) *J. Biol. Chem.* 257, 3987–3994.
126. Kitani, T., & Fujisawa, H. (1983) *J. Biol. Chem.* 258, 235–239.
127. Grimshaw, C.E., Henderson, G.B., Soppe, G.G., Hansen, G., Mathur, E.J., & Huennekens, F.M. (1984) *J. Biol. Chem.* 259, 2728–2733.
128. Caron, M.G., Srinivasan, Y., Pitha, J., Kociolek, K., & Lefkowitz, R.J. (1979) *J. Biol. Chem.* 254, 2923–2927.
129. Deutsch, D.G., & Mertz, E.T. (1970) *Science* 170, 1095–1096.
130. Chibber, B.A., Deutsch, D.G., & Mertz, E.T. (1974) *Methods Enzymol.* 34, 424–432.
131. Raeber, A.J., Riggio, G., & Waser, P.G. (1989) *Eur. J. Biochem.* 186, 487–492.
132. Moomaw, J.F., & Casey, P.J. (1992) *J. Biol. Chem.* 267, 17438–17443.
133. Pfeuffer, E., Dreher, R.M., Metzger, H., & Pfeuffer, T. (1985) *Proc. Natl. Acad. Sci. U.S.A.* 82, 3086–3090.
134. Pang, I.H., & Sternweis, P.C. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 7814–7818.
135. Manenti, S., Sorokine, O., Van Dorsselaer, A., & Taniguchi, H. (1992) *J. Biol. Chem.* 267, 22310–22315.
136. Pettersson, I., Kusche, M., Unger, E., Wlad, H., Nylund, L., Lindahl, U., & Kjellen, L. (1991) *J. Biol. Chem.* 266, 8044–8049.
137. Chang, G.G., Wang, J.K., Huang, T.M., Lee, H.J., Chou, W.Y., & Meng, C.L. (1991) *Eur. J. Biochem.* 202, 681–688.
138. Cheng, Q., Finkel, D., & Hostetter, M.K. (2000) *Biochemistry* 39, 5450–5457.
139. Sugden, B., & Keller, W. (1973) *J. Biol. Chem.* 248, 3777–3788.
140. Hsu, Y.P., & Kohlhaw, G.B. (1980) *J. Biol. Chem.* 255, 7255–7260.
141. Cvetanovic, M., Moreno de la Garza, M., Dommes, V., & Kunau, W.H. (1985) *Biochem. J.* 227, 49–56.
142. Payne, M.E., Schworer, C.M., & Soderling, T.R. (1983) *J. Biol. Chem.* 258, 2376–2382.
143. Kadonaga, J.T., & Tjian, R. (1986) *Proc. Natl. Acad. Sci. U.S.A.* 83, 5889–5893.
144. Reardon, J.E. (1990) *J. Biol. Chem.* 265, 7112–7115.
145. Amaya, Y., Yamazaki, K., Sato, M., Noda, K., Nishino, T., & Nishino, T. (1990) *J. Biol. Chem.* 265, 14170–14175.
146. Pringle, J.R. (1970) *Biochem. Biophys. Res. Commun.* 39, 46–52.
147. Nealon, K., Nicholl, I.D., & Kenny, M.K. (1996) *Nucleic Acids Res.* 24, 3763–3770.
148. Nicholl, I.D., Nealon, K., & Kenny, M.K. (1997) *Biochemistry* 36, 7557–7566.
149. Durchschlag, H., Biedermann, G., & Eggerer, H. (1981) *Eur. J. Biochem.* 114, 255–262.
150. Tanford, C. (1962) *Adv. Protein Chem.* 17, 69–165.
151. Tanford, C., & Hauenstein, J.D. (1956) *J. Am. Chem. Soc.* 78, 5288–5291.
152. Qin, B.Y., Bewley, M.C., Creamer, L.K., Baker, H.M., Baker, E.N., & Jameson, G.B. (1998) *Biochemistry* 37, 14014–14023.
153. Carr, C.W. (1953) *Arch. Biochem. Biophys.* 46, 417–423.
154. Carr, C.W. (1956) *Arch. Biochem. Biophys.* 62, 476–484.
155. Matthew, J.B., Hanania, G.I., & Gurd, F.R. (1979) *Biochemistry* 18, 1928–1936.

156. Isupov, M.N., Antson, A.A., Dodson, E.J., Dodson, G.G., Dementieva, I.S., Zakomirdina, L.N., Wilson, K.S., Dauter, Z., Lebedev, A.A., & Harutyunyan, E.H. (1998) *J. Mol. Biol.* 276, 603–623.
157. Stout, T.J., Graham, H., Buckley, D.I., & Matthews, D.J. (2000) *Biochemistry* 39, 8460–8469.
158. Wedekind, J.E., Trame, C.B., Dorywalska, M., Koehl, P., Raschke, T.M., McKee, M., FitzGerald, D., Collier, R.J., & McKay, D.B. (2001) *J. Mol. Biol.* 314, 823–837.
159. Tanford, C., Swanson, S.A., & Shore, W.S. (1955) *J. Am. Chem. Soc.* 77, 6414–6421.
160. Tanford, C. (1961) *Physical Chemistry of Macromolecules*, Wiley, New York.
161. Berne, B.J., & Pecora, R. (1976) *Dynamic Light Scattering: With Applications to Chemistry, Biology, and Physics*, Wiley, New York.
162. Roche, T.E., Powers-Greenwood, S.L., Shi, W.F., Zhang, W.B., Ren, S.Z., Roche, E.D., Cox, D.J., & Sorensen, C.M. (1993) *Biochemistry* 32, 5629–5637.
163. Chien, W.J., Cheng, S.F., & Chang, D.K. (1998) *Anal. Biochem.* 264, 211–215.
164. Haner, R.L., & Schleich, T. (1989) *Methods Enzymol.* 176, 418–446.
165. Tiselius, A., & Svensson, H. (1940) *Trans. Faraday Soc.* 36, 16–22.
166. Manning, G.S. (1981) *J. Phys. Chem.* 85, 1506–1515.
167. Debye, P., & Huckel, E. (1923) *Z. Physik* 24, 305–325.
168. Henry, D.C. (1931) *Proc. R. Soc. London, A* 133, 106–129.
169. Brown, R.A., & Timasheff, S.N. (1959) in *Electrophoresis* (Bier, M., Ed.) pp 317–367, Academic Press, New York.
170. Duke, J.A., Bier, M., & Nord, F.F. (1952) *Arch. Biochem. Biophys.* 40, 424–436.
171. Borza, D.B., Tatum, F.M., & Morgan, W.T. (1996) *Biochemistry* 35, 1925–1934.
172. Carbeck, J.D., & Negin, R.S. (2001) *J. Am. Chem. Soc.* 123, 1252–1253.
173. Pauling, L., & Itano, H.A. (1949) *Science* 110, 543–548.
174. Velick, S.F. (1949) *J. Phys. Colloid Chem.* 53, 135–149.
175. Longworth, L.G. (1959) in *Electrophoresis* (Bier, M., Ed.) pp 137–177, Academic Press, New York.
176. Tiselius, A. (1937) *Trans. Faraday Soc.* 33, 524–531.
177. Morris, C.J.O.R. (1966) in *Protides of the Biological Fluids* (Peeters, H., Ed.) Vol. 14, pp 543–561, Elsevier, Amsterdam.
178. Ferguson, K.A. (1964) *Metabolism* 13, 985–1002.
179. Philippov, P.P., Shestakova, I.K., Tikhomirova, N.K., & Kochetov, G.A. (1980) *Biochim. Biophys. Acta* 613, 359–369.
180. Stahl, P.D., & Touster, O. (1971) *J. Biol. Chem.* 246, 5398–5406.
181. Hames, D. (1990) in *Gel Electrophoresis of Proteins* (Hames, B. D., & Rickwood, D., Eds.) pp 1–147, Oxford University Press, Oxford, U.K.
182. Ornstein, L. (1964) *Ann. N.Y. Acad. Sci.* 121, 321–349.
183. Davis, B.J. (1964) *Ann. N.Y. Acad. Sci.* 121, 404–427.
184. Kohlrausch, F. (1897) *Ann. Phys. Chem.* 62, 209–239.
185. Laemmli, U.K. (1970) *Nature* 227, 680–685.
186. Kyte, J., & Rodriguez, H. (1983) *Anal. Biochem.* 133, 515–522.
187. Jovin, T.M. (1973) *Biochemistry* 12, 871–879.
188. Chrambach, A., Jovin, T.M., Svendsen, P.J., & Rodbard, D. (1976) in *Methods of Protein Separation* (Catsimpoilas, N., Ed.) Vol. 2, pp 27–144, Plenum Press, New York.
189. Cohn, E.J., Green, A.A., & Blanchard, M.H. (1937) *J. Am. Chem. Soc.* 59, 509–517.
190. Ruch, F.E., & Vagelos, P.R. (1973) *J. Biol. Chem.* 248, 8086–8094.
191. Katze, J.R., & Konigsberg, W. (1970) *J. Biol. Chem.* 245, 923–930.
192. Henderson, G.B., & Snell, E.E. (1973) *J. Biol. Chem.* 248, 1906–1911.
193. Zampighi, G., Kyte, J., & Freytag, W. (1984) *J. Cell Biol.* 98, 1851–1864.
194. Oakley, B.R., Kirsch, D.R., & Morris, N.R. (1980) *Anal. Biochem.* 105, 361–363.
195. Switzer, R.C.R., Merril, C.R., & Shifrin, S. (1979) *Anal. Biochem.* 98, 231–237.
196. Kolhouse, J.F., Utley, C., & Allen, R.H. (1980) *J. Biol. Chem.* 255, 2708–2712.
197. Karawya, E., Swack, J.A., & Wilson, S.H. (1983) *Anal. Biochem.* 135, 318–325.
198. Schachter, H., Sarney, J., McGuire, E.J., & Roseman, S. (1969) *J. Biol. Chem.* 244, 4785–4792.
199. Li, J.J., Ross, C.R., Tepperman, H.M., & Tepperman, J. (1975) *J. Biol. Chem.* 250, 141–148.
200. Morley, C.G., & Stadtman, T.C. (1970) *Biochemistry* 9, 4890–4900.
201. Yu, C., Gunsalus, I.C., Katagiri, M., Suhara, K., & Takemori, S. (1974) *J. Biol. Chem.* 249, 94–101.
202. Takahashi, S., Kuzuyama, T., Watanabe, H., & Seto, H. (1998) *Proc. Natl. Acad. Sci. U.S.A.* 95, 9879–9884.
203. Olsen, A.S., & Milman, G. (1974) *J. Biol. Chem.* 249, 4030–4037.
204. Scott, W.A., & Tatum, E.L. (1971) *J. Biol. Chem.* 246, 6347–6352.
205. Warnick, G.R., & Burnham, B.F. (1971) *J. Biol. Chem.* 246, 6880–6885.
206. Fernandez-Sorensen, A., & Carlson, D.M. (1971) *J. Biol. Chem.* 246, 3485–3493.
207. Reed, B.C., & Rilling, H.C. (1975) *Biochemistry* 14, 50–54.
208. Beytia, E., Dorsey, J.K., Marr, J., Cleland, W.W., & Porter, J.W. (1970) *J. Biol. Chem.* 245, 5450–5458.
209. Fenton, W.A., Hack, A.M., Willard, H.F., Gertler, A., & Rosenberg, L.E. (1982) *Arch. Biochem. Biophys.* 214, 815–823.
210. Arnold, W.J., & Kelley, W.N. (1971) *J. Biol. Chem.* 246, 7398–7404.
211. Norton, I.L., Pfuderer, P., Stringer, C.D., & Hartman, F.C. (1970) *Biochemistry* 9, 4952–4958.
212. Mihalik, S.J., McGuinness, M., & Watkins, P.A. (1991) *J. Biol. Chem.* 266, 4822–4830.
213. Ohshita, T., Sakuda, H., Nakasone, S., & Iwamasa, T. (1989) *Eur. J. Biochem.* 179, 201–207.
214. O'Farrell, P.H. (1975) *J. Biol. Chem.* 250, 4007–4021.
215. Traniello, S., Melloni, E., Pontremoli, S., Sia, C.L., & Horecker, R.L. (1972) *Arch. Biochem. Biophys.* 149, 222–231.
216. Kennedy, S.C., Rauner, R., & Gawron, O. (1972) *Biochem. Biophys. Res. Commun.* 47, 740–745.

217. Gennis, L.S. (1976) *Proc. Natl. Acad. Sci. U.S.A.* 73, 3928–3932.
218. Kirschner, K., Eigen, M., Bittman, R., & Voigt, B. (1966) *Proc. Natl. Acad. Sci. U.S.A.* 56, 1661–1667.
219. Sloan, D.L., & Velick, S.F. (1973) *J. Biol. Chem.* 248, 5419–5423.
220. Mockrin, S.C., Byers, L.D., & Koshland, D.E., Jr. (1975) *Biochemistry* 14, 5428–5437.
221. Weber, K., Pringle, J.R., & Osborn, M. (1972) *Methods Enzymol.* 26C, 3–27.
222. Lorand, L. (1970) *Methods in Enzymology*, Vol. 19, Academic Press, New York.
223. Lorand, L. (1976) *Methods in Enzymology*, Vol. 45, Academic Press, New York.
224. Lorand, L. (1981) *Methods in Enzymology*, Vol. 80, Academic Press, New York.
225. North, M.J. (1989) in *Proteolytic Enzymes: A Practical Approach* (Benyon, R. J., & Bond, J. S., Eds.) pp 105–124, IRL Press, New York.
226. Mackall, J.C., Lane, M.D., Leonard, K.R., Pendergast, M., & Kleinschmidt, A.K. (1978) *J. Mol. Biol.* 123, 595–606.
227. Grant, G.A., Keefer, L.M., & Bradshaw, R.A. (1978) *J. Biol. Chem.* 253, 2724–2726.
228. Suzuki, H., Li, S.C., & Li, Y.T. (1970) *J. Biol. Chem.* 245, 781–786.
229. Dixon, M., & Webb, E.C. (1964) *Enzymes*, pp 794–808, Longmans, Green & Co., London.
230. Cannata, J.J. (1970) *J. Biol. Chem.* 245, 792–798.
231. Shiokawa, H., & Noda, L. (1970) *J. Biol. Chem.* 245, 669–673.
232. Iwai, K., & Taguchi, H. (1974) *Biochem. Biophys. Res. Commun.* 56, 884–891.
233. Monteilhet, C., & Blow, D.M. (1978) *J. Mol. Biol.* 122, 407–417.
234. McPherson, A. (1990) *Eur. J. Biochem.* 189, 1–23.
235. Pjura, P.E., Lenhoff, A.M., Leonard, S.A., & Gittis, A.G. (2000) *J. Mol. Biol.* 300, 235–239.
236. Nunn, R.S., Artymiuk, P.J., Baker, P.J., Rice, D.W., & Hunter, C.N. (1995) *J. Mol. Biol.* 252, 153.

Chapter 2

Electronic Structure

When proteins are submitted to chemical analysis, they are found to be composed of 20 amino acids: aspartic acid, asparagine, threonine, serine, glutamine, glutamic acid, proline, glycine, alanine, cysteine, valine, methionine, isoleucine, leucine, tyrosine, phenylalanine, lysine, histidine, tryptophan, and arginine. Each protein has different relative amounts of each of these amino acids. The amino acids a protein contains are coupled together in a particular order to create polymers 50–5000 amino acids in length, referred to as polypeptides. To understand the structure of molecules of protein, one must understand the amino acids, the order in which they are connected, and the way that these long polymers are folded up to produce the native conformation of the molecule. The first level of understanding is grounded in a firm knowledge of the bonding and molecular structure of small molecules. The second level of understanding requires a description of the complete covalent structure of the polymers composing proteins. The third level of understanding proceeds from crystallographic molecular models of proteins that are the products of X-ray crystallography.

It is remarkable that each molecule of a particular protein, if it has not been heterogeneously posttranslationally modified, has the same covalent structure and that when it is in its natural environment, the polypeptides from which it is composed assume the same few conformations even though the complete molecule of the protein is large. These two properties are foreign to a synthetic chemist. Molecules produced synthetically are either precise but small or large but heterogeneous. Large heterogeneous polymers produced synthetically seldom have defined structures. Yet a molecule of protein is made from atoms held together by the same covalent chemical bonds holding together the smaller molecules to which one is already accustomed. All of the rules of bonding exerted with such inescapability in small molecules are as inescapable in a molecule of protein.

The covalent bonds holding the atoms together in any molecule are pairs of electrons confined to molecular orbitals. The molecular orbitals are either localized σ molecular orbitals or delocalized π molecular orbitals. A distinction between these two types of molecular orbitals is crucial to an understanding of bond lengths, bond angles, and rotational motions about bonds.

In addition to the covalent bonds, molecules of pro-

tein are filled with lone pairs of electrons. Because σ lone pairs of electrons are the only valence electrons that do not participate in covalent bonds and because there are also lone pairs of electrons participating in π molecular orbitals, to understand the details of molecular structure one must be able to distinguish localized σ lone pairs of electrons from delocalized π lone pairs of electrons. The distinction between these two types of electrons is reflected in their basicity, their ability to house a proton.

Each lone pair of electrons in a molecule is a potential base, and each hydrogen in a molecule is a potential acid. Which lone pair will act as a base is determined by the acid dissociation constant for its conjugate acid, and which hydrogen will act as an acid is determined by its own acid dissociation constant. Every lone pair is basic and every hydrogen is acidic, but most lone pairs are such weak bases and most hydrogens are such weak acids that their basicity or acidity can be ignored. To understand the atomic structure of a molecule of protein, the significant acids and bases within it must be identified and categorized. It is also necessary to distinguish an acid dissociation, in which a proton leaves the molecule, from a tautomerization, in which protons redistribute among lone pairs of electrons within the molecule.

The chemical capacities available to a protein are a reflection of the amino acids from which it is constructed. Each of the 20 amino acids has its own peculiar set of chemical capacities. These are mixed in a unique way by the amino acid sequence and the resulting native structure to produce those of the particular protein, but to understand the mixture, the properties of the ingredients must be understood. These properties include the bonding and acid–base behavior of each of the 20 side chains of the amino acids. With the exception of the regular polyamide backbone of the polymer, the covalent bonds, acidic hydrogens, and basic lone pairs of electrons that fill a molecule of protein are contributed by these side chains.

π and σ

Molecules, including proteins, are arrays of atomic nuclei required to maintain particular distances and angular dispositions relative to each other by electrons confined to particular regions of space known as

orbitals. Every electron in a molecule is confined to a specific orbital, and almost every orbital is occupied by two electrons. Each orbital is either confined exclusively to one nucleus or distributed between or among particular nuclei. The electrons, in their occupation of these orbitals, create the covalent structure of the molecule. The electrons present in a molecule can be divided into three categories, core electrons, π electrons, and σ electrons, that reflect the degree to which they are confined and that define their chemical reactivity.

Core electrons are the electrons that are immediately adjacent to a nucleus. Aside from hydrogen, almost all of the atoms present in molecules of protein are either carbon, oxygen, or nitrogen. Each of these atoms has two core electrons spherically confined about the nucleus. Occasionally, sulfur or phosphorus occurs in a protein, and these atoms each have 10 core electrons. Because they are confined close to the nucleus, the core electrons provide the greatest electron density and are the prominent features in a map of electron density. They are, however, chemically inert.

Valence electrons are the outermost electrons surrounding each atom. All of the chemistry of a molecule, which is the consequence of its chemical bonds and its sites of reactivity, results from these valence electrons. Unless one electron is missing, as in the case of a radical, or two electrons are momentarily missing, as in a carbocation, every carbon, nitrogen, oxygen, sulfur, or phosphorus in a molecule of protein can be formally associated with eight valence electrons. By convention, these octets are assigned by Lewis structures. This formalism divides valence electrons into **bonding electrons** and **lone pairs of electrons** and assigns **formal charge** to certain atoms. An example would be the Lewis structure of the model compound for glutamic acid in a polypeptide, *N*-acetylglutamate α -amide (Figure 2-1A). The intent of a Lewis structure is to count valence electrons.

A pair of bonding electrons occupies a **bonding molecular orbital** that is formed from the overlap of two or more atomic orbitals, each contributed by a different atom in the molecule. These bonding electrons must be clearly distinguished as occupants of either π molecular orbitals, forming π bonds, or σ molecular orbitals, forming σ bonds.

The overlap of two or more adjacent and parallel p atomic orbitals on two or more adjacent atoms in a molecule creates a **system of π molecular orbitals**. Two adjacent p atomic orbitals can overlap only above and below the line of centers between the two atoms from which they are contributed (Figure 2-2). This geometry has two consequences: it prevents rotation about axes connecting the nuclei of adjacent atoms and it permits a series of overlaps to occur simultaneously. Because rotation is prevented, structures containing a system of π molecular orbitals are rigid. The fact that a series of overlaps can occur permits the electrons occupying a system of π molecular orbitals to be delocalized.

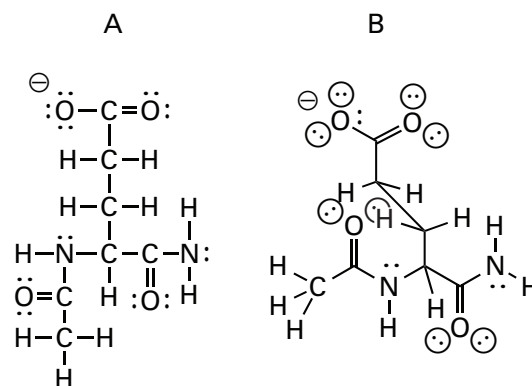


Figure 2-1: Two ways of representing the electronic structure of *N*-acetylglutamate α -amide. (A) In the Lewis dot formula, each main atom is surrounded by an octet of electrons and the total number of electrons represented equals the sum of the number of valence electrons contributed by each neutral atom plus the elementary molecular charge. The negative sign surrounded by a circle locates formal charge. (B) In a σ - π stereochemical representation distinguishing types of electrons, a σ bond is designated by a line, a localized σ lone pair of electrons is designated by two dots surrounded by a circle, a π bond is indicated by a second or third line between two atoms, and a π lone pair of electrons is shown by two uncircled dots. The atoms are arranged in space to represent the tetrahedral or trigonal geometry dictated by their respective hybridizations.

Delocalization of a pair of electrons occupying one π molecular orbital in such a system results from the fact that each π molecular orbital is a linear combination of the p orbitals that overlap. Each π molecular orbital is spread over and shared by every atom that contributed a p orbital to the system unless a node is located at that atom. When a pair of electrons occupies a π molecular orbital, it cannot be assigned to a particular atom, notwithstanding the formal requirement of the Lewis dot structure that it be so localized for the purposes of book-keeping. Confusion between actuality and accounting sometimes leaves the impression that π electrons are localized.

An example of a combination of p atomic orbitals is the system of π molecular orbitals that forms when four parallel p orbitals mix (Figure 2-2). The number of π molecular orbitals that result from any combination of this type is always equal to the number of p orbitals that have mixed; in this case there are four π molecular orbitals in the system. Each p orbital can be mixed in one of two **phases**, and adjacent p orbitals can be either in phase, in which case they overlap—a favorable interaction—or out of phase, in which case a node—an unfavorable interaction—occurs between them. A **node** is a position at which the phase inverts. In a linear system such as the one shown in Figure 2-2, the number of nodes increases by one for each molecular orbital in the series.

Each of these four π molecular orbitals in Figure 2-2 has an **energy level** associated with it that is equal to the energy one electron would experience were it confined

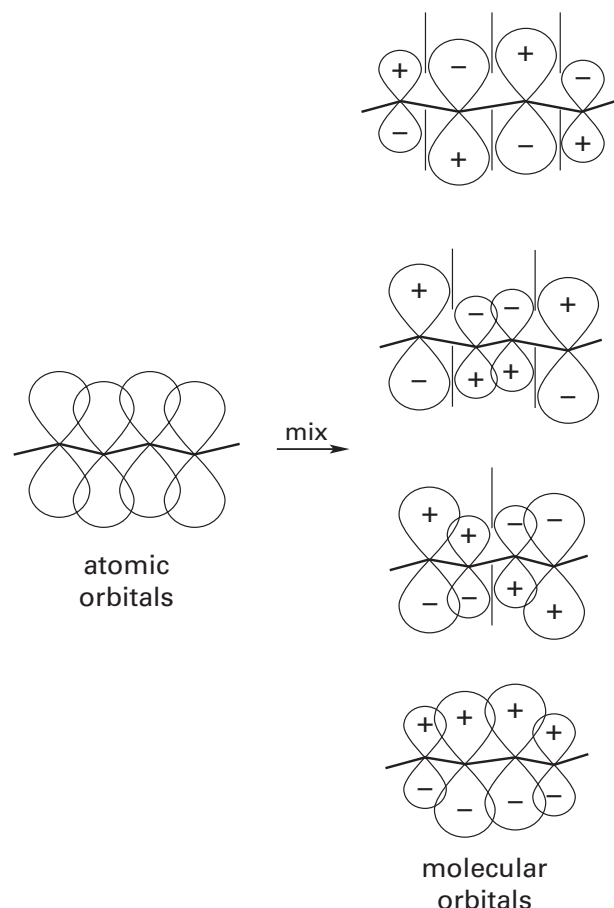


Figure 2-2: A π molecular orbital system formed by the combination of four parallel p atomic orbitals on four adjacent atoms held together by three σ bonds. The four p orbitals overlap, and four linear combinations of these four atomic orbitals, the four π molecular orbitals, are permitted. In the combination of lowest electronic energy, all four p orbitals overlap in phase (indicated arbitrarily with + and -). In each of the higher π molecular orbitals, nodes are present, on either side of which the constituent p orbitals have opposite phase so overlap is antibonding. The four individual π molecular orbitals are arranged in order of increasing electronic energy from bottom to top. In all linear π molecular orbital systems, such as the one represented, the number of nodes increases by one upon moving to the next higher energy level. The nodes are evenly distributed from one end of the structure to the other. In this π molecular orbital system formed from four p atomic orbitals there are usually four electrons occupying the two π molecular orbitals of lowest energy. The sizes of the atomic orbitals approximately represent the magnitude of their contribution to each π molecular orbital.

within that orbital. In a π molecular orbital system formed entirely from p orbitals on atoms of the same element, such as that of the four carbon atoms of butadiene or that of the six carbon atoms in benzene, these energy levels are distributed symmetrically above and below the potential energy an electron would have in one of the isolated p orbitals from which the system has been created. The more nodes the molecular orbital has, the higher its energy and the less likely that an electron will occupy it. A π molecular orbital is designated as

bonding, nonbonding, or antibonding depending on whether its energy level is less than, equal to, or greater than the energy level of an isolated p orbital, respectively. If the π molecular orbital is bonding, the equivalent of a covalent bond is formed because a pair of electrons has a lower energy in the molecular orbital than it would have were it split between two isolated atoms. In π molecular orbital systems formed from three or more parallel p orbitals, that covalent bond is spread over the atoms contributing the p orbitals, notwithstanding the impression often left by the Lewis structure that it is the second localized bond in a double bond.

The general structure of each π molecular orbital is determined only by the number of p orbitals that have mixed together and their connectivity; however, the nature of the atom—carbon, nitrogen, or oxygen—that has contributed each of the p orbitals does affect the shape and energy of the molecular orbital through coulomb effects. These effects are most easily understood as perturbations of the symmetric π molecular orbital system that would be formed from the same number and arrangement of carbon atoms by the fact that some of the atoms are of other elements. A **coulomb effect** is the distortion of the system of symmetric π molecular orbitals that would exist if all of the atoms were carbons. It is caused by the electronegativity and electron deficiency of the atom other than carbon that the π molecular orbital system actually contains. A coulomb effect causes the region of a bonding or nonbonding π molecular orbital over a more electronegative atom, such as oxygen, or a more electron-deficient atom, usually nitrogen, to swell at the expense of the region or regions over the less electronegative atoms, usually carbons.

The number of π molecular orbitals in a given system is determined solely by the number of p orbitals that have been mixed together, but the number of those molecular orbitals that are occupied by pairs of electrons is determined by other properties of the molecule as well. The two decisions, how many p orbitals have combined and how many π electrons have occupied the system of π molecular orbitals, are made by examining all valid **resonance structures** for the molecule. Drawing resonance structures is nothing more than making this decision. Any electrons that are active participants in resonance have been explicitly designated as π electrons by the person who drew those resonance structures, and any atom the bonding of which changes among the resonance structures has been explicitly designated as an atom that has contributed a p orbital to the system of π molecular orbitals. All double and triple bonds in a molecule are necessarily participants in systems of π molecular orbitals.

The **amide**, which is of wide biochemical relevance, is a simple example of this process of designation (Figure 2-3). The chemical properties of an amide are usually explained by drawing two resonance structures.¹ These

two resonance structures state that each of the three atoms, the oxygen, the carbon, and the nitrogen, contributes a p orbital to the system of π molecular orbitals because their bonding changes between the two Lewis structures of the resonance pair. The resonance structures state that the system of π molecular orbitals contains four π electrons because two of the pairs of electrons shift between the two structures. When three adjacent p orbitals are mixed, three π molecular orbitals are created (Figure 2-3). That four electrons occupy these three molecular orbitals places one pair in each of the two molecular orbitals of lowest energy. If coulomb effects are disregarded for the moment, the two electrons

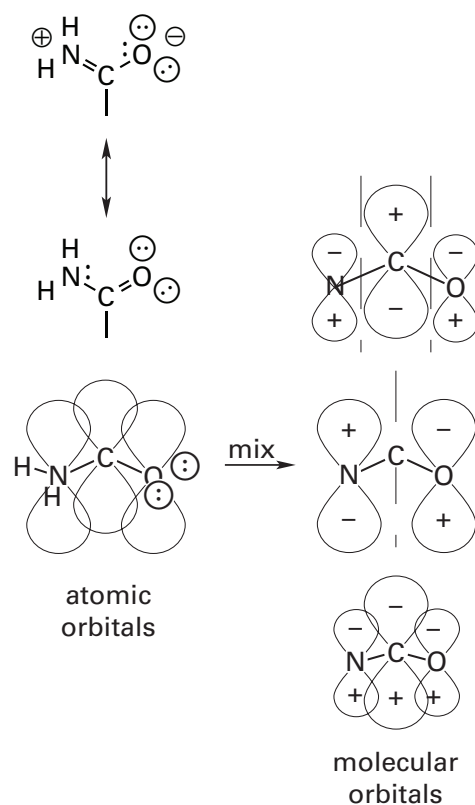


Figure 2-3: Electronic structure of an amide. In the upper left quadrant of the figure the two resonance structures for the amide are presented. The molecular orbital system in the center of the figure is formed from the linear combination of three p atomic orbitals: one from nitrogen, one from carbon, and one from oxygen. They are combined to produce three π molecular orbitals presented in order of increasing energy. To emphasize the symmetry of these π molecular orbitals, they have been drawn as if there were no coulomb effect. In the combination of lowest energy, all three of the constituent p orbitals overlap in phase. In linear π molecular orbital systems with an odd number, n , of atoms, the number of nodes in the central nonbonding molecular orbital is equal to $\frac{1}{2}(n - 1)$. To achieve a symmetric distribution of nodes in the central nonbonding molecular orbital, there are lobes on the two end atoms and nodes on every other atom in between, alternating lobe, node, lobe, node, and so forth. In the idealized three-atom system displayed here, there is a lobe at nitrogen, a node at carbon, and a lobe at oxygen in the central nonbonding π molecular orbital.

in the lowest bonding level would have half of their density distributed over carbon, one-quarter over oxygen, and one-quarter over nitrogen. The two electrons in the middle nonbonding level would have half of their density distributed over nitrogen and half over oxygen.

If the four π electrons were removed from the three atoms of the amide, the carbon and the oxygen would each have formal charges of +1 but the nitrogen would have a formal charge of +2, making it electron-deficient relative to the other two. If two pairs of π electrons occupy the two undistorted π molecular orbitals of lowest energy, oxygen would end up with a formal charge of $-\frac{1}{2}$; carbon, 0; and nitrogen, $+\frac{1}{2}$. This is the distribution of charge designated by the two resonance structures. Usually the resonance structures provide information about the distribution of electrons in the **highest occupied molecular orbital** or the distribution of electron deficiency in the **lowest unoccupied molecular orbital**. In the case of the amide, the resonance structures indicate that the pair of electrons in the highest occupied molecular orbital can occupy locations only over the nitrogen and the oxygen. This example illustrates the fact that resonance structures and molecular orbitals should agree in their assessment of electron distribution.

In Figure 2-3 and the description just presented, it was assumed that there was no coulomb effect; in other words, that all three atoms in the σ structure had the same electronegativity and formal charge. In a real amide, oxygen is the most electronegative atom and nitrogen is electron-deficient. The resulting coulomb effects cause the bonding molecular orbital of lowest energy to be skewed so that oxygen ends up with more electron density than carbon, rather than less, and the nonbonding molecular orbital of intermediate energy to be skewed so that nitrogen ends up with more electron density than oxygen. The node at carbon in the ideal nonbonding π molecular orbital shifts toward oxygen or toward nitrogen depending on whether or not the oxygen is hydrogen-bonded.²

Resonance theory has always incorporated the fact that the several structures drawn do not have independent existence, but occasionally, by mistake, it is implied that they do.³ In the extreme, the **double-headed arrow** of resonance becomes replaced with the two arrows of a chemical equilibrium, a mistake that engenders serious confusion.⁴ To avoid such confusion, a double-headed arrow should be used only to indicate resonance, never to indicate an equilibrium, and the two arrows of a chemical equilibrium should never be used to indicate resonance. That only one, undivided system of π molecular orbitals represents the resonance hybrid is a reaffirmation of the absence of independent existence. Unfortunately, while π molecular orbitals present a more accurate picture of the molecular structure and avoid the confusion with equilibrium, they do not have the accounting capability of formal resonance structures,

and each view, whether molecular orbitals or resonance structures, has its appropriate use.

The first decision that must be made about the electronic structure of any molecule is the location of all systems of π molecular orbitals. Any carbon, nitrogen, or oxygen that has contributed a $2p$ atomic orbital to a system of π molecular orbitals has only two other $2p$ atomic orbitals remaining to hybridize with its lone $2s$ atomic orbital, but any carbon, nitrogen, or oxygen that is not involved in a system of π molecular orbitals has three $2p$ atomic orbitals to hybridize with its $2s$ atomic orbital. It is these hybrids between s atomic orbitals and p atomic orbitals that overlap to form **σ bonds**. These σ bonds lie along the line of centers between the two respective atoms that they connect, and they are localized. Because they are localized, they are usually stronger covalent bonds than π bonds, and as a result every pair of atoms joined by one or more than one covalent bond must be joined by one σ bond. These σ bonds form the molecular skeleton defining the structure of the molecule, in particular its bond angles. This skeleton is the **σ structure** of the molecule. Each σ bond is also an occupied molecular orbital, but this realization is not informative in issues of molecular structure. In the particular instance of molecules in biological situations, when an atom has contributed one p orbital to a system of π molecular orbitals, it will almost always be hybridized [p , sp^2 , sp^2 , sp^2]. At that atom in the σ structure, the molecule will be planar, and the σ covalent bonds and σ lone pairs will radiate within that plane in three directions from the atom at approximately 120° angles. When an atom has not contributed a p orbital to a system of π molecular orbitals, it will almost always be hybridized [sp^3 , sp^3 , sp^3 , sp^3]. At that atom in the σ structure, σ covalent bonds or σ lone pairs will radiate in four directions tetrahedrally, at angles of approximately 109.5° .

Because the σ structure incorporates these bond orders and bond angles, it dictates the details of molecular structure. These details cannot be appreciated until decisions on hybridization can be made correctly. To pursue an earlier example, the oxygen, carbon, and nitrogen of an amide are each contributing a p orbital to the system of π molecular orbitals, and each is hybridized [p , sp^2 , sp^2 , sp^2]. In the σ structure each of these three atoms and all of the σ bonds and σ lone pairs of electrons radiating from them are in a plane, and each bond angle is approximately 120° (Figure 2-3).

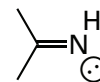
Lone pairs of electrons are identified by writing a Lewis structure of the molecule. Thereafter, it is conventional to ignore them, on the assumption that everyone knows that they are there. This assumption is somewhat vain; it seems to say that if you do not always realize that they are there, you are not someone. Because lone pairs of electrons are of paramount importance in biochemistry and because an understanding of a biologically important molecule is incomplete if ever they are forgotten, it is safer to include them explicitly in any structure

drawn, especially if they contribute to what is being discussed.

Because of electron repulsion, a lone pair of electrons on any oxygen or nitrogen unconjugated to a system of π molecular orbitals will occupy one of the sp^3 orbitals of that atom. This σ lone pair of electrons resides at one of the tetrahedral vertices of the atom. A **σ lone pair of electrons** is a lone pair confined to a single atom because it resides within a hybridized or unhybridized atomic orbital that does not overlap with any other atomic orbital from another atom. A σ lone pair of electrons is designated in a σ - π stereochemical representation by enclosing it within a circle (Figures 2-1B and 2-3) to symbolize its confinement. A **σ - π stereochemical representation** (Figure 2-1B) is a drawing of the molecule that indicates the bond angles and angles of σ lone pairs and distinguishes σ lone pairs of electrons from π lone pairs of electrons.

If an oxygen or nitrogen containing a lone pair of electrons is sterically able to rotate until that lone pair is parallel to an immediately adjacent system of π molecular orbitals and sterically able to rehybridize to sp^2 at its three remaining bonded positions, the lone pair of electrons is capable of entering the system of π molecular orbitals. For the lone pair of electrons to do this, the atom carrying it must rehybridize. This rehybridization requires sufficient energy to overcome the electron repulsion that originally placed the lone pair in an sp^3 orbital. The favorable energy resulting from the delocalization of the lone pair into the system of π molecular orbitals must exceed this deficit. If it does, the lone pair of electrons becomes a delocalized **π lone pair of electrons**, occupying a π molecular orbital spread over two or more atoms. It is so designated in a drawing by not enclosing it within a circle (Figures 2-1B and 2-3) to indicate its unconfinement.

When either oxygen or nitrogen has contributed only one of its p orbitals to a system of π molecular orbitals and is left with three valence orbitals, one $2s$ orbital and two $2p$ orbitals, it is usually assumed that they mix to form three sp^2 orbitals that lie together within a plane normal to the system of π molecular orbitals and are arrayed at 120° angles. If there are two or three covalent σ bonds to the heteroatom, the hybridization is usually [p , sp^2 , sp^2 , sp^2] because sp^2 orbitals provide maximum overlap in a σ bond. Thus a single lone pair left on a nitrogen that has contributed only one p orbital and one of its valence electrons to a system of π molecular orbitals and also participates in two σ bonds is always a σ lone pair in an sp^2 orbital, and it is designated as such by surrounding it with a circle. An example of such a lone pair is the lone pair on a nitrogen in an imine:



2-1

The situation becomes ambiguous, however, in the case of an oxygen that has contributed a p orbital and one valence electron to a system of π molecular orbitals, participates in one σ bond, and remains with two lone pairs of electrons. An example of such an oxygen would be an **acyl oxygen** or the oxygen of a carbonyl (Figure 2–4). The possibility arises that such an oxygen is hybridized [p, p, sp, sp]. In this case, one lone pair would occupy an sp orbital in line but opposite to the σ bond between the carbon and the oxygen, and the other lone pair would occupy a p orbital normal to both the π bond and the axis of the two sp orbitals (Figure 2–4A). Indeed, there is evidence from ultraviolet spectra and mass spectra of isolated carbonyl compounds that this occurs. The alternative possibility is that oxygen is hybridized [p, sp^2, sp^2, sp^2] and that both lone pairs are in sp^2 orbitals (Figure 2–4B). The decision between these two alternatives is not an insignificant one, for oxygens that have contributed one p orbital and one valence electron to a system of π molecular orbitals and participate in only one σ bond are by far the majority of the oxygen atoms in a molecule of protein. In a hydrogen-bonding environment, such as the water in which all biochemistry occurs, it appears that these oxygens place their two lone pairs in two sp^2 orbitals. This follows from the fact that, in crystallographic molecular models of small molecules in which an N–H forms a hydrogen bond with such a carbonyl or acyl oxygen, the nitrogen–hydrogen σ bond of the N–H usually points to the location where an sp^2 lone pair of electrons would be located.⁵ On the basis of this observation, it will be assumed that acyl oxygens are hybridized [p, sp^2, sp^2, sp^2], and their two lone pairs will both be designated as sp^2 by enclosing them in circles at 120° angles to the carbon–oxygen bond (Figure 2–1B).⁶ These are σ lone pairs of electrons, they lie within a plane shared with the carbon–oxygen σ bond and normal to the plane of the carbon–oxygen π bond (Figure 2–4B).

The σ structure of a molecule is the basic skeleton producing the σ bonds, the bond angles, and the fixed positions of the localized σ lone pairs. The π electrons are spread over this skeleton above and below the atoms contributing the p orbitals. Therefore, neither the bond angles of the molecule, which are defined by hybridiza-

tion, nor the positions of σ lone pairs, which are localized, can be affected by resonance. It necessarily follows that when one draws two or more resonance structures, one must make certain that the same σ structure is present in each resonance structure and that only the disposition of π electrons differs among them. The best way to ensure this is to draw a σ structure for the second resonance structure identical to the σ structure of the first resonance structure before putting in the π electrons, and to draw an identical σ structure for each of the successive resonance structures before putting in the π electrons. Always include all σ lone pairs oriented as the hybridization of each atom requires. After the set of valid resonance structures has been exhausted, look closely at any lone pair that did not participate and decide if it might not be a σ lone pair. If it is not completing an aromatic complement or being withdrawn by an adjacent π bond, it is probably a σ lone pair in a σ orbital confined to only the one atom.

When the atoms contributing the p orbitals to a system of π molecular orbitals form an unbroken ring rather than being branched or linearly arrayed, the possibility of aromaticity arises. In a continuous **ring of p orbitals** of any size, the energy levels of the individual π molecular orbitals are arrayed in a peculiar pattern. The π molecular orbital with the lowest energy is always the completely overlapping ring of p atomic orbitals in phase with no nodes other than the one at the nuclear plane. This π molecular orbital is occupied by two electrons. If coulomb effects were disregarded, the other bonding π molecular orbitals in the ring would always come in pairs that have identical energies. Because of Hund's rule, no such pair of orbitals can be filled with electrons to form a stable closed shell until four electrons have been provided simultaneously. These two features, the one continuous ring occupied by a pair of π electrons and the pairs of orbitals of higher energy occupied by quartets of π electrons, define an aromatic system of π molecular orbitals. An **aromatic π molecular orbital system** is an unbroken ring of parallel p orbitals occupied by 2, 6, 10, 14, or 18 π electrons.

From these rules it is clear that a phenyl ring is aromatic, but it is the **aromatic nitrogen heterocycles** such as pyridine and pyrrole

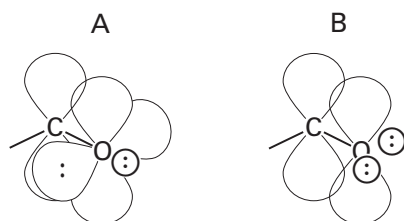
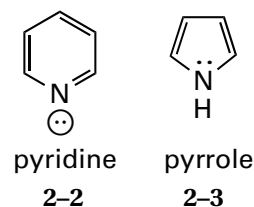


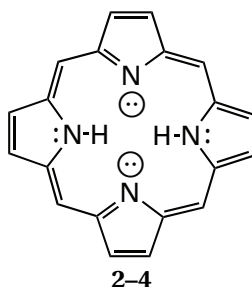
Figure 2–4: Two alternative hybridizations for an oxygen in a carbonyl. (A) One lone pair is in an sp orbital collinear with the carbon–oxygen bond, and the other is in the p orbital orthogonal to the double bond. (B) Both lone pairs are in sp^2 orbitals in the σ plane.



that are more interesting examples. Pyridine is a neutral, six-membered ring with one nitrogen. Each carbon contributes one valence electron to the π system, so nitrogen can contribute only one to complete the sextet of the

aromatic system. This leaves a neutral nitrogen with two remaining valence electrons that end up as a lone pair in the σ structure confined to an sp^2 orbital. Pyrrole, however, is a five-membered ring. Each carbon again contributes one valence electron to the system of π molecular orbitals, and nitrogen provides the two required to complete the sextet required for the aromatic system. Nitrogen is left with one valence electron and forms a covalent N–H bond to finish the neutral molecule. Pyridinyl and pyrrolyl nitrogens appear throughout aromatic heterocycles. A nitrogen can be identified as one or the other by whether one or two of its valence electrons are used to complete the aromatic system of 6, 10, 14, or 18 π electrons.

An interesting heterocycle that serves as an example of the application of these considerations is porphine:



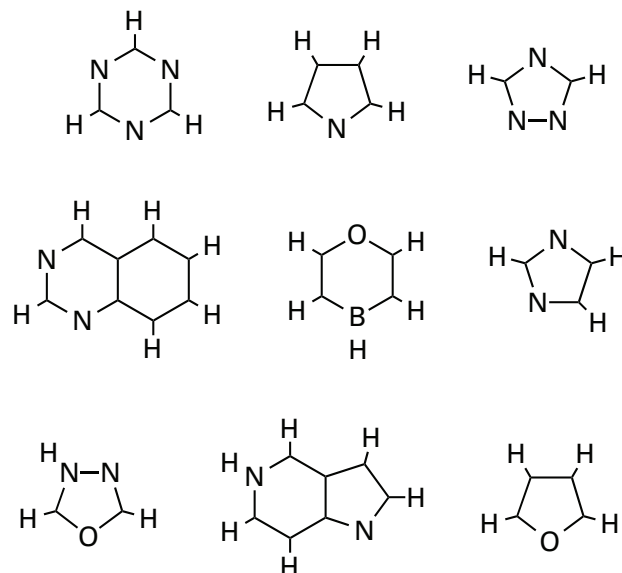
This is the simplest porphyrin; more elaborate porphyrins are components of the heme-containing coenzymes. Hidden within this molecule is an unbroken ring of 16 carbon atoms and two nitrogen atoms, each contributing a p orbital and creating a system of π molecular orbitals containing 18 π electrons. Therefore, porphine is aromatic. For this to be possible, two of the nitrogens must be pyridine nitrogens and each of them contributes a p orbital and one π electron. The other two nitrogens do not participate in the aromatic ring but nevertheless reside immediately adjacent to it and inside of it. They each retain a lone pair of electrons. Each of those lone pairs resembles the π lone pair on the nitrogen of aniline, but each is located endocyclically rather than exocyclically. This requires that each of them, as with the nitrogen in aniline, have three covalent σ bonds, hence the two central hydrogens.

Behind the distinctions among systems of π molecular orbitals, σ bonds, and σ lone pairs is the concept of **orthogonality**. Each system of π molecular orbitals, each σ bond, and each σ lone pair of electrons is orthogonal to every other system of π molecular orbitals, σ bond, and σ lone pair in the molecule. As such, to a first approximation, each is an independent moiety that does not share electrons with the others. It is this compartmental quality of bonding that permits each of these positions to be chemically distinct and have its own properties. It is this fact, rather than a desire to categorize, that renders these distinctions important. They must be clearly made in any drawing of the molecule.

The chemical properties of σ and π lone pairs of electrons are remarkably different. This difference is most clearly expressed in their behavior as bases, and it is the basicity of a lone pair of electrons that, in questionable cases, indicates whether it is a σ or π lone pair of electrons. When the basicity of the lone pair is relied upon as a criterion, a proton is being used to probe its availability. Lone pairs of electrons in π systems are far less basic than those in σ orbitals because σ lone pairs of electrons are localized and directionally oriented by the atomic orbital in which they are confined, whereas π lone pairs of electrons are delocalized and immersed within the system of π molecular orbitals.

Problem 2-1: Draw σ - π stereochemical structures as in Figure 2-1B for the *N*-acetyl α -amides of aspartate, asparagine, glutamine, proline, methionine, tyrosine, tryptophan, phenylalanine, histidine, and arginine.

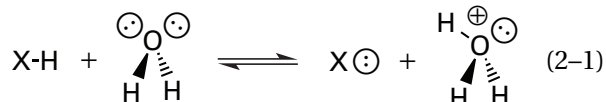
Problem 2-2: The following skeleton structures are various heterocycles. None is intended to be a radical, all have pairwise filled molecular and atomic orbitals, and none has a total of more than one positive or negative elementary charge. No π electrons or lone pairs of electrons are shown, and all atoms are shown. All of the compounds are aromatic.



- Decide how many π electrons each heterocycle contains to make it aromatic.
- Draw resonance forms that indicate the distribution of these π electrons. There may be only one.
- Complete the octet for every nitrogen and oxygen by adding lone pairs.
- Assign formal charges in each resonance form.
- Draw the σ structure of each heterocycle, including all lone pairs, and assign hybridization to each atom.

Acids and Bases

The quantitative measure of the basicity of a lone pair of electrons is the microscopic acid dissociation constant of its conjugate acid. In this way all lone pairs of electrons are related to the lone pair on a molecule of water. The reaction that defines a **microscopic acid dissociation** for a particular proton in a molecule is



The **central atom** in a microscopic acid dissociation is the atom directly bonded to the proton that dissociates.* The lone pair on the resulting conjugate base is usually localized on the central atom (as represented in Reaction 2-1) when it is oxygen, nitrogen, or sulfur but usually delocalized when it is carbon. In a microscopic acid dissociation, the **acid** is a position within the molecule from which a proton can dissociate to produce a lone pair of electrons, and the **base** is a lone pair of electrons with which a proton can associate. Because the reaction occurs in aqueous solution, a bare proton is transferred between the lone pair of the base and a lone pair on a molecule of water and back again. Every acid is always present in solution with a finite concentration of its conjugate base, and every base is always present in solution with a finite concentration of its conjugate acid.

The equilibrium constant for Reaction 2-1 is

$$K_{\text{eq}} = \frac{[\text{H}_3\text{O}^+][\text{X}]}{[\text{H}_2\text{O}][\text{HX}]} \quad (2-2)$$

Because $[\text{H}_2\text{O}] = 55 \text{ M}$ at all times, this term is passed to the left, and for convenience $[\text{H}_3\text{O}^+]$ is written as $[\text{H}^+]$.** These substitutions produce the definition of the microscopic acid dissociation constant:

$$K_a \equiv \frac{[\text{H}^+][\text{X}]}{[\text{HX}]} \quad (2-3)$$

A **microscopic acid dissociation constant** is the acid dissociation constant of a particular proton in a polyprotic acid. An acid dissociation constant is usually presented as a $\text{p}K_a$, where $\text{p}K_a \equiv -\log K_a$, solely for convenience. A theoretical justification of this practice is that the $\text{p}K_a$ is directly proportional to the change in free energy for Reaction 2-1. The larger the $\text{p}K_a$, the less likely is

Reaction 2-1 in the direction written and the more likely is it in the opposite direction. Because water is the same in all acid dissociations, the difference in $\text{p}K_a$ between two acids is proportional to the free energy for transferring a proton from the one acid to the conjugate base of the other. The smaller the $\text{p}K_a$, the more acidic is the acid and the less basic, or less available, is the lone pair of electrons on its conjugate base, and vice versa. There are several properties of the position from which the proton dissociates that affect the value of its microscopic $\text{p}K_a$.

The atomic number of the central atom from which the proton dissociates and on which the lone pair remains has a profound effect (Table 2-1). Within the same period of the periodic table, as **electronegativity** increases to the right, for example, carbon, nitrogen, oxygen, the atom is more capable of supporting the lone pair, and the acidity increases. Atoms in lower periods hold a lone pair of electrons in a larger atomic orbital, making it easier to support. For example, a proton on sulfur is more acidic than one on oxygen. Because a localized σ lone pair of electrons on carbon is such a strong base, the only time that there is a lone pair associated with carbon in biochemical situations is when it is a delocalized π lone pair of electrons. Because nitrogen and oxygen are more electronegative elements than carbon, delocalized π lone pairs of electrons associated with these elements are rarely bases in biochemical situations, and bases on these atoms are almost always localized σ lone pairs of electrons.

The **successive creation of negative elementary charge** on the same polyprotic acid causes each dissoci-

Table 2-1: Electronic Properties Affecting Values of the Acid Dissociation Constant

effect of identity of central atom on acidity ¹¹			
CH_4	$<$	NH_3	$<$
OH_2	$<$	SH_2	
$\text{p}K_a = 48$		$\text{p}K_a = 38$	$\text{p}K_a = 15.7$
		$\text{p}K_a = 7.0$	
effect of creation of charge on acidity ¹²			
PO_4H_3	$>$	PO_4H_2^-	$>$
PO_4H^{2-}			
$\text{p}K_a = 2.1$		$\text{p}K_a = 7.2$	$\text{p}K_a = 12.7$
$^+\text{H}_3\text{NCH}_2\text{CH}_2\text{NH}_2$	$<$	$^+\text{H}_3\text{NCH}_2\text{CH}_2\text{NH}_3^+$	
$\text{p}K_a = 9.98$		$\text{p}K_a = 7.52$	
effect of hybridization of the central atom on acidity ¹¹⁻¹³			
$\text{HC}\equiv\text{CH}$	$>$	$\text{H}_2\text{C}=\text{CH}_2$	$>$
H_3CCH_3			
$\text{p}K_a = 25$		$\text{p}K_a = 44$	$\text{p}K_a = 50$
$\text{HC}\equiv\text{NH}^+$	$>$	pyridine	$>$
H_3CNH_3^+			
$\text{p}K_a = -10$		$\text{p}K_a = 5.2$	$\text{p}K_a = 10.6$
$\text{CH}_3\text{HC}=\text{OH}^+$	$>$	$\text{CH}_3\text{CH}_2\text{OH}_2^+$	
$\text{p}K_a = -6$		$\text{p}K_a = -2$	
effect of induction on acidity ¹²			
$\text{CF}_3\text{CH}_2\text{OH}$	$>$	$\text{CHF}_2\text{CH}_2\text{OH}$	$>$
$\text{CH}_2\text{FCH}_2\text{OH}$	$>$	$\text{CH}_3\text{CH}_2\text{OH}$	
$\text{p}K_a = 12.4$		$\text{p}K_a = 13.1$	$\text{p}K_a = 14.2$
		$\text{p}K_a = 16.0$	
$\text{H}_5\text{C}_2\text{OOCCH}_2\text{NH}_3^+$	$>$	$\text{H}_5\text{C}_2\text{OOCCH}_2\text{H}_4\text{NH}_3^+$	$>$
$\text{H}_5\text{C}_2\text{OOCCH}_3\text{H}_6\text{NH}_3^+$			
$\text{p}K_a = 7.7$		$\text{p}K_a = 9.1$	$\text{p}K_a = 9.7$

* The central atom is not in the center of the molecule; it is just the atom that is central to the acid dissociation.

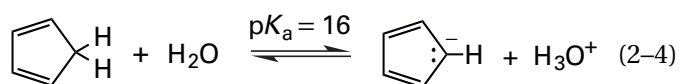
** In fact, the designation H_3O^+ is as misleading as H^+ because the dissociated proton in water is shared by four molecules of water^{7,8} as the cation H_9O_4^+ to 21 molecules of water^{9,10} as the cation $\text{H}_{43}\text{O}_{21}^+$.

ation of a proton to be more difficult than the previous one, and the successive creation of positive charge on the same polybasic molecule causes each association of a proton to be more difficult than the previous one (Table 2-1). The farther apart the charges that are created, however, the narrower are the increments in pK_a . For example, the difference in the two values of pK_a for 1,3-diaminopropane (1.98) is smaller than the difference in the two values of pK_a for 1,2-diaminopropane (2.87).

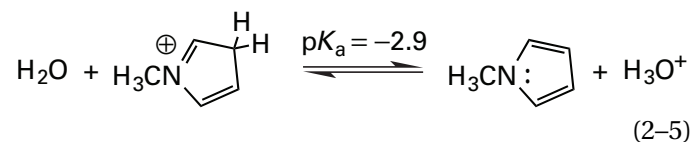
The **hybridization** of the central atom affects its pK_a considerably (Table 2-1). All localized σ lone pairs of electrons are in hybrid orbitals formed by mixing one s orbital with one, two, or three p orbitals, as indicated by the designations sp , sp^2 , and sp^3 , respectively. The fewer the number of p orbitals in the mixture, the greater the fraction of the s orbital distributed into each hybrid and the more s character the hybrid will have. The more s character there is to the orbital, the closer the lone pair of electrons is held next to the nucleus, the less extension of electron density along any particular axis will be displayed, and the less basic will be the orbital.

Electronegative or electropositive atoms adjacent to the central atom also have a significant but less remarkable effect on acidity (Table 2-1). These withdraw or donate electrons by **induction** through σ bonds and decrease or increase the basicity of the lone pair of electrons accordingly.

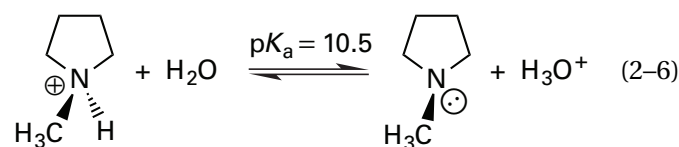
Because the conjugate acid has a single, σ covalent bond between the central atom and the hydrogen, the pair of electrons that has been protonated in its creation must already be or must become σ electrons. If they were a σ lone pair of electrons before the proton was added, **rehybridization** of the central atom is not usually involved in the protonation. If, however, they were a π lone pair of electrons prior to the protonation, the favorable delocalization energy that they gained within the π system is eliminated as the σ bond is formed. The greater this delocalization energy, the more free energy will be required to protonate the lone pair of electrons and, because this free energy is supplied by the concentration of protons, the smaller will be the pK_a of the conjugate acid. It is this fact that causes the pK_a of the conjugate acid to register the **degree of delocalization** of a pair of π electrons or in other words the strength of their π character. This ability is most clearly exemplified in the protonation of aromatic compounds where the lone pair of electrons is delocalized over the aromatic ring. For example, cyclopentadiene is a strong carbon acid¹¹



compared to propane ($pK_a = 50$)¹¹ because protonation of the lone pair of electrons on the conjugate base of the former destroys the aromaticity of the anion. Likewise, *N*-methylpyrrole is a weak base¹²



compared to *N*-methylazacyclopentane¹²



because protonation of the former destroys its aromaticity. In this case, the comparison is a minimum estimate of the difference in pK_a resulting from aromatic delocalization because pyrrole protonates on carbon rather than on nitrogen. The pK_a for the conjugate acid protonated on nitrogen must be much lower than -2.9 .

There are lone pairs of electrons in nonaromatic configurations, the acid-base behavior of which reveals the degree of their π character. The pK_a associated with the lone pair of electrons on aniline ($pK_a = 4.6$)¹² can be compared with the pK_a associated with the lone pair of electrons on cyclohexylamine ($pK_a = 10.6$).¹² The significant difference in basicity demonstrates that the lone pair in aniline is a π lone pair of electrons conjugated to the neighboring π system of the phenyl ring. The pK_a ($pK_a = -6$)¹⁴ associated with the lone pair of electrons on the nitrogen in an amide (Figure 2-3) is even lower than that of the lone pair on aniline and indicates that it is even more delocalized. This is not surprising since the phenyl ring of aniline is otherwise involved in its own aromaticity and the oxygen of the amide has a strong coulomb effect in the lowest occupied molecular orbital.²

Other examples of the use of a proton to evaluate the π character of a lone pair of electrons occur in carbon acids. The pK_a of a methyl group in propene ($pK_a = 43$)¹¹ is much lower than that in propane ($pK_a = 50$)¹¹ because the lone pair produced upon the dissociation of a proton from propene conjugates with the neighboring π system of the alkene. The analogous lone pair on carbon in the conjugate base of acetaldehyde ($pK_a = 17.6$)¹⁵ is even less basic because, as occurs with an amide, an oxygen is located two atoms away and exerts a strong coulomb effect. When the π system of the conjugate base is extended from three to five atoms in length, as in the conjugate base of 2,4-dioxopentane ($pK_a = 9$), the lone pair of electrons in the conjugate base becomes even more delocalized and less accessible to protonation.

In making such comparisons, care must be taken to avoid confounding the reasons for the changes in pK_a . A common confusion is that between the effects of hybridization and conjugation. One reason that the acetate anion, the conjugate base of acetic acid ($pK_a = 4.75$),¹² is a weak base compared to the ethoxide

anion, the conjugate base of ethanol ($pK_a = 16$),¹² is that the basic lone pairs of electrons in the acetate anion are hybridized sp^2 (Figure 2-5) rather than sp^3 . The system of π molecular orbitals of the acetate anion, composed of four π electrons in a three-atom system (Figure 2-3), does not provide a pair of electrons to be protonated, notwithstanding any drawing suggesting this to be the case. It is a σ lone pair of electrons orthogonal to the system of π molecular orbitals that is protonated, and acetate anion cannot be used as an example of the decrease in basicity that results when the lone pair of electrons created upon the departure of the proton is conjugated to a π system.

There is an indirect effect of conjugation on the acidity of a carboxylic acid such as acetic acid. When one of the σ lone pairs on the acetate anion is protonated or alkylated, the functional group is no longer symmetric, and the oxygen that has been so modified becomes more electronegative. This change withdraws more electron density onto the protonated or alkylated oxygen, as indicated by the resonance structures:

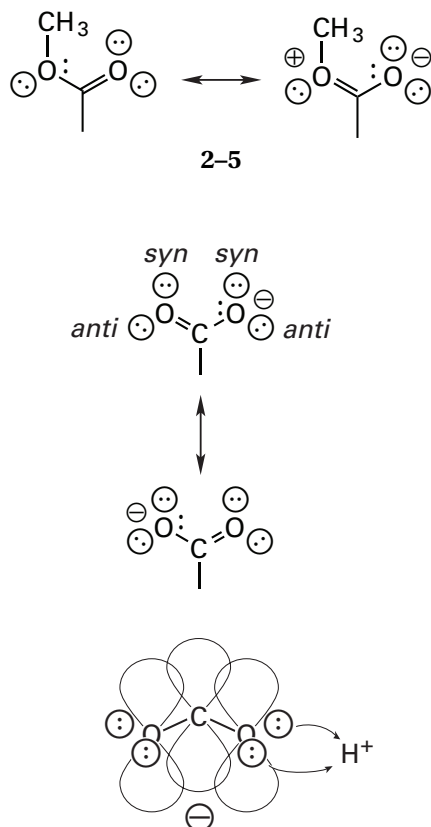
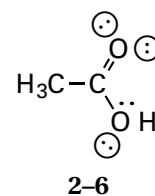


Figure 2-5: Resonance structures (top) and atomic orbital overlap (bottom) in a carboxylate anion. The resonance structures indicate that one pair of electrons from each oxygen is delocalized and that the π molecular orbital system results from the overlap of three p atomic orbitals, one from the central carbon and one from each of the two oxygens. Each of these three atoms is hybridized [p , sp^2 , sp^2 , sp^2]. As indicated schematically, the two σ lone pairs of electrons in the molecular plane are the bases that can associate with a proton, not the delocalized lone pairs of the π molecular orbital system. The *syn* and *anti* lone pairs are labeled.

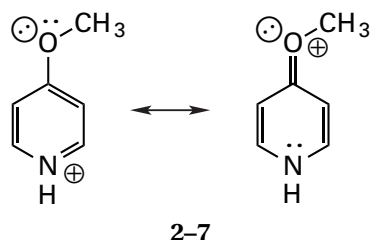
The separation of charge in the structure on the right is the reason that the lone pair of electrons on the alkylated oxygen or protonated oxygen is less delocalized than the π lone pair of electrons on an unalkylated or unprotonated oxygen in the carboxylate anion. Nevertheless, the bond between the protonated or alkylated oxygen and the acyl carbon retains some of the double-bond character indicated by the less advantageous form on the right. This is manifested in the almost 120° angle (116.5°) between an alkyl carbon and an acyl carbon at the oxygen of an ester and a shortening of the bond between the oxygen of an ester and the acyl carbon by 0.09 nm relative to carbon–oxygen bonds between sp^2 carbons and oxygens in aryl and vinyl compounds.¹⁶ Therefore, an ester or the conjugate acid of a carboxylic acid retains the overlap of the system of π molecular orbitals, but the overlap is considerably weakened relative to the unalkylated or unprotonated anion. During protonation of a carboxylate anion, the delocalization in the orthogonal π system is considerably diminished, and this effect destabilizes the conjugate acid and lowers the pK_a . A similar but less pronounced effect of a decrease in delocalization upon protonation occurs with phenol. In the case of phenol, the conjugation in the anionic conjugate base is weaker than that in the anionic conjugate base of a carboxylic acid because the elementary negative charge is distributed over the oxygen and three carbons. Consequently, the effect of diminishing this conjugation upon protonation is less, and phenol is a weaker acid than acetic acid.

The acetate anion illustrates another property of a system of π molecular orbitals—its ability to **redistribute charge**. The elementary negative charge in the acetate anion is shared between the two oxygens because the system of π molecular orbitals is spread over all three atoms. The two electrons in the highest occupied molecular orbital, which account for the negative charge of the functional group, can reside only over the two oxygens, as there is a node over carbon (Figure 2-3). When one of the oxygens becomes protonated, the π system redistributes and more π electron density is shifted over the oxygen that has become protonated because its coulomb effect has increased. This shift in distribution of charge is reflected in the resonance structure chosen for portraying the conjugate acid:

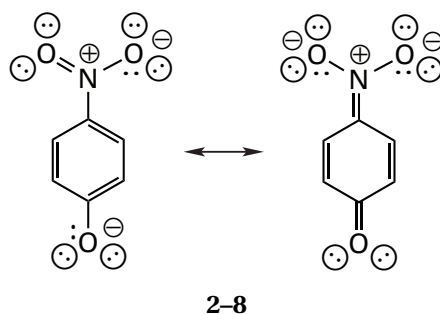


A similar ability of a π system to redistribute charge is reflected in the lower acidity of *p*-methoxypyridinium ($pK_a = 6.67$) compared to pyridinium ($pK_a = 5.17$). This difference results from the ability of the electron density

of the methoxy substituent to push into the π system through the conjugation represented by the resonance structure



so that the elementary positive charge on the nitrogen is delocalized. In the opposite sense, an example of a shift of electron density away from the central atom occurs in the *p*-nitrophenolate anion, whose associated pK_a is 7.2, compared to the phenolate anion, whose associated pK_a is 10.0. This can be explained by the resonance structure

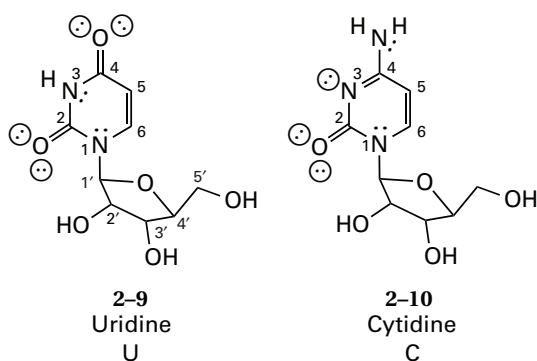


In each of these examples, the redistribution of charge among electronegative atoms is accomplished by the highest occupied molecular orbital of the π system, which is spread over the whole molecule.

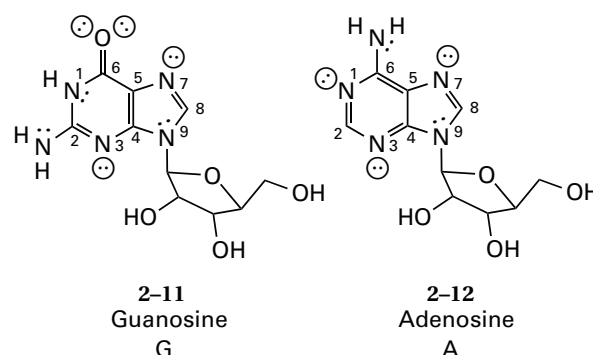
The microscopic pK_a of an acid–base is determined by a combination of all of these properties: the electronegativity and hybridization of the central atom, any creation of charge, the inductive effect, any delocalization of the lone pair of electrons, and any redistribution of charge.

The biological molecules that illustrate most extensively the various aspects of bonding and acid–bases discussed so far are the **bases of the nucleosides**.

Pyrimidines



Purines



Each of these nucleosides, **uridine**, **cytidine**, **guanosine**, and **adenosine**, is composed of the base itself, **uracil**, **cytosine**, **guanine**, and **adenine**, respectively, and a ribosyl group attached to N1 or N9 of that base. The nucleoside bases are hybrid structures of aromatic heterocycles and amides. The most aromatic base is adenine. It is susceptible to electrophilic aromatic substitution at carbons 2 and 8 but is also susceptible to nucleophilic substitution at carbon 6 in reactions that resemble acyl exchange. The most amidic base is uracil. It is unambiguously an *N*-acyl-*N'*-alkenyl-*N'*-ribosylurea. The carbon–carbon double bond in uracil has almost olefinic character. It is susceptible to addition reactions, unlike the system of π molecular orbitals in an aromatic compound, which would be susceptible only to substitution.

The nucleoside bases in adenosine, guanosine, and cytidine have exocyclic nitrogens resembling the nitrogen in aniline. The lone pairs of electrons on these nitrogens are even more delocalized than the one on the nitrogen in aniline ($pK_a = 4.6$) because the pK_a for the conjugate acid of each of these nitrogens in the nucleoside bases¹⁷ is less than or equal to -2 , similar to that for *N*-protonated urea ($pK_a \leq -4$). Therefore, each of these exocyclic nitrogens is planar and trigonal, as is always depicted in drawings of the base pairs.

The nucleoside bases in uridine, cytidine, and guanosine have exocyclic oxygens resembling the oxygen in an amide. The values of pK_a for the conjugate acids of these exocyclic oxygens are 0.5, <4.2 , and <1.6 , the upper limits being the values of pK_a for the *N*-protonated tautomers. These values can be compared to -0.7 , the pK_a for the oxygen of acetamide.¹⁸ The values of pK_a for these oxygens in the iminolic tautomers of these three bases, estimated from the measured¹⁹ or theoretical^{20,21} values for the equilibrium constants between the amidic and iminolic tautomers, are 5, 4, and -3 for uridine, cytidine, and guanosine, well below the value of 10 for the pK_a of phenol. These values of pK_a as well as the fact that the iminol tautomers are far less stable than the amidic tautomers are the justification for depicting these oxygens as acyl oxygens.

There are two types of calculations performed with acid–bases. The pH of a solution to which a weak acid or

weak base has been added can in theory be calculated, and the ratio of the molar concentrations of conjugate acid and conjugate base in a solution of a given pH can in practice be calculated.

The calculation of the **pH of a solution** upon the addition of an acid–base is an exercise in simultaneous equations. The problem takes the form “Calculate the pH of a solution to which 0.1 mol of sodium acetate has been added for every liter.” The equations always used are the **conservation of mass**

$$[\text{HOAc}] + [\text{OAc}^-] = 0.1 \text{ M} \quad (2-7)$$

where HOAc is acetic acid and OAc^- is acetate anion; the **conservation of charge**

$$[\text{OAc}^-] + [\text{OH}^-] = [\text{Na}^+] + [\text{H}^+] \quad (2-8)$$

where $[\text{Na}^+] = 0.1 \text{ M}$; the **acid dissociation constant** or constants

$$K_a = \frac{[\text{OAc}^-][\text{H}^+]}{[\text{HOAc}]} \text{ M} \quad (2-9)$$

where $\text{p}K_a = 4.75$ and $K_a = 1.78 \times 10^{-5} \text{ M}$; and the **water constant**

$$[\text{H}^+][\text{OH}^-] = 10^{-14} \text{ M}^2 \quad (2-10)$$

These comprise four—or if necessary more, depending on the number of dissociation constants the acid has— independent simultaneous equations with four, or if necessary more, unknowns. In the case of acetate, the four unknowns are $[\text{H}^+]$, $[\text{OH}^-]$, $[\text{OAc}^-]$, and $[\text{HOAc}]$. These four equations with four unknowns can be readily solved for $[\text{H}^+]$ ($1.33 \times 10^{-9} \text{ M}$) if the assumption is made that $[\text{H}^+]$ in Equation 2–10 is negligible relative to the other terms.

The value of this exercise is that the creation of the simultaneous equations and the cancellation of certain terms to avoid a cubic or quadratic equation requires an understanding of the acid–base chemistry that is occurring in the solution. For example, one is required to know that the only ions that can be present are H^+ , Na^+ , OAc^- , and OH^- and that sodium acetate is a base so the concentration of protons in the final solution will be small.

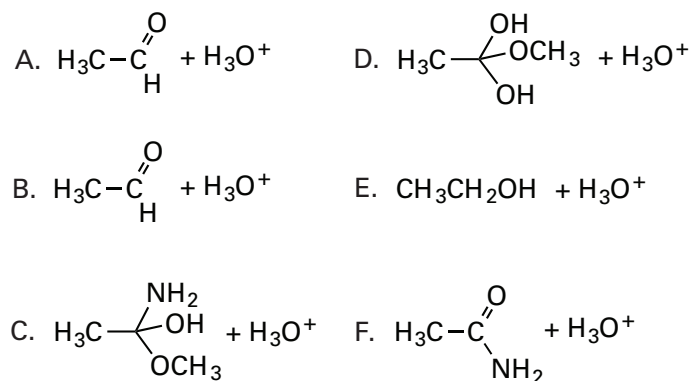
The calculation of the concentrations of a conjugate acid and its conjugate base at a given pH fulfills one of two purposes. First, if the solution contains an acid–base of experimental interest, such as an acid–base in a molecule of protein, this calculation will provide the molar concentrations of the conjugate acid and the conjugate base of that acid–base. Second, if a particular buffer is used to stabilize the pH at a particular value, this calculation can be used to determine the concentrations of conjugate acid

and conjugate base required for the buffer. A **buffer** is a solution of an acid and its conjugate base, both present at high enough concentrations so that the acid can neutralize bases added to the solution and the base can neutralize acids added to the solutions, and between them they can keep the pH of the solution constant.

A problem concerning a buffer can be stated “Calculate the number of moles of acetic acid and sodium acetate that are present in 2.00 L of an 0.1 M solution of acetate plus acetic acid at pH 5.5.” This problem requires only two simultaneous equations, Equations 2–7 and 2–9. Because $[\text{H}^+]$ is given as $3.16 \times 10^{-6} \text{ M}$, there are only two unknowns, $[\text{HOAc}]$ and $[\text{OAc}^-]$, which are 0.015 and 0.085 M, respectively. The answer is 0.03 mol of acetic acid and 0.17 mol of sodium acetate.

The quantitative behavior of the concentrations of the conjugate acid and conjugate base of each acid–base is described by a **titration curve** (Figure 2–6) that relates the fraction of the acid–base in the form of the conjugate acid or in the form of the conjugate base to the pH of the solution. This can be presented as the fraction itself (Figure 2–6A), as is usually done, but this presentation leaves the erroneous impression that the fraction of acid goes to zero about 2 pH units above the $\text{p}K_a$ and the fraction of base goes to zero about 2 pH units below the $\text{p}K_a$. This misimpression is corrected by examining the logarithms of the fractions as a function of pH (Figure 2–6B). It can be seen that finite fractions of both acid and base are still present at high and low pH, respectively. At a distance of 2 pH units above the $\text{p}K_a$, 1% of the acid–base is in the form of the acid, and this percentage drops off by a factor of 10 for every rise of unit of pH but never reaches zero. The importance of this point is that often only one species of the acid–base participates in a chemical reaction, yet the reaction will occur quite well at a pH where the reactive species is present at only 1% or 0.1% or 0.01% or less of the total acid–base. Protonation and deprotonation are extremely rapid, and as the minor but reactive species is consumed in the reaction, it is continuously replaced.

Problem 2–3: Complete the following acid–base equilibria. Draw the structures of the conjugate base and the acid in σ – π stereochemical representation (Figure 2–1B).



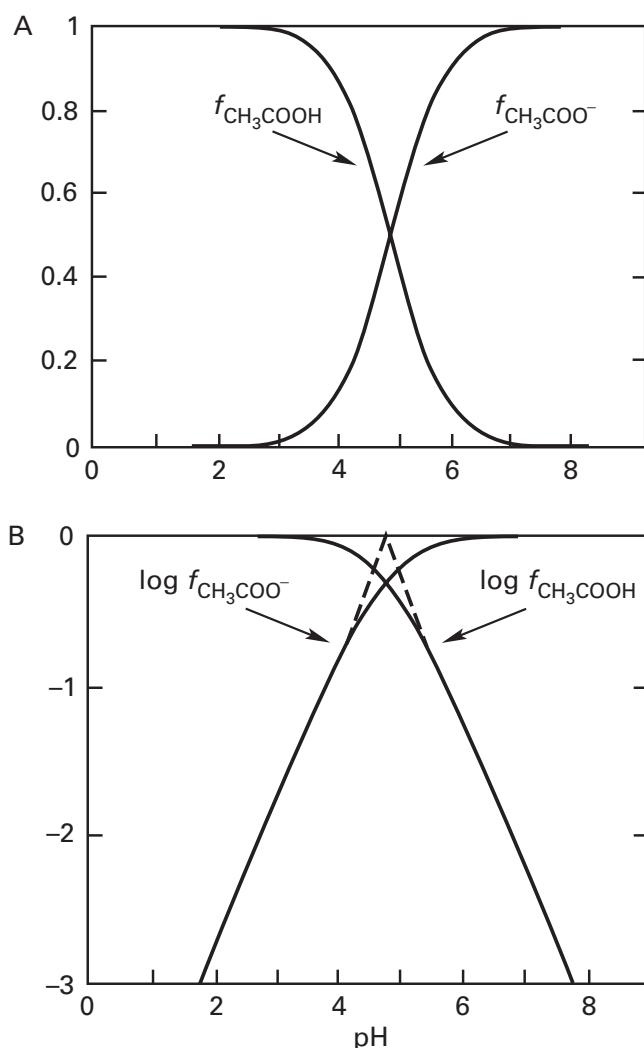


Figure 2-6: Titration curves for acetic acid. From the acid dissociation constant for acetic acid ($pK_a = 4.75$), the fraction of the acid–base present in solution as the conjugate acid ($f_{\text{CH}_3\text{COOH}}$) or as the conjugate base ($f_{\text{CH}_3\text{COO}^-}$) as a function of pH, or the logarithms of these values (B) can be plotted as a function of pH. When the pH is greater than the pK_a by 2 units, the concentration of acetic acid decreases by a factor of 10 for each increase in pH of 1 unit. When the pH is less than the pK_a by 2 units, the concentration of acetate ion decreases by a factor of 10 for each decrease in pH of 1 unit.

Problem 2-4: Write the acid–base equilibrium to which each of the following values of pK_a refer. Write each as a chemical reaction, and draw out the structures of the conjugate acid and the base in σ – π stereochemical representation. In tables¹² of values of pK_a , the name of the compound and the value of the pK_a are all that one is given, so it will be necessary for you to compare each of these values of the pK_a with those for acids and bases about which you are certain to judge whether the pK_a is for the molecule as named or for one of its conjugate bases or one of its conjugate acids.

compound	pK_a
1-amino-2-bromoethane	8.49
1-aminoethane	10.56
2,2,2-trifluoroethanol	12.43
ethanethiol	10.50
3-hydroxypropyne	13.55
diethylamine	10.98
ethanol	–2, 16
2-hydroxyethanethiol	9.5
2-aminoethanethiol	8.6, 10.75
2-chloroethanol	14.31
morpholine	8.36
2,2-dichloroethanol	12.89
diallylmethylamine	8.79
diethyl ether	–3.5
1-aminobutane	10.59
2-hydroxyethanamine	9.50
allylmethylamine	10.11
pyrrolidine	11.27
piperidine	11.22
piperazine	5.68, 9.82
pyridine	5.14
imidazole	7.05, 14.52
pyrimidine	1.10
isoquinoline	5.14
pyrazole	2.48
aniline	4.62
<i>o</i> -chloroaniline	2.62
<i>m</i> -chloroaniline	3.32
<i>p</i> -chloroaniline	3.81
<i>p</i> -methylaniline	5.07
<i>p</i> -methoxyaniline	5.29
<i>p</i> -nitroaniline	1.02
phenol	9.95
<i>p</i> -(trimethylammonio)phenol	8.0
<i>o</i> -chlorophenol	8.48
<i>m</i> -chlorophenol	9.02
<i>p</i> -chlorophenol	9.38
<i>p</i> -methylphenol	10.19
<i>p</i> -methoxyphenol	10.20
<i>p</i> -nitrophenol	7.14
2-aminobutanoic acid	2.27, 9.68
<i>N</i> -ethylmorpholine	7.70
1-aminonaphthalene	3.40
2-thioethanesulfonate	7.5
ethyl acetate	25
1-chloro-2-propanone	16.5
$\text{CH}_3\text{COCH}(\text{C}_2\text{H}_5)\text{CO}_2\text{C}_2\text{H}_5$	12.7
nicotine	3.13, 8.02
<i>p</i> -hydroxyaniline	5.50, 10.30
1-amino-2,2,2-trifluoroethane	5.7
$\text{CH}_3\text{C}(\text{NH})\text{NH}_2$	12.52
trichloroacetic acid	0.65
fumaric acid	3.03, 4.52
thiazole	2.44
methoxyacetic acid	3.53
thiourea	–0.96

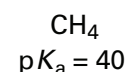
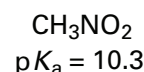
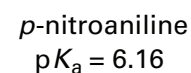
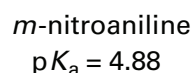
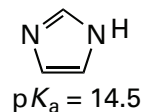
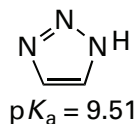
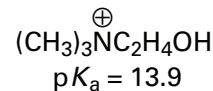
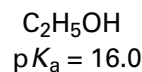
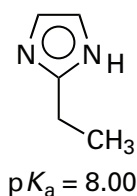
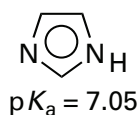
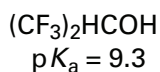
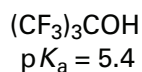
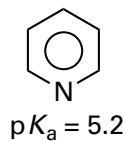
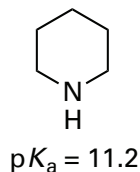
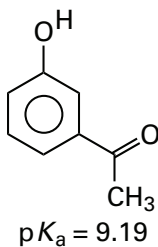
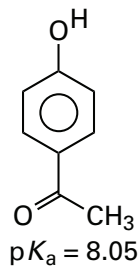
Problem 2-5: The following zwitterionic acid–bases are used widely for buffering solutions of protein.²² Write the full structures of both the acid and the conjugate base for the acid–base equilibrium to which the pK_a refers. In what range of pH would each of these acid–bases buffer?

buffer	p <i>K</i> _a
<i>N</i> -(2-sulfoethyl)morpholine (MES)	6.1
1,4-bis(2-sulfoethyl)piperazine (PIPES)	6.8
<i>N,N</i> -bis(2-hydroxyethyl)-2-aminoethanesulfonic acid (BES)	7.1
<i>N</i> -(3-sulfopropyl)morpholine (MOPS)	7.2
<i>N</i> -(2-sulfoethyl)-2-amino-1,3-dihydroxy-2-hydroxymethylpropane (TES)	7.4
1-(2-hydroxyethyl)-4-(3-sulfoethyl)piperazine (HEPES)	7.5
1-(2-hydroxyethyl)-4-(3-sulfopropyl)piperazine (EPPS)	8.0
<i>N</i> -[2-hydroxy-1,1-bis(hydroxymethyl)ethyl]glycine (Tricine)	8.1
<i>N,N</i> -bis(2-hydroxyethyl)glycine (Bicine)	8.3
<i>N</i> -(3-sulfopropyl)-2-amino-1,3-dihydroxy-2-hydroxymethylpropane (TAPS)	8.4
<i>N</i> -(2-sulfoethyl)cyclohexylamine (CHES)	9.3

Why is HEPES more acidic than EPPS?

Problem 2-6: From the following list, select the reason for the difference in p*K*_a between the two molecules in each pair presented below.

- (A) hybridization
- (B) electronegativity
- (C) π donation
- (D) σ donation
- (E) π withdrawal
- (F) σ withdrawal-induction
- (G) aromaticity



Problem 2-7: What two effects in combination cause the p*K*_a of the methyl ester of 2-methoxypropenoic acid (−3.37) to be 0.9 unit less than that of dimethylether (−2.5)?^{13,23}

Problem 2-8: What are the exact pHs of the following solutions?

- 10^{−2} M acetic acid
- 10^{−2} M imidazolium acetate
- 5 × 10^{−2} M sodium dihydrogen phosphate
- 5 × 10^{−2} M aniline
- 10^{−3} M pyridinium chloride
- 10^{−2} M *p*-nitroanilinium chloride
- 10^{−2} M morpholine
- 5 × 10^{−2} M sodium 2,2-difluoroethoxide

Problem 2-9: Calculate the concentration of imidazolate anion in a 0.1 M solution of imidazole at pH 9.52.

Problem 2-10: Determine the molar concentrations of each species of the weak acids and weak bases in the following solutions.

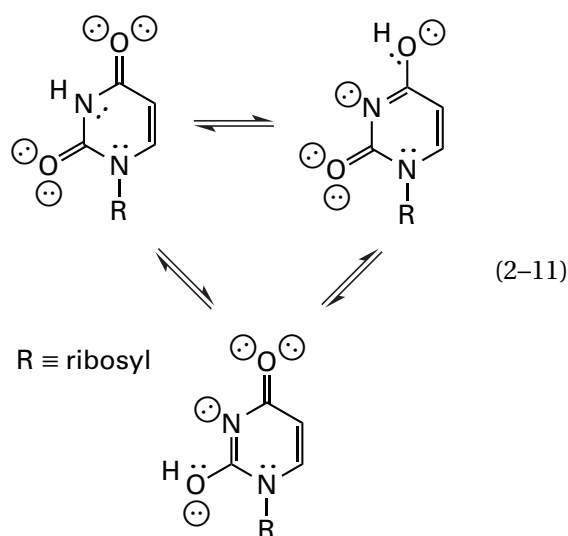
solute and concentration	pH
0.4 M 1-aminobutane	6.5
0.2 M 1-aminobutane	11.0
0.05 M <i>p</i> -chlorophenol	12.0
0.01 M <i>p</i> -chlorophenol	7.3
0.01 M <i>p</i> -methylaniline	5.0
0.001 M <i>p</i> -methylaniline	2.0
0.03 M 2-aminoethanethiol	9.2
0.08 M 2-aminoethanethiol	5.0
0.05 M morpholine	3.5
0.002 M piperazine	7.5
0.03 M ethanol	6.4
0.03 M diethyl ether	8.0
0.03 M 3-hydroxypropyne	4.0

Problem 2-11: From the following information calculate the pH of the final solutions.

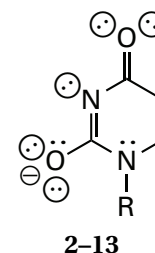
buffer species	concentration of buffer (M)	initial pH	amount of NaOH added (mol L ⁻¹)
imidazole	0.1	6.70	0.02
imidazole	0.03	6.50	0.02
phosphate	0.01	6.80	0.005
phosphate	1.0	6.35	0.1
borate	0.2	9.50	0.002
borate	0.15	8.40	0.05
imidazole	0.1	6.50	0.01
imidazole	0.05	7.00	0.02
phosphate	0.2	7.20	0.05
phosphate	0.3	6.20	0.15
borate	0.05	9.40	0.001
borate	0.02	8.60	0.01

Tautomers

One isomer is a **tautomer** of another isomer if the only difference between them is the position of a proton. There are several tautomers of uridine (2-9):

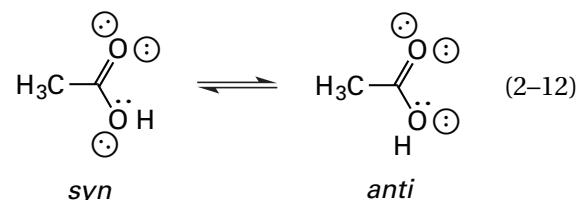


Each of these three tautomers of uridine is a distinct molecule with distinct chemical properties. It can be converted to another tautomer in the set by the removal of its acidic hydrogen by a base in solution, almost always a molecule of water (Reaction 2-1), and the readdition of another proton in the solution to another lone pair of electrons in the conjugate base. None of the interconversions between two of these tautomers can result from the intramolecular transfer of a proton because none of the lone pairs of electrons in the molecules is disposed properly for such an intramolecular transfer. Because they are acid-base reactions and because the conjugate base common to all of them



is a stable, anionic molecule, these interconversions are rapid. In water, the tautomer of uridine that is normally written, the one in which the proton occupies the nitrogen, is the dominant one, exceeding in concentration the other two combined by a factor of more than 4000.¹⁹

By formal definition, two otherwise identical isomers are tautomers of each other only when the tautomeric proton sits on two different atoms in the two isomers, as in the case of the three tautomers of uridine in Equation 2-11. If the proton sits on two different lone pairs on the same atom, the two isomers are, by formal definition, **conformational isomers** of each other. An example of two such conformational isomers would be the *syn* and *anti* conformations of acetic acid:



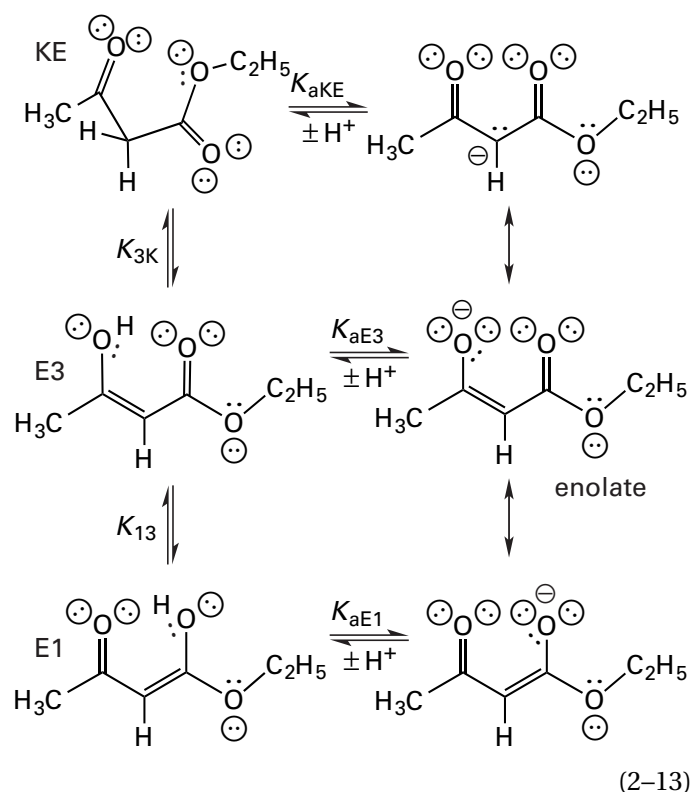
Because the barrier to rotation around the carbon-oxygen bond is large²⁴ due to the conjugation in the acid (2-5) and because protons shuttle on and off the oxygens rapidly due to the aqueous solution, neither of these two conformational isomers probably exists long enough to convert to the other by rotation about the carbon-oxygen bond. If this is the case, each of these isomers over its lifetime is a distinct molecule, each has distinct chemical properties, each is almost always converted to the other isomer only by the removal of its acidic hydrogen by a base in solution and the readdition of a proton to one of the two lone pairs of electrons that have the opposite orientation in the acetate anion (Figure 2-5), and the two isomers are in fact, if not by definition, tautomers of each other.

In the gas phase, the *syn* isomer of acetic acid is about 20 kJ mol⁻¹ more stable than the *anti* isomer.²⁵⁻²⁸ This difference in stability results from steric repulsion between the methyl group and the proton,^{25,27} from the repulsion between the dipole of the carbon-oxygen double bond and the dipole of the oxygen-hydrogen bond in the *anti* conformation, and from the fact that the unfavorable electron repulsion between the two *syn* lone pairs of electrons (Figure 2-5) is relieved when one of them is protonated but not when an *anti* lone pair is protonated.²⁹ Although the high relative permittivity of

water should damp both the dipolar repulsion and the electron repulsion,³⁰ it has been proposed that the difference in stability between these two tautomers is the same in water as in the gas phase.³¹ If this were the case, the microscopic acid dissociation constant for the *anti* isomer should be about 3000 times larger than that for the *syn* isomer; or in other words, the *syn* lone pairs of electrons should be 3000 times more basic than the *anti*. There is, however, experimental evidence suggesting that in water the difference in basicity between the *syn* and *anti* lone pairs is much less significant.³²

If the rotation about the carbon-oxygen bond in each of the two tautomers of uridine in which the oxygens are protonated (Equation 2-11) is also sufficiently hindered that neither interconverts significantly by rotation around the carbon-oxygen bond during its lifetime, then there would be *syn* and *anti* conformations of each of them that would be in fact two tautomers of each of them. In this case, the five actual tautomers of uridine would be the five molecules resulting from the protonation in turn of the five respective σ lone pairs on anion 2-13. As the protons shuffle, the σ structure of the uridine remains constant, and a proton is simply found on a different σ lone pair of electrons.

In some sets of tautomers, however, **rehybridization** of the atoms in the acid-base occurs during tautomerization. Such rehybridization is required to take place when one of the lone pairs that is protonated is a π lone pair in the intermediate base. The usually cited example of this is that of the keto and enol tautomers of a carbonyl compound such as ethyl acetoacetate:

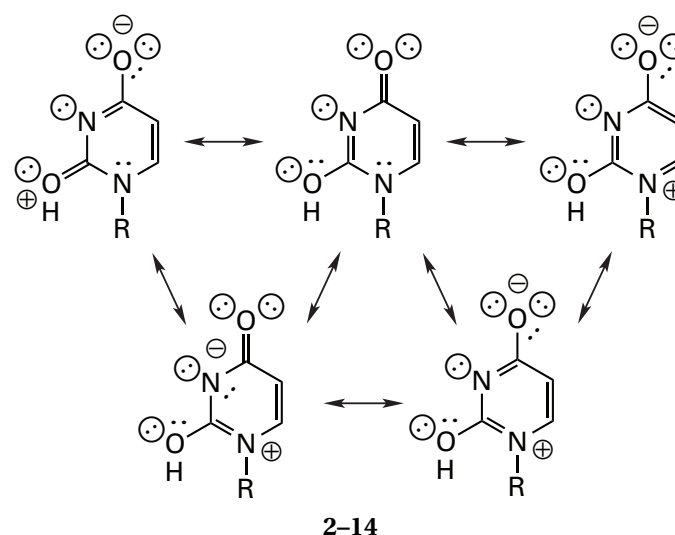


The most stable form of the acid, the ketoester (KE), is in equilibrium with two tautomers, the enol at carbon 3 (E3) and the enol at carbon 1 (E1). The common conjugate base of all three of these tautomers is the enolate anion (enolate). The enolate anion has a five-atom system of π molecular orbitals, and each of the five atoms of the system is hybridized [p , sp^2 , sp^2 , sp^2]. Two of the six π electrons of the anion, however, must be protonated at carbon to form the ketoester, an event requiring rehybridization at that central carbon.

In the case of the two enols of ethyl acetoacetate, in contrast to the tautomers of acetic acid or the tautomers of uridine, the proton can be readily transferred intramolecularly between the two oxygens. In fact, in either enol the proton forms a hydrogen bond to the adjacent carbonyl oxygen. These comparisons illustrate the specific geometric requirements for **intramolecular proton transfer**. Not counting the proton transferred, efficient intramolecular proton transfer requires that a ring of five or six atoms can be formed.

There are three aspects of the situation that must be clearly distinguished from each other. One is the set of tautomers itself (Equations 2-11 and 2-13). The second is the resonance structures that can be written for each member of the set of tautomers. The third is the microscopic acid dissociations of the individual tautomers.

Each of the tautomers in the set can often be drawn as a subset of **resonance structures**. For example, just one of the tautomers of uridine can be examined in this way:



The resonance structures, as distinct from the tautomers themselves, are not independent molecules and do not have independent existences. In such a subset of resonance structures, as is always required, no σ bond or σ lone pair has engaged in the exercise because they are all orthogonal to the π electrons that are being shifted. The resonance structures designate which electrons are

the π electrons and which atoms—in the case of uridine, all of them—are contributing p orbitals to the system of π molecular orbitals. Each of the three tautomers of uridine can be submitted to this treatment to generate three subsets of resonance structures. It becomes clear that if the hierarchy of tautomers and resonance structures is not always clearly recognized, significant confusion ensues.

Because it is a tautomer, any one of the tautomers in a set can simply lose a proton in a microscopic acid dissociation that produces its conjugate base. Although the conjugate base may itself be a member of a set of tautomers existing at its level of protonation, in the examples discussed so far, none of the conjugate bases have had acidic protons. For example, the enolate is the common conjugate base produced upon the dissociation of a proton from any one of the three tautomers of ethyl acetoacetate (Equation 2-13). The ratios between the concentrations of each of the pairs of the members of a set of tautomers is independent of the pH of the solution because a proton appears on neither side of any chemical equation interconverting the two. For example, as the pH increases, the molar concentration of the enolate increases according to a function of the same form as that displayed for the conjugate base in Figure 2-6, and the sum of the molar concentrations of the three tautomers decreases accordingly, but the ratio between their concentrations remains unaltered at all values of pH, even when the conjugate base accounts for almost all of the molecules in the solution. In the case of ethyl acetoacetate, these ratios are defined by the three **equilibrium constants among the tautomers**:

$$K_{1K} = \frac{[KE]}{[E1]} \quad K_{3K} = \frac{[KE]}{[E3]} \quad K_{31} = \frac{[E1]}{[E3]} = \frac{K_{3K}}{K_{1K}} \quad (2-14)$$

To treat this situation quantitatively, a distinction must be made between the microscopic dissociation constants and the macroscopic dissociation constant. The microscopic dissociation constants involved are those for the dissociation of each tautomer:

$$K_{aE1} = \frac{[H^+][\text{enolate}]}{[E1]} \quad K_{aE3} = \frac{[H^+][\text{enolate}]}{[E3]} \quad K_{aKE} = \frac{[H^+][\text{enolate}]}{[KE]} \quad (2-15)$$

In contrast to such relationships, a **macroscopic acid dissociation constant** is an acid dissociation constant in which all tautomers with the same number of protons

are considered to be indistinguishable. As a result, the molar concentrations of all tautomers with the same number of protons must be summed, and only those undivided sums can appear in the expression defining a macroscopic acid dissociation constant. The expression for the macroscopic dissociation constant of ethyl acetoacetate is

$$K_{aEAA} = \frac{[H^+][\text{enolate}]}{[KE] + [E3] + [E1]} = 2.1 \times 10^{-11} \text{ M} \quad (2-16)$$

Were there more than one tautomer of the enolate, the molar concentrations of all these tautomers would be summed and that sum would be multiplied by $[H^+]$ in the numerator.

It is the macroscopic pK_a that is measured during the titration of an acid–base because such a measurement makes no distinction among all of the tautomers yielding a proton at a particular pH or among all of the tautomers produced upon the surrender of the proton. All that is measured is the consumption of hydroxide ions or protons by the solution. A tautomeric acid behaves as if it were a simple acid with an acid dissociation constant equal to its macroscopic acid dissociation constant. The total concentrations of conjugate bases and conjugate acids behave as if they were the concentrations of one simple base and one simple acid. Because only the macroscopic pK_a is the result of an acid–base titration and because measurements of the ratios of tautomers or their microscopic acid dissociation constants are more difficult, it is always the macroscopic pK_a that appears in a table. The tabulated value¹² for the macroscopic pK_a of ethyl acetoacetate (pK_{aEAA}) is 10.68.

By simple manipulation it can be shown that

$$\frac{1}{K_{aEAA}} = \frac{1}{K_{aKE}} + \frac{1}{K_{aE1}} + \frac{1}{K_{aE3}} \quad (2-17)$$

This relationship demonstrates that the macroscopic acid dissociation constant is a function of all of the microscopic acid dissociation constants, in particular if the magnitudes of the microscopic acid dissociation constants are all similar to each other. In the case of ethyl acetoacetate, however, the ketoester is much less acidic than the enols ($K_{aKE} < K_{aE3} < K_{aE1}$). As a result, $K_{aEAA} \cong K_{a2}$.

It is possible to calculate a microscopic acid dissociation constant from the macroscopic dissociation constant and the equilibrium constants among the tautomers (Equations 2-14). For example, Equation 2-17 can be rearranged and Equations 2-14 and 2-15 can be used to obtain

$$K_{aE3} = K_{aEAA} (1 + K_{3K} + K_{31}) \quad (2-18)$$

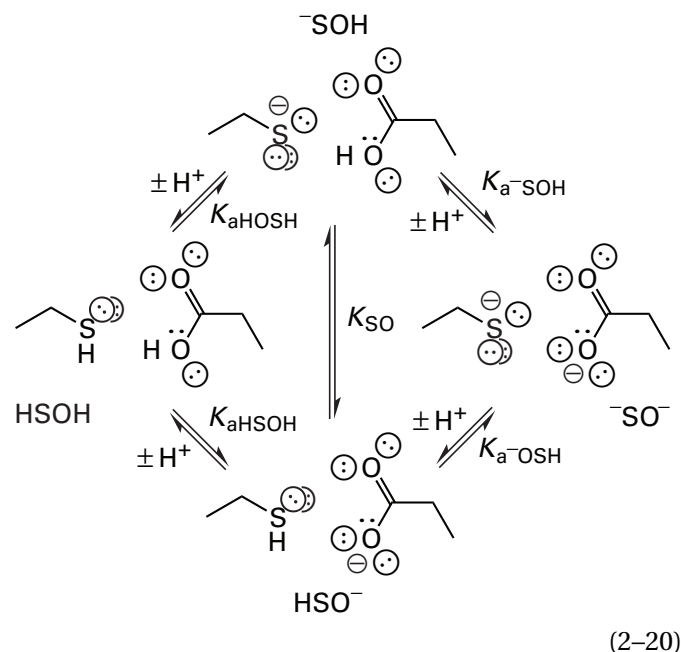
If it is assumed that the enol at carbon 3 is the more stable ($K_{31} < 1$) and that the measured equilibrium constant between the enols and the ketoester (250)¹¹ is approximately K_{3K} , then the pK_a for the microscopic acid dissociation of the enol at carbon 3 is approximately 8.3.

As Equation 2-18 suggests, the ratios among the tautomers can also be calculated from their microscopic acid dissociation constants. In fact, all of the equilibrium constants governing the tautomers and the conjugate base of ethyl acetoacetate are dependent upon each other, or linked. The linkage is reflected in the relationships

$$K_{1K} = \frac{K_{aE1}}{K_{aKE}} \quad K_{3K} = \frac{K_{aE3}}{K_{aKE}} \quad K_{31} = \frac{K_{aE3}}{K_{aE1}} \quad (2-19)$$

The equalities of Equations 2-19 simply state that the ratio between the concentrations of any two tautomers is equal to the inverse of the ratio of their respective microscopic acid dissociation constants, which makes chemical sense. The stronger the bond between the heteroatom and the proton, the smaller will be its intrinsic acid dissociation constant but the greater its relative concentration.

A molecule of protein has a large number (>100) of acidic protons and basic lone pairs distributed over the side chains of its amino acids. As a result it is a waste of time even to imagine all the tautomers of that protein that are present in solution, but usually these acid-bases on the side chains are separated widely enough from each other that each behaves as an independent acid-base and can be treated as such. Occasionally, however, two or three amino acids are not only of functional significance, so attention is paid to them, but also close enough to each other that tautomers and the distinction between macroscopic acid dissociation constants and microscopic acid dissociation constants become important.³³ In thioredoxin from *Escherichia coli*, there is an aspartate (Aspartate 26) close enough in the structure to a cysteine (Cysteine 32) that their acid dissociations become linked.³³ When both are protonated or both are unprotonated, there are no tautomers; but when one is protonated and the other is not, there are two tautomers, one in which the proton is on the aspartic acid and the cysteinate is in the form of the anionic base, and the other in which the proton is on the cysteine and the aspartate is the anionic base:



The equilibrium constant for the tautomerization (K_{SO}) is defined with the thiol-carboxylate as product and the thiolate-carboxylic acid as reactant. The linkage relationships are

$$K_{SO} = \frac{[\text{HSO}^-]}{[\text{⁻SOH}]} = \frac{K_{a\text{HSOH}}}{K_{a\text{HOSH}}} = \frac{K_{a\text{⁻SOH}}}{K_{a\text{⁻OSH}}} \quad (2-21)$$

and the relationships between the macroscopic dissociation constants and the microscopic dissociation constants are³⁴

$$K_{a1} = \frac{([\text{HSO}^-] + [\text{⁻SOH}])[\text{H}^+]}{[\text{HSOH}]} = K_{a\text{HOSH}} + K_{a\text{HSOH}} \quad (2-22)$$

and

$$\frac{1}{K_{a2}} = \frac{[\text{HSO}^-] + [\text{⁻SOH}]}{[\text{⁻SO}^-][\text{H}^+]} = \frac{1}{K_{a\text{⁻OSH}}} + \frac{1}{K_{a\text{⁻SOH}}} \quad (2-23)$$

The equation describing the titration curve for the cysteine is

$$f_{\text{cysteinate}} = \frac{K_{a\text{HOSH}}([\text{H}^+] + K_{a\text{⁻SOH}})}{K_{a\text{HOSH}}([\text{H}^+] + K_{a\text{⁻SOH}}) + [\text{H}^+](\text{[H}^+] + K_{a\text{HSOH}})} \quad (2-24)$$

where $f_{\text{cysteinate}}$ is the fraction of the cysteine that is the anionic base. It is possible to **walk through the titration**

curve. Assume that the first and second macroscopic acid dissociation constants are well separated, that the respective pairs of microscopic acid dissociation constants are close together ($K_{aHOSH} \cong K_{aHSOH} > K_{a-OSH} \cong K_{a-SOH}$), that the initial pH is low, and that the titration is performed by adding hydroxide ion. As the concentration of protons decreases into the range of the first macroscopic acid dissociation constant, the inequality $[H^+] \geq K_{aHOSH} \cong K_{aHSOH} > K_{a-OSH} \cong K_{a-SOH}$ holds and

$$f_{\text{cysteinate}} \cong \frac{K_{aHOSH}}{(K_{aHOSH} + K_{aHSOH}) + [H^+]} \quad (2-25)$$

This equation describes a normal titration curve for a conjugate base (Figure 2-6A) with a macroscopic acid dissociation constant equal to $K_{aHOSH} + K_{aHSOH}$, the sum of the two lower microscopic dissociation constants, which is the macroscopic dissociation constant K_{a1} , and that reaches a plateau at

$$f_{\text{cysteinate}} = \frac{K_{aHOSH}}{K_{aHOSH} + K_{aHSOH}} = \frac{K_{a-OSH}}{K_{a-OSH} + K_{a-SOH}} = \frac{1}{1 + K_{SO}} \quad (2-26)$$

which is the fraction of cysteinate in the tautomeric mixture. The plateau is reached when $K_{aHOSH} \cong K_{aHSOH} > [H^+] > K_{a-OSH} \cong K_{a-SOH}$.

As $[H^+]$ is decreased further during the titration into the range of the second macroscopic dissociation constant and on above it, the inequality $K_{aHOSH} \cong K_{aHSOH} > K_{a-OSH} \cong K_{a-SOH} \cong [H^+]$ holds and

$$f_{\text{cysteinate}} \cong \frac{\frac{K_{a-OSH}}{K_{a-OSH} + K_{a-SOH}} (K_{a-SOH} + [H^+])}{[H^+] + \frac{K_{a-OSH} K_{a-SOH}}{K_{a-OSH} + K_{a-SOH}}} \quad (2-27)$$

which is the equation for a normal titration curve beginning at the tautomeric fraction (Equation 2-26), having a macroscopic acid dissociation constant equal to the term $K_{a-SOH} K_{a-OSH} (K_{a-OSH} + K_{a-SOH})^{-1}$, which is the macroscopic dissociation constant K_{a2} , and reaching a final level at which all of the cysteine is unprotonated (the fully ionized form on the right of Equation 2-20).

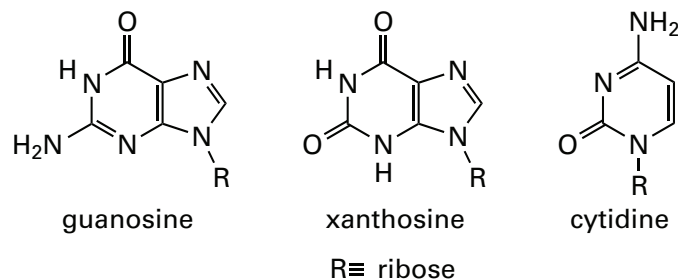
The titration curve for the cysteine of thioredoxin conforms to these expectations.^{33,35-37} The values observed for the macroscopic acid dissociation constants are $pK_{a1} = 7.2$ and $pK_{a2} = 9.5$ and the tautomeric ratio, as determined by the level of the plateau observed

in the titration curves (Equation 2-26), is 1.3. These values, when inserted into the equations, give microscopic acid dissociation constants for the cysteine and the aspartic acid of $pK_{aHOSH} = 7.6$, $pK_{aHSOH} = 7.5$, $pK_{a-OSH} = 9.2$, and $pK_{a-SOH} = 9.1$. The titration curve for the aspartic acid has tautomeric ratios and macroscopic acid dissociation constants of about 1.3, 7.2, and 9.5, as expected.^{33,35,36}

From the microscopic acid dissociation constants, it can be seen that when the aspartic acid is protonated, the thiol is a much better acid ($pK_{aHOSH} = 7.6$) than when the aspartate is unprotonated and anionic ($pK_{a-OSH} = 9.2$). Because of the linkage, the same difference in microscopic acid dissociations is necessarily seen for the aspartic acid ($\Delta pK_a = 1.6$) when the cysteine is the neutral thiol or the anionic thiolate. These differences make electrostatic sense because it should be significantly more difficult to produce two adjacent negative charges than a single negative charge. For example, the pK_a of the first macroscopic acid dissociation of succinic acid is 1.29 units less than that of the second. Before the ionization of these two acid-bases in thioredoxin was analyzed in terms of tautomeric equilibria and microscopic acid dissociation constants,³³ there was considerable confusion as to what was happening.^{35,37,38}

Glutamate 172 and Glutamate 78 in the *endo*-1,4- β -xylanase from *Bacillus circulans* are close enough to each other in the native protein to be linked by a tautomeric equilibrium.³⁹ The microscopic pK_a of Glutamate 172 when Glutamate 78 is the neutral acid is 5.5, but when Glutamate 78 is the anionic carboxylate, it is 6.7.

Problem 2-12:



- (A) Draw complete σ structures for the above heterocycles in the above tautomeric forms including all σ lone pairs. Draw them with proper bond angles. Abbreviate the ribose as R.
- (B) Indicate which protons are involved in tautomeric shifts between which lone pairs of electrons. Draw some of the tautomeric forms of these neutral molecules.
- (C) How many π electrons are there in each compound?
- (D) The macroscopic values of pK_a for guanosine are 1.6, 9.2, and 12.5; those for xanthosine are 0.0, 5.5,

and 13.0; and those for cytidine are 4.2 and 12.5. Draw vertically chemical equations for the two or three acid dissociations that have these values of pK_a and horizontally next to the molecule in each level of protonation draw two of its tautomers.

- (E) How many of the tautomers at each level of protonation are insignificant because they require separation of charge?
- (F) Draw the σ structure of a tautomer of xanthine that could substitute for adenine in the A–T base pair.

Problem 2–13: Derive Equation 2–17.

Problem 2–14: If

$$f_{\text{cysteinate}} = \frac{[\text{SOH}^-] + [\text{SO}^{2-}]}{[\text{HSOH}] + [\text{HSO}^-] + [\text{SOH}^-] + [\text{SO}^{2-}]}$$

where the four species are as labeled in Equation 2–20, derive Equations 2–22, 2–23, 2–24, 2–25, 2–26, and 2–27. The values observed for the macroscopic acid dissociation constants for the tautomeric equilibrium of Equation 2–20 are $pK_{a1} = 7.2$ and $pK_{a2} = 9.5$ and the tautomeric ratio is 1.3. Calculate the values of the four microscopic dissociation constants.

Problem 2–15: Write a set of linked equilibria resembling the one in Equation 2–20 for the tautomeric equilibrium and four microscopic acid dissociations of 1,5-dimethyl-4-mercaptoimidazole. The two macroscopic acid dissociations of 1,5-dimethyl-4-mercaptoimidazole have values of pK_a equal to 2.3 and 10.3.⁴⁰ Assume that the microscopic acid dissociation constant between 1,5-dimethyl-4-mercaptoimidazolium cation and its neutral conjugate base has the same value as the macroscopic acid dissociation constant for *S*-methyl-1,5-dimethyl-4-mercaptoimidazole ($pK_a = 6.0$), and calculate values for the other three microscopic acid dissociation constants and the tautomeric equilibrium constant for 1,5-dimethyl-4-mercaptoimidazole. At pH 7, what is the major form of the molecule present in the solution?

Problem 2–16: The macroscopic acid dissociation constants for succinic acid are $10^{-4.19}$ and $10^{-5.48}$. What are its four microscopic acid dissociation constants as values of pK_a ?

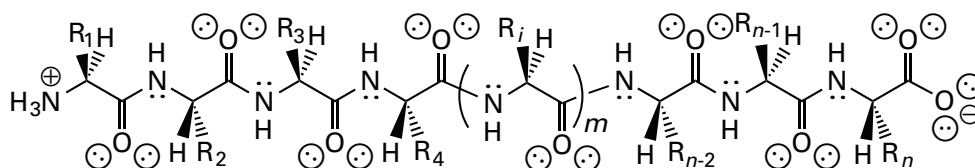
Amino Acids

The fundamental, covalent scaffold of a molecule of protein is the polypeptide (see 2–15 below).

A **polypeptide** is a long (50–5000 amino acids) linear polymer, the monomers of which are L- α -amino acids. Because a protein constructed entirely of D- α -amino acids is functionally indistinguishable from its biological enantiomer,⁴¹ the original choice of L- α -amino acids was arbitrary. The covalent bonds that link the amino acids together to form a polypeptide, which are referred to as **peptide bonds**, are those of amides. Because a molecule of water is lost between two amino acids when a peptide bond is formed, the amino acids, when they are incorporated into a polypeptide, should be referred to as **amino acid residues**. Every polypeptide has the same backbone of peptide bonds and α carbons with an end at which an unbonded primary amine is usually located, the **amino terminus**, and an end at which an unbonded carboxylic acid usually is located, the **carboxy terminus**. At the values of pH usually encountered in living organisms (pH 7–8), the amino terminus ($pK_a = 8.0$)¹² should be partially protonated and cationic, and the carboxy terminus ($pK_a = 3.3$)¹² should be unprotonated and anionic. The rhythm of a polypeptide is N, C α , CO, N, C α , CO, N, C α , CO.

In a polypeptide, the amino acid residues each contribute a **side chain** (R_i in 2–15) to the covalent structure. It is the order in which these side chains appear along the polymer that defines the protein. Conceptually, the contribution of an amino acid residue to the structure of the protein can be divided into its α -imido nitrogen, its α carbon, and its α -acyl carbon and oxygen on the one hand and its side chain on the other. The former always provide the same six atoms (with the minor exception of a carbon for an α hydrogen in the case of proline) to the backbone of the polypeptide. The structure of this backbone can be treated as if it were an independent molecule, albeit a long tortuous polymer, and the six atoms contributed to it by each amino acid as if they were separate from each side chain. The side chains themselves can be treated as separate entities by detaching them, in the imagination, from their respective α carbons and replacing what was the bond to the α carbon with a hydrogen. In this way, a model compound for the amino acid side chain is created.⁴²

A **model compound** for a particular amino acid residue in a polypeptide would be a small molecule that



2–15

incorporates the structure of the side chain and any additional structural elements necessary to duplicate the properties of the amino acid residue that are of interest. The model compounds in which the α carbon is replaced only by a hydrogen are simple, readily available chemicals. For example, in this set, the model compound for glutamic acid would be propionic acid; that for histidine, 4-methylimidazole. Another set of model compounds that has been used is the *N*-acetyl α -amides of the amino acids.⁴³ *N*-Acetylaspartic acid α -amide and *N*-acetylglutamic acid α -amide (Figure 2-1) are members of this set.

Unfortunately, the free amino acids themselves are poor model compounds for the amino acid residues in a polypeptide. This arises from the fact that, at all values of pH, they are either zwitterionic or bear net charge. Their solubilities, acid-base behavior, and ability to participate in noncovalent interactions are dominated by the carboxylate anion and ammonium cation they contain. Contrary to original expectations, an understanding of the properties of proteins depends little on an understanding of the properties of the amino acids themselves, while an examination of the structures of the amino acid side chains and the behavior of uncharged model compounds for them does provide essential information.

One of the more important properties of an amino acid residue incorporated in a polypeptide is the value of the **pK_a of its side chain**. The amino acid side chains that contain acid-bases are those of aspartate, asparagine, serine, threonine, glutamate, glutamine, cysteine, tyrosine, histidine, lysine, tryptophan, and arginine (Table 2-2). The *N*-acetyl α -amides of glutamate and aspartate have been useful in examining the electronic effects of the peptide bonds that surround the α carbon on the acid dissociation constants of the amino acid side chains in a polypeptide.⁴³ The α carbon in an *N*-acetyl α -amide, which is transmitting inductively the significant electron-withdrawing capacity of both the carboxamido and the acylimido groups that are attached to it, seems to have about the same electron-withdrawing capacity as a hydroxyl, a cyanomethyl, a chloromethyl, or a bromomethyl group. This conclusion follows from the fact that replacing the α carbon of *N*-acetylaspartic acid α -amide or *N*-acetylglutamic acid α -amide with any of these functional groups produces little change in the pK_a of the respective carboxylic acid, but the completely aliphatic model compounds, acetic acid and propionic acid, respectively, are significantly less acidic than the respective *N*-acetyl α -amides (Table 2-2). These four more common electron-withdrawing substituents can be used to estimate the inductive effect of the polypeptide on the acid dissociations of the various acid-bases on other amino acids (Table 2-2).

These values of pK_a for the side chains have been shown to be accurate when the amino acid is in a polypeptide if the polypeptide is in the form of a struc-

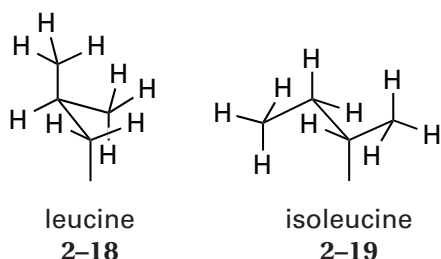
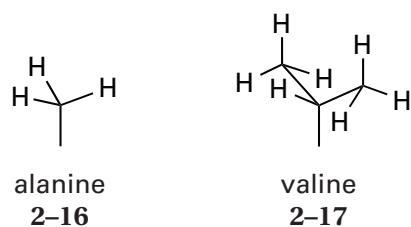
Table 2-2: Acid Dissociation Constants for Model Compounds for the Amino Acid Residues^a

amino acid residues	model compound	pK _a
aspartic acid	in polypeptide (estimate) ^b	4.0
	<i>N</i> -acetylaspartic acid α -amide ⁴³	4.0
	Gly-Gly-Asp-Gly-Gly ⁴⁴	3.9
	3-chloropropionic acid ¹²	4.1
	hydroxyacetic acid ¹²	3.8
	3-bromopropionic acid ¹²	4.0
	3-cyanopropionic acid ¹²	4.0
	acetic acid ¹²	4.75
glutamic acid	in polypeptide (estimate)	4.3
	<i>N</i> -acetylglutamic acid α -amide ⁴³	4.3
	Gly-Gly-Glu-Gly-Gly ⁴⁴	4.1
	glutamic acid (microscopic pK _a neutral) ⁴³	4.5
	4-chlorobutyric acid ¹²	4.5
	3-hydroxypropionic acid ¹²	4.5
	4-bromobutyric acid ¹²	4.6
	4-cyanobutyric acid ¹²	4.4
	propionic acid ¹²	4.9
serine	in polypeptide (estimate)	-3, 14.2
	2-chloroethanol ¹²	14.3
	2-bromoethanol ¹²	14.4
	2-cyanoethanol ¹²	14.0
	ethanol ^{11,12,45}	2, 15.9
	methanol ¹²	15.5
threonine	in polypeptide (estimate)	-3, 15
cysteine	in polypeptide (estimate)	8.7
	glutathione ⁴⁶	8.7
	cysteine (microscopic pK _a neutral) ⁴⁷	9.1
	ethanethiol ¹²	10.5
	2-mercaptoethanol ¹²	9.5
	ethyl mercaptoacetate ¹²	8.0
tyrosine	in polypeptide (estimate)	9.8
	polytyrosine ⁴⁸	9.5
	tyrosine (microscopic pK _a neutral) ⁴⁹	9.8
	4-(hydroxymethyl)phenol ¹²	9.8
	phenol ¹²	10.0
histidine	4-methylphenol ¹²	10.2
	in polypeptide (estimate) ⁵⁰	6.6, 14
	Pro-His-glycinamide ⁴⁷	6.4
	histidine hydantoin ⁵¹	6.4
	histidine (microscopic pK _a neutral) ⁵²	6.0
	<i>N</i> -acetyl-L-histidine methylamide ⁵⁰	6.5
	Gly-Gly-His-Gly-Gly ⁵⁰	6.7
	Gly-His-Gly ⁵⁰	6.6
	imidazole ¹²	7.1, 14.5
lysine	4-methylimidazole ⁵²	7.5
	in polypeptide (estimate)	10.5
	Gly-Gly-Lys-Gly-Gly ⁵³	10.5
	1-amino-5-hydroxypentane ¹²	10.5
	Ala-Lys-(Ala) _n (<i>n</i> = 1, 3) ¹²	10.5
glutamine	1-aminopentane ¹²	10.6
	in polypeptide (estimate)	-1, 17
asparagine	acetamide ¹⁸	-0.7, 17
arginine	in polypeptide (estimate)	-1.4, 16
	in polypeptide (estimate)	13
tryptophan	<i>N</i> -methylguanidine ¹²	13.4
	in polypeptide (estimate)	17
	indole ¹²	16.9

^aThe values for the pK_a of the model compounds are from the noted sources for 25°C. ^bThe estimate for each amino acid is based entirely on the values tabulated for the model compounds.

tureless random coil⁴⁷ and if that amino acid does not have an immediate neighbor in the polypeptide with an ionized side chain. When the polypeptide folds to form a globular protein, however, significant shifts in the values of pK_a for its side chains occur.^{36,38,54,55} Neighboring charged functional groups in the compact folded structure can affect the pK_a of a particular acid–base. An adjacent anion makes it harder to remove a proton from an acid and raises its pK_a (Equation 2–20). An adjacent elementary positive charge makes it easier to remove a proton from an acid and lowers its pK_a . If, upon the folding of the protein, the acid–base finds itself in an aprotic environment, secluded from water, the more charged form of the acid–base will be less stable relative to the less charged form than it would be in water. This shifts the pK_a in the direction favoring the less charged form of the acid–base. A simple paradigm for such an effect would be the shift in the pK_a of acetic acid in dimethyl sulfoxide, a relatively polar but aprotic solvent, to 12.9 from its value of 4.75 in water, which occurs because the anionic conjugate base is poorly solvated by the dimethyl sulfoxide relative to the solvation provided by water. For all of these reasons, when the polypeptide folds to form the native structure, the values for the pK_a of the various amino acids shift away from their ideal values.

Alanine (A, Ala), valine (V, Val), leucine (L, Leu), and isoleucine (I, Ile) have unsaturated alkyl groups as side chains:*

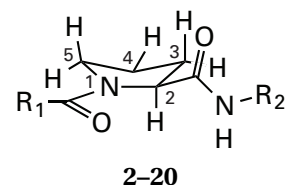


All of their carbons are hybridized sp^3 . Because alkyl groups are sterically more bulky than functional groups containing atoms hybridized [p , sp^2 , sp^2 , sp^2], steric considerations are more important in examining the

* The drawings of all of the side chains in this section, except for proline, are for the entire functional group that is attached through a carbon–carbon bond to the respective α -carbon in the backbone of the polypeptide. The open bond in each drawing indicates the point of this attachment.

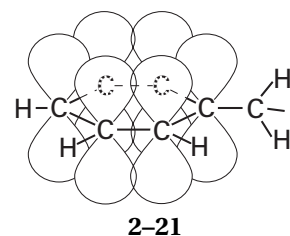
structures of these four amino acids than with most of the others. The view down every carbon–carbon bond should be staggered, and methyl or other alkyl groups should be *anti* to each other in the most stable conformers.

Proline (P, Pro) and **glycine (G, Gly)** are amino acid residues the effect of which on the polypeptide is almost entirely steric. Glycine has no side chain at all, merely a hydrogen, and as such can occupy positions in the native structure of a protein that are cramped. A proline, because it is a ring



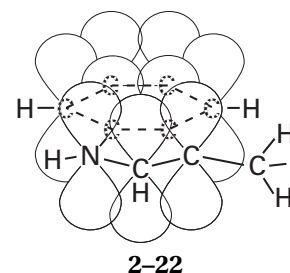
forces the polypeptide to assume particular orientations.

Phenylalanine (F, Phe) is aromatic by virtue of its phenyl ring:

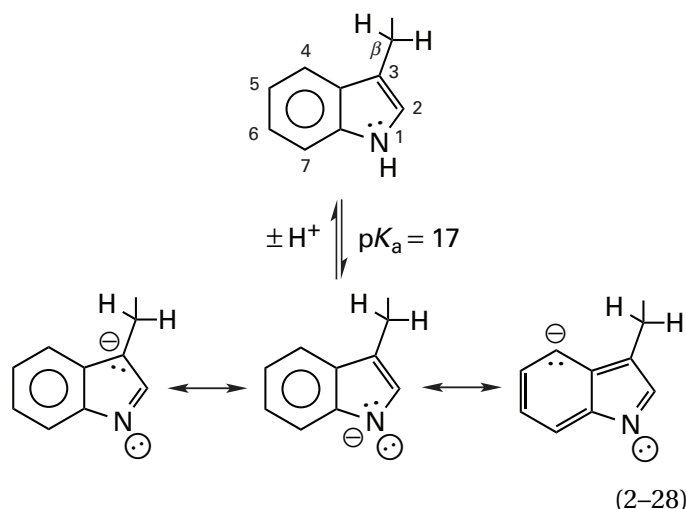


The six π electrons are delocalized above and below the plane of the ring in three bonding molecular orbitals over the six carbons that contribute the six p orbitals. This causes the σ structure of the ring to be planar, and it is sandwiched between two circular clouds of π electrons. A phenylalanine side chain absorbs ultraviolet light ($\lambda_{\max} = 253 \text{ nm}$, $\epsilon = 1550 \text{ M}^{-1} \text{ cm}^{-1}$), and its absorption spectrum displays the usual fine structure seen in unadorned alkylbenzenes.

The side chain of **tryptophan (W, Trp)** is an indole, which is a benzopyrrole. The indole is entirely aromatic, consisting of an unbroken ring of nine atoms each contributing a p orbital, and the aromatic system of π molecular orbitals contains 10 π electrons:

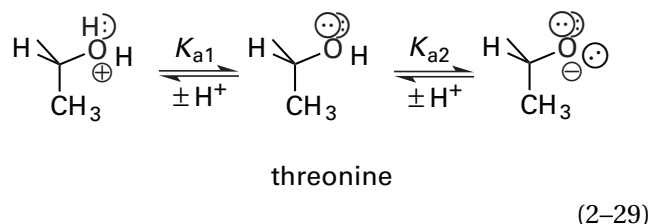


The hydrogen on the pyrrole nitrogen of indole ($pK_a = 17.0$)¹² is even less acidic than the hydrogen on a molecule of water ($pK_a = 15.7$):



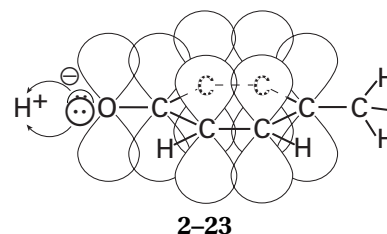
When it departs as a proton, the lone pair left behind is an sp^2 lone pair, and the negative formal charge can be distributed by the system of π molecular orbitals over all nine atoms in the ring. This delocalization is greater in extent than the delocalization available to pyrrole, which is somewhat less acidic ($pK_a = 17.5$) than indole ($pK_a = 16.9$).¹² The indolyl group of tryptophan is planar with hydrogens directed outward along its edge and clouds of π electrons above and below the σ plane. It has the strongest ultraviolet absorption of any functional group in an amino acid ($\lambda_{\max} = 281 \text{ nm}$, $\epsilon = 5690 \text{ M}^{-1} \text{ cm}^{-1}$)^{56,57} and is the principal contributor to the absorbance of protein at 280 nm (Figures 1-6 and 1-10).

The side chains of **serine (S, Ser)** and **threonine (T, Thr)** are primary and secondary aliphatic alcohols resembling ethanol and 2-propanol, respectively, except that they are more acidic because of the electron withdrawal of the immediately adjacent polypeptide. Their oxygens are hybridized sp^3 and have two σ lone pairs that can act as bases as well as an acidic hydrogen:

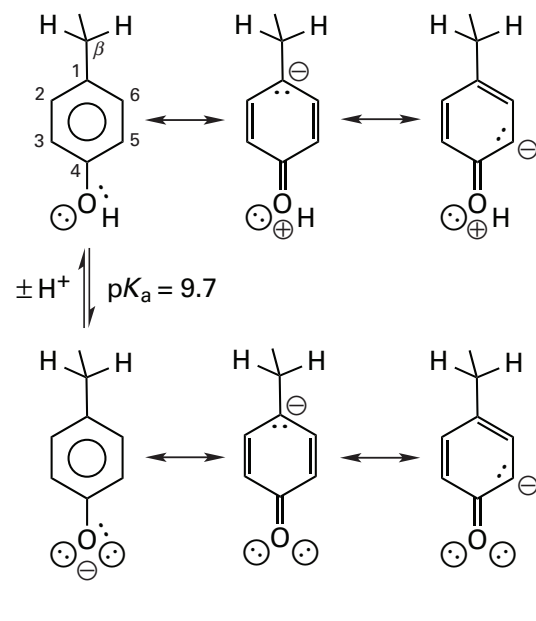


The values of pK_a for these acid-base reactions can be estimated (Table 2-2) from a series of alcohols containing appropriate electron-withdrawing substituents.

Tyrosine (Y, Tyr) resembles phenylalanine because it is aromatic and serine because it has a hydroxyl group. As a phenol, however, its properties are distinct from either. Tyrosine ($pK_a = 9.7$) is more acidic than serine ($pK_a = 14.2$) because of the ability of the neighboring π system to delocalize the excess electron density of the anion:

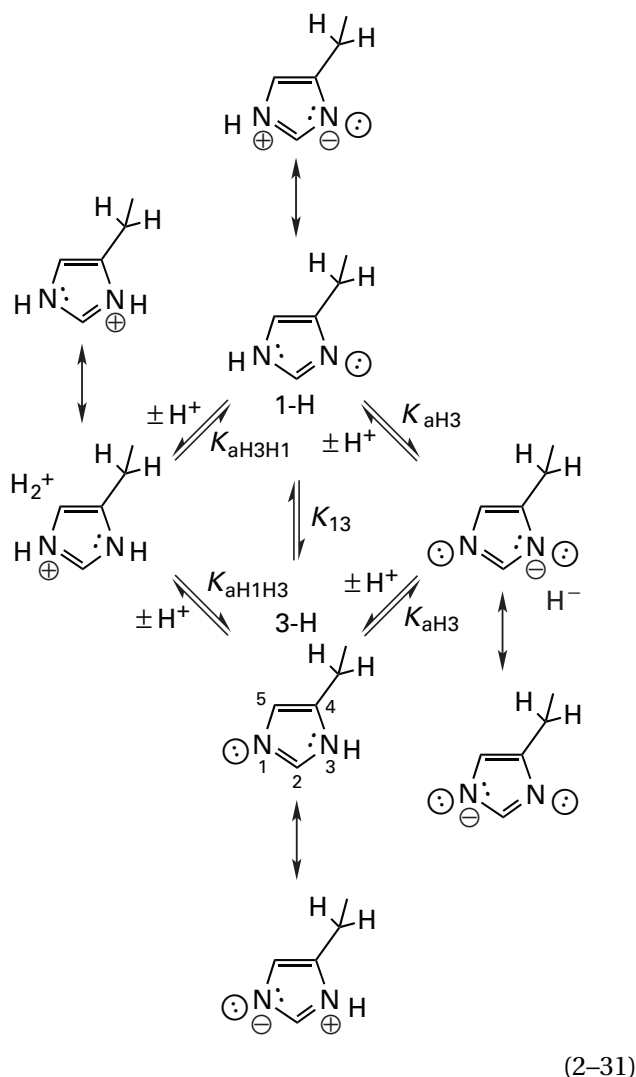


The six p orbitals from the ring and the one p orbital from the exocyclic oxygen that overlap in the anion are distributed above and below the plane of the ring. As indicated schematically in 2-23, the two σ lone pairs on the oxygen of the anion are in the plane of the ring at angles of 120° to the carbon-oxygen bond and are the only bases on the anion that associate with a proton. To the extent that one of the lone pairs of electrons on the conjugate acid is delocalized,



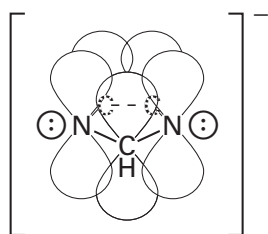
the lowered pK_a of the hydroxyl reflects the lowered pK_a of an sp^2 oxygen-hydrogen bond. To the extent that the lone pair is not so delocalized in the conjugate acid as it is in the anion, the lowered pK_a reflects the stability of the anion relative to the neutral acid resulting from the ability of the system of π molecular orbitals to spread its excess electron density over one oxygen and three carbons. Because of this increase of delocalization in the anion, a significant change in the ultraviolet spectrum of a tyrosine side chain occurs when the acid ($\lambda_{\max} = 275 \text{ nm}$, $\epsilon = 1410 \text{ M}^{-1} \text{ cm}^{-1}$) becomes the conjugate base ($\lambda_{\max} = 293 \text{ nm}$, $\epsilon = 2380 \text{ M}^{-1} \text{ cm}^{-1}$) upon acid dissociation.⁵⁶ It is for this reason that proteins absorb more light at 280 nm when the pH of the solution is raised.

The side chain of **histidine (H, His)** is also an aromatic acid-base by virtue of its imidazolyl group (Equation 2-31):



The neutral imidazolyl group is an aromatic heterocycle containing one pyrrole nitrogen and one pyridine nitrogen, which together contribute three valence electrons to an aromatic ring formed from five p orbitals. Six π electrons are located in the three bonding molecular orbitals of the aromatic system. The six π electrons remain in this aromatic system of π molecular orbitals at all times in all three protonation states but fluidly redistribute in response to changes in coulomb effects as the nitrogens gain or lose protons at their σ lone pairs.

The anion of the parent compound, **imidazole**



2-24

is the best place to begin. All atoms are hybridized [p , sp^2 , sp^2 , sp^2], the two nitrogens each have a σ lone pair of electrons located in the plane of the ring, and both are electronically equivalent. The excess electron density associated with the formal negative charge is distributed by the π system over both nitrogens, and resonance structures can be drawn to show this (Equation 2-31). The first proton adds to one of the two σ lone pairs in the imidazolate anion to form an sp^2 covalent bond in an acid-base reaction with a macroscopic $pK_{a2} = 14.5$. Thus, the imidazolate anion is less basic than the pyrrolate anion ($pK_a = 17.5$)¹² because its system of π molecular orbitals can spread the excess electron density over two nitrogens, but the adenosinate anion ($pK_a = 12.5$) is less basic than the imidazolate anion because its system of π molecular orbitals can spread the excess electron density over three nitrogens.

In neutral imidazole, the two nitrogens are necessarily nonequivalent because one has a proton attached to it. The proton causes its nitrogen to be more electronegative, and the lobes of the π molecular orbitals at this location swell accordingly. This is represented in the resonance structures by placing a π lone pair of electrons over this nitrogen, but such formalism is not meant to imply that this becomes a basic position or that this nitrogen rehybridizes or that the ring is no longer aromatic. The only base on a neutral imidazole is the σ lone pair on the other nitrogen, and it gains a proton in an acid-base reaction with a macroscopic $pK_{a1} = 6.6$ when the base is a histidine side chain in a polypeptide (Table 2-2) or $pK_{a1} = 7.1$ when the base is imidazole itself. The imidazolium cation ($pK_{a1} = 7.1$) is less acidic than the pyridinium cation ($pK_a = 5.2$) because its system of π molecular orbitals can spread its electron deficiency over two nitrogens.

The two ring nitrogens in a histidine side chain, unlike the two in imidazole, are not stereochemically equivalent to each other because of the substitution at carbon 4. The behavior, as a function of pH, of the chemical shifts of the nuclear magnetic resonances of the various carbon-13 nuclei of the imidazole ring in histidine has been compared to their behavior in 1-methylhistidine and 3-methylhistidine. It was concluded from these observations⁵⁸ that in aqueous solution the ratio between the two neutral tautomers (Equation 2-31), 1-protiohistidine and 3-protiohistidine, is 4:1. Assume that the same ratio obtains for histidine in a polypeptide,⁵⁹ and let H_2^+ be the cation of a histidine side chain in an unfolded polypeptide and 1-H and 3-H be the two tautomers (Equation 2-31). If

$$K_{13} = \frac{[3-H]}{[1-H]} = 0.25 \quad (2-32)$$

and

$$K_{aH3H1} = \frac{[1-H][H^+]}{[H_2^+]} \quad (2-33)$$

and

$$K_{aH1H3} = \frac{[3-H][H^+]}{[H_2^+]} \quad (2-34)$$

then the ratio of the two microscopic dissociation constants is 4 as it should be.⁵⁹ If the macroscopic dissociation constant⁵⁰

$$K_{a1} = \frac{([1-H] + [3-H])[H^+]}{[H_2^+]} = K_{aH3H1} + K_{aH1H3} = 10^{-6.6} \quad (2-35)$$

then

$$K_{aH3H1} = 0.8K_{a1} = 10^{-6.7} \quad (2-36)$$

and

$$K_{aH1H3} = 0.2K_{a1} = 10^{-7.3} \quad (2-37)$$

If a histidine in a protein is fully accessible to the aqueous phase, either one of its two protons can dissociate, and the imidazolium cation of that histidine side chain has two microscopic acid dissociation constants, $pK_{aH3H1} = 6.7$ and $pK_{aH1H3} = 7.3$. All of these relationships can be presented graphically (Figure 2-7).

It is important to distinguish carefully between the use of the macroscopic and microscopic acid dissociation constants. If the imidazolium of a histidine side chain is on the surface of a molecule of protein and free to rotate around the carbon-carbon σ bonds connecting it to the polypeptide so that both nitrogen-hydrogens can participate freely in acid dissociations, the macroscopic pK_a will dictate its acid-base behavior as it did in the model compounds. If the imidazolium is held in place by the neighboring amino acids in such a way that the surroundings have the same polarity as water and such that one of its acidic hydrogens is always engaged in a hydrogen bond with an acceptor that resembles water closely, the single remaining site available for acid dissociation would display its respective microscopic acid dissociation constant.⁵⁹ A decision to use the macroscopic pK_a or the microscopic pK_a implies that the respective situation has been assumed to occur.

The side chains of **aspartic acid (D, Asp)** and **glutamic acid (G, Glu)** are simple carboxylic acids (Figure 2-5 and Table 2-2). The side chains of **asparagine (N,**

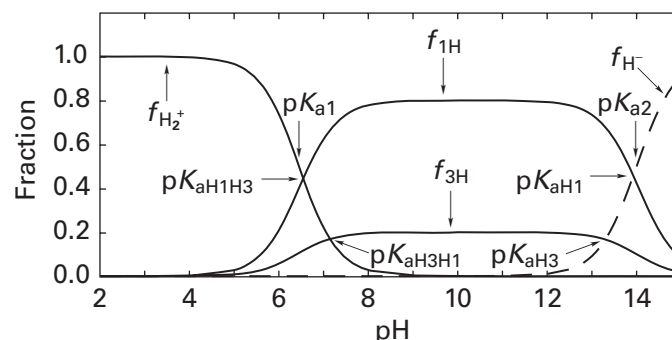
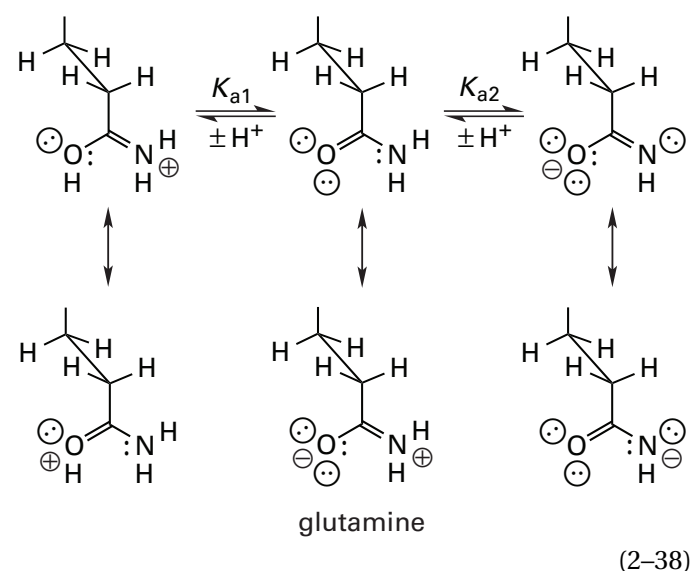


Figure 2-7: Titration curves for the three conjugate acids of histidine (Equation 2-31): H_2^+ , 1-H, and 3-H. The titration curves for histidine graphically illustrate the equations (Equations 2-32 to 2-37) governing the tautomerization. The ratio of the two tautomers remains constant over the entire range. The value of each microscopic pK_a is defined by the intersection between the curve representing the concentration of the respective tautomer and the curve representing the concentration of its nontautomeric conjugate base or conjugate acid. As a result, the microscopic acid dissociation constant for a reaction producing a tautomer from a nontautomer is always less than the corresponding macroscopic acid dissociation constant, and the microscopic acid dissociation constant for the reaction in which a tautomer dissociates to form a nontautomer is always greater than the corresponding macroscopic acid dissociation constant. The pK_a for each of the two macroscopic dissociations coincides with the pH at which half of the histidine is in the form of the nontautomer, the H_2^+ cation or the H^- anion, respectively.

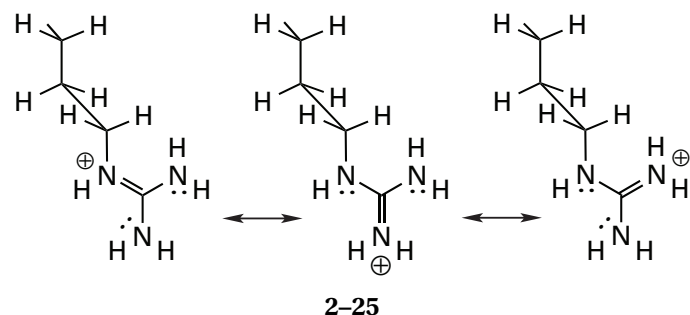
Asn) and **glutamine (Q, Gln)** are the primary amides of these two carboxylic acids. Primary amides participate in two acid-base reactions:



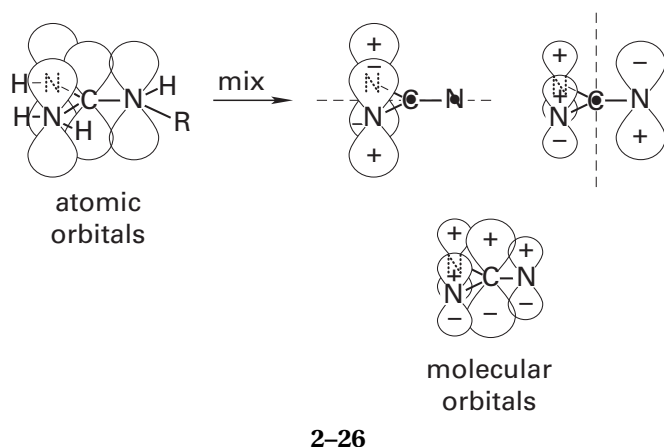
As is the case with imidazole, two protons are removed successively from two heteroatoms separated by one carbon and connected to each other by a system of π molecular orbitals. The acid dissociations also proceed from a cation through a neutral form to an anion. The values for the two acid dissociation constants, however,

are more widely separated from each other ($pK_{a1} = -0.7$ and $pK_{a2} = 17$) than the two for imidazole ($pK_{a1} = 7.1$ and $pK_{a2} = 14.5$).

The side chain of **arginine (R, Arg)** contains a cation



that slightly resembles the protonated amides of glutamine and asparagine but has a system of π molecular orbitals larger by one atom, a nitrogen:



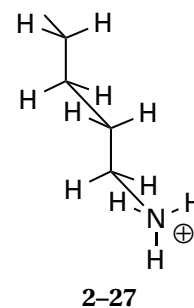
The functional group in arginine is a guanidinium cation, and it is composed from four atoms, three nitrogens and a central carbon, in the shape of a Y. Each of the four atoms contributes one p orbital, and the four mix to produce the four π molecular orbitals, one bonding (ψ_1) and two nonbonding (ψ_2 and ψ_3) molecular orbitals shown, as well as a fourth antibonding orbital (ψ_4) not shown.

The **guanidinium cation** has six π electrons distributed in pairs among the bonding molecular orbital and the two nonbonding molecular orbitals above and below the plane of the ring. The two highest occupied nonbonding molecular orbitals are responsible for distributing two pairs of electrons evenly over the three nitrogen atoms as is described by the resonance structures of 2-25. This causes the one elementary positive charge to be divided evenly among the three nitrogens. An arginine side chain ($pK_a = 13$) is less acidic than a histidinium side chain ($pK_a = 6.4$) because the elementary positive charge is distributed by the system of π molecular orbitals over three nitrogens rather than over two.

The guanidinium of an arginine side chain (2-25)

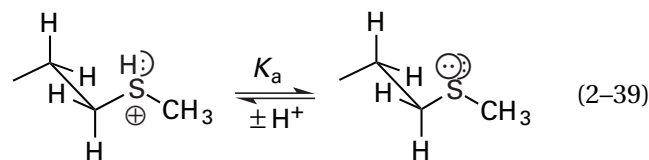
defines a plane in which its central carbon, three nitrogens, five hydrogens, and the δ carbon all reside. The hydrogens bristle from the three nitrogens at 120° angles around the periphery, and the flat clouds of π electrons sandwich the σ structure from above and below. The entire structure bears a net elementary positive charge that is neutralized by removing a proton from any one of the three nitrogens.

The side chain of **lysine (K, Lys)**, the other strongly basic amino acid side chain ($pK_a = 10.5$), is a simple primary ammonium cation at neutral pH:



With four carbons, it has the longest linear alkane chain among the amino acids. The conformation of lowest free energy is all *anti* as shown. The introduction of a *gauche* conformation at any of the carbon-carbon bonds requires about 4 kJ mol^{-1} standard free energy.

The side chain of **cysteine (C, Cys)** is fairly acidic ($pK_a = 8.7$). The thiolate anion that results from the acid dissociation (Figure 2-8), although it is not a strong base, is a strong nucleophile because sulfur is an element of the third row. The side chain of **methionine (M, Met)**, although it contains a thioether, resembles in its properties the side chains of the amino acids that are purely alkanes, but it is linear rather than branched. The sulfur of methionine is large and electron-rich but not very basic; the pK_a for its conjugate acid¹¹ is about -9.



Methionine is, however, nucleophilic. At low pH, only methionine and cysteine react with alkylating electrophiles.⁶⁰

A drawback of methionine and cysteine, both to a protein in its normal environment and when the protein is studied in the laboratory, is the susceptibility of the sulfur they contain to oxidation by reaction with molecular oxygen, peroxides, or other oxidants. These reactions produce, in addition to disulfides, various oxides of sulfur (Figure 2-8). These are sulfoxides and sulfenic acids, sulfones and sulfinic acids, and sulfonic acids. To understand the bonding in these various products, the best way to begin is to examine **sulfate**:

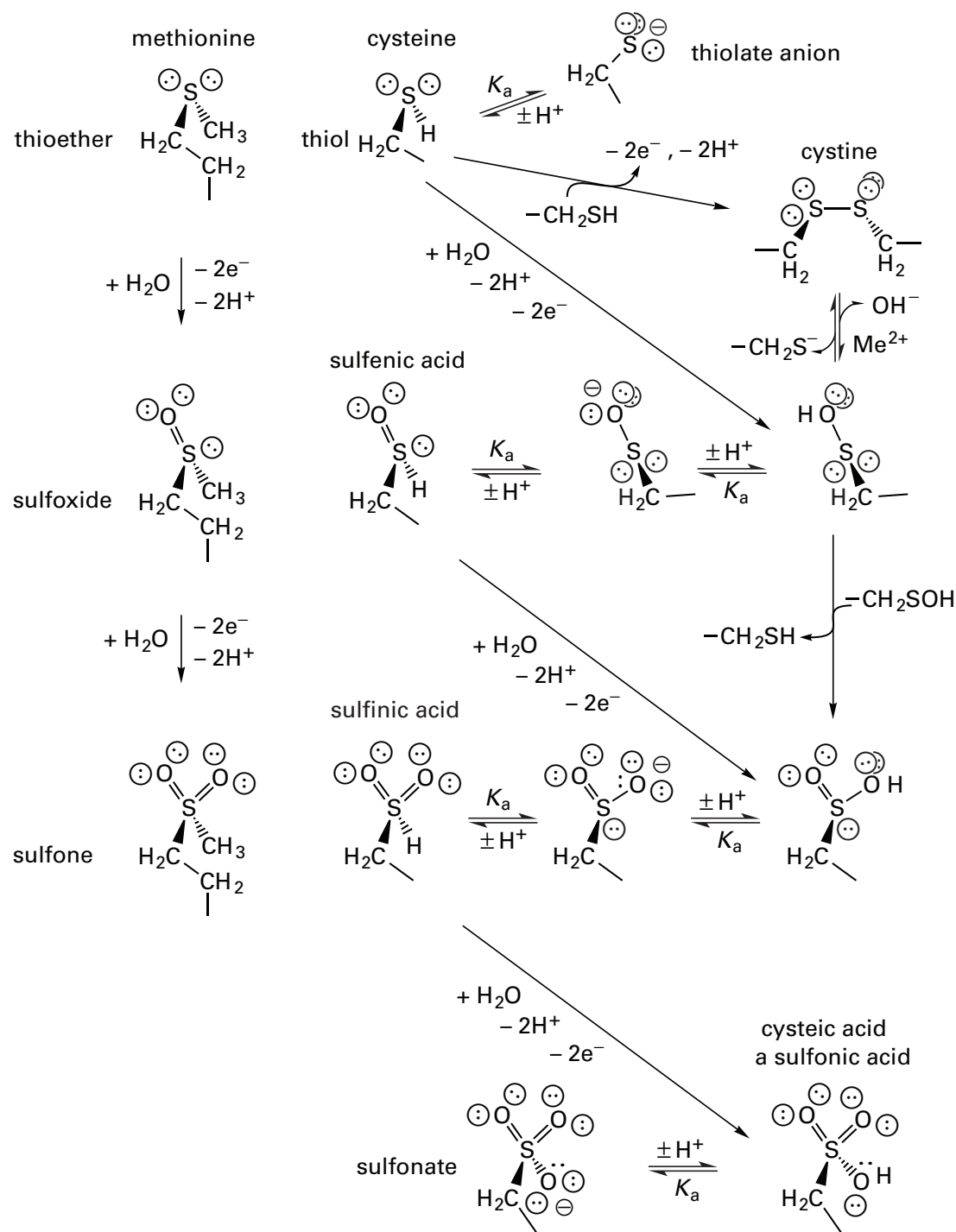
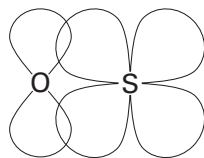


Figure 2-8: Products of the oxidation of cysteine and methionine side chains and their conjugate acids and bases. When a cysteine side chain is oxidized by the removal of two electrons, the sulfenic acid is formed, and when a methionine side chain is oxidized by the removal of two electrons, the sulfoxide is formed. One of the tautomers of a sulfenic acid is the lower homolog of a sulfoxide. Cystine is the disulfide of two cysteines formed either by their direct oxidation or by the reaction of the sulfenic acid of one cysteine with the thiol of another cysteine. When a cysteine side chain is further oxidized by the removal of two more electrons, the sulfinic acid is formed, and when a methionine side chain is further oxidized by the removal of two more electrons, the sulfone is formed. One of the tautomers of a sulfinic acid is the lower homolog of a sulfone. Cysteine can be further oxidized by the removal of two more electrons to produce the sulfonic acid.



2-28

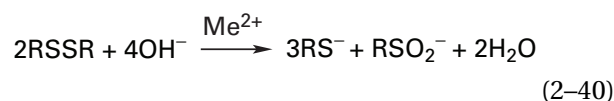
The sulfate dianion is perfectly tetrahedral so sulfur is hybridized [sp^3 , sp^3 , sp^3 , sp^3] to provide the atomic orbitals that overlap to produce the four σ bonds. Every sulfur–oxygen bond is the same length so each oxygen must be electronically identical. The sulfur–oxygen bonds are quite short so they must possess double-bond character. Sulfur has expended its s and p orbitals on the four σ bonds, but as an element of the third period, it has vacant, accessible $3d$ orbitals that can be involved in overlap with adjacent $2p$ orbitals on the oxygens to form dp π bonds. The lobes on a $3d$ orbital are of the proper size to accomplish such an overlap:



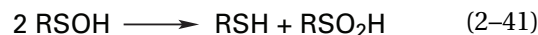
2-29

All of these features are indicated by the six resonance structures of sulfate, one of which is 2-28. The double bonds in these resonance structures indicate the dp π overlaps, not pp π overlaps as are indicated by the double bonds in resonance structures for molecules containing only elements of the second period. Because they are dp π overlaps, the octet rule can be violated in resonance structures involving sulfur. In all of the oxides of sulfur (Figure 2-8), between the tetrahedral sulfur and the various oxygens, there are the σ bonds and dp π bonds.

A **sulfenic acid**, the first oxidation level of a thiol, would be the monothio analog of a peroxide just as a **disulfide** is the dithio analog of a peroxide. One of the tautomers of a sulfenic acid would be the hydrogen analog of a sulfoxide. A **sulfoxide** is the first oxidation level of a thioether. Sulfoxides are stable oxides of sulfur, but sulfenic acids have not been isolated because they are so unstable. They have been postulated to exist as intermediates in the cleavage of disulfides produced by hydroxide in the presence of catalytic amounts of metal ions:



It has been proposed⁴⁶ that the sulfenic acids that are intermediates in this reaction would disappear as the result of their immediate disproportionation,



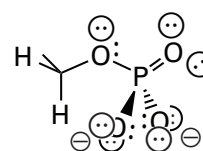
in a reaction homologous to but much more rapid than the disproportionation of peroxides. This reaction, however, requires the collision of two sulfenic acids. A sulfenic acid at a cysteine in a molecule of a native protein can be sterically prevented from such a collision. A cysteine that is buried in the structure of amidophosphoribosyltransferase from *E. coli* is a mixture of its sulfenic acid and its sulfinic acid, each protected and stabilized in turn by the protein that surrounds it.⁶¹ In the absence of such protection, however, a cysteine in the form of a sulfenic acid would be both a strong reductant and a strong oxidant and rapidly susceptible to further oxidation and reduction.

Sulfinic acids, the next level in the oxidation of thiols, are stable compounds that can be isolated. One of the tautomers of a sulfinic acid is the hydrogen analog of a sulfone. **Sulfones**, the next level in the oxidation of thioethers, are also stable oxides of sulfur.

The **sulfonate** is the last oxidation state available to an alkylthiol. The sulfonate of cysteine is **cysteic acid**. Sulfonates are also quite stable. Methionine and cysteine are often purposely converted to methionine sulfone and cysteic acid to make them stable to further oxidation.⁶²

Oxidations such as those outlined in Figure 2-8 often occur adventitiously and can introduce charge heterogeneity into a protein or peptide owing to the formation of cysteic acid. Such oxidations can also cause functional damage to a protein. It is the adventitious oxidation of a methionine in α_1 -antitrypsin caused by cigarette smoke that destroys the function of this protein and produces emphysema.⁶³

Phosphoserine, phosphothreonine, and phosphotyrosine*

phosphoserine
2-30

are formed by posttranslational modification. The phosphate is attached to the side chain of the amino acid as a monoester of phosphoric acid.

The model for the bonding in these phosphorylated amino acids is the trianion of **phosphate**, PO_4^{3-} . The bonding in the phosphate trianion is similar to that in

* Unfortunately, the prefix for the $-\text{PO}_3^{2-}$ functional group officially sanctioned by IUPAC for use by organic chemists is “phosphono”, but the prefix for the same functional group officially sanctioned by IUPAC for use by biochemists is “phospho”. Consequently, some confusion can arise.

sulfate dianion with $dp \pi$ bonds, formed by the overlap of d orbitals on phosphorus and p orbitals on oxygen. These bonds are indicated by the four resonance structures for the phosphate trianion, indicating the equivalence of all of the bonds between phosphorus and oxygen. The σ lone pairs of electrons in the unperturbed trianion must be distributed around each oxygen in such a way that the tetrahedral symmetry of the entire anion is retained. This symmetry is, however, readily perturbed because the $dp \pi$ bonds are polarized owing to the difference in electronegativity between phosphorus and oxygen. For example, the donors of hydrogen bonds are oriented at randomly assumed angles around each of the three equivalent oxygens in the hydrogen phosphate dianion bound to phosphate-binding protein from *E. coli* as if there were no incontrovertible geometry for the lone pairs on each of them.⁶⁴

The acid–base properties of inorganic phosphate (Table 2–1) and monoesters and diesters of phosphoric acid reflect this ability of the system of the $dp \pi$ molecular orbitals to spread negative charge over two or more oxygens because the acid dissociation constants (Table 2–1) are much closer together than one might expect for a series of steps that each increase the negative charge number of a small acid–base by 1 unit. The acid dissociation constants for an alkyl monoester of phosphoric acid ($pK_{a1} = 1.7$ and $pK_{a2} = 6.7$)¹² and for a dialkyl diester of phosphoric acid ($pK_a = 1.5$)¹² are close to those of phosphoric acid itself, but sugar phosphates, and presumably also serine phosphate and threonine phosphate, are more acidic ($pK_{a1} = 0.9$ and $pK_{a2} = 6.1$)¹² because of inductive electron withdrawal.

Problem 2–17:

- At pH 7.0, what fraction of the lysine in the peptide Gly-Pro-Lys-Ala-Thr would be in the neutral nucleophilic form? What fraction at pH 12?
- The ϵ -amino group of lysine in a polypeptide reacts readily with acetic anhydride. Write a mechanism for this reaction.
- At pH 12, 10 °C, and 0.1 M KCl, the lysine in the above pentapeptide would react with acetic anhydride at a rate of $1.3 \times 10^5 \text{ M}^{-1} \text{ min}^{-1}$ (k_N). Write a kinetic mechanism for this reaction at any pH that involves only this rate constant and the acid dissociation constant K_{aK} of the lysine, and solve it for the initial velocity (v_i) of the reaction between lysine and acetic anhydride. Assume that the acid dissociation equilibrium is rapid compared to k_N .

Problem 2–18: In the peptide $\text{CH}_3\text{CO-Gly-Glu-Gly-His-NH}_2$, which acid–bases would be titrating in the region between pH 2 and 11? What are the approximate values of each pK_a ? Plot as a function of pH the fraction of each of the three major ionic forms of the peptide present in solution.

Problem 2–19: Two compounds (A and B) have been isolated from a protein by enzymatic hydrolysis. Both have the composition $\text{C}_5\text{H}_{10}\text{N}_2\text{O}_3$. The titration behavior of the compounds is the following:

compound A		compound B	
pK_{a1}	3.85	pK_{a1}	2.15
pK_{a2}	8.25	pK_{a2}	9.19

After acid hydrolysis for 20 h in 6 M HCl, both compounds have the composition $\text{C}_5\text{H}_9\text{NO}_4$ and the following titration behavior:

compound A'		compound B'	
pK_{a1}	2.16	pK_{a1}	2.16
pK_{a2}	4.32	pK_{a2}	4.32
pK_{a3}	9.95	pK_{a3}	9.95

- What are compounds A and B?
- Explain their behavior on titration.

Problem 2–20: Draw a linkage relationship between the microscopic acid dissociation constants of glycine and its two tautomers in the form of Equation 2–20. The values of pK_a for the two macroscopic acid dissociation constants of glycine are $pK_{a1} = 2.34$ and $pK_{a2} = 9.6$. The macroscopic pK_a for glycolic acid is 3.82 and that for acetic acid is 4.75. Estimate the equilibrium constant between the two tautomers of glycine and its four microscopic equilibrium constants.

References

- Bennet, A.J., Wang, Q.P., Slebockatilk, H., Somayaji, V., Brown, R.S., & Santarsiero, B.D. (1990) *J. Am. Chem. Soc.* 112, 6383–6385.
- Wang, Y., Purrello, R., Georgiou, S., & Spiro, T.G. (1991) *J. Am. Chem. Soc.* 113, 6368–6377.
- Kuhn, H., Eggert, L., Zabolotsky, O.A., Myagkova, G.I., & Schewe, T. (1991) *Biochemistry* 30, 10269–10273.
- Wall, M.A., Socolich, M., & Ranganathan, R. (2000) *Nat. Struct. Biol.* 7, 1133–1138.
- Taylor, R., Kennard, O., & Versichel, W. (1983) *J. Am. Chem. Soc.* 105, 5761–5766.
- Jelsch, C., Teeter, M.M., Lamzin, V., Pichon-Pesme, V., Blessing, R.H., & Lecomte, C. (2000) *Proc. Natl. Acad. Sci. U.S.A.* 97, 3171–3176.
- Jiang, J.C., Wang, Y.S., Chang, H.C., Lin, S.H., Lee, Y.T., & Niedner-Schatteburg, G. (2000) *J. Am. Chem. Soc.* 122, 1398–1410.
- Eigen, M. (1964) *Angew. Chem., Int. Ed. Engl.* 3, 1–19.
- Yang, X., & Castleman, A.W., Jr. (1989) *J. Am. Chem. Soc.* 111, 6845–6846.
- Wei, S., Shi, Z., & Castleman, A.W., Jr. (1991) *J. Chem. Phys.* 94, 3268–3270.
- March, J. (1985) *Advanced Organic Chemistry: Reactions, Mechanisms, and Structure*, 3rd ed., pp 220–223, Wiley, New York.
- Jencks, W.P., & Regenstein, J. (1976) in *Handbook of Biochemistry and Molecular Biology*, 3rd Edition:

- Physical and Chemical Data* (Fasman, G.D., Ed.) Vol. I, pp 305–351, CRC Press, Cleveland, OH.
13. Taft, R.W., Gal, J.F., Geribaldi, S., & Maria, P.C. (1986) *J. Am. Chem. Soc.* 108, 861–863.
 14. Fersht, A.R. (1971) *J. Am. Chem. Soc.* 93, 3504–3515.
 15. Capon, B., & Zucco, C. (1982) *J. Am. Chem. Soc.* 104, 7567–7572.
 16. Zacharias, D.E., Murray-Rust, P., Preston, R.M., & Glusker, J.P. (1983) *Arch. Biochem. Biophys.* 222, 22–34.
 17. Abrams, W.R., & Kallen, R.G. (1976) *J. Am. Chem. Soc.* 98, 7789–7792.
 18. Cox, R.A., Druet, L.M., Klausner, A.E., Modro, T.A., Wan, P., & Yates, K. (1981) *Can. J. Chem.* 59, 1568–1573.
 19. Schollhorn, H., Thewalt, U., & Lippert, B. (1989) *J. Am. Chem. Soc.* 111, 7213–7221.
 20. Sambrano, J.R., de Souza, A.R., Queralt, J.J., & Andres, J. (1976) *Chem. Phys. Lett.* 317, 437–443.
 21. Gorb, L., & Leszczynski, J. (1998) *J. Am. Chem. Soc.* 120, 5024–5032.
 22. Good, N.E., Winget, G.D., Winter, W., Connolly, T.N., Izawa, S., & Singh, R.M. (1966) *Biochemistry* 5, 467–477.
 23. Kresge, A.J., Liebovitch, M., & Sikorski, J.A. (1992) *J. Am. Chem. Soc.* 114, 2618–2622.
 24. Peterson, M.R., & Csizmadia, I.G. (1979) *J. Am. Chem. Soc.* 101, 1076–1079.
 25. Miyazawa, T., & Pitzer, K.S. (1959) *J. Chem. Phys.* 30, 1076–1086.
 26. Allinger, N.L., & Chang, S.H.M. (1977) *Tetrahedron* 33, 1561–1567.
 27. Blom, C.E., & Gunthard, H.H. (1981) *Chem. Phys. Lett.* 84, 267–271.
 28. Hocking, W.H. (1976) *Z. Naturforsch., A* 31A, 1113–1121.
 29. Li, Y., & Houk, K.N. (1989) *J. Am. Chem. Soc.* 111, 4505–4507.
 30. Jung, M.E., & Gervay, J. (1991) *J. Am. Chem. Soc.* 113, 224–232.
 31. Gandour, R.D. (1981) *Bioorg. Chem.* 10, 169–176.
 32. Tadayoni, B.M., Parris, K., & Rebek, J., Jr. (1989) *J. Am. Chem. Soc.* 111, 4503–4505.
 33. Chivers, P.T., Prehoda, K.E., Volkman, B.F., Kim, B.M., Markley, J.L., & Raines, R.T. (1997) *Biochemistry* 36, 14985–14991.
 34. Edsall, J.T., Martin, R.B., & Hollingworth, B.R. (1958) *Proc. Natl. Acad. Sci. U.S.A.* 44, 505–518.
 35. Jeng, M.F., Holmgren, A., & Dyson, H.J. (1995) *Biochemistry* 34, 10101–10105.
 36. Qin, J., Clore, G.M., & Gronenborn, A.M. (1996) *Biochemistry* 35, 7–13.
 37. Takahashi, N., & Creighton, T.E. (1996) *Biochemistry* 35, 8342–8353.
 38. Jeng, M., & Dyson, H.J. (1996) *Biochemistry* 35, 1.
 39. McIntosh, L.P., Hand, G., Johnson, P.E., Joshi, M.D., Korner, M., Plesniak, L.A., Ziser, L., Wakarchuk, W.W., & Withers, S.G. (1996) *Biochemistry* 35, 9958–9966.
 40. Holler, T.P., & Hopkins, P.B. (1988) *J. Am. Chem. Soc.* 110, 4837–4838.
 41. Zawadzke, L.E., & Berg, J.M. (1992) *J. Am. Chem. Soc.* 114, 4002–4003.
 42. Wolfenden, R.V., Cullis, P.M., & Southgate, C.C. (1979) *Science* 206, 575–577.
 43. Nozaki, Y., & Tanford, C. (1967) *J. Biol. Chem.* 242, 4731–4735.
 44. Keim, P., Vigna, R.A., Morrow, J.S., Marshall, R.C., & Gurd, F.R. (1973) *J. Biol. Chem.* 248, 7811–7818.
 45. Ballinger, P., & Long, F.A. (1960) *J. Am. Chem. Soc.* 82, 795–798.
 46. Calvin, M. (1954) in *Glutathione* (Colowick, S., Lazarow, A., Racker, E., Schwartz, D. R., Stadtman, E., & Waelsch, H., Eds.) p 9, Academic Press, New York.
 47. Tanford, C. (1962) *Adv. Protein Chem.* 17, 69–165.
 48. Tanford, C. (1968) *Adv. Protein Chem.* 23, 121–282.
 49. Martin, R.B., Edsall, J.T., Wetlaufer, D.B., & Hollingsworth, B.R. (1958) *J. Biol. Chem.* 233, 1429–1435.
 50. McNutt, M., Mullins, L.S., Raushel, F.M., & Pace, C.N. (1990) *Biochemistry* 29, 7572–7576.
 51. Lennette, E.P., & Plapp, B.V. (1979) *Biochemistry* 18, 3933–3938.
 52. Edsall, J.T., & Wyman, J. (1958) *Biophysical Chemistry*, Vol. I, Academic Press, New York.
 53. Keim, P., Vigna, R.A., Nigen, A.M., Morrow, J.S., & Gurd, F.R. (1974) *J. Biol. Chem.* 249, 4149–4156.
 54. Westheimer, F.H. (1995) *Tetrahedron* 51, 3–20.
 55. Stites, W.E., Gittis, A.G., Lattman, E.E., & Shortle, D. (1991) *J. Mol. Biol.* 221, 7–14.
 56. Gratzer, W.B., & Minalyi, E. (1976) in *Handbook of Biochemistry and Molecular Biology, 3rd Edition: Proteins* (Fasman, G.D., Ed.) Vol. I, pp 186–191, CRC Press, Cleveland, OH.
 57. Edelhoch, H. (1967) *Biochemistry* 6, 1948–1954.
 58. Reynolds, W.F., Peat, I.R., Freedman, M.H., & Lyster, J.R., Jr. (1973) *J. Am. Chem. Soc.* 95, 328–331.
 59. Matthew, J.B., & Richards, F.M. (1982) *Biochemistry* 21, 4989–4999.
 60. Gundlach, H.G., Moore, S., & Stein, W.H. (1959) *J. Biol. Chem.* 234, 1761–1764.
 61. Muchmore, C.R., Krahn, J.M., Kim, J.H., Zalkin, H., & Smith, J.L. (1998) *Protein Sci.* 7, 39–51.
 62. Hirs, C.H.W. (1967) *Methods Enzymol.* 11, 59–62.
 63. Johnson, D., & Travis, J. (1979) *J. Biol. Chem.* 254, 4022–4026.
 64. Luecke, H., & Quiocho, F.A. (1990) *Nature* 347, 402–406.

Chapter 3

Sequences of Polymers

By direct chemical analysis of purified proteins, it has been shown that they are composed of linear polymers of amino acids, referred to as polypeptides. These polymers are formed by a ribosome that reads the messenger RNA and converts the sequence of codons into a sequence of amino acids coupled covalently together in the dictated order. Every polypeptide begins its existence as a single polymer of amino acids of a precise length coupled in a precise order. By and large, this polymer of amino acids is conserved in the mature protein. On its way to maturity, however, various alterations can occur. One class of such alterations is the one that includes changes to the sequence of the amino acids. Short segments of amino acids are often removed from the amino-terminal or carboxy-terminal ends of the protein or cut out of the middle, leaving a broken chain. If such an alteration occurs, it causes the actual amino acid sequence of the polypeptide in a mature protein to differ from the sequence encoded in the messenger RNA.

The sequence of the amino acids in a mature polypeptide can be determined directly, but this is rarely done anymore. It is far easier to sequence the messenger RNA for the protein and translate the sequence of nucleotides into a sequence of amino acids. Because an amino acid sequence determined today is almost always the one encoded by the messenger RNA, alterations in the amino acid sequence of a protein that occur naturally during its maturation often escape detection initially. Eventually, however, most are detected, for example, as unexpected behavior of the protein upon electrophoresis or an incorrect mass on mass spectrometry, and then the sequence of the mature protein must be defined by direct analysis. This direct analysis always relies heavily on the knowledge of the amino acid sequence encoded by the messenger RNA because the lion's share of the original amino acid sequence usually remains in the polypeptides forming the mature protein.

As part of the process that produces a mature protein, other changes are often made to the constituent polypeptides. These changes are either chemical modifications of the amino acids themselves or the attachment of other compounds to the amino acids. For the most part, these posttranslational modifications are unpredictable, and each presents a challenge in analytical chemistry. There is a series of such modifications, however, that consists of the addition of oligosaccharides to particular amino acids, and these modifications are

defined by the sequences in which the sugars are linked together in these oligomers.

With the exception of the unexpected posttranslational modifications, which are relatively infrequent, defining the covalent structure of a mature polypeptide is an exercise in the sequencing of polypeptides, nucleic acids, and oligosaccharides.

Sequencing of Polypeptides

Each naturally occurring polypeptide (2–15) has its own length and its own amino acid sequence. The **amino acid sequence** is the order in which the side chains of the amino acids (R_i in 2–15) are arranged along the polymer. The continuous lengths of the polypeptides found in molecules of protein, and hence the lengths of their unique sequences, can be quite long. For example, human apolipoprotein B100 is 4560 amino acids (aa) long,¹ human mucin MUC2 is 5159 aa long,² and human cardiac titin is 26,926 aa long.³ The amino acid sequence of a given polypeptide is written as a word, each of whose letters stands for an amino acid. The word begins at the amino terminus, ends at the carboxy terminus, and is usually spelled correctly.

The amino acid sequence of a polypeptide determines which protein it will become. Bovine pancreatic ribonuclease can be defined as the protein produced in the pancreas of a steer that can cleave ribonucleic acid at random along its length in a reaction that leaves the phosphate on the 2'- and 3'-positions of the products, or it can be defined as the folded polypeptide, 124 amino acids long, with the amino acid sequence KETAAAKFER-QHMDSSTSAASSSNYCNQMMKSRNLTKDRCKPVNTFV-HESLADVQAVCSQKNVACKNGQTNCYQSYSTMSITDC-RETGSSKYPNCAYKTTQANKHIIIVACEGNPYVPVHFDASV. That the sequence is sufficient to define ribonuclease has been demonstrated by total synthesis.⁴ A similar demonstration was made for the peptidase from human immunodeficiency virus.⁵

The complete amino acid sequences of polypeptides were, in the past, determined directly. The amino acids in a polypeptide can be removed in single steps from the amino-terminal end by the **Edman degradation**⁶ (Figure 3–1).^{*} The strategy of the Edman degradation relies upon

^{*} From here on only those lone pairs involved in each step of a chemical mechanism will be drawn.

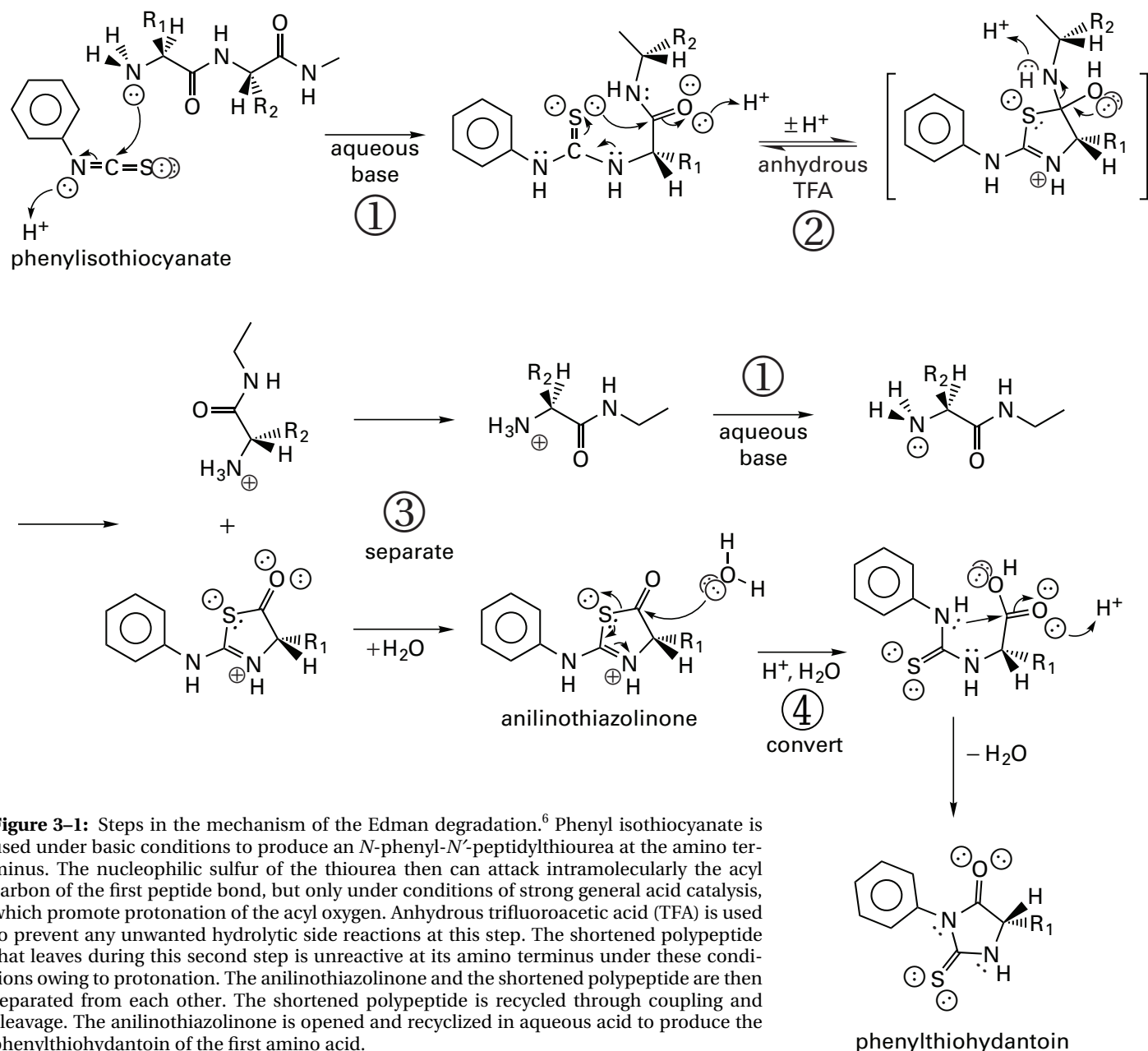


Figure 3-1: Steps in the mechanism of the Edman degradation.⁶ Phenyl isothiocyanate is used under basic conditions to produce an *N*-phenyl-*N'*-peptidylthiourea at the amino terminus. The nucleophilic sulfur of the thiourea then can attack intramolecularly the acyl carbon of the first peptide bond, but only under conditions of strong general acid catalysis, which promote protonation of the acyl oxygen. Anhydrous trifluoroacetic acid (TFA) is used to prevent any unwanted hydrolytic side reactions at this step. The shortened polypeptide that leaves during this second step is unreactive at its amino terminus under these conditions owing to protonation. The anilinothiazolinone and the shortened polypeptide are then separated from each other. The shortened polypeptide is recycled through coupling and cleavage. The anilinothiazolinone is opened and recycled in aqueous acid to produce the phenylthiohydantoin of the first amino acid.

the separation of the chemistry into two discrete, controlled steps (labeled ① and ② in Figure 3-1) that permit the removal of one amino acid at a time from the polypeptide as the **phenylthiohydantoin**. The phenylthiohydantoin from each step can be positively identified on chromatography by adsorption.⁷

Only in fortuitous circumstances, however, can the Edman degradation be run for more than 20 or 30 cycles. The necessity for two steps in each cycle as well as the step separating the shortened polypeptide from the thiazolinone, none of which can be performed in 100% yield, causes the cumulative yield of phenylthiohydantoin to decrease inexorably and noise to increase apace. Side reactions such as random hydrolysis of the polypeptide and cyclization of amino-terminal glutamines to

pyrrolidones⁸ also increase noise and lower yield, respectively. Methods for sequencing a polypeptide from its carboxy terminus⁹ and alternative methods for sequencing one from its amino terminus¹⁰ have been described. So far, the former have been far less reliable than the Edman degradation and the latter have been supplanted by automated machines the chemistry of which relies on the Edman degradation.⁷ These machines are able to provide a sequence from tens of picomoles of a peptide, but they have not overcome the inherent shortcomings of the chemistry.

In its present applications, the **automated Edman degradation** is performed on peptides or polypeptides noncovalently¹¹ attached to thin membranes of glass fiber⁷ or poly(vinylidene difluoride).¹² Because the pep-

tide remains bound to a solid phase, the reagents, in solution or as gases, can be sequentially applied to and removed from the peptide efficiently. It is also possible to transfer polypeptides that have been separated by electrophoresis onto these supports and then submit them to sequencing.¹²⁻¹⁴

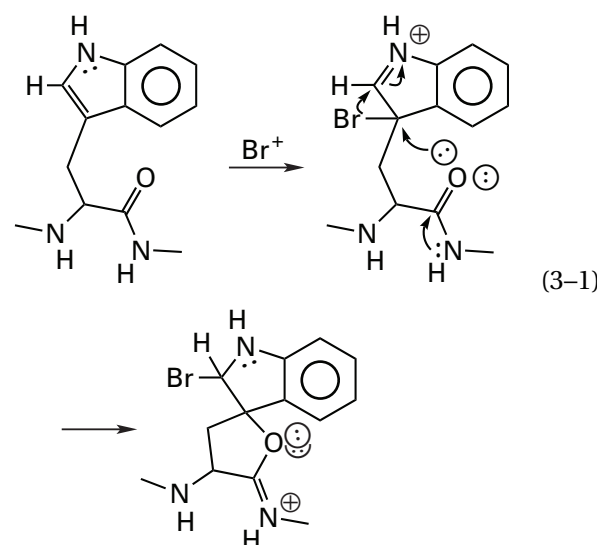
Because polypeptides cannot be sequenced in their entirety by the Edman degradation, they are cleaved into pieces, or **peptides**, that can be. This cleavage can be performed with **endopeptidases** that hydrolyze the peptide bonds at the locations of specific amino acid residues in the sequence (Figure 3-2).¹⁵⁻²⁵ All of these enzymes, except the papain from *Zingiber officinale*, have been used to digest long polypeptides specifically during elucidations of their complete amino acid sequences. Because these enzymes cleave only peptide bonds adjacent to specific amino acids, high yields of a reasonable number of peptides, each with a specific sequence, can be obtained from a long polypeptide.

If polypeptides are to be cleaved by endopeptidases, they must be unfolded or **denatured**. A folded, compact molecule of protein is usually resistant to digestion by endopeptidases for steric reasons. Although the most common way to denature a protein to prepare it for digestion is to precipitate it irreversibly at high temperature, this approach can fail. If it does, denaturing the protein that is to be cleaved without simultaneously denaturing the endopeptidase, which is itself a protein, requires some strategy. Usually the chemical modification of one type of amino acid in the polypeptide while it is unfolded in a solution of a salting-in solute such as urea is sufficient to prevent it from refolding after the denaturant is removed. The carboxymethylation of cysteines with 2-iodoacetate, after the cystine side chains in the protein have been reduced,²⁶ and the maleylation of lysines^{17,27} are examples of this strategy. When proteins that are normally embedded in biological membranes are removed from the membrane, their polypeptides often remain soluble and unfolded and can be cleaved with endopeptidases.²⁸ Some endopeptidases are themselves quite stable and will function in solutions of denaturants sufficient to unfold the protein to be cleaved.

At times it is useful to cleave a polypeptide at only one or two specific locations in its sequence so that long fragments can be isolated from it. The most common way that this is done is to take advantage of the resistance of the native, properly folded protein to digestion by endopeptidases. The consequence of this resistance is that when a properly folded protein is treated with an endopeptidase such as trypsin or chymotrypsin, often only one or two of its peptide bonds are exclusively hydrolyzed, and this hydrolysis produces the long fragments desired. Because this is completely the result of steric effects, no control over the location of the sites of cleavage, other than that exerted by the intrinsic specificity of the endopeptidase, can be exercised.

Polypeptides can also be cleaved chemically. The

paradigm of **chemical cleavages** is that produced to carboxy-terminal sides of methionines by **cyanogen bromide** (Figure 3-3).²⁹ Several other chemical cleavages of more limited usefulness have been developed. 2-Nitro-5-thiocyanatobenzoate induces cleavage on the amino-terminal side of cysteine residues (Figure 3-4), but the yield is less than quantitative and the amino terminus of the carboxy-terminal product is blocked.³⁰ Cleavage at tryptophan residues can be performed chemically with brominating agents under heterolytic conditions.³¹



This reaction proceeds through a bromonium cation that results from insertion of Br^+ into the olefin between carbons 2 and 3 of the indole to create an electrophilic center. A nucleophilic attack of the acyl oxygen five atoms away then occurs as in the cleavage with cyanogen bromide. The resulting iminolactone hydrolyzes as it does in the cleavage with cyanogen bromide to release a fragment with a free amino terminus from the carboxy-terminal side of the tryptophan. The olefin between carbons 2 and 3 in indole is an easily brominated position, and the mildest brominating agent capable of reacting at this location should be used under the mildest conditions to avoid widespread bromination of the polypeptide elsewhere.³²

A chemical cleavage that can produce large fragments from a polypeptide is the cleavage that occurs preferentially at the peptide bond between an aspartate and a proline under mildly acidic conditions (Figure 3-5).³³ This **cleavage with acid** results from intramolecular attack of the carboxylate anion of the 3-carboxy group of the aspartate on its own acyl carbon, the acyl oxygen of which has been protonated, to produce, upon departure of the amide nitrogen of the proline, an anhydride, which is subsequently hydrolyzed.³⁴ The cleavage occurs preferentially at proline because the amine in the initial tetravalent intermediate is by far the poorer leaving group, but proline, because it is a hindered secondary amine, is the best leaving group of all the amino acids.

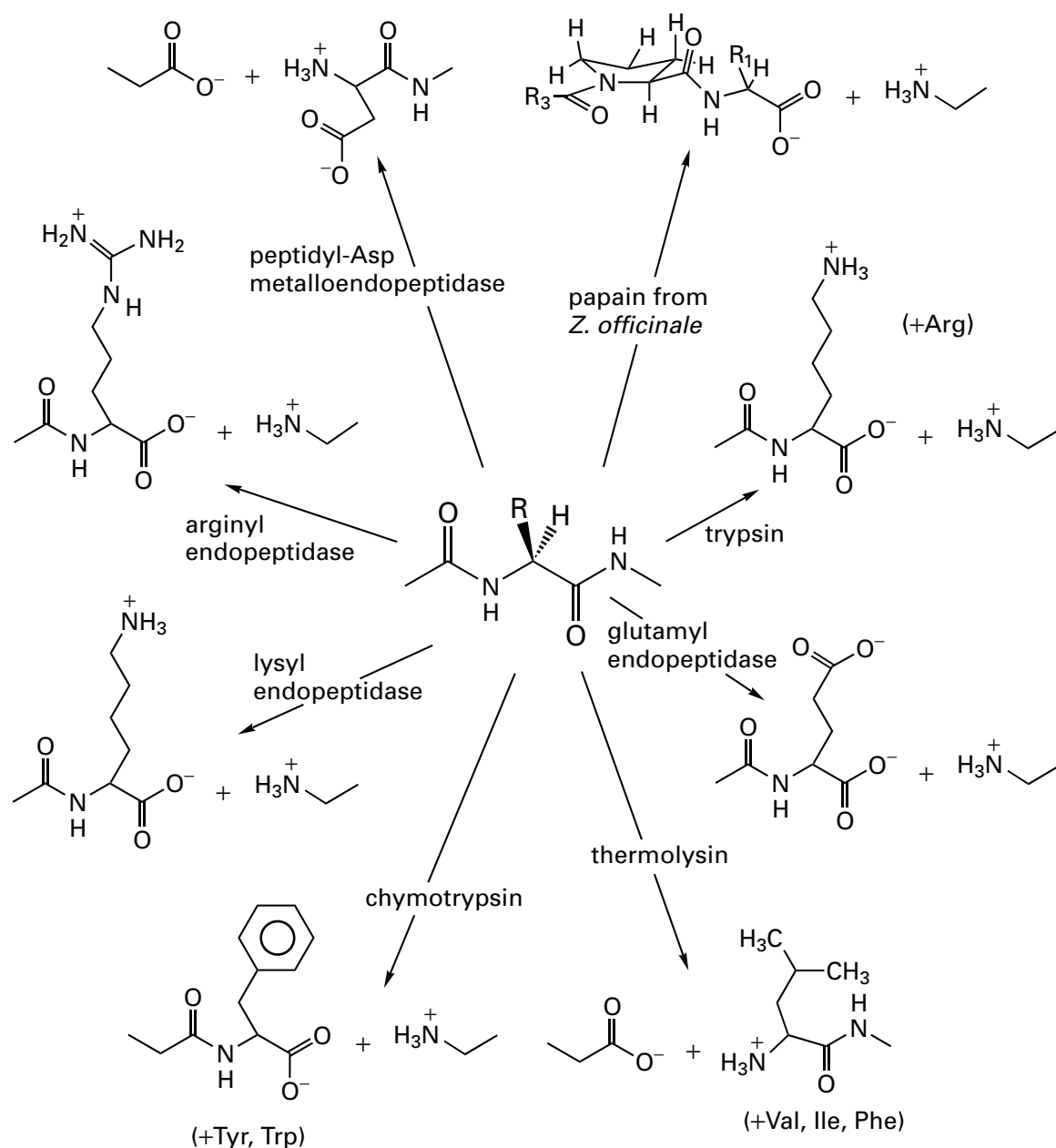


Figure 3-2: Specific cleavage of a polypeptide with endopeptidases. Pancreatic **trypsin** hydrolyzes the peptide bonds on the carboxy-terminal sides of lysine and arginine residues with high specificity to produce a series of peptides. Each of these peptides has the respective lysine or arginine at its carboxy terminus.¹⁵ The lysine side chains can be rendered incapable of being recognized by trypsin by modification with succinic anhydride, maleic anhydride,¹⁷ or citraconic anhydride.¹⁶ The latter two modifications are reversible, and the lysines can be regenerated, after cleavage with trypsin, to yield a series of unmodified peptides the carboxy-terminal residues of which are the respective arginines. **Glutamyl endopeptidase** (Glu-C) from the bacterium *Staphylococcus aureus*, strain V8, hydrolyzes polypeptides with high specificity at the peptide bonds on the carboxy-terminal sides of glutamate residues.¹⁸ Under the proper conditions, the same enzyme also can be made to hydrolyze the bonds on the carboxy-terminal side of aspartate residues. **Thermolysin**, an endopeptidase from the bacterium *Bacillus thermoproteolyticus*, hydrolyzes polypeptides at peptide bonds on the amino-terminal sides of leucine, isoleucine, valine, phenylalanine, methionine, and occasionally alanine and tyrosine.¹⁹ Pancreatic **chymotrypsin** usually catalyzes the hydrolysis of the amide bonds on the carboxy-terminal sides of phenylalanine, tyrosine, and tryptophan.²⁰ **Lysyl endopeptidase** (Lys-C) from either of the bacteria *Achromobacter lyticus*²¹ or *Lysobacter enzymogenes*²² hydrolyzes polypeptides with high specificity at the peptide bonds on the carboxy-terminal sides of lysines. **Arginyl endopeptidase** (Arg-C) from murine submaxillary gland hydrolyzes polypeptides at the peptide bonds on the carboxy-terminal sides of arginines.²³ **Peptidyl-Asp metalloendopeptidase** (Asp-N) from the bacterium *Pseudomonas fragi* hydrolyzes polypeptides at the peptide bonds on amino-terminal sides of aspartate residues²⁴ and, occasionally, glutamate residues. Papain from *Zingiber officinale* hydrolyzes peptides at the next peptide bond beyond the one to the carboxy-terminal side of proline residues with little preference for the amino acids immediately adjacent to the peptide bond that is cleaved.²⁵

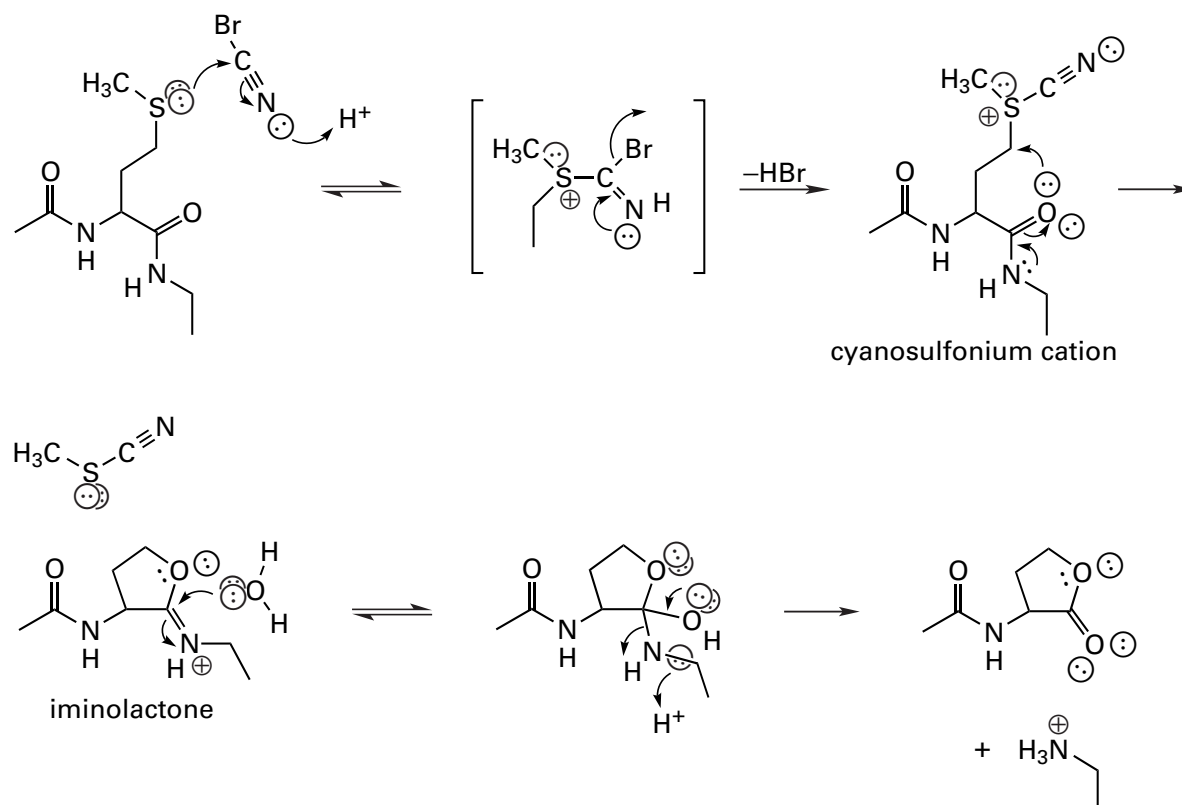


Figure 3-3: Mechanism of cyanogen bromide cleavage of a polypeptide on the carboxy-terminal side of a methionine. At acidic pH, a methionine side chain, because it is not protonated, remains nucleophilic enough to react in an acyl exchange reaction with cyanogen bromide to produce a cyanosulfonium cation. This cationic center causes the carbon of the adjacent methylene to be electrophilic. This electrophile is five atoms away from the weakly nucleophilic acyl oxygen of the same amino acid, and an intramolecular, nucleophilic substitution ensues. The conjugate acid of the iminolactone formed in this nucleophilic substitution is susceptible to hydrolysis under the acidic conditions. This hydrolysis produces a mixture of the lactone and the open γ -hydroxycarboxylic acid of homoserine at the carboxy terminus of the resulting peptide.

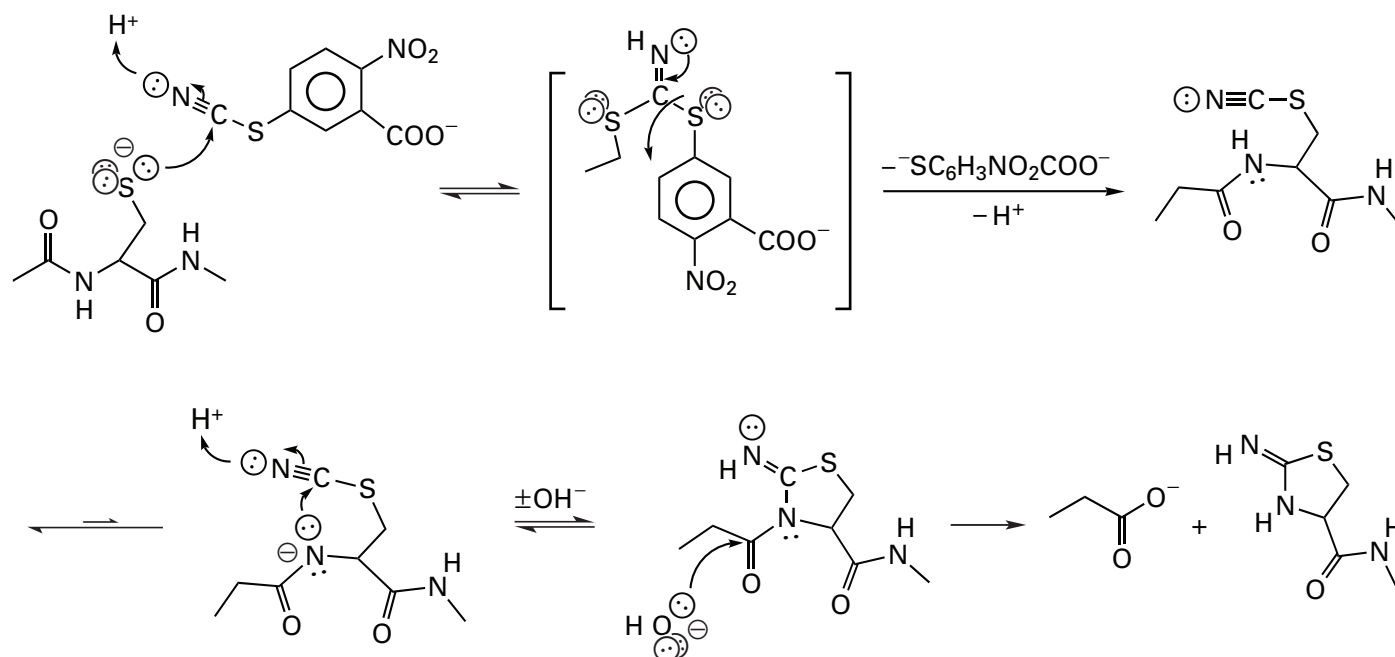


Figure 3-4: Cleavage of a polypeptide to the carboxy-terminal side of cysteine by cyanylation with 2-nitro-5-thiocyanatobenzoate.

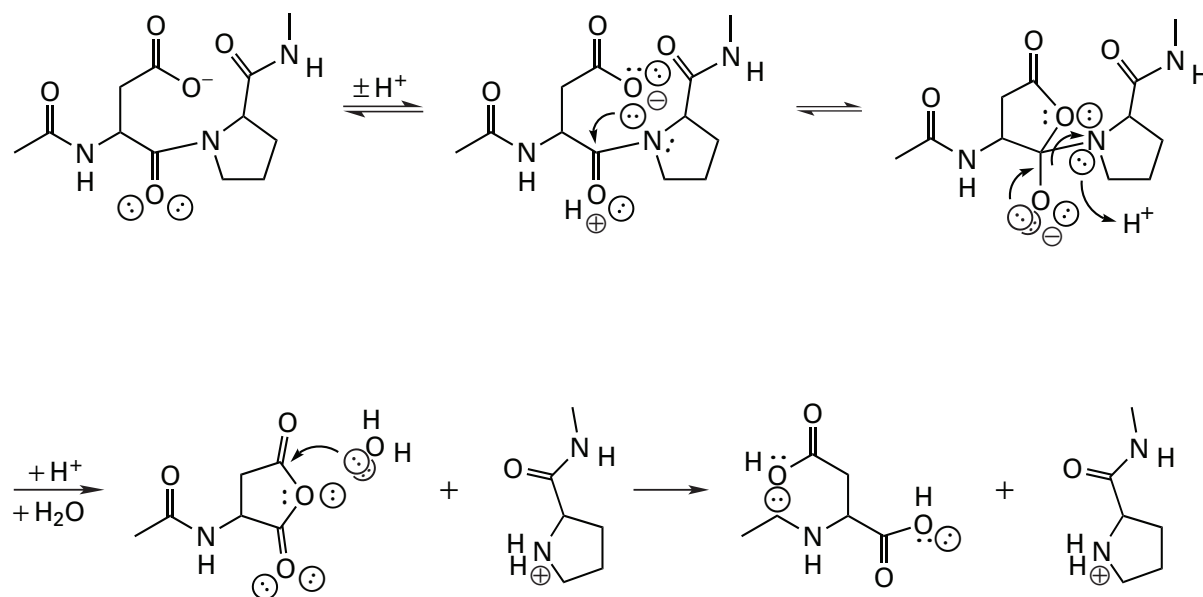


Figure 3-5: Cleavage of a polypeptide at the peptide bond between an aspartate and a proline under acidic conditions.

The treatment with acid can be prolonged intentionally to produce cleavage to the carboxy-terminal side of many more of the aspartates in the polypeptide.^{35,36} Preferential chemical **cleavage between an asparagine and a glycine** can be produced with hydroxylamine at alkaline pH and elevated temperature.^{37,38} Both the cleavage between aspartate and proline and the cleavage between asparagine and glycine produce large fragments of the polypeptide because the frequency at which aspartylprolyl and asparaginylglycyl positions occur within the amino acid sequence of a protein is low.

Each of these enzymatic or chemical cleavages produces a particular set of peptides from a given polypeptide, and the complex mixtures that result must be separated by **chromatography**. Chromatography by molecular exclusion can be used to separate the mixture into groups of peptides of different lengths (Figure 3-6).³⁹ The larger peptides from this first step can be further separated on chromatography by ion exchange with matrices of cellulose or dextran.⁴⁰ Because these larger peptides often aggregate or precipitate, these columns are generally run in solutions of trifluoroacetic acid⁴¹ or formic acid⁴² or denaturants such as urea. At high or low pH, the net charges on all of the peptides are negative or positive, respectively, and aggregation is discouraged by mutual electrostatic repulsion. Large peptides can also be made more soluble by modification of all the lysine side chains with citraconic anhydride to increase their net negative charge at neutral and basic pH.⁴²

The smaller peptides, either those isolated first on chromatography by molecular exclusion or those in the whole digest, can be separated on chromatography by cation exchange with sulfonated polystyrene^{15,43} or **high-pressure liquid chromatography** by reverse-phase adsorption under acidic conditions on alkylated silica gel (Figure 3-7).^{36,41} The latter method can also be used to sep-

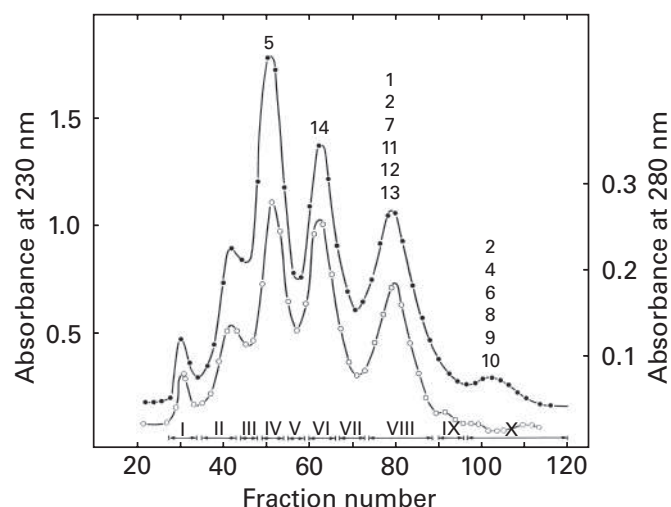


Figure 3-6: Separation of peptides produced by cleavage of S-carboxymethylated human phosphoglycerate kinase with cyanogen bromide.³⁹ The protein (50 mg) was dissolved in 70% formic acid and solid cyanogen bromide was added to a final concentration of 20 mg mL⁻¹. After 24 h, the solution was frozen and the water and cyanogen bromide were removed by sublimation. The cyanogen bromide fragments (50 mg) were applied to a column (1.9 cm × 150 cm) of Sephadex G-75 run in 0.2 M ammonium bicarbonate. The fractions of the effluent were monitored by absorbance at 230 nm (●) and 280 nm (○). Pools (I-X) were made as indicated. The numbers indicate which fragments, identified later in other separations, were in each pool. Reprinted with permission from ref 39. Copyright 1980 *Journal of Biological Chemistry*.

arate large peptides such as cyanogen bromide fragments.⁴¹ The resolution obtained with either chromatography by cation exchange or high-pressure chromatography by adsorption are similar, but the latter has become the method of choice because of its rapidity, the continuous spectrophotometric monitoring it permits, and its adaptability to samples containing small quanti-

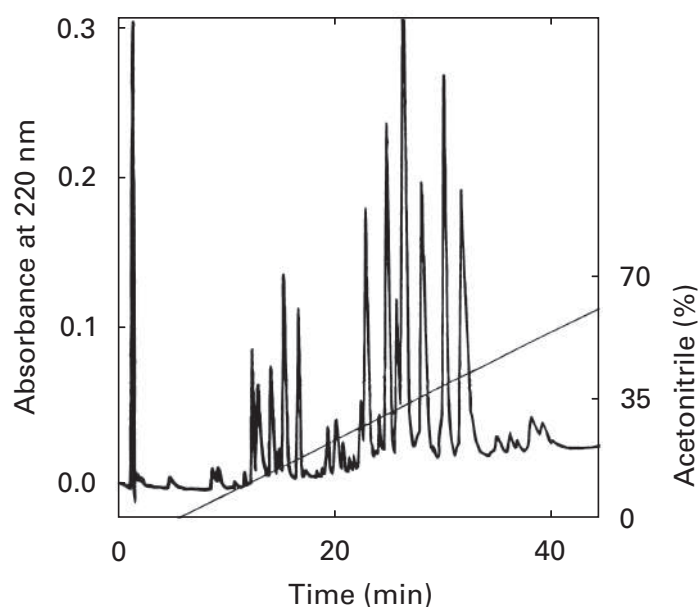


Figure 3-7: Separation of peptides from cytochrome *c* peroxidase on chromatography by adsorption.³⁶ The hemoprotein cytochrome *c* peroxidase from *Paracoccus denitrificans* was dissolved in 8 M urea containing HgCl_2 , and after 20 h, the heme was separated from the protein by molecular exclusion chromatography performed in 5% formic acid. The solvent was evaporated and the resulting solid protein (30 nmol) was suspended in 0.1 M ammonium hydrogen carbonate. Lysyl endopeptidase from *L. enzymogenes* (30 μg) was added to the suspension, and after 4 h at 37°C, the solution had clarified. The sample was evaporated to dryness and redissolved in a dilute solution of trifluoroacetic acid. The peptides were injected onto a column (0.46 \times 25 cm) of a reverse-phase chromatographic medium of octadecylated silica equilibrated with 0.1% trifluoroacetic acid. The peptides were eluted with a linear gradient between 0.1% trifluoroacetic acid and 70% acetonitrile, 0.1% trifluoroacetic acid (solvent B).⁴¹ Peptides were detected by their absorbance at 220 nm. Peaks were pooled as indicated. Reprinted with permission from ref 36. Copyright 1997 American Chemical Society.

ties of peptide. In all cases, the art of the chromatography lies in choosing solvents and buffers that will dissolve the peptides, meet the demands of the chromatographic process chosen for the separation, and be easily removed from the peptides after they have been separated.

Once the peptides have been purified, their **amino acid composition** can be determined by hydrolysis,³⁶ performed under vacuum in 6 M HCl, followed by quantitative cation-exchange chromatography with sulfonated polystyrene (Figure 1-3). In this way, if the peptide is pure and not too long, the amount of each amino acid it contains can be determined. Usually, however, the peptides are sequenced directly because procedures for sequencing by automated Edman degradation⁷ have become more sensitive than procedures for amino acid analysis.

Exopeptidases, such as carboxypeptidase A,⁴⁴ carboxypeptidase B,⁴⁵ serine-type carboxypeptidase,⁴⁶ or leucyl aminopeptidase,⁴⁷ can be used to assist in determining or confirming the sequence of a peptide. These

enzymes remove amino acids one at a time from one or the other of the ends of the peptide. Because the shortened peptide released as a product by one of these enzymes immediately becomes a reactant for the next cleavage, these digestions do not release the amino acids from the respective end in a stepwise fashion as does the Edman degradation but by a progressive process,⁴⁸ and absolute information about sequence beyond three or four residues from the end is rarely obtained with one of these enzymes alone.

A strategy similar to those just described has been developed for the mass spectrometric analysis of mixtures of peptides produced by digesting a protein.⁴⁹

A **mass spectrometer** is an instrument that separates a population of ionic molecules in the gas phase in the order of their mass to charge number ratio (m/z). The ionic molecules, after they have been separated by the mass spectrometer, can be registered individually by a detector to produce a mass spectrum (Figure 3-8),⁴⁹ which records the amount of each ion in the sample as a function of its mass to charge number ratio. A mass spectrometer can also be used to select only ions of a particular mass to charge number ratio, which can then be directed into another instrument. **Quadrupole mass spectrometers** and **ion-trap mass spectrometers** separate the ionic molecules by passing them through specifically designed, oscillating electric fields. **Time-of-flight (TOF) mass spectrometers** accelerate the entire population of ionic molecules in a uniform electric field and then pass them through a vacuum chamber. Because $e_a E_x x = \frac{1}{2} m z^{-1} v^2$ and the electric field (E_x) accelerates all of the ionic molecules over the same distance (x), the time it takes each of them to arrive at the end of the chamber is proportional to the square root of its mass to charge ratio.

There are currently three ways to transfer a biological molecule such as a peptide, an oligosaccharide, a nucleic acid, or a molecule of protein from the aqueous solution in which it is normally found to the gas phase in the form of a **monodisperse vapor** of individual, ionized molecules.

The first method is to pass a dilute aqueous solution containing the macromolecule through an electrospray atomizer,⁵⁰ which produces a mist or **electrospray** so fine that each macromolecule finds itself in its own droplet. The solvent in the droplet evaporates and leaves the intact macromolecule in the gas phase bearing one or more of the elementary positive charges or negative charges that were generated on the surface of the droplet by the atomizer. For proteins and oligosaccharides, the atomizer is usually polarized to produce positive ions, while for nucleic acids, which are already negatively charged in solution, it is polarized to produce negative ions. The elementary charges generated by the atomizer are the result of an excess or a deficit of protons, just as charge is produced on a macromolecule in solution (Figure 1-11). Electrospray produces a family of ions from each macromolecule, each one of the ions differing

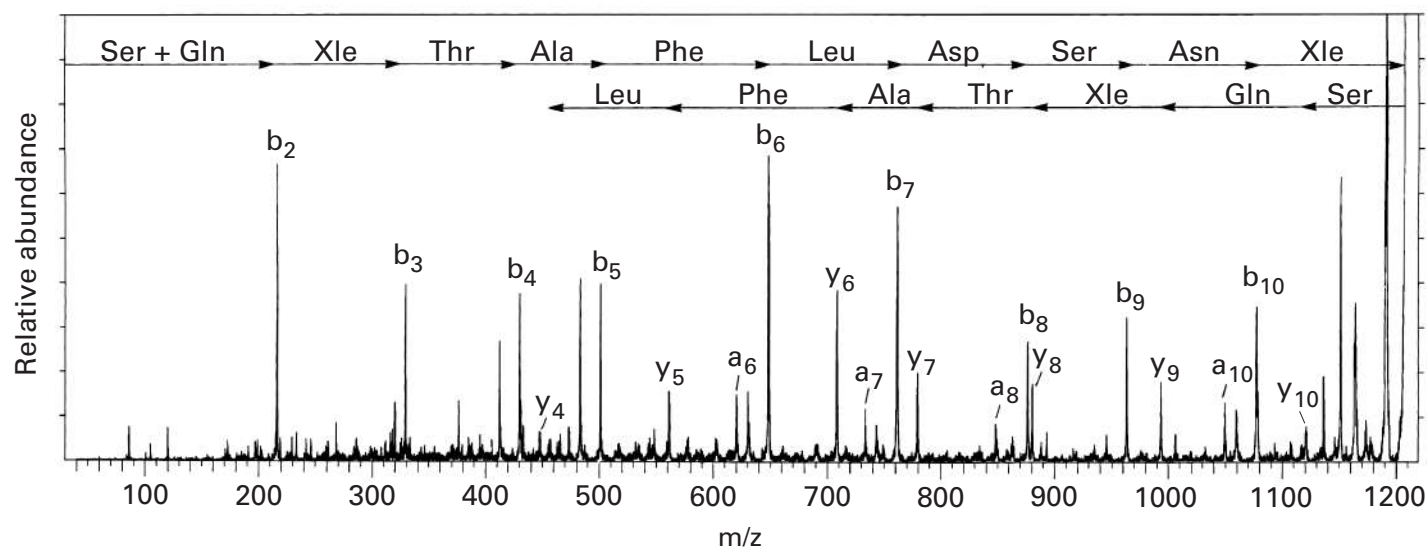


Figure 3–8: Mass spectrometry of a tryptic peptide from thioredoxin.⁴⁹ Thioredoxin from *Chromatium vinosum* was dissolved in 6 M guanidinium chloride and 0.1 M tris(hydroxymethyl)aminomethane, pH 8.5. The cysteines in the protein were reduced with dithiothreitol, and the resulting cysteines were alkylated with iodoacetamide. The product was separated from the small molecules by molecular exclusion chromatography and evaporated to dryness, and a portion (50 nmol) of the dry powder was suspended in 0.1 M ammonium hydrogen carbonate and 0.1 mM CaCl₂. Bovine pancreatic trypsin (12 μ g) was added to the suspension and the digestion proceeded for 2 h at 37°C. The peptides produced by the digestion were collected by evaporating the solvent, and they were redissolved in a dilute solution of trifluoroacetic acid and injected into a column of octadecylated silica equilibrated with 0.05% trifluoroacetic acid. They were eluted with a linear gradient from 0% to 50% acetonitrile in 0.05% trifluoroacetic acid. The peptides in one of the 10 pools of peaks from this chromatographic step were vaporized by fast-atom bombardment from a matrix of glycerol and passed into a tandem mass spectrometer. The beam of monocationic ($M + H^+$) peptide ions of mass 1208.2 Da was selected, fragmented by a beam of helium atoms of high kinetic energy, and passed into the second mass spectrometer. The abundances of the various fragments produced are displayed as a function of their mass. The fragment patterns are labeled as in Equation 3–2, and the amino acids, identified by the distances in mass units between each of the steps, are indicated above the respective steps. Fragments are produced by cleavage at successive points from each end of the peptide. Reprinted with permission from ref 49. Copyright 1987 American Chemical Society.

from the others in the number of elementary charges that it bears. For example, ions of cytochrome *c* ($n_{aa} = 104$) carrying between 11 and 21 elementary positive charges were generated by such an atomizer.⁵⁰

The other two methods rely on the initial monodispersion of the individual macromolecule into a solid glass or liquid of low volatility referred to as a **matrix**. The matrix is formed from a small molecule such as nicotinic acid, a solid, or glycerol, a liquid. An aqueous solution of the macromolecule, at a low molar concentration, and the molecule that will form the matrix itself, at a high molar concentration, is applied to a solid surface, and the water is evaporated to produce the dilutely occupied matrix.

There are two ways to shatter the matrix and in the process eject the macromolecules within it into the gas phase. A beam of neutral argon atoms of high kinetic energy can be directed onto the matrix (**fast-atom bombardment**, FAB), and the explosive collisions of these atoms with the matrix vaporize the macromolecules as a mixture of mainly neutral intact molecules, monoprotonated neutral molecules (monocations), and singly unprotonated neutral molecules (monoanions) dispersed in the gas phase^{51,52} Alternatively, a neodymium-Yag laser emitting light of wavelength 266 nm, which is absorbed by the molecules of the matrix, for example, nicotinic acid, can be directed onto the sample (**matrix-**

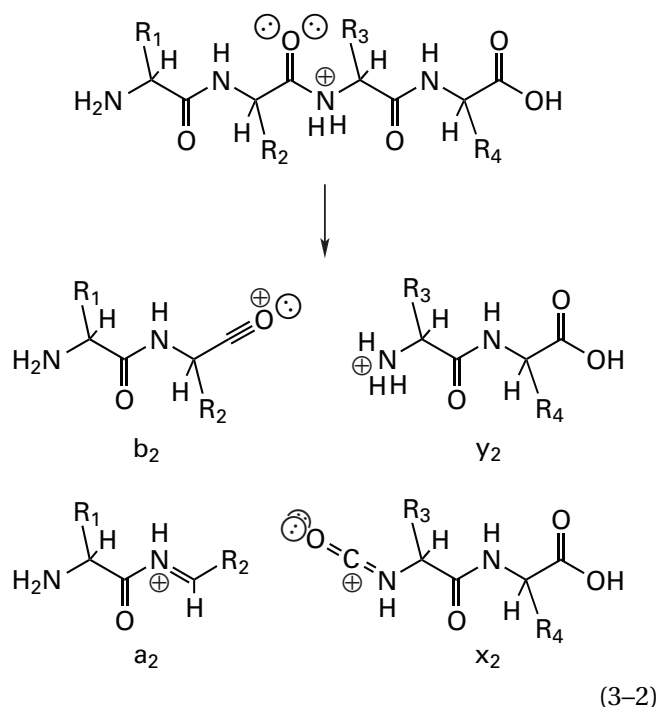
assisted-laser-desorption ionization, MALDI).⁵³ The heat evolved from the absorption of the light by the matrix produces an explosive vaporization of its top layer, ejecting the macromolecules into the gas phase mostly as monoactions⁵³ and presumably neutral molecules and monoanions as well.

Electrospray and fast-atom bombardment are both continuous processes that produce a continuous flux of ionic molecules. This continuous flux can be directed into a quadrupole mass spectrometer to produce continuous streams of separated ionic molecules. Matrix-assisted-laser-desorption ionization, however, is accomplished with short pulses of the laser (<10 ns) to avoid overheating the sample. As a result, the source emits short pulses of ionic molecules, and each pulse contains only a few of each of the individual ionized molecules. In such a situation, a time-of-flight mass spectrometer, which requires a pulsed source, is usually used to separate the gaseous ionic molecules.

Each of these three procedures has its advantages and, unfortunately, its disadvantages. To its detriment, electrospray produces a mass spectrum in which each macromolecule is represented by an envelope of many individual peaks. For example, the envelope for α -amylase contained more than 30 peaks, each representing a molecule of α -amylase with a particular number

(30–60) of elementary positive charges.⁵⁰ These large numbers of peaks generated from each molecule complicate the analysis of mixtures of molecules such as the mixtures of peptides obtained from endopeptidolytic digestion of a protein. Electrospray, however, is the mildest method for producing a high yield of a vapor of ionized molecules, and large molecules of protein can be vaporized. Matrix-assisted-laser-desorption ionization and fast atom bombardment both produce mainly mono-cations or monoanions of a molecule, thereby providing only one unambiguous molecular ion for each molecule. The former technique is able to vaporize significantly larger molecules (up to 200,000 Da) than the latter (up to 20,000 Da),⁵⁰ but the former has the disadvantage that the yield of ions for each pulse is low. Nevertheless, mass spectrometry has become a routine procedure, and in the sequencing of peptides it is rapidly supplanting chemical sequencing based on the Edman degradation.

When mass spectrometry is applied to dissecting proteins and sequencing the resulting peptides,⁴⁹ the polypeptide of the protein is first digested with an endopeptidase. Usually the digest is then separated with one chromatographic step (Figure 3–7). Each pool of a peak from the chromatogram is subjected to vaporization, and the flux of cations produced is directed into a **tandem mass spectrometer**. In addition to being able to register the mass of each peptide in the pool, the first mass spectrometer of the tandem can choose the stream of only one of the ionic molecules and hence only one of the monocationic peptides. This beam of purified peptide cations is passed through an orthogonal beam of helium atoms of high kinetic energy that cleave the molecules of peptide by collision-induced dissociation (CID) into characteristic fragments:



These **fragment ions** are then passed into the second mass spectrometer of the tandem, which can be either a quadrupole mass spectrometer or a time-of-flight mass spectrometer. The resulting pattern of masses that is observed (Figure 3–8) is a set of four separate arrays (a_1 – a_n , b_1 – b_n , y_1 – y_n , and x_1 – x_n), one from each type of fragmentation (Equation 3–2). The number of mass units between each step in each of these arrays provides the sequence of the peptide. In this procedure, the first mass spectrometer performs the separation of the peptides in each chromatographic pool that would normally be performed by subsequent steps of chromatography, and the second mass spectrometer performs the sequencing that would normally be performed by automated Edman degradation.

If only the identity of the polypeptide is desired, not its complete sequence, it is possible to slice a band containing that polypeptide from a polyacrylamide gel, digest it with trypsin, and introduce the entire digest into a tandem mass spectrometer without performing the initial chromatography. Peptide ions that are well resolved by the first mass spectrometer can be selected for fragmentation, and the pattern of the fragments obtained provides the amino acid sequence of those peptides.⁵⁴ In this way, a protein appearing on an electrophoretogram can be positively identified from the amino acid sequences of many of its constituent peptides.

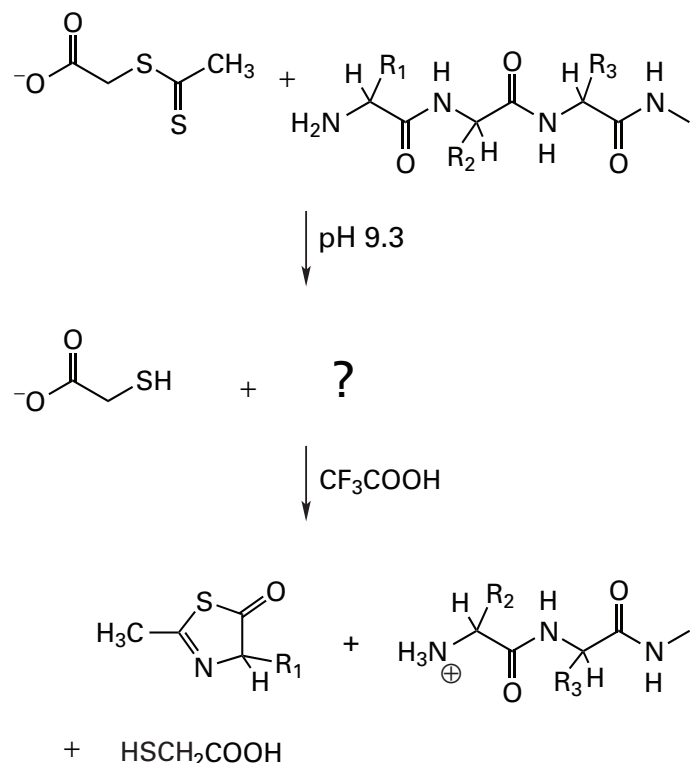
The grand strategy for determining the complete sequence of a polypeptide directly is to separate and sequence all of the peptides from one particular cleavage, to cleave the protein at a set of different locations, to identify all of the peptides in this second set that contain the points of cleavage for the first set, and to sequence these overlapping peptides to learn the order in which the first peptides are arranged in the intact polypeptide. The dramatic epics,^{55–62} in each of which this strategy was applied to another protein and its sequence was revealed, are now seldom produced.³⁶ The expectation and excitement surrounding each of them is only dimly remembered. In their place are myriads of short essays that present the sequences of often long polypeptides. This flood of information has been possible because the sequences of polypeptides are now determined by sequencing DNA complementary to the messenger RNA that encodes them.

Suggested Reading

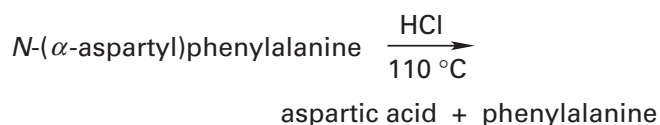
- Suzuki, N., & Wood, W.A. (1980) Complete primary structure of 2-keto-3-deoxy-6-phosphogluconate aldolase, *J. Biol. Chem.* 255, 3427–3435.
- Johnson, R.S., & Biemann, K. (1987) The primary structure of thioredoxin from *Chromatium vinosum* determined by high-performance tandem mass spectrometry, *Biochemistry* 26, 1209–1214.

Problem 3–1: Write a complete mechanism for the following chemical reaction.¹⁰ Draw in important lone pairs

and indicate the combination of nucleophiles and electrophiles with arrows. Use protons where appropriate. For what purpose is this chemical reaction used? Write the step-by-step cycle for using this reaction to accomplish this purpose.



Problem 3-2: Write a complete mechanism for this reaction:



Problem 3-3: A cyanogen bromide fragment has been purified from a digest of certain protein. Consider the following information. The compositions shown in parentheses are those obtained following complete acid hydrolysis in 6 M HCl, 110 °C, for 24 h.

- Complete acid hydrolysis
(1) (E, F, 2 G, homoserine (Hse), K, L, R, S, V)
- Amino terminus
(2) (V)
- Amino acid composition of peptides from tryptic digest
(3) (E, G, R, V)
(4) (F, G, K, S)
(5) (Hse, L)

(D) Edman degradation

peptide	cycle	
	1	2
(3)	V	E
(4)	F	S

(E) Reaction with 2,3-butanedione, followed by tryptic digest

(6) (E, F, 2 G, K, R, S, V)

(7) (Hse, L)

What is the sequence of the fragment? With which amino acid side chain does 2,3-butanedione react?

Problem 3-4: Deduce the sequence of an unknown peptide from the following information.

- Amino acid composition of intact peptide
(A, 2 E, G, L, K, R, 2 S, T)
- Tryptic peptides
(1) (E, T)
(2) (G, K, S)
(3) (A, E, L, R, S)
- Trypsin followed by one cycle of Edman degradation yields the phenylthiohydantoin of S, A, and T.
- Peptides produced by digestion with thermolysin
(4) (A, G, K, 2 S)
(5) (2 E, L, R, T)
- At pH 8.0, tryptic peptide 3 moved on electrophoresis with a positive charge

Problem 3-5: Deduce the sequence of a peptide from the following information.

- Tryptic peptides
(1) (A, E, F)
(2) (Q, S, R, V)
(3) (H, K, V)
- Carboxypeptidase A
(4) A then F and E
- Modification with methyl acetimidate followed by trypsin
(5) (H, K, Q, R, S, 2 V)
(6) (A, E, F)
- Amino-terminal amino acids
peptide (1), F
peptide (2), V
peptide (3), H

(E) Edman degradation

peptide	cycle		
	1	2	3
(2)	V	S	Q

Problem 3-6: What are the expected masses of the 28 cations produced by fragmentation of the protonated gaseous cation ($M + H^+$) of the peptide GGEVATK?⁴⁹

Cloning, Sequencing, Expressing, and Mutating of Deoxyribonucleic Acids

Nucleic acids are linear polymers (see 3-1 below) the monomers of which are **nucleoside 5'-monophosphates**. The covalent bonds that link the monomers together to form the polymer are the oxygen-phosphorus bonds that connect the 3'-hydroxyl group of one nucleoside and the 5'-phosphoryl group of the next. Each of these bonds produces a diester of the respective phosphoryl group (a phosphodiester linkage). Nucleic acids are divided structurally and biologically into ribonucleic acids (RNA), which have a 2'-hydroxyl group on each of their furanosyl rings as in 2-9 to 2-12, and deoxyribonucleic acids (DNA), which are unsubstituted at the 2'-position of their furanosyl rings as in 3-1. Aside from this distinction, every nucleic acid has the same polymer backbone. One end of a molecule of single-stranded nucleic acid is a phosphorylated 5'-hydroxyl group (5'-phosphate); the other end is a 3'-hydroxyl group. The 5'-phosphate and 3'-hydroxyl group are the **5'-end** and **3'-end**, respectively, of the polymer. At the pH usually encountered in living organisms (pH 7-8), the oxygens on each of the phosphoryl diesters in the backbone of a nucleic acid are unprotonated and each monomer bears a full elementary negative charge except for the monomer at the 5'-end, the phosphate of which bears an average of between 1.5 and 2 elementary negative charges, depending on the exact pH.

There are four nucleoside 5'-monophosphates incorporated into a particular nucleic acid as it is synthesized biologically. These four nucleoside 5'-monophosphates are distinguished by the heterocyclic bases they contain (R_i in 3-1). Cytosine (C) is the base in the nucleosides cytidine (2-10) and 2'-deoxycytidine,

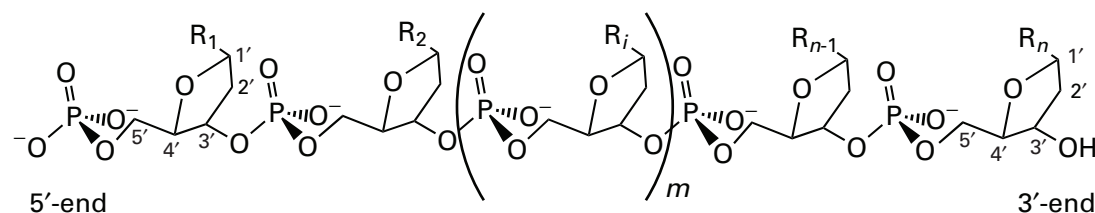
guanine (G) is the base in the nucleosides guanosine (2-11) and 2'-deoxyguanosine, and adenine (A) is the base in the nucleosides adenosine (2-12) and 2'-deoxyadenosine. The ribonucleoside 5'-monophosphates are incorporated into RNA, and the 2'-deoxyribonucleoside 5'-monophosphates are incorporated into DNA. Uracil (U) is incorporated into RNA on the 5'-monophosphate of the nucleoside uridine (2-9). Uridine, however, is converted by dehydroxylation and methylation into **thymidine**, the 2'-deoxyribonucleoside of 5-methyluracil, before its 5'-monophosphate is incorporated into DNA. The base 5-methyluracil is called **thymine** (T).

Within each nucleoside, the base is attached to its respective ribose or 2'-deoxyribose in an azacetal (*N*-glycosidic) linkage (see 2-9 to 2-12) between a pyridine nitrogen or an imidazole nitrogen of the pyrimidine or purine, respectively, and the aldehydic carbon at position 1 in the furanose ring. In the unpolymerized nucleoside phosphates, a monophosphate, diphosphate, or triphosphate group is found on the 5'-carbon. A nucleoside 5'-monophosphate, 5'-diphosphate, or 5'-triphosphate is referred to as a **nucleotide**.

Each nucleic acid has its own length and its own **sequence** in which the nucleotide bases, R_i in 3-1, are arranged. The sequence of a nucleic acid is written as a word, each of whose letters stands for the base of the respective nucleotide. Unless otherwise noted, the word begins at the 5'-end and ends at the 3'-end.

Deoxyribonucleic acid usually and ribonucleic acid often occur as double helices. In a **double helix**, two molecules of nucleic acid, running in opposite directions, are wrapped around each other (Figure 3-9). The bases in the core of the double helix are paired, adenine next to thymine and guanine next to cytosine. Because the positions in the sequence of the one strand of DNA occupied by deoxyadenosine, deoxyguanosine, thymidine, and deoxycytidine are paired with positions in the sequence of the other strand of DNA occupied by thymidine, deoxycytidine, deoxyadenosine, and deoxyguanosine, respectively, the sequence of one strand read 5' → 3' **complements** the sequence of the other strand read 3' → 5' (Figure 3-10). For example, the sequence -AGCAGA- complements the sequence -TCTGCT-.

A polypeptide can be cleaved at specific sites with a particular endopeptidase (Figure 3-2), and DNA can be cleaved at specific sites with **site-specific deoxyribonucleases** (restriction enzymes). Just as trypsin or ther-



3-1

molysin catalyzes hydrolysis of amide bonds in a protein only next to particular amino acids to produce specific peptides, site-specific deoxyribonucleases catalyze the hydrolysis of double-helical DNA only at phosphate diesters within particular sequences of nucleotides (Figure 3–10). The particular sequences of nucleotides and the associated points of cleavage are known as **restriction**

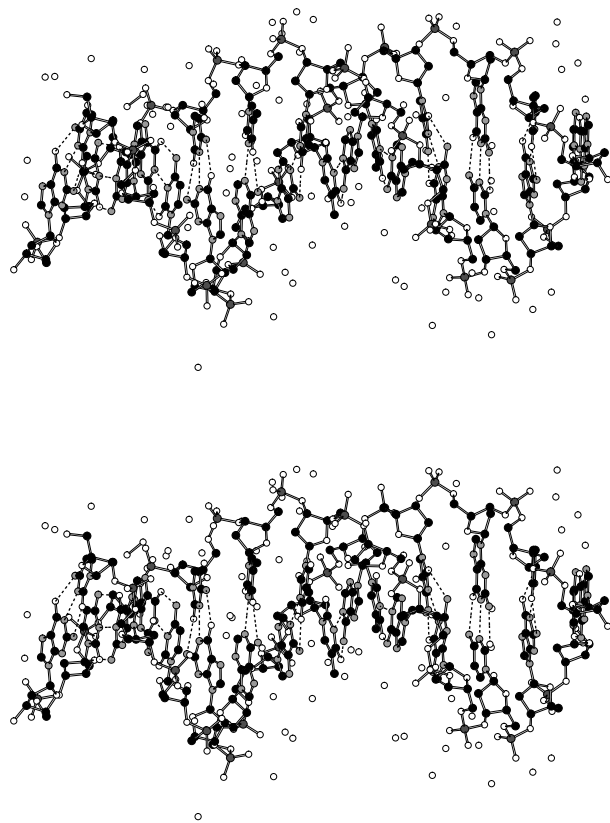
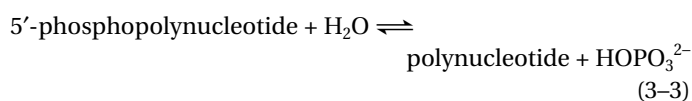


Figure 3-9: Double-helical DNA in the standard B conformation.⁵⁴⁵ A segment of single-stranded DNA with the self-complementary sequence CGCGAATTCGCG was chemically synthesized. When dissolved in solution, the individual molecules paired up and formed double-helical segments of DNA containing two identical strands running antiparallel to each other. The double-helical dimers were crystallized and a crystallographic molecular model was generated. A stereo view of the model is presented in the figure. Atoms of oxygen are white; atoms of nitrogen, light gray; atoms of phosphorus, dark gray; and atoms of carbon, black. The unattached white circles are the oxygens of molecules of water that assume fixed positions in the crystal. The dashed lines represent hydrogen bonds.

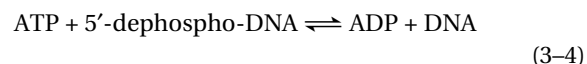
sites, and the fragments of DNA produced by these cleavages are known as **restriction fragments**. Unlike the situation in dissecting proteins, a much larger number of site-specific deoxyribonucleases⁶³ are available, the specificities of which vary in their complexity. The particular sequence recognized by a given site-specific deoxyribonuclease can be anywhere from four to 12 nucleotides long. The longer the sequence recognized, the less frequently will it occur in the DNA, and the longer will be the restriction fragments produced. By using the appropriate site-specific deoxyribonuclease and carrying the digestion to the appropriate degree of completion, restriction fragments can be obtained of a desired size range containing within their population the complete sequence of the original DNA, just as a digest of a protein contains within its population of peptides the complete sequence of the protein.

Site-specific deoxyribonucleases produce restriction fragments with **blunt ends** or **sticky ends**. If the particular enzyme used cleaves phosphodiester linkages in the two strands that are directly opposite each other, the two new ends it produces will both be completely double-stranded and blunt. If the particular enzyme used cleaves phosphodiester linkages on the two strands that are offset relative to each other (Figure 3–10), a short segment of single-stranded DNA will protrude from each of the new ends. Because they were before the cleavage, these two segments will necessarily be complementary to each other in sequence, will adhere to each other when they come in contact, and, consequently, are sticky.

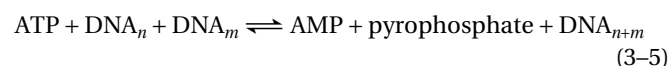
In addition to site-specific deoxyribonucleases, there are several other enzymes that are used to manipulate DNA (Figure 3–10). The phosphate on the 5'-end of a nucleic acid can be removed with polynucleotide 5'-phosphatase



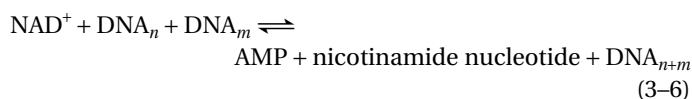
and the phosphate can be added back to the 5'-hydroxyl group of DNA, usually as a radioactive [³²P]phosphate, with polynucleotide 5'-hydroxyl-kinase:



Both DNA ligase (ATP)



and DNA ligase (NAD⁺)



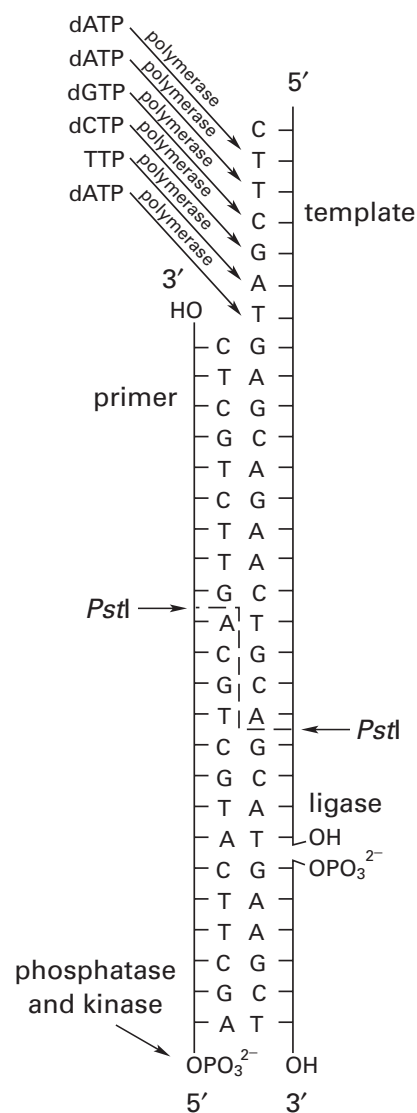
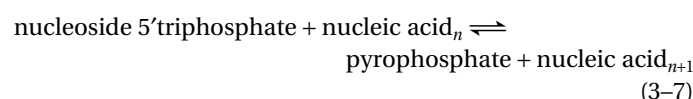


Figure 3–10: Enzymes that are used to manipulate DNA. A double-stranded segment of DNA is represented diagrammatically. In the core of the double helix, the bases are paired adenine next to thymine and cytosine next to guanine. The two antiparallel strands each have a 3'-end, at which there is usually a free 3'-hydroxyl, and a 5'-end, at which there is usually a phosphorylated 5'-hydroxyl group. **Polynucleotide 5'-phosphatase** (phosphatase) is used to remove the phosphoryl group from the 5'-hydroxyl group. **Polynucleotide 3'-hydroxyl-kinase** (kinase) is used to phosphorylate a dephosphorylated 5'-hydroxyl group, usually with radioactive [γ - 32 P]ATP. Either **DNA ligase (ATP)** or **DNA ligase (NAD $^{+}$)** (ligase) can be used to join the 3'-hydroxyl group and the 5'-phosphate at a single-strand break in the DNA between two nucleotides. The two nucleotides to be joined are usually held immediately adjacent to each other by their pairing with partners on the adjacent unbroken strand. **Site-specific deoxyribonucleases** are used to cleave double-stranded DNA on both strands. The cleavages occur within a sequence specific to the particular site-specific deoxyribonuclease chosen. The site-specific deoxyribonuclease *Pst*I, used as an illustration, cleaves each strand between the A and the G in the sequence –CTGCAG– to produce the two complementary, sticky 3'-ends –TGCA. **DNA-directed DNA polymerase** (polymerase) elongates a strand of DNA, the primer, at its 3'-end by successively adding the nucleotides that pair with the consecutive bases on the opposite strand of the DNA, the template. The reactant for each step of the elongation is the deoxyribonucleoside triphosphate of the appropriate base.

specific deoxyribonuclease. The junctions that are effected by this procedure are at random, but the desired product in which the restriction fragment of DNA has been inserted into the middle of the other molecule of DNA are selected from the overall population. The site-specific deoxyribonuclease is often chosen because it produces blunt ends, which can be ligated at random with other blunt ends. If the restriction fragment to be inserted has been produced with one site-specific deoxyribonuclease and the molecule of DNA into which it is to be inserted has been cleaved with another, the sticky ends are usually removed and the pieces ligated at their blunt ends.

Polymerases are enzymes that synthesize a new strand of nucleic acid in consecutive steps:



repair single-stranded breaks in a double helix. The complementarity of the bases around the break usually juxtapose the two ends to be ligated.

Site-specific deoxyribonucleases and DNA ligases are used to **insert** one molecule of DNA into another molecule of DNA. The molecule to be inserted has been prepared by cleaving a longer molecule of DNA with a particular site-specific deoxyribonuclease, usually one that produces sticky ends. The molecule of DNA into which the restriction fragment is to be inserted is then cleaved with the same site-specific deoxyribonuclease to produce a break with the same sticky ends as those on the restriction fragment to be inserted. The two preparations of DNA are then mixed, and the various sticky ends, for example, the two 3' sticky ends, –TGCA, produced by *Pst*I (Figure 3–10), spontaneously pair up. The pairs of resulting offset breaks in the two strands of the double helices are then repaired with DNA ligase (Figure 3–10) to produce a new unbroken molecule of DNA in which the restriction fragment has been inserted into the other molecule of DNA at the restriction site specific to the site-

Polymerases usually require a particular arrangement of double-helical nucleic acid. There must be a shorter strand of nucleic acid, the **primer**, associated in a double helix through complementary base pairing with a longer strand of nucleic acid. The longer strand of nucleic acid, the **template**, must extend beyond the 3'-end of the primer. The polymerase elongates the primer from its 3'-end by adding a nucleotide at each step that is complementary to the adjacent base on the template. **DNA-directed DNA polymerase** synthesizes a single strand of DNA that is complementary to a template of DNA and that remains associated with the template of DNA in a double helix. **RNA-directed DNA polymerase** synthesizes a single strand of DNA that is complementary to a template of RNA and that remains associated with the template of RNA in a double helix. The enzyme pairs A with U and T with A. **DNA-directed RNA polymerase** synthesizes a single strand of RNA that is complementary to a template of DNA but that does not remain associated with the template. It pairs A with T and U with A. The DNA polymerases use the 2'-deoxyribonucleoside

triphosphates as reactants; the RNA polymerases use, the ribonucleoside triphosphates.

Polypeptides are synthesized biologically by ribosomes that translate the sequence of nucleotides in a single-stranded **messenger RNA** (mRNA) into a sequence of amino acids in a polypeptide. The two corresponding words written in the respective sequences are in the same language, the language of the structure of the protein, and they have the same spelling, but the alphabets are different. The alphabet of the polypeptide sequence consists of the 20 amino acids; the alphabet of the messenger RNA consists of triplets of nucleotides. The correspondence between the letters in the two alphabets is known as the **genetic code**. Each triplet specifies a particular amino acid, and the triplets are sequentially arranged in the same order as the amino acids of the protein encoded by the message (Figure 3–11). Because the sequence of the nucleotides, however, is continuous and does not indicate how they are grouped as triplets, there are three ways to divide any sequence of nucleotides into triplets, or three distinct **reading frames**, only one of which encodes the sequence of the protein. If the sequence and the correct reading frame of a messenger RNA have been determined, it can be immediately translated on paper into the sequence of the polypeptide which it encodes.

Messenger RNA is synthesized by DNA-directed RNA polymerase from a gene in the double-helical DNA of the genome of the organism. Its sequence matches that of one of the strands of DNA in the double helix, the **sense strand**, except that uridine monophosphate replaces thymidine monophosphate. During the synthesis of messenger RNA, the other strand of the DNA, the **antisense strand**, serves as the template (Figure 3–11). The sequence of the sense strand of a prokaryotic gene is identical to that of the messenger RNA transcribed from it, and the sequence of the protein encoded by that sense strand can be read directly from the sequence of the genomic DNA. The genomic DNA of eukaryotes, however, contains introns. An **intron** is a segment of unrelated DNA that has been inserted during evolution into the genomic DNA of the eukaryote and that interrupts the sequence on the sense strand that encodes the protein. These introns are **spliced out** of the messenger RNA

before it is read by the ribosome. Although they cause no problems for the organism, introns make it difficult if not impossible to read the sequence of a eukaryotic protein from the sequence of the gene that encodes the sequence of that protein. Consequently, it is the messenger RNA for a eukaryotic protein that must be sequenced.

Almost every protein molecule present at a particular time in a living cell is being continuously produced by ribosomes from messenger RNA molecules, and it follows that if a protein is found in a eukaryotic cell or tissue, the messenger RNA encoding it should be there as well. Messenger RNA can be isolated as a complex mixture of all of the messages normally being expressed in a particular tissue. This isolation is assisted by the fact that all eukaryotic messenger RNAs have a segment of poly(adenosine monophosphate) about 200 bases in length at their 3'-ends. Affinity adsorption with a stationary phase to which poly(thymidine monophosphate) has been attached covalently is used to separate the messenger RNA from all of the other RNA in the homogenate.

The stratagem devised to obtain the nucleic acid sequence of a particular single-stranded messenger RNA in this purified mixture is to transcribe all of the single-stranded messenger RNAs in the mixture into a mixture of double-helical DNAs of the same respective sequences, separate these molecules of DNA biologically, select the DNA derived from the messenger RNA of interest, and sequence that DNA. Deoxyribonucleic acid that has the same sequence in one of its two complementary strands as the sequence of a messenger RNA is referred to as **complementary DNA** (cDNA). Messenger RNA is transcribed into complementary DNA in the laboratory by first using RNA-directed DNA polymerase to synthesize single-stranded DNA complementary in sequence to the messenger RNA. The single strands of DNA end up in hybrid double helices with the messenger RNAs. The RNA is then removed by digesting it with RNase, and then DNA-directed DNA polymerase is used to synthesize the complements to the single strands of DNA. Each strand of this newly synthesized DNA remains associated with its template in a double helix. In its sense strands, this double-helical DNA contains the original sequences of the messenger RNAs. One advantage of the complementary

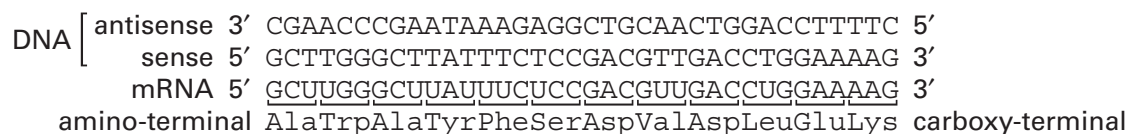


Figure 3–11: Relationships between the sense strand and the antisense strands of a segment of double-helical DNA and the messenger RNA and between the messenger RNA and the amino acid sequence encoded by that messenger RNA. In the messenger RNA, the amino acid sequence of the protein is encoded by triplets of bases. Each triplet of bases is a letter in the alphabet of the messenger RNA. The genetic code is the correspondence between each of these triplets and the amino acid it encodes. The amino acid is the letter in the alphabet of the protein. The messenger RNA is synthesized by DNA-directed RNA polymerase using the antisense strand of the DNA as a template and has the same sequence of nucleotides as the sense strand, except that uridine monophosphate is in place of each thymidine monophosphate. When RNA-directed DNA polymerase synthesizes a single strand of complementary DNA using messenger RNA as a template, that single strand of DNA has the same sequence as the antisense strand from which the messenger RNA was synthesized.

DNA derived from a particular tissue is that it catalogs all of the genes that are being expressed in that tissue.

To **clone** a particular segment of DNA is to insert that DNA, usually present in a complex mixture of other DNAs, into the DNA of a bacteriophage or a bacterium and then isolate a population of identical bacteria or identical bacteriophage, all of which carry just that one segment of DNA. A bacteriophage is a virus that infects and replicates within a bacterium. For the purposes of this discussion, the segment of DNA to be cloned is a segment encoding a protein of interest. In the **cloning** of eukaryotic DNA encoding a protein, complementary DNA is usually the starting point because of the problem of the introns in the genomic DNA. Complementary DNA is also advantageous because tissues producing significant amounts of the protein can be chosen as the source for the messenger RNA, a strategy that increases the chances of finding its complementary DNA. In the cloning of prokaryotic DNA encoding a protein of interest, genomic DNA is usually the starting point⁶⁴ because it is already double-helical DNA, and in prokaryotes there are no problems with introns. The genomic DNA of the bacterium is cut into restriction fragments long enough to contain all or most of the gene encoding the protein.

The complementary DNAs or fragments of genomic DNA in one of these complex mixtures are then usually incorporated into a library in which they can be stored, replicated, and screened. A **library** is a large population of bacteriophage or bacteria, each of which contains within its DNA one of the complementary DNAs or fragments of genomic DNA from the original mixture just as the usual library contains a large population of different books. Each piece of foreign DNA is integrated into the DNA of one of the bacteriophage or one of the bacteria in the library in such a way that it is replicated along with its genomic DNA, ensuring that all of the progeny of that one bacteriophage or bacterium contain the inserted complementary DNA or genomic DNA. Each of the fragments of foreign DNA is inserted into the same location in the DNA of the bacteriophage or bacteria in the library.

If the library consists of **bacteriophage**, the foreign DNA is inserted at the same site in the genomic DNA of each bacteriophage. These genomic DNAs containing the inserts can be biologically replicated to a high concentration by infecting a suspension of bacteria, usually *Escherichia coli*, with the bacteriophage.

Complementary DNAs or fragments of genomic DNA are incorporated into a population of **bacteria** by first inserting them into plasmids. A **plasmid** is a circular molecule of double-stranded DNA that is able to replicate independently of the chromosome in a bacterium. The species of bacteria usually used to carry a plasmid is *E. coli*. In addition to the inserted DNA, each of the various plasmids used for cloning contains a gene causing the bacterium that carries it to be resistant to a particular antibiotic. Consequently, when the plasmids have been incorporated into a population of bacteria, only the

bacteria carrying one of the plasmids will replicate in the presence of the antibiotic. In the library, each antibiotic-resistant bacterium contains a plasmid and most of the plasmids contain a copy of one of the original complementary DNAs or fragments of genomic DNA. Not only do plasmids and bacteriophage permit the inserted DNA to be replicated as they themselves are replicated, they also provide a way of storing the inserted DNA, because once it has been incorporated into the bacteriophage or its plasmid has been incorporated into a bacterium, it is stable for long periods of time if the bacteriophage or bacterium is stored in its dormant state.

Occasionally, the messenger RNA in a tissue producing mainly one protein is so enriched for the messenger RNA encoding that particular protein that the most of the individuals in the library carry complementary DNA for that one messenger RNA, and one of these can be picked out from the rest directly.⁶⁵ Usually, however, the library has to be screened to find an individual carrying the desired complementary DNA or fragment of genomic DNA. To **screen** a library is to isolate bacteriophage or bacteria that carry complementary DNA or fragments of DNA encoding one particular protein from the vast majority of the bacteriophage or bacteria that carry complementary DNA or fragments of genomic DNA encoding other proteins.

When the library is stored in bacteriophage, a continuous lawn of a particular bacterium growing on an agar plate is infected with a dilute solution of those bacteriophage carrying the inserted DNA. Small circular holes or **plaques** appear in the lawn. Each plaque results from the infection and lysis by bacteriophage of the bacteria that had been happily growing within the lawn. All of the bacteriophage in one of the plaques are offspring of a single bacteriophage from the original solution that fell upon the lawn at the position of the center of the plaque and then replicated outward by consecutive infections of the bacteria. Ultimately, each plaque contains millions of the progeny of that one bacteriophage, and each one of the progeny contains only the one inserted DNA its common ancestor contained.

When the library is stored in plasmids in a population of bacteria, a suitably diluted suspension of those bacteria is spread on a plate of agar containing the antibiotic. Only bacteria containing plasmids can grow, and each of these replicates until a round **colony** of bacteria appears on the plate at the location where the original one fell. Each of the bacteria in the colony contains a plasmid because it survived the antibiotic, and each of the plasmids within the same colony contains the same segment of inserted complementary DNA or genomic DNA because all the bacteria are offspring of the original one.

Each plaque or each colony contains copies of a different complementary DNA or fragment of genomic DNA or lacks an insert. The trick is to discover which of the plaques or colonies, respectively, clearly visible to the

naked eye but numbering in the thousands to hundreds of thousands, happens to contain the complementary DNA or genomic DNA that encodes the protein of interest.

The most rapid and unambiguous method of screening is to synthesize chemically or biologically a fragment of radioactive single-stranded or double-stranded DNA, referred to as a **probe**, the sequence of one strand of which encodes the amino acid sequence of the protein of interest (Figure 3-11). When the double-helical DNA in a plaque or a colony containing that particular short nucleic acid sequence is heated so that it unwinds and becomes single-stranded DNA, the sequence on the antisense strand or the sequences on the sense and the antisense strands that are complementary to the sequence or the two sequences of the probe will become accessible for hybridization. **Hybridization** is the formation in the laboratory of double-helical DNA from two complementary single strands of DNA. Because hybridization is usually performed in a complex mixture of single-stranded DNAs such as the denatured DNA from a plaque or colony, the mixture is cooled slowly or **annealed**, to give the pairs of complementary single-stranded molecules of DNA enough time to find each other and form a double helix. If the clone contains sequences of DNA that are complementary to those of the probe, those sequences, after the DNA has been denatured and probe has been added, will hybridize with the sequences of the probe to form short segments of double-helical DNA, and in this way the probe is captured. This trapping of the probe makes the plaque or the colony containing the desired complementary DNA radioactive, marking the position of the bacteriophage or bacteria carrying the desired complementary DNA and allowing that one plaque or colony to be isolated.

An example^{66,67} will illustrate this screening procedure. Factor VIII is one of the proteins that are together responsible for the cascade of events leading to the clotting of the plasma of mammalian blood. Human Factor VIII was digested with trypsin, and the peptides that resulted from the digestion were separated⁶⁶ on chromatography by adsorption. Several of these peptides were resolved cleanly from their neighbors, and they were submitted to Edman degradation. The amino acid sequence determined for one of these peptides was AWAYFSDVDLEK. A segment of radioactive, single-stranded DNA with the nucleic acid sequence CTTTTCCAGGTCAACGTCGGAGAAATAAGCCCAAGC (Figure 3-11), one of the many possible antisense sequences to that encoding the peptide, was synthesized chemically to act as a probe. Long restriction fragments (15 kb) of human genomic DNA were inserted into bacteriophage λ Charon, and these bacteriophage were used to produce plaques on lawns of *E. coli*. The DNA in the plaques was then denatured. During subsequent annealing and hybridization, the radioactive probe was captured by the denatured, single-stranded DNA in 15

plaques out of the 500,000 screened for DNA containing a nucleic acid sequence that would capture the probe.⁶⁷ Bacteriophage from each of these 15 clones were separately grown on a large scale, and the inserted DNA was cut out of the DNA of the bacteriophage that had been carrying it with site-specific deoxyribonucleases.

The **polymerase chain reaction**⁶⁸ can be used to produce probes for screening plaques or colonies or even a segment of the DNA encoding a significant portion of the protein of interest. This is a method for replicating to a high concentration only a specific segment from any source of DNA. To replicate only a particular segment of DNA in a complex mixture or within a much longer molecule of DNA by the polymerase chain reaction, all that is required is that the segment of double-stranded DNA to be replicated is flanked on either side by known sequences of nucleotides. Two short primers of single-stranded DNA are synthesized, one complementary to the flanking sequence at one end of the segment to be replicated and the other complementary to the flanking sequence at the other end. These two primers for the two ends, however, must be complementary to the sequences on opposite strands of the initial double-stranded DNA. The initial double-stranded DNA is melted, and the two primers are hybridized. DNA-directed DNA polymerase is then used to elongate from the 3'-end of each primer (Figure 3-10). This produces two copies of duplex DNA over the segment of interest. The new DNA is melted and rehybridized with the same two primers and elongation is performed again to produce four copies of duplex DNA for the segment of interest and so forth. If the heat-stable DNA-directed DNA polymerase from *Thermus aquaticus*⁶⁹ or *Pyrococcus furiosus*⁷⁰ is used for the elongation, new polymerase does not have to be added after each melting cycle. After repeated cycles of melting, annealing, and elongation, essentially all of the newly synthesized DNA is a copy of the double-stranded segment of the original DNA between and including the sequences of the priming DNA, and the concentration of this segment increases exponentially with each step.

An example of the use of this procedure is the synthesis of a probe for screening a library containing the gene for extensin from *Volvox carteri*.⁷¹ The amino-terminal sequence of the protein, AVSYSVSVYNNIAVT-GAP-, and the sequence of a tryptic peptide from the protein, IDPPSNFGNLPVK, were used to guide the synthesis of two primers, GT(T/C/A/G)TA(T/C)AA(T/C)-AA(T/C)AT(T/C/A)GC and GG(G/T)AGGTT(T/C/A/G)-CCGAA(G/A)TT, where letters in parentheses indicate that two or more nucleotides were coupled in that step of the synthesis to allow for the redundancy of the genetic code. When complementary DNA from sperm packets of *V. carteri* was amplified with these primers in a polymerase chain reaction, a segment of 410 bp of double-stranded DNA was produced beginning with the sequence GTCTACAACAACATCGC- and ending with the

sequence –AACTTTGGCAACCTGCC on its sense strand. This sequence encoded 136 aa of the amino acid sequence of the protein. This segment of DNA was inserted into a plasmid, replicated with radioactive precursors, and used successfully as a probe to screen a library of genomic DNA from *V. carteri*. By use of this probe, a clone of bacteria was identified that carried a plasmid containing a segment of complementary DNA 1392 bp long, encoding 464 aa from the amino acid sequence of extensin.

The complementary DNA or the fragment of genomic DNA encoding the protein of interest that has been produced by replicating the bacteriophage or bacterium identified by the screen, or the segment of DNA encoding a portion of the protein that has been amplified by the polymerase chain reaction, can be quite long, from thousands to tens of thousands of nucleotides. The sequence of a particular piece of single-stranded DNA can be read only to a certain length (300–400 nucleotides). Therefore, long DNAs must be cleaved into smaller restriction fragments with site-specific deoxyribonucleases, just as polypeptides have to be cleaved into peptides before they can be sequenced. By trial and error, a pattern of restriction fragments ideally suited to the demands of sequencing can be prepared.

The shorter double-helical restriction fragments produced from a longer double-helical DNA are rapidly separated by **preparative electrophoresis on gels of agarose**. They are usually visualized by use of fluorescent dyes. Their length can be estimated from their electrophoretic mobilities. The order in which a given set of restriction fragments is arranged in the original DNA is determined by **restriction mapping**. To produce a restriction map of a large piece of DNA, it is cleaved separately with several site-specific deoxyribonucleases. The restriction fragments produced in each of these separate digestions are isolated and assigned a length by electrophoresis. Each of these restriction fragments of DNA is then submitted to digestion by the other sets of site-specific deoxyribonucleases, and the shorter restriction fragments that result are separated and assigned a length. This dissection is continued until the restriction fragments observed, which are designated by the pedigree of the cleavages that produced them, are consistent with only one distribution of restriction sites through the original piece of long DNA as well as being of the desired length. This unique distribution of restriction sites, the restriction map, orders the different restriction fragments that have been obtained relative to the complete sequence.

An example will serve to illustrate the complete process.⁷² A clone containing the complementary DNA encoding the α polypeptide of the murine nicotinic acetylcholine receptor within the tetracycline-resistant plasmid pBR322 (Figure 3–12)⁷² was identified by screening. The cloned complementary DNA was cut from the plasmid as an intact double-helical polymer with the

site-specific deoxyribonuclease *Pst*I, which cleaves at CTGCA↓G. This DNA was digested with the following site-specific deoxyribonucleases: *Alu*I, which cleaves at the nucleic acid sequence AG↓CT; *Taq*I, which cleaves at T↓CGA; *Hpa*II, which cleaves at C↓CGG; *Hae*III, which cleaves at GG↓CC; *Rsa*I, which cleaves at GT↓AC; and *Hinc*II, which cleaves at GTPy↓PuAC, where Py is either pyrimidine and Pu is either purine.

The pattern of restriction fragments obtained when these enzymes were used in various combinations was consistent with only one restriction map (Figure 3–12). For example, the *Hpa*II restriction fragment between positions 478 and 1063 would give three restriction fragments about 60, 240, and 280 base pairs in length upon digestion with site-specific deoxyribonuclease *Alu*I. The order in which these three subfragments occur in the *Hpa*II restriction fragment could be determined by gathering the following observations. Deoxyribonuclease *Taq*I would cut only the *Alu*I restriction fragment that is about 280 base pairs in length to yield the same restriction fragment, about 120 base pairs long, that it would produce from one end of the *Hpa*II restriction fragment. Deoxyribonuclease *Hinc*II would cut only the *Alu*I restriction fragment that is about 240 base pairs in length to give a restriction fragment about 140 base pairs in length. This restriction fragment, together with the *Alu*I restriction fragment about 60 base pairs in length, would form the restriction fragment about 200 base pairs in length produced during the digestion of the *Hpa*II restriction fragment with deoxyribonuclease *Hinc*II alone.

When restriction fragments of a convenient size had been produced from this complementary DNA encoding the α polypeptide of the murine acetylcholine receptor, a group of single-stranded DNAs within the set were chosen for sequencing (arrows in Figure 3–12). These single-stranded DNAs were subcloned in the single-stranded bacteriophage M13, and each was submitted to sequencing from its 5'-end.

The property of denatured, single-stranded nucleic acids that allows them to be sequenced is that they behave with extraordinary **regularity upon electrophoresis**. For example, when 4.5S ribosomal RNA from the chloroplasts of spinach,⁷³ which is 107 bases long, is elongated with RNA ligase (ATP) from T4 bacteriophage by one nucleotide at its free 3'-hydroxyl group by use of [5'-³²P]cytidine 3',5'-bisphosphate and then submitted to partial alkaline hydrolysis, a random mixture of fragments of all possible lengths and all possible beginning and ending points within the sequence is produced. Only those fragments that begin at the original 3'-end, however, are radioactive. In the case of the 4.5S rRNA, these formed a set of 108 unique fragments that were of all the possible lengths between 1 and 108 nucleotides. When this mixture was submitted to electrophoresis under denaturing conditions on a gel cast from 12% acrylamide and the radioactive components

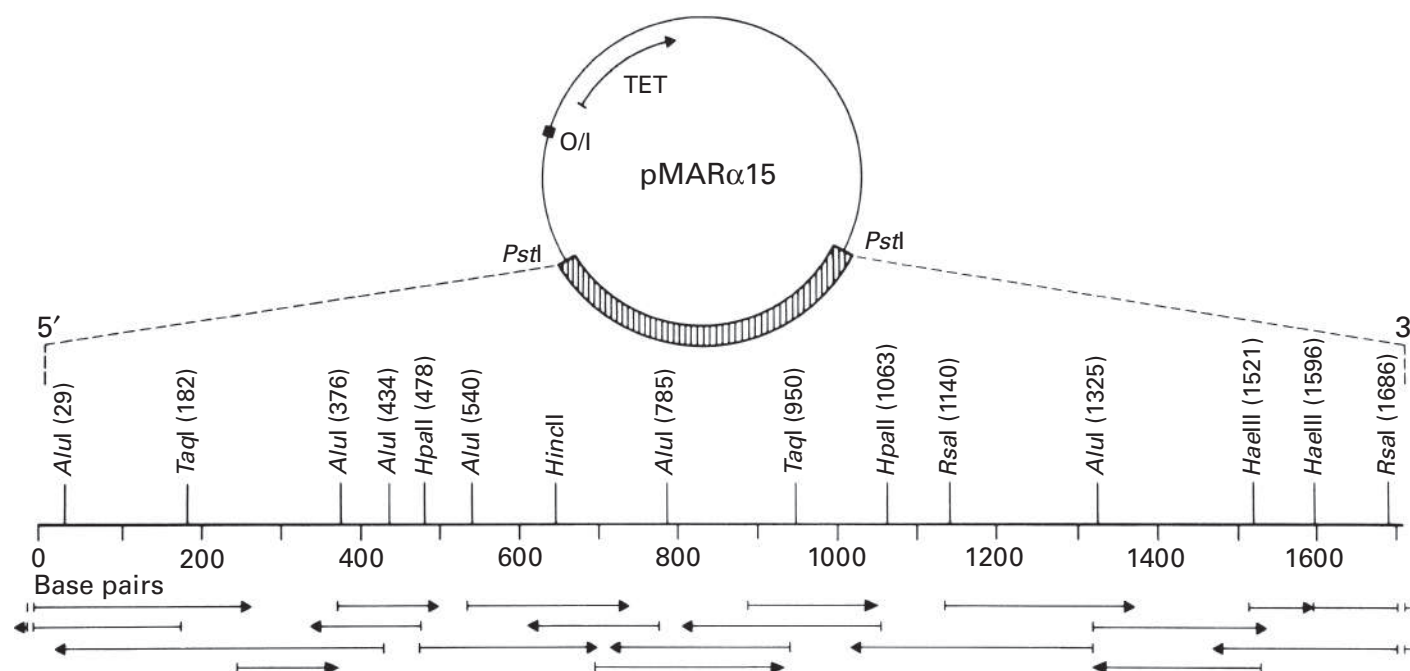


Figure 3-12: Restriction map of a fragment of DNA cut out of a plasmid.⁷² A large fragment of complementary DNA (17 kb) was removed with the site-specific deoxyribonuclease *PstI* from the circular plasmid pMARα15, which had been originally constructed from the circular plasmid pBR322. The plasmid contained a gene for resistance to the antibiotic tetracycline (TET) so that only bacteria carrying the plasmid would grow on a medium containing tetracycline. The origin of replication for the plasmid is indicated (O/I). The plasmid pMARα15 was isolated during a screening procedure for complementary DNA encoding the α-polypeptide of the murine acetylcholine receptor. The fragment of complementary DNA was purified by electrophoresis and submitted to a series of digestions with the noted site-specific deoxyribonucleases. The patterns of fragments established the restriction map displayed. The arrows below the restriction map indicate which restriction fragments were submitted to sequencing from which 5'-end. The positions in the nucleic acid sequence cleaved by each site-specific deoxyribonuclease are identified by numbers in parentheses. Reprinted with permission from ref 72. Copyright 1985 Oxford University Press.

were located by placing the polyacrylamide gel on a photographic film, a regular array of bands, referred to as a **ladder**, could be observed (Figure 3-13).⁷³ Each of these bands, with one interesting exception that will be discussed later, represents a single-stranded RNA that begins at the labeled 3'-end of the original 4.5S rRNA, because it is radioactive and is one nucleotide longer than the nucleic acid in the band below it in the figure.

The ability of electrophoresis on polyacrylamide gels to separate nucleic acids only on the basis of their length arises from the properties of these polymers and the nature of the electrophoresis. The free electrophoretic mobility of denatured single-stranded DNA at $I_c = 0.01$ M, pH 7.5, and 0°C is $(1.82 \pm 0.02) \times 10^{-4} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ and does not vary⁷⁴ with its length. The free electrophoretic mobility of denatured RNA under the same conditions is the same, $(1.77 \pm 0.05) \times 10^{-4} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$, and it also shows no tendency to vary with length.⁷⁴ The electrophoretic mobilities of single-stranded DNA and RNA on polyacrylamide gels also conform to Equation 1-81,⁷⁵ and the free mobilities extrapolated from their behavior on polyacrylamide gels are in reasonable agreement with those measured directly.⁷⁶ Because their free electrophoretic mobilities are all the same, it is only the resistance posed by the polyacrylamide, $\exp(-K_{r,i}T_a)$, that separates the nucleic acids of the various lengths. It is not

surprising that this sieving, accomplished at the molecular level by the strands of polyacrylamide, should be a regular, continuous, monotonic function of the lengths of the nucleic acids (Figure 3-13).

Suppose that a single-stranded deoxyribonucleic acid, labeled at its 5'-end by phosphorylation with [^{32}P]phosphate, has been cleaved in a low yield and randomly on the 5'-side of each of the deoxyguanosines in its sequence. This partial cleavage will have produced a series of radioactive fragments of different length, each of which ends at a nucleotide whose only distinction is that it preceded a deoxyguanosine in the original sequence. When the products of this **partial cleavage** are submitted to electrophoresis, a series of radioactive bands will appear the mobilities of which correspond to only those rungs in the ladder the 3'-terminal nucleotide of which precedes a deoxyguanosine. The knowledge that the cleavage occurred only at deoxyguanosines and the position of the products in the ladder identifies the relative positions of every deoxyguanosine in the original sequence.

Suppose further that four samples have been prepared from the original single-stranded deoxyribonucleic acid such that they contain radioactive fragments, all of which begin at the original 5'-end because they were made radioactive by phosphorylating only that

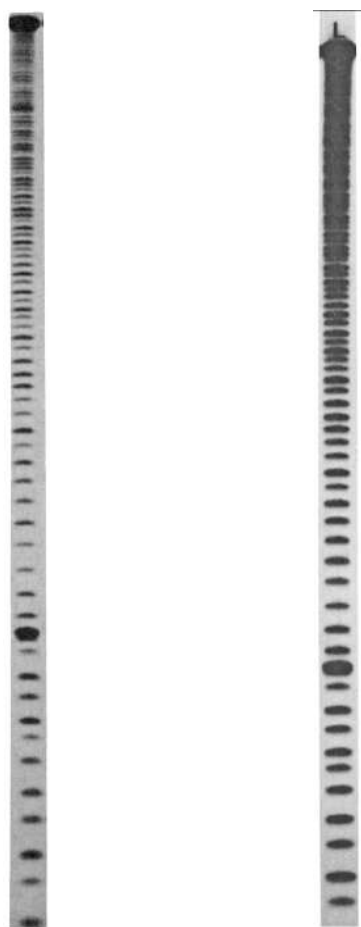


Figure 3-13: Separation of fragments of end-labeled RNA by electrophoresis on gels of polyacrylamide.⁷³ 4.5S Ribonucleic acid, isolated from spinach chloroplasts, was labeled at its 3'-end with [5'-³²P]cytidine 3',5'-bisphosphate by T4 RNA ligase (ATP). The end-labeled RNA was partially digested under alkaline conditions and then submitted to electrophoresis on slabs of 12% polyacrylamide cast in a buffered solution of 7 M urea. The two lanes in the figure were loaded with different amounts of sample and were run for different lengths of time. Reprinted with permission from ref 73. Copyright 1982 *Journal of Biological Chemistry*.

location, but which end at every nucleotide preceding a deoxyadenosine in one sample, preceding a deoxyguanosine in another sample, preceding a deoxycytidine in a third sample, or preceding a thymidine in the fourth sample. When these samples are submitted to electrophoresis, side by side, every band in the ladder will be represented in the four lanes, but each band in the ladder will be found in only one of the lanes. As one scanned the pattern, from the bands of greatest mobility to the bands of least mobility, one would encounter each band of the ladder in its proper succession. The lane in which each successive band was found would have been determined by the identity of the nucleotide that follows its actual 3'-terminal nucleotide in the complete sequence of the original DNA. By starting with the band of greatest mobility and noting its lane and the lane in which each successive band of lower mobility occurs,

one would be reading the sequence of the DNA in the direction from 5' to 3'.

The strategy for sequencing DNA illustrated by this simplified situation requires that a set of **end-labeled fragments** of single-stranded DNA be produced. Each of these fragments must have as its 5'-terminus the same nucleotide in the nucleic acid sequence to be determined, but this does not have to be the actual 5'-terminus of the original piece of DNA. For example, this result could be achieved by cleaving all of the molecules of the original DNA at the same nucleotide with a site-specific deoxyribonuclease. Every position in the portion of the complete nucleic acid sequence to be determined from the four lanes of a particular polyacrylamide gel must be represented by a labeled fragment that ends at this position and that has been produced in sufficient yield to be visualized. The observer must have enough information about each fragment visualized to associate its 3'-terminus with a particular nucleoside, deoxyadenosine, deoxyguanosine, deoxycytidine, or thymidine. In practice, this information is either that the 3'-terminus of a particular fragment precedes a particular nucleotide in the complete nucleic acid sequence or that its 3'-terminus is a particular nucleotide. There are two methods, chemical and enzymatic, for producing such a set of fragments. Neither corresponds exactly to the simplified illustration just described, but both satisfy the requirements of the strategy.

In the **chemical method** of Maxam and Gilbert,⁷⁷ reagents that take advantage of the hybrid nature of the nucleotide bases, which are partly aromatic heterocycles and partly acyl derivatives, are used to cleave chemically the single-stranded DNA, labeled at its 5'-terminus, at locations occupied by a particular base. The chemical cleavages used are based on reactions previously developed to remove selectively either purine bases or pyrimidine bases from DNA. Such reactions are depurinations⁷⁸ or depyrimidinations,⁷⁹ respectively. Reagents are used to depurinate the single-stranded DNA preferentially at deoxyguanosine^{80,81} or deoxyadenosine^{77,78} or to depyrimidinate single-stranded DNA selectively at both deoxycytidine and thymine^{79,82,83} or preferentially at deoxycytidine.⁷⁷ A position in DNA that has lost its nucleoside base by depurination or depyrimidination is susceptible to hydrolysis in base^{79,84} while normal DNA is not. In preparation for sequencing, the DNA is partially depurinated or depyrimidinated, respectively, at locations the identity of which has been controlled by the conditions of these reactions and is then hydrolyzed at each of these locations by treatment with base to produce fragments that have as their 3'-terminus a nucleotide that preceded, in the original nucleic acid sequence, a target for the depurination or depyrimidination.

In the **enzymatic method** of Sanger, Nicklen, and Coulson,⁸⁵ the properly terminated fragments required for the electrophoresis are made by synthesizing com-

plementary strands of DNA by use of the single-stranded DNA to be sequenced as a template in four separate elongations catalyzed by DNA-directed DNA polymerase (Figure 3-10). The nucleotides inserted by the polymerase are present in solution as their activated 5'-triphosphates. In the original method, the newly synthesized polymer of DNA is made radioactive by including [α - 32 P]dATP in the synthetic mixture. The successive fragments that have at their 3'-end only a particular nucleotide are produced by including a small amount of 3'-deoxythymidine triphosphate, 2',3'-dideoxycytidine triphosphate, 2',3'-dideoxyguanosine triphosphate, or 2',3'-dideoxyadenosine triphosphate, each in one of the four elongations, along with the thymidine triphosphate, 2'-deoxycytidine triphosphate, 2'-deoxyguanosine triphosphate, and 2'-deoxyadenosine triphosphate present in all of them. Occasionally, a **2',3'-dideoxynucleotide** is incorporated into one of the growing polymers by the DNA-directed DNA polymerase, and its incorporation terminates polymerization because that polymer then lacks the 3'-hydroxyl group necessary for further elongation. In this way fragments satisfying two of the requirements for electrophoretic sequencing are produced.

The last requirement, that every fragment have as its 5'-terminus the same position in the complete sequence, is satisfied by taking advantage of the requirement of DNA-directed DNA polymerase for a primer to provide a 3'-hydroxyl group from which the new strand can be elongated. To initiate the reaction, a primer that is complementary to a segment of the DNA to be sequenced is annealed to the template to provide the necessary 3'-hydroxyl group. Because the DNA-directed DNA polymerase starts at the primer when it synthesizes a complementary, radioactive single strand of DNA, the sequence of the primer can be chosen so that the newly synthesized DNA will begin at a particular point in the sequence of the template. The complementary sequence to which the primer is annealed can be a short piece of DNA of known sequence that has been deliberately attached to the 3'-end of the DNA to be sequenced,⁸⁶ or it can be any internal sequence for which a complementary fragment of single-stranded DNA happens to be available.⁸⁵ Often this complementary fragment is a probe that had been made for purposes of screening. It is also possible to use an oligonucleotide that has the same sequence as a segment near the 3'-end of the longest single-stranded fragment that provided readable nucleotide sequence in the last set of polyacrylamide gels, to extend the sequencing of the template further to its 5'-end. In this way, one can walk along a long template and read its entire sequence.

The polyacrylamide gels that result from the application of these two methods, the chemical and the enzymatic, are similar in appearance (Figure 3-14A, B).^{77,85} Sequence is read from the bottom (shortest fragments) to the top (longest fragments), 5' to 3'. In the chemical

method the sequence of the original single-stranded DNA is being read. In the enzymatic method the sequence of the complement of the original single-stranded DNA is being read. Since DNA is normally double-helical with two antiparallel strands of complementary sequence, either sequence is formally the sequence of the DNA, as long as the correct direction (5' \rightarrow 3') is assigned to the sequence by the observer.

These original methods, the chemical and the enzymatic, were both based on the use of fragments of nucleic acid made radioactive by incorporating [32 P]phosphate (Figure 3-14A, B), but in the **automated DNA sequencers** currently in use, **end-labeled fluorescent fragments** of nucleic acid are used. Although chemical methods have been developed⁸⁷ that may eventually be more efficient, the current automated procedures for sequencing DNA are based on the original enzymatic method of Sanger, Nicklen, and Coulson.⁸⁵ The products of the terminations by the dideoxynucleotides are all separated together on the same gel of polyacrylamide, which is continuously scanned by a fluorometer.⁸⁸ The products from the four respective termination reactions are end-labeled with four different fluorescent dyes that can be distinguished by the fluorometer on the basis of the colors of their fluorescence. The separate fluorescent tags are applied one of two ways.

Synthetic derivatives of 2',3'-dideoxyadenosine triphosphate, 2',3'-dideoxyguanosine triphosphate, 3'-deoxythymidine triphosphate, and 2',3'-dideoxycytidine triphosphate have been prepared that each have a different fluorescent dye covalently attached to their heterocyclic bases.^{89,90} When these derivatives are used to terminate the single-stranded fragments and thereby label them at their 3'-ends, the fluorometer can distinguish strands of DNA terminated at deoxyadenosines, deoxyguanosines, thymidines, or deoxycytidines from each other by their differences in fluorescence.

Alternatively, four distinguishable fluorescent dyes can be attached separately to the 5'-ends of four identical samples of the primer that will be used,⁸⁸ and a different one of the resulting fluorescent primers can be used in each of the four termination reactions. When the separate dideoxy terminations have been completed, the products of the four reactions are mixed. When the mixture is separated by electrophoresis, the fluorometer distinguishes each strand by the color of the fluorescence emitted by the fluorescent dye on its 5'-end, which identifies the termination mixture in which it arose.

The DNA-directed DNA polymerase used in these automated sequencers is an improved version. The original enzyme used, the Klenow fragment of DNA-directed DNA polymerase from *E. coli*, terminates the elongation at each position with a different yield that can vary significantly (Figure 3-14B). This variability can result in uncertainty in reading the sequence, especially when it is to be read by a machine. DNA-directed DNA polymerase from bacteriophage T7 produces a much more uniform

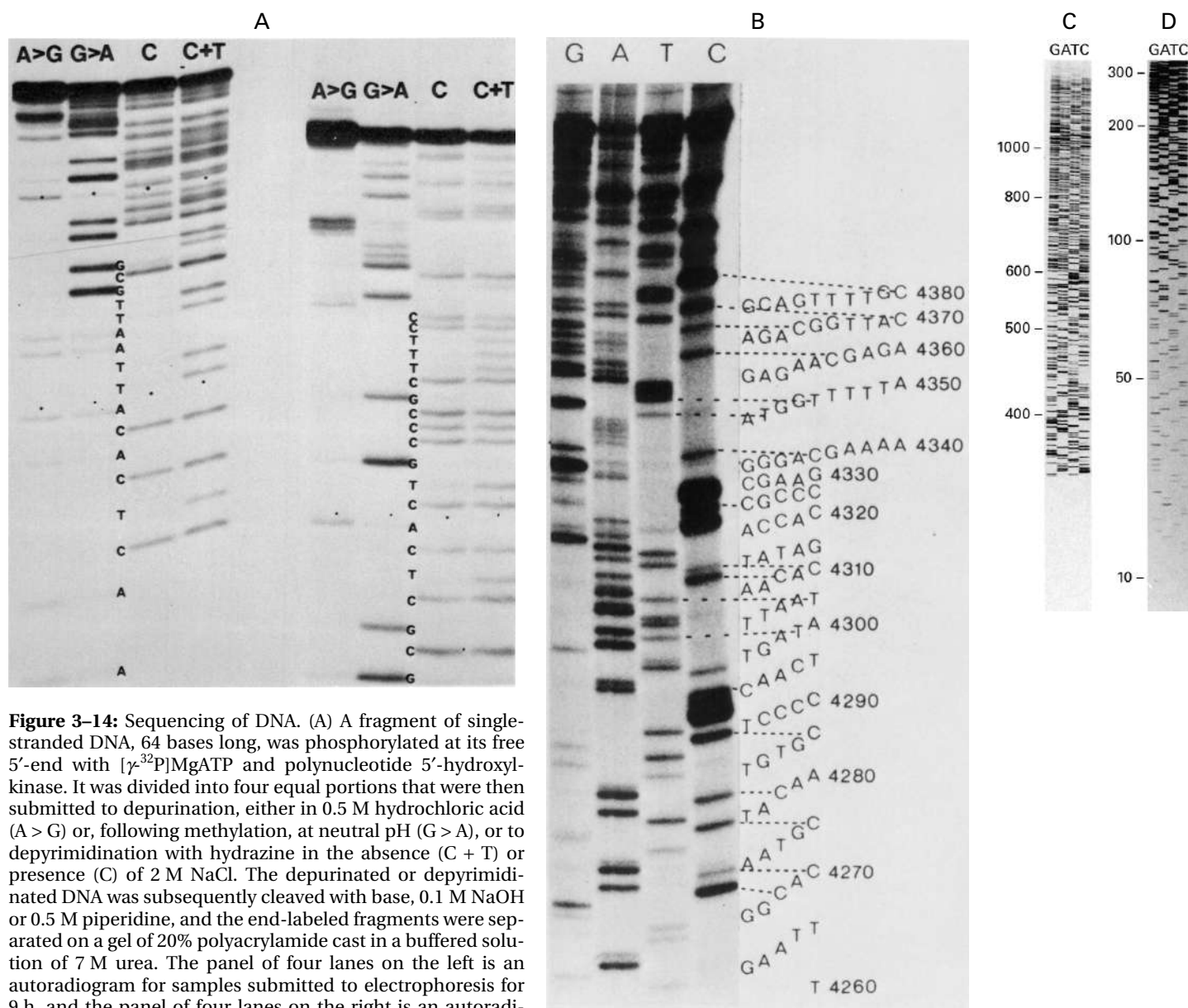


Figure 3-14: Sequencing of DNA. (A) A fragment of single-stranded DNA, 64 bases long, was phosphorylated at its free 5'-end with $[\gamma\text{-}^{32}\text{P}]\text{MgATP}$ and polynucleotide 5'-hydroxyl-kinase. It was divided into four equal portions that were then submitted to depurination, either in 0.5 M hydrochloric acid ($\text{A} > \text{G}$) or, following methylation, at neutral pH ($\text{G} > \text{A}$), or to depyrimidination with hydrazine in the absence ($\text{C} + \text{T}$) or presence (C) of 2 M NaCl. The depurinated or depyrimidinated DNA was subsequently cleaved with base, 0.1 M NaOH or 0.5 M piperidine, and the end-labeled fragments were separated on a gel of 20% polyacrylamide cast in a buffered solution of 7 M urea. The panel of four lanes on the left is an autoradiogram for samples submitted to electrophoresis for 9 h, and the panel of four lanes on the right is an autoradiogram for identical samples submitted to electrophoresis for 33 h. The sequence noted on the autoradiogram on the left, $-\text{TTTCCCGACTGG}-$, continues on the autoradiogram on the right.⁷⁷ Reprinted with permission from ref 77. Copyright 1977 National Academy of Sciences. (B) A fragment of DNA complementary to a short segment of the nucleic acid sequence of the single-stranded DNA from the bacteriophage ϕX174 was annealed to the template as a primer. The initiation complex was divided into four equal portions. Elongation of the primer in these separate samples was performed with DNA-directed DNA polymerase in the presence of 2',3'-dideoxyguanosine triphosphate (lane G), 2',3'-dideoxyadenosine triphosphate (lane A), 3'-deoxythymidine triphosphate (lane T), and 2',3'-dideoxycytosine triphosphate (lane C). Each sample contained a small amount of $[\alpha\text{-}^{32}\text{P}]\text{MgATP}$ to render the newly synthesized DNA radioactive. Following these respective reactions, each sample was digested with the *Hae*III site-specific deoxyribonuclease, which cut the DNA within the primer, so that all of the newly synthesized DNA would start with the same nucleotide at the 5'-end. The single-stranded, radioactive fragments of DNA in each sample were separated by electrophoresis on a slab of 12% polyacrylamide.⁸⁵ Reprinted with permission from ref 85. Copyright 1977 National Academy of Sciences. (C, D) A segment of single-stranded DNA 2707 bases long from the bacteriophage T7 was cloned in bacteriophage M13. The single-stranded genome of the M13 bacteriophage carrying the insert was used as template, and a short segment of synthetic DNA complementary to a region adjacent to the insert was annealed to the template as a primer. This initiation complex was labeled radioactively by mixing it with low concentrations (0.3 μM) of dGTP, TTP, dCTP, and $(\alpha\text{-}^{35}\text{S})\text{thio}(\text{dATP})$, as well as DNA-directed DNA polymerase from bacteriophage T7. After 5 min at room temperature, this mixture was divided into four equal portions to which were added high concentrations (150 μM) of dATP, dGTP, dCTP, and TTP. To each portion was added one of the dideoxynucleotides at 15 μM : 2',3'-dideoxyguanosine triphosphate (lanes G), 2',3'-dideoxyadenosine triphosphate (lanes A), 3'-deoxythymidine triphosphate (lanes T), or 2',3'-dideoxycytidine triphosphate (lanes C). The reaction was initiated by adding a high concentration of DNA-directed DNA polymerase from T7 bacteriophage. After 5 min at 37°C, the four samples were prepared for electrophoresis, the single-stranded radioactive fragments of DNA in each sample were separated on slabs of 7% polyacrylamide run for 12 h (C) or 2 h (D), and the gels were submitted to autoradiography. The numbers to each side indicate the lengths of the single-stranded fragments.⁹¹ Reprinted with permission from ref 91. Copyright 1987 held by authors.

yield of fragments terminated by 2',3'-dideoxynucleotides (Figure 3-14C, D).⁹¹ Certain mutants of the enzyme from *E. coli* are even more reliable than their parent.⁹²

The electrophoretic separations presented in Figure 3-13 illustrate an important artifact common to all methods of sequencing DNA. The thick band at the eleventh rung of the ladder on the electrophoretogram to the right is a **compression**. Within that one band, single-stranded RNAs from 24 to 27 nucleotides in length comigrate. The band above the compression is the single-stranded RNA 28 nucleotides long; and that below, 23 nucleotides long.⁷³ A compression usually occurs when the 3'-end of the fragment of single-stranded nucleic acid is rich in G and C, and it is caused by structures formed intramolecularly among these bases by the usual GC pairing.⁹¹ When a compression is not recognized as such by a person or by a machine, it is mistaken for a normal band representing a single polynucleotide and the sequence read will be missing one or more nucleotides. It is possible to eliminate such compressions by using 2'-deoxyinosine triphosphate instead of 2'-deoxyguanosine triphosphate in the elongation mixtures of the enzymatic method.⁹¹

Once the sequence of a segment of prokaryotic genomic DNA or eukaryotic complementary DNA is in hand, the complete sequence of the protein can be read from the open reading frame that encodes it. The sequence of nucleotides in the messenger RNA that encodes a polypeptide begins with the **initiation codon** -AUG-, which can be recognized in the sequence by its proximity to sequences encoding a binding site for a ribosome, and ends with a **termination codon**, -UAA-, -UGA-, or -UAG-. An **open reading frame** (ORF) is any sequence of nucleotides that begins with an initiation codon and ends with a termination codon. Because a sequence of DNA does not indicate the reading frame used to synthesize the protein nor which of the two complementary strands is the sense strand (Figure 3-11), there are six possible sequences of triplets that could encode the protein. These six different reading frames in any sequence of nucleotides obtained experimentally each contain open reading frames, and the open reading frame encoding an actual protein can be recognized only if some information about the protein is known, for example, its length or sequences from some of its peptides.

In the case of the α polypeptide of murine nicotinic acetylcholine receptor, each sequence of an individual single strand of DNA from the restriction mapping (Figure 3-12) began at the 5'-end of one of the two complementary strands of a double-helical fragment and was read as far as was possible. With the exception of two short segments, each region of the sequence was read at least twice. Together, all of these individual sequences produced the complete sequence of the complementary DNA (Figure 3-15).⁷² Of the six reading frames in the

completely sequenced double-helical complementary DNA, the one containing the open reading frame encoding the sequence of the α polypeptide of murine nicotinic acetylcholine receptor was easily identified by locating the one that encoded the amino acid sequences on which the probes used to screen the clones were based.

In the case of human Factor VIII, when the sequences of the different segments of DNA identified by the screen were translated into amino acid sequences, each segment was found to contain an overlapping region of the same open reading frame that encoded the sequence AWAYFSDVDLEK. The amino acid sequences of four of the other peptides of factor VIII that had been submitted to Edman degradation could also be found in the translation of this one complete open reading frame found in the nucleic acid sequences of the overlapping clones. Comparisons like these between directly determined amino acid sequences of a particular protein or its experimentally determined composition of amino acids^{93,94} and the amino acid sequence translated from an open reading frame in the complementary DNA or genomic DNA are often used to substantiate the identification of the complementary DNA or genomic DNA as that encoding a particular protein.

The sequences of peptides from the protein permit the proper open reading frame to be assigned, but if bases have been omitted by mistake during the sequencing of the nucleic acid, these omissions can produce a frameshift. A **frameshift** is the inadvertent shift into another reading frame as the sequence of nucleotides is being divided into triplets. It is caused by the omission of $3n + 1$ or $3n + 2$ bases in the sequence. One common mistake leading to a frameshift is the omission of several bases in the sequence that results when a compression goes unrecognized.⁹⁵ It is not unusual for an initial DNA sequence to contain errors, often quite a few,⁹⁶ but they are usually recognized and corrected by sequencing both the complementary strands or during further examinations of the protein it encodes.^{97,98}

When amino acid sequences of proteins were determined directly, the proteins chosen for sequencing were usually enzymes, and the sequences obtained were unremarkable. When the explosion of amino acid sequences from sequencing DNA commenced, so many more proteins were being sequenced that peculiar ones were discovered. The most obvious peculiarities were **enrichments** in a particular amino acid, such as a cell wall protein (465 aa) containing 60% glycine,⁹⁹ a 127 aa segment of chicken vitellogenin containing 75% serine,¹⁰⁰ a 655 aa segment of spider dragline silk containing 47% glycine and 28% alanine,¹⁰¹ a 246 aa segment of murine MP-2 containing 46% proline and 17% glutamine,¹⁰² a 71 aa segment of histidine-proline-rich glycoprotein containing 50% histidine,¹⁰³ a 30 aa segment of the Abdominal-B domain of the bithorax complex of *Drosophila melanogaster* containing 22 glut-

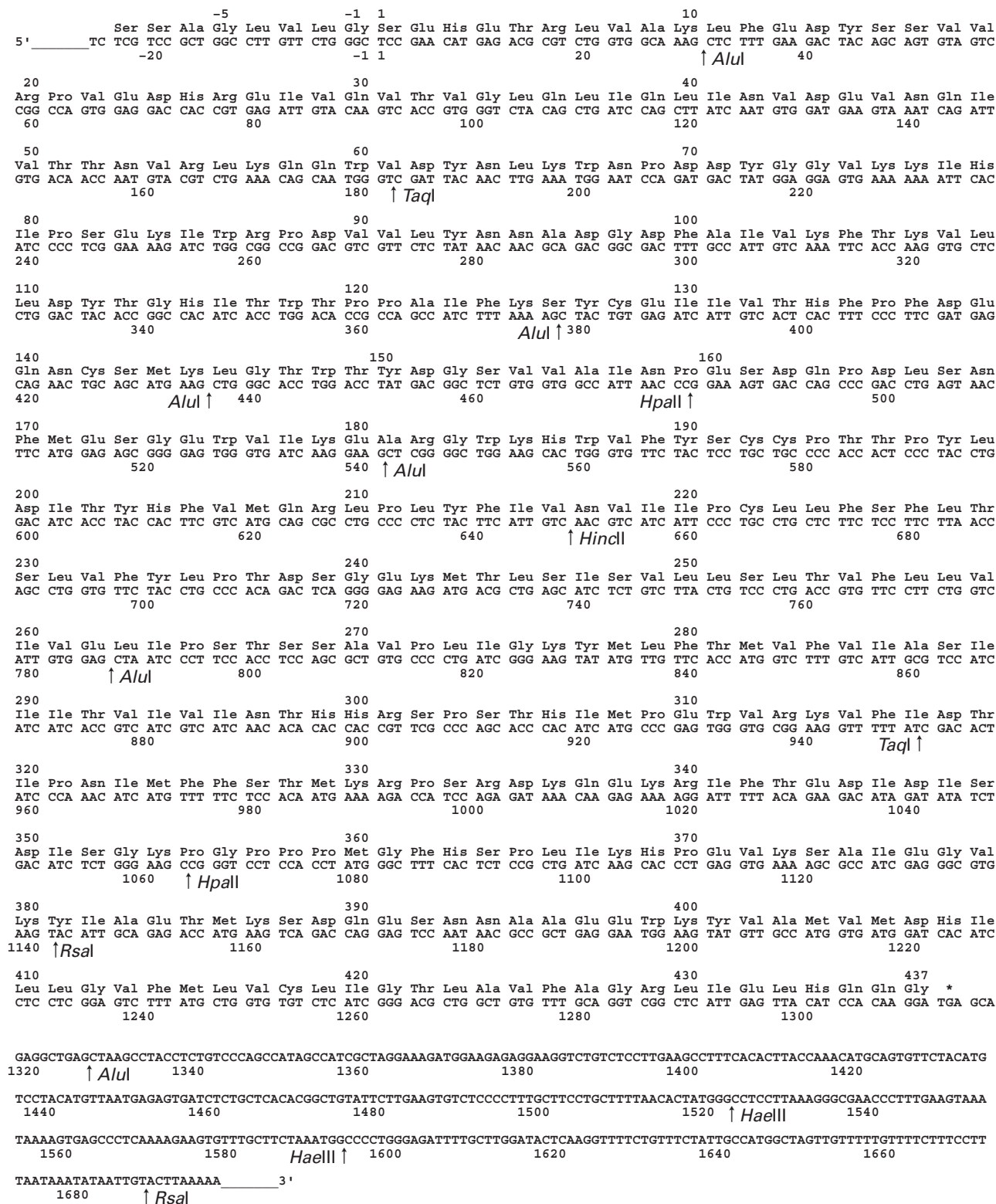


Figure 3–15: Nucleic acid sequence and deduced amino acid sequence for the α polypeptide of murine nicotinic acetylcholine receptor.⁷² The nucleotides are presented in the 5' to 3' direction for the coding strand. Both sequences are numbered starting with the first amino acid in the mature protein. The first eight amino acids in the presented sequence are removed posttranslationally. The initiation codon for translation was not on the cloned piece of complementary DNA. The asterisk marks the codon at which translation is terminated. The restriction sites that produced the restriction map (Figure 3–12) are identified in the nucleic acid sequence. The *Pst*I restriction sites situated at the two ends of the insert that were used to remove complementary DNA from the plasmid (Figure 3–12) are lost during the insertion of the restriction fragments of the complementary DNA into the M13 bacteriophage, but the sequence shown begins just after the initial *Pst*I site and ends just before the final *Pst*I site. Reprinted with permission from ref 72. Copyright 1985 Oxford University Press.

amines,¹⁰⁴ or a stretch of up to 37 successive glutamines in normal human HD protein.¹⁰⁵⁻¹⁰⁸ A protein induced by abscisic acid in maize has a segment 66 aa long containing only arginine, tyrosine, and glycine in which the sequence GYG₂G repeats 7 times,¹⁰⁹ and the repeating unit of rat filaggrin (406 aa) containing 15% glutamine but no asparagine, 14% arginine but no lysine, 2% isoleucine but no leucine, no cysteine or methionine, and no tryptophan or phenylalanine.^{110,111} In many of these enriched proteins, strings of 5-20 aa in length containing only one amino acid are common. Most of these proteins form flexible polymeric solids with physical properties appropriate to their function, and their peculiar sequences provide an intuition of the way in which they accomplish their roles.

The large majority of proteins, however, have sequences that in isolation say nothing about their structure or function, but the large number of sequences that have become available as a result of sequencing DNA permit proteins to be grouped into **families**. By the structure and function of its relatives, the structure and function of an unknown protein can often be revealed. The most obvious instance of such grouping of amino acid sequences is when a protein isolated from one organism is recognized to be the same protein, albeit with a somewhat different sequence, as one that is found in other species.¹¹² The most dramatic examples of such comparisons of amino acid sequences are those in which a protein responsible for one function is isolated and found to be identical to a protein isolated from the same species on the basis of its responsibility for another function.^{113,114}

The amino acid sequences of all the proteins in many prokaryotic species are now available. **Complete genomic sequences** have been determined for a number of those that are widely used in experimentation. These genomic sequences permit the amino acid sequence of a prokaryotic protein to be obtained without the need for sequencing its DNA. If enough amino acid sequence has been obtained from a protein to construct a probe to screen a library, one has enough amino acid sequence to search by computer the complete genome for the species from which the protein was isolated, identify the gene encoding that protein, and thereby obtain its complete amino acid sequence. Unfortunately, the existence of introns in the genes encoding eukaryotic proteins usually makes it impossible to obtain an accurate amino acid sequence of an unknown protein from the genomic sequence of the eukaryote from which it is derived. If, however, a set of already known amino acid sequences of proteins closely related to the protein of interest can be assembled, these sequences can be often be used to define the boundaries of the introns in the gene encoding the protein of unknown sequence.¹¹⁵ If the boundaries of the introns can be defined, the sequence of the protein can be read from the genomic DNA.

Once complementary DNA, if the protein is eukaryotic, or genomic DNA, if the protein is prokaryotic, encoding the complete, uninterrupted amino acid sequence of a protein has been cloned, it is possible to use that DNA to direct the production of the protein. This strategy provides a convenient and abundant source of the protein. It is also possible to cut out any piece of that DNA with the proper site-specific deoxyribonucleases and direct the production of the fragment of the protein encoded by the resulting fragment of the DNA. This strategy provides precisely designed pieces of the protein.

An **expression system** is any process by which foreign DNA encoding a protein of interest or a portion of that protein has been incorporated into a population of living cells and those cells have been induced to transcribe that foreign DNA into messenger RNA and translate that messenger RNA into usable quantities of the protein for which it encodes. The cells expressing the protein are usually not of the same species or even kingdom of the species from which the protein was first purified. *Escherichia coli*, a bacterium, is the organism most widely used to express proteins, often those from animals. For example, complementary DNA encoding the 5-aminolevulinate synthase from *Mus musculus* has been expressed in cells of *E. coli*, and when these bacterial cells were harvested, 50% of the protein in them was murine 5-aminolevulinate synthase. Each liter of culture medium yielded bacterial cells from which 5 mg of the pure enzyme could be isolated.¹¹⁶ Because the expression system is usually unrelated to the organism from which the DNA to be expressed was originally derived, that expression system is usually unable to splice introns out of the messenger RNA encoded by genomic DNA. Consequently, if a eukaryotic protein or a portion of a eukaryotic protein is to be expressed, its complementary DNA is used.

An **expression vector** is a molecule of DNA into which the DNA encoding the protein to be expressed is inserted and which compels the cells of the expression system to express the protein. When proteins are expressed in *E. coli*, the expression vectors are usually plasmids. There are many plasmids used as expression vectors, but each of them usually contains a gene conveying resistance to an antibiotic so that only cells carrying the plasmid will grow. The insertion is performed at a restriction site that occurs only once in the sequence of the DNA for the plasmid. The insertion is performed by cleaving that site with the appropriate site-specific deoxyribonuclease, adding the fragment of DNA encoding the protein, and ligating the pieces of DNA. The expression vector has been designed so that this restriction site for insertion is immediately adjacent to sequences of DNA that enforce the transcription and translation of the inserted DNA. To guarantee high levels of transcription, there is a **strong promoter**, for example, a T3 promoter, a T7 promoter, a *lacZ* promoter, an alka-

line phosphatase promoter, a *tacII* promoter, or a *trc* promoter. These promoters are segments of DNA that serve as unusually active sites for the initiation of the synthesis of messenger RNA by DNA-directed RNA polymerase. The DNA preceding the point of insertion on the plasmid must also have sequences necessary for the active translation of the messenger RNA into protein.

The DNA inserted into the **restriction site on the expression vector** is often complementary DNA or genomic DNA that has just been used for sequencing, and that DNA is cut out of the bacteriophage or plasmid in which it was screened and amplified. Occasionally, the inserted complementary DNA is from an organism the codon usage of which is so different from that of *E. coli* that poor expression occurs because of this mismatch. One solution to this problem is to synthesize the complementary DNA with compatible codons.¹¹⁷ The insertion into the expression vector is accomplished most effectively if the fragment has sticky ends that are compatible with the restriction site on the expression vector. One way this is accomplished is to use primers for the polymerase chain reaction that contain the sequences of DNA necessary to anneal to complementary sequences of the DNA at the beginning and end of the coding sequence but in addition contain sequences of DNA for the appropriate endonucleolytic cleavage sites^{64,118,119} and even sequences necessary for translation.¹²⁰ The final DNA produced in the polymerase chain reaction will incorporate these additional sequences even though they did not exist in the initial DNA used as the template. If the complementary DNA encodes a segment of amino acid sequence that is normally removed from the native protein by a posttranslational process absent from *E. coli*, the portion of the DNA encoding that segment sometimes has to be removed before a fully functional protein can be expressed.¹²¹

A piece of DNA encoding another amino acid sequence is often inserted ahead of the DNA encoding the protein of interest. For example, a portion of DNA encoding a strong promoter as well as a short segment of the protein that promoter usually controls, such as a segment of β -galactosidase or the λ cII protein, can be placed in front of the DNA to be expressed to guarantee that it is produced efficiently. In this instance, a stop codon followed by a start codon can be inserted between the two coding regions so that the fragment of DNA promoting transcription is not translated attached to the protein being expressed. It has been found in many instances, however, that **fusion proteins**, proteins in which the protein of interest is coupled during translation to another complete protein such as glutathione transferase, β -galactosidase, or ubiquitin, are expressed in much higher yield than the unfused, intact protein of interest. Often this is due to the fact that the fusion protein resists the endopeptidases of the *E. coli*^{122,123} that would otherwise degrade the protein of interest. To iso-

late the protein of interest without the associated fusion protein, an amino acid sequence is often introduced between the two proteins that is a target for an endopeptidase of stringent specificity, such as activated factor Xa or renin, so that the unwanted portion can be removed by cleavage with that endopeptidase.

A fusion protein can also be one between the protein of interest and a portion of a protein such as an enterotoxin that contains a signal for secretion from *E. coli*. In this case, the protein produced ends up in the medium rather than in the cells. In one instance, however, expression of a protein that is normally excreted from *E. coli* was toxic to the cells at the levels produced, and the sequences signalling excretion had to be removed to keep the protein inside the cells.¹²⁴

One problem with expression of a foreign protein in *E. coli* is its precipitation to form large **inclusion bodies**. In this precipitated form, the protein being expressed is inactive and indistinguishable from any other precipitated protein. It is often possible, however, to dissolve these precipitates in a solution of a salting-in solute such as urea or guanidinium chloride and renature functionally active, fully soluble protein from this solution.

Proteins can be expressed in cells other than those of *E. coli*. Expression plasmids containing promoters active in *Saccharomyces cerevisiae*¹²⁵ that can be incorporated into cells of this species of **yeast** are available for expressing proteins.¹²⁶ One of the difficulties of expressing animal proteins in bacteria or fungi is that these cells are unable to perform normal **posttranslational modifications**. An animal protein that is normally modified posttranslationally is usually expressed in animal cells capable of such modifications. One such animal system that provides high yields of protein is cells of the insect *Spodoptera frugiperda*. These **insect cells**, grown in culture, can be infected with virions containing an expression vector constructed from viral DNA of the nuclear polyhedrosis virus *Autographa californica*¹²⁷ just as a culture of *E. coli* can be infected with bacteriophage λ . If the DNA encoding the protein of interest is inserted at a point in the viral genome under the control of the promoter for the viral coat protein, high yields of the expressed protein are produced. Even higher yields can be produced if larvae (caterpillars) of *Trichoplusia ni* are infected with such a virus.¹²⁸ These insect expression systems produce proteins with many of the normal posttranslational modifications of animals.¹²⁹

To ensure that posttranslational modifications of mammalian proteins that are foreign to insects are correctly made or to express a mammalian protein in the biological context of a mammalian cell, proteins are often expressed in cultured **mammalian cells** by use of an expression vector carrying a promoter from an animal virus, such as cytomegalovirus or simian virus. Such expression vectors can be inserted into the genomic DNA

of an animal cell such as Chinese hamster ovary cells or murine L cells by transfection.

When a protein is expressed in any of these expression systems, the final product of the expression is usually a pellet of cells that is then homogenized, producing a complex mixture of proteins. Even if the expression has been so successful that the protein that has been expressed accounts for the majority of the protein in this mixture, it still must be purified. This **purification** is usually performed by the standard procedures because they are simple to implement, but it is possible to design the expressed protein to ease its purification. For example, the protein can be expressed with a string of six histidines attached at its carboxy terminus or amino terminus. An affinity adsorbent to which Ni^{2+} has been attached through a covalently bound iminodiacetic acid¹³⁰ binds such **histidine tails** with high specificity, and the expressed protein can be eluted, often in pure form, with a gradient of imidazole.¹³¹ It is also possible to purify expressed proteins specifically if they have been designed to contain a short amino acid sequence on one of their termini recognized by a specific **immunoglobulin** immobilized on a solid phase. Fusion proteins between the protein of interest and **glutathione transferase** can be purified by using an affinity adsorbent on which glutathione has been covalently attached and eluting with glutathione. All of these strategies require that a short sequence of amino acids or even another protein be fused with the protein of interest, but if a short sequence recognized by a stringent endopeptidase is incorporated between the two, the protein of interest can be released in its unmodified form by digestion.

One advantage of expressing a protein in a system in which it is produced as a major fraction of the cellular protein or it has been tagged for affinity adsorption is that its purification often requires fewer steps than purification from its natural source. Because the steps of a purification are often accompanied by slow degradation of the protein, the fewer the steps, the more homogeneous will be the final purified protein. Crystals are more readily obtained from a protein the purification of which has been simple and rapid. For this reason, if they are available in high yield, expressed proteins are usually used in **crystallographic studies** in preference to the same proteins purified from natural sources. Often, however, expressing a protein in cells, even in *E. coli*, provides far less of the purified protein than can be obtained by starting with 10 kg of liver, heart, blood, or skeletal muscle. In such instances, if all that is desired is the pure protein, using an expression system is inefficient and costly. If, however, one experimental goal is to mutate specific amino acids in the sequence of the protein, an expression system is unavoidable.

Site-directed mutation^{132,133} converts one particular amino acid in the sequence of a polypeptide into another of the 20 amino acids. It is also possible to delete

amino acids from the sequence of a polypeptide or insert extra amino acids at a particular location with this technique. The method requires that the complementary DNA or genomic DNA for the protein of interest has been cloned and that the encoded protein can be expressed, in quantities sufficient for the contemplated experiments. The site-directed mutation is incorporated into the DNA, and the mutated DNA is used to direct the production of the modified polypeptide in which one particular amino acid has been deliberately changed. For example, a collection of 13 mutated versions of the lysozyme from T4 bacteriophage, in which Threonine 157 had been changed to 13 of the other 19 amino acids, was produced by site-directed mutation. Each of these 13 different proteins was obtained as a pure crystalline product in quantities sufficient for crystallographic analysis.¹³⁴

A site-directed mutation can be introduced into a particular segment of DNA by annealing a short piece of synthetic DNA, the **mutagenic oligonucleotide**, to one of the two strands of the unmutated DNA to form a short section of double-helical DNA in which one or more of the nucleotide bases are mismatched.¹³² The mutagenic oligonucleotide is designed so that the desired mismatches occur in the middle of the duplex formed by the annealing and there are sufficiently long regions of complementary nucleotide sequence on each flank to guarantee that a stable and specific duplex is formed. The original way this was accomplished is the following.

A restriction fragment of the DNA encoding the protein of interest and containing the site to be mutated is inserted into the genome of an M13 bacteriophage, a bacteriophage that carries its genome as single-stranded DNA. Infection of a suspension of *E. coli* with the altered bacteriophage produces virus particles containing the enlarged genome on a closed, single-stranded circle of DNA.¹³⁵ Closed, single-stranded circles containing the strand of the inserted DNA complementary to the mutagenic oligonucleotide are selected¹³³ for hybridization. The mutagenic oligonucleotide is complementary to sequences on this single-stranded DNA except at the central, mismatched positions, chosen to produce the desired change in a particular codon. For example, the deoxyribonucleotide sequence –CTCTACTGCGGGTT–TG– occurs in DNA encoding the sequence of tyrosyl-tRNA synthetase from *Bacillus stearothermophilus*. It encodes the amino acid sequence –LYCGF–, which contains amino acids 33–37 in the sequence of the intact protein. The mutagenic oligonucleotide –CAAACCCGC–CGTAGAG– was chemically synthesized.¹³⁶ It is complementary to the coding sequence of the unmutated complementary DNA except at its tenth residue, which is a C instead of the complementary A. When it was annealed to a single-stranded, circular DNA containing DNA with the unmutated sequence, it formed a short self-complementary segment of double-stranded DNA in which its C was mismatched with the T of the unmutated

sequence. It was this mismatch that eventually produced the mutated DNA with the sequence –CTCTACG-GCGGGTTTG–, encoding the mutated protein sequence, –LYGGF–.

The short mutagenic oligonucleotide sits upon the single-stranded, circular M13 DNA as a primer offering a free 3'-hydroxyl group. This hydroxyl is used to initiate the synthesis of DNA by DNA-directed DNA polymerase.^{132,133} The enzyme synthesizes a single strand of DNA upon the circular template until it comes around the circle to the 5'-end of the mutagenic oligonucleotide where it stops. The newly synthesized, single-stranded circle is then closed with DNA ligase to produce a closed, double-stranded circle of DNA, completely complementary except at the designed mismatch. This double-stranded circular DNA is then replicated in a suspension of *E. coli*. Half of the resulting viral DNA should contain the mutated sequence of the segment of the inserted DNA because it is the progeny of the single strand into which the mutagenic oligonucleotide was incorporated originally.

Plaques produced by the viruses are screened to locate ones producing the mutated DNA,¹³³ double-stranded DNA is produced from one of these mutants and amplified, and the desired restriction fragment containing the mutation is isolated and reintroduced into the original DNA to create full-length DNA incorporating the mutation. The mutant protein expressed from this full-length, mutant DNA should contain the designated substitution. For example, in the case of the mutated tyrosyl-tRNA synthetase, it was shown by direct sequencing of the purified protein that it had a glycine rather than a cysteine at position 35.¹³⁶ That the modification has occurred, however, is usually verified by sequencing the mutated DNA rather than the protein itself.

Several **improvements** in the original method for site-directed mutation just described have been made. The most important is the adaptation of the procedure so that double-stranded plasmids, rather than single-stranded M13 DNA, can be mutated directly.¹³⁷ Another improvement has been the development of strategies permitting the removal of the parental unmutated strands of DNA that served as the template for the mutation so that all of the newly synthesized DNA carries the mutated sequence,^{138–141} increasing the percentage of the product that bears the mutation. A related method that also selects for DNA bearing the mutation is to use two primers, one that mutates the position of interest and the other that mutates a unique restriction site on the plasmid outside of the DNA inserted into it. In this way only the DNA containing the desired mutation, which also has the mutated restriction site, is immune to cleavage at the restriction site.¹⁴² Finally, the PCR method has been applied to produce mutated DNA.^{143–145} Because of its importance, many different procedures are now available for site-directed mutation, and each investigator believes that the one she is using is the best.

Site-directed mutations can also be produced by insertion of **cassettes** of synthetic double-stranded DNA into a particular complementary DNA. In this method, preexisting or purposely designed restriction sites for site-specific deoxyribonucleases that flank the region to be mutated are chosen. These restriction sites are designed or chosen so that the piece of double-stranded DNA produced by the site-specific deoxyribonucleases is short and has single-stranded, sticky ends, such as those produced by *Pst*I (CTGCA↓G). A double-stranded segment of DNA is synthesized so that it has the appropriate sticky ends and incorporates complementary nucleotide sequences that encode the desired mutation. This is the cassette, which is then inserted into the hole in the original complementary DNA produced by the site-specific deoxyribonucleases. The advantage of the cassette is that the mutation is produced directly by insertion of synthetic double-stranded DNA. The disadvantage is that two complementary pieces of synthetic single-stranded DNA have to be synthesized. Nevertheless, mutation with cassettes has particular advantages when sets of mutants are prepared in which all of the possible 19 substitutions need to be made at a particular location.¹⁴⁶ A similar but much more ambitious strategy is to synthesize fragments of DNA that when ligated together constitute the entire coding sequence for a protein. In this way a mutation can be introduced anywhere by synthesizing the corresponding fragment that has the altered sequence at the position to be mutated and ligating it with the remaining unmutated fragments.¹⁴⁷

One of the supposed drawbacks of site-directed mutation is that only the 19 other natural α amino acids are available for substitution at the mutated site. It is rather easy to synthesize an α amino acid. A large number are available commercially and if one that has been drawn on a piece of paper is not available commercially, it can usually be synthesized. It is now possible to replace an amino acid at any position in a polypeptide with any one of these **unnatural amino acids**. To do this, advantage is taken of the fact that there are three stop codons for translation: UAA (ochre), UAG (amber), and UGA. A rare tRNA, the amber suppressor tRNA, reads the codon UAG and normally inserts phenylalanine at that position. The triplet encoding the chosen amino acid in the coding sequence of the protein is mutated by usual site-directed mutation to TAG, and an amber suppressor tRNA (tRNA_{CUA}) to which the unnatural amino acid to be inserted has been synthetically attached is used to effect the desired substitution in a cell-free system for transcription and translation.^{148,149} The requirements for chemically synthesizing the derivative of the suppressor tRNA and the low yields of protein from the cell-free translation system have limited the application of these procedures, but in at least one instance protein sufficient for crystallographic studies has been prepared.¹⁵⁰

Suggested Reading

Chaiyen, P., Ballou, D.P., & Massey, V. (1997) Gene cloning, sequence analysis, and expression of 2-methyl-3-hydroxypyridine-5-carboxylic acid oxygenase, *Proc. Natl Acad. Sci. U.S.A.* 94, 7233–7238.

Foulon, V., Antonenkov, V.D., Croes, K., Waelkens, E., Mannaerts, G.P., VanVeldhoven, P.P., & Casteels, M. (1999) Purification, molecular cloning, and expression of 2-hydroxyphytanoyl-CoA lyase, a peroxisomal thiamine pyrophosphate-dependent enzyme that catalyzes the carbon-carbon bond cleavage during α -oxidation of 3-methyl-branched fatty acids, *Proc. Natl Acad. Sci. U.S.A.* 96, 10039–10044.

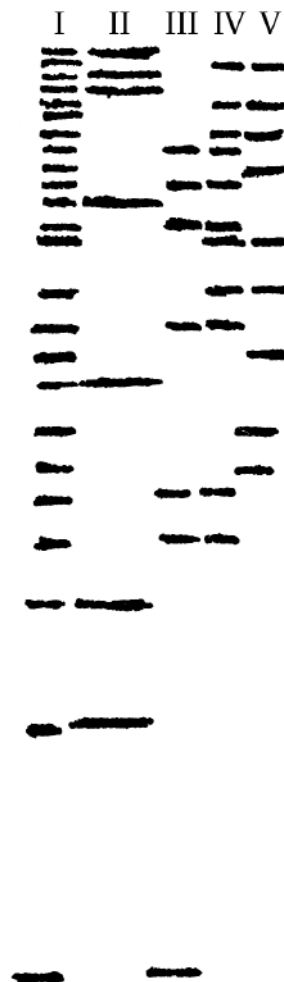
Problem 3-7: Draw the complete structures of 2'-deoxyadenosine 5'-triphosphate (dATP), 2'-deoxyguanosine 5'-triphosphate (dGTP), thymidine 5'-triphosphate (dTTP), 2'-deoxycytidine 5'-triphosphate (dCTP), and the single-stranded deoxyribonucleic acid AGTC.

Problem 3-8: Write out, in the three-letter abbreviations for the amino acids, the amino acid sequences of the amino terminus and the internal tryptic peptide of extensin from *V. carteri* that guided the synthesis of the two primers used to make the probe by the polymerase chain reaction. Look up the genetic code. Below these two amino acid sequences, write out the nucleic acid sequences of the 5'- and 3'-ends of the sense strand of the segment of 410 bp amplified by the polymerase chain reaction aligned by the genetic code with the respective sequences of the amino acids. Below these two sequences of nucleic acid, write out their complementary sequences. Below each of the appropriate positions in these two blocks of aligned sequences, write out all of the redundant codons that the probe was designed to include. How many polynucleotides of different sequence resulted from each of the two syntheses?

Problem 3-9: A fragment of single-stranded RNA, 488 nucleotides long, was obtained from one of the ribosomal RNAs of rat liver, 28S rRNA, by treatment with α -sarcin.¹⁵¹ It was treated with alkaline phosphatase to remove any phosphate from its 5'-end and then with [γ -³²P]MgATP and T4 polynucleotide 5'-hydroxyl-kinase to attach a radioactive phosphate to its 5'-end. The sample was then split into five separate portions. They were treated with the following reagents, respectively:

- (I) NaOH
- (II) ribonuclease T₁, which cleaves on the 3'-side of G
- (III) ribonuclease U₂, which cleaves on the 3'-side of A
- (IV) ribonuclease PhyM, which cleaves on the 3'-sides of A and U
- (V) ribonuclease BC, which cleaves on the 3'-sides of U and C

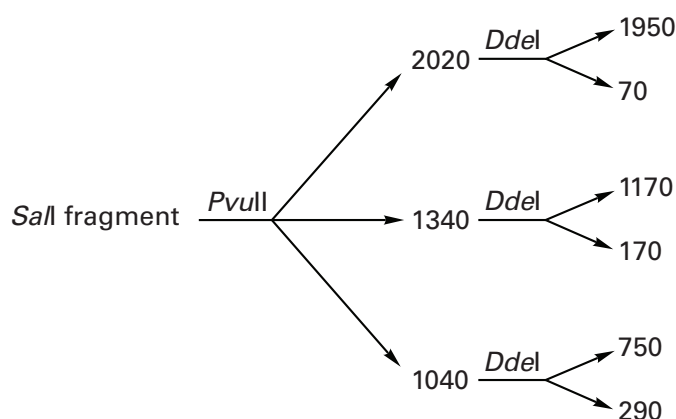
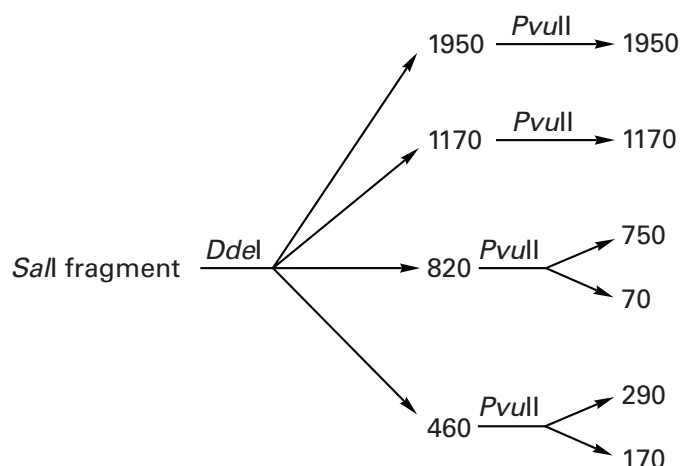
The alkaline hydrolysis (I) and the enzymatic digestions (II–V) were carefully controlled so that only a small amount of cleavage occurred at each sensitive position. The five mixtures were then placed in adjacent lanes on a polyacrylamide gel and submitted to electrophoresis followed by autoradiography. A tracing of that autoradiogram is presented below. An autoradiogram only registers radioactive fragments.



Each lane is labeled with the appropriate roman numeral. The most rapidly migrating bands were mononucleotides.

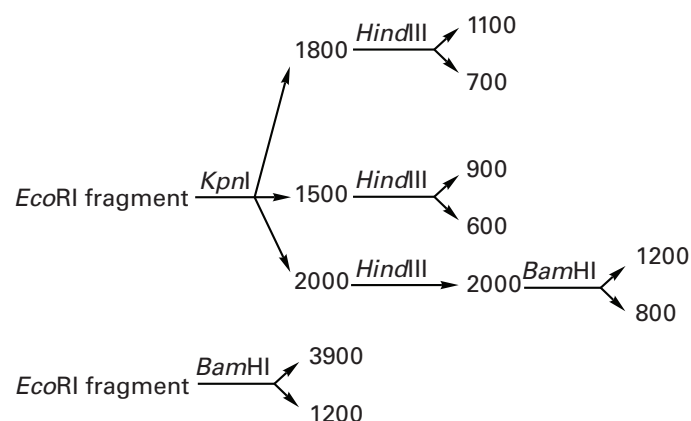
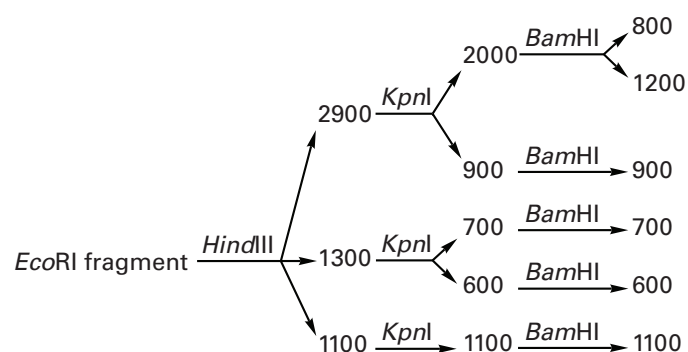
- (A) Starting with the nucleotide on the 5'-end, write the sequence of the α -sarcin fragment covered by the gel. Indicate clearly 5' \rightarrow 3' polarity.
- (B) Look carefully at the gel and then give a reason for including the digest in lane I.

Problem 3-10: A piece of double-stranded DNA about 4360 base pairs in length was produced by the site-specific deoxyribonuclease *SalI*. When this was digested with the site-specific deoxyribonucleases *DdeI* and *PvuII*, the fragments described in the diagram below were obtained. The numbers are the approximate lengths of the fragments.



Construct a restriction map.

Problem 3-11: A piece of double-stranded DNA about 5300 base pairs in length has been produced by the action of the site-specific deoxyribonuclease *EcoRI*. When this fragment was digested with the site-specific deoxyribonucleases *HindIII*, *KpnI*, and *BamHI*, the following results were obtained. The numbers are the approximate lengths of the fragments.

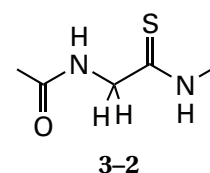


Construct a restriction map.

Posttranslational Modification

With the exception of the evanescent N^α -formyl group on its amino terminus and perhaps the 21st primary amino acid, selenocysteine,¹⁵² the infant polypeptide as it emerges from the peptidyltransferase site on the ribosome is a polymer containing only the 20 natural amino acids. Each amino acid is coupled to its neighbors by the amides of the peptide backbone, and the amino acids are arranged in the sequence encoded by the particular messenger RNA. It is this covalent structure and only this covalent structure that can be read by the investigator from the sequence of the messenger RNA or genomic DNA. The covalent structures of many proteins, however, do not remain in this untouched state but are biologically modified. A **posttranslational modification** is any change in the covalent structure of a polypeptide that occurs after its emergence from the ribosome.

Although a thiopeptide bond



has been observed at Glycine 445 of coenzyme-B sulfoethylthiotransferase,^{153,154} most posttranslational modifications of the polypeptide backbone result from endopeptidolytic cleavage or covalent rearrangements.

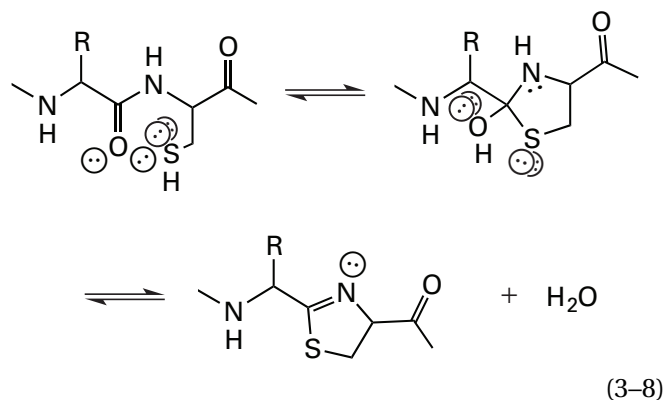
Modifications of the original covalent structure of the polypeptide are performed naturally by cellular **endopeptidases**. Such normal editing of the amino acid sequence of the protein must be distinguished from artifactual degradation by endopeptidases that can occur, for example, during the purification of a protein. In the course of a normal, natural modification, the polypeptide of a particular protein is cleaved internally, either as a mechanism for controlling its enzymatic activity or for architectural purposes. An example of the former is the

activation of endopeptidases in the pancreatic secretions or the serum by internal cleavages by endopeptidases.¹⁵⁵ An example of the latter is the trimming of folded proinsulin to produce insulin. As in the production of insulin from proinsulin, a number of other hormones are produced by endopeptidic cleavage at –Lys–Lys– or –Arg–Lys– positions in the sequence of longer precursors.¹⁵⁶ For example, corticotrophin, β -lipotropin, γ -lipotropin, β -endorphin, α -melanocyte-stimulating hormone, and γ -melanocyte-stimulating hormone are all cut from the same precursor 265 aa in length.^{157,158} Following the initial endopeptidolytic, posttranslational cleavage, the new amino terminus and carboxy terminus can be further digested by exopeptidases.¹¹¹

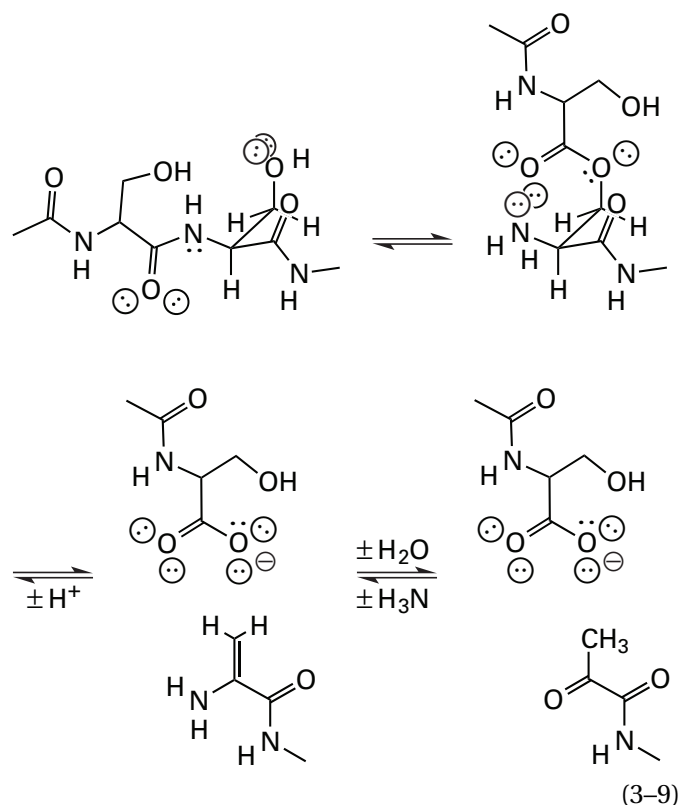
Almost all of the proteins of animals are posttranslationally shortened by the removal of one or more of the amino acids from their amino terminus, but some proteins have particular segments removed from their amino termini as they are passed from one compartment in the cell to another compartment. These amino-terminal **signal sequences**¹⁵⁹ address the proteins to the proper locations, and their removal is presumably involved in keeping them there. These successive removals of portions of the amino-terminal sequence have led to the terms pre-proprotein and proprotein.

There is a set of posttranslational modifications involving cysteines, serines, threonines, asparagines, and aspartates that result in **rearrangements in the covalent structure of the polyamide backbone** of a protein or **self cleavage** of its backbone. These five amino acids promote these modifications because they place either a nucleophile or an electrophile four atoms away from either an electrophilic acyl carbon or a nucleophilic amide nitrogen, respectively. Thus, the chemistry involved is the chemistry of five-membered heterocycles. Almost all of these posttranslational modifications are catalyzed intramolecularly by the protein itself.

Cysteines, serines, and threonines have their nucleophilic oxygens or sulfurs four atoms away from the acyl carbon of their amino-terminal neighbor. One example of a consequence of this spacing is the posttranslational modifications that produce thiazolines



and oxazolines in microsin.^{160,161} Another is the self-catalyzed posttranslational modification that cleaves the polypeptides of human *S*-adenosylmethionine decarboxylase¹⁶² between Glutamine 67 and Serine 68', histidine decarboxylase from *Lactobacillus*^{163,164} between Serine 81 and Serine 82, and aspartate 1-decarboxylase from *E. coli*¹⁶⁵ between Glycine 24 and Serine 25, in each case producing a pyruvated amino terminus.

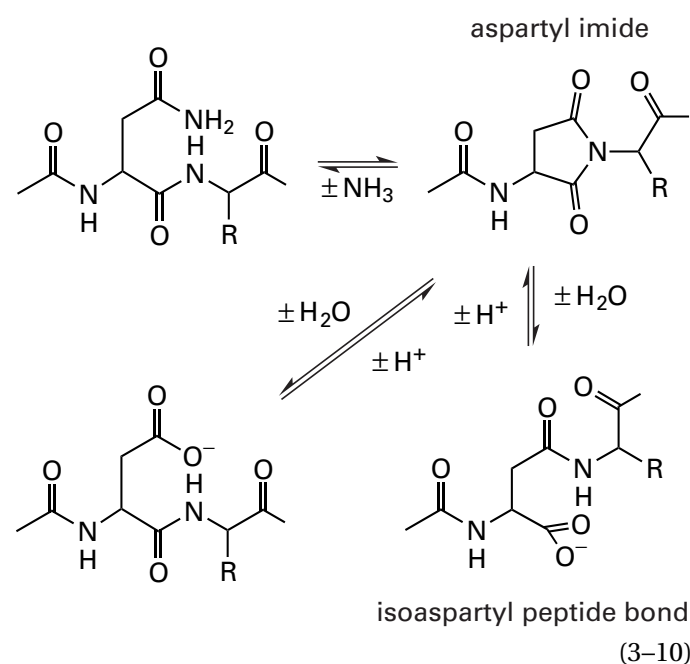


The first step in this reaction is a five-membered tetravalent intermediate as in the first step of Equation 3-8, but the amine leaves the intermediate rather than water to produce the intermediate ester, which has been observed crystallographically.^{162,165} This step is an example of an **N \rightarrow O acyl migration**. The next step in this reaction utilizes the superior ability of carboxylate as a leaving group to effect the dehydration ultimately producing the pyruvyl group. The oxygen of the original serine ends up in the carboxylate of the new carboxy terminus produced in the reaction.¹⁶⁶ An α -ketobutyryl group¹⁶⁷ is found as an acyl substituent at the amino terminus of one of the two polypeptides composing threonine dehydratase, and it presumably arises by a similar mechanism (Equation 3-9) from a threonine in the protein rather than a serine.

An N \rightarrow O acyl migration also occurs in the self-catalyzed¹⁶⁸ cleavage of the peptide bond on the amino-terminal side of Threonine 206 in human *N*⁴-(β -*N*-acetylglucosaminy)-L-asparaginase.¹⁶⁹ In this case, the ester produced by the migration (Equation 3-9), rather than providing a leaving group, hydrolyzes to

produce the break in the polypeptide between amino acids 205 and 206. In hedgehog protein from *D. melanogaster*, the thioester resulting from an **N → S migration** of the polypeptide at Cysteine 258 is transesterified onto the hydroxyl group of cholesterol, which takes the place of the water that would hydrolyze the ester.¹⁷⁰ In the process, the polypeptide is cleaved between Glycine 257 and Cysteine 258, and the cholesterol ends up as a posttranslational modification esterified to the new carboxy terminus.

An **asparagine or aspartic acid** places an electrophile four atoms away from the amide nitrogen of its carboxy-terminal neighbor. This can lead to the production of an aspartyl imide, an isoaspartyl peptide bond, or an aspartate where there was an asparagine.^{171,172}



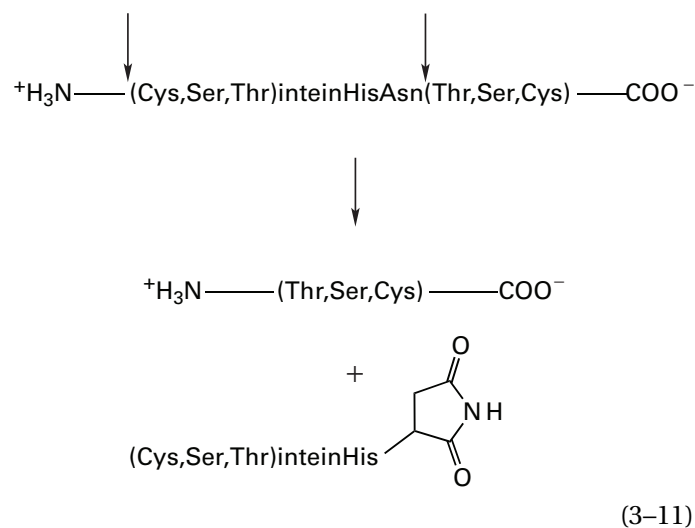
It is the production of aspartyl imides between asparagines and adjacent glycines ($R = H$), promoted by alkaline pH and elevated temperature, that is thought to be the first step in the chemical cleavage of a polypeptide by hydroxylamine.³⁷ The preference, in this case, for asparaginylglycyl peptide bonds is thought to be due to a steric effect on the initial cyclization that is minimized when the amido nitrogen that attacks the acyl carbon of the side chain of the asparagine is that of a glycine. It is thought that the aspartyl imide is then cleaved by the nucleophilic hydroxylamine to produce the chemical cleavage used experimentally to produce large fragments of a polypeptide.³⁸

The formation of aspartyl imides, aspartates, and isoaspartyl peptide bonds also seems to occur spontaneously but slowly at many of the asparagines, as well as at aspartic acids,¹⁷¹ in most proteins in their natural environment to produce a low level of isoaspartyl peptide bonds,¹⁷³ which are more stable than normal aspartyl or

asparaginyl peptide bonds.¹⁷² Both an aspartylimide and an isoaspartyl peptide bond have been observed crystallographically at the position of Aspartate 101 in hen lysozyme, which precedes Glycine 102,¹⁷⁴ and an isoaspartyl peptide bond has been observed crystallographically at the position of Asparagine 67 in bovine pancreatic ribonuclease, which precedes Glycine 68.¹⁷⁵ Because the hydrolysis of an aspartyl imide can lead to the replacement of an asparagine with an aspartate still in a normal peptide bond, this reaction may be responsible for the deamidation observed at particular sites in some proteins.^{176,177} Because the aspartylimide racemizes more rapidly at its α carbon than does either of the amides,¹⁷² this process also introduces D-aspartates into the polypeptide.

Both the unnatural isoaspartyl peptide bonds and the D-aspartates are recognized by a repair enzyme that methylates their free carboxylates. This **methylation** reinitiates the formation of the aspartyl imide, which can spontaneously racemize and hydrolyze to produce L-aspartate in a normal peptide bond, thus repairing the problem.¹⁷⁸⁻¹⁸¹ Only a fraction of the imide racemizes before it hydrolyzes, and when it hydrolyzes the isoaspartyl peptide bond is the favored product, but if only the D-aspartates and the isoaspartates are methylated and if they are recycled often enough, significant repair can be accomplished.¹⁸⁰⁻¹⁸²

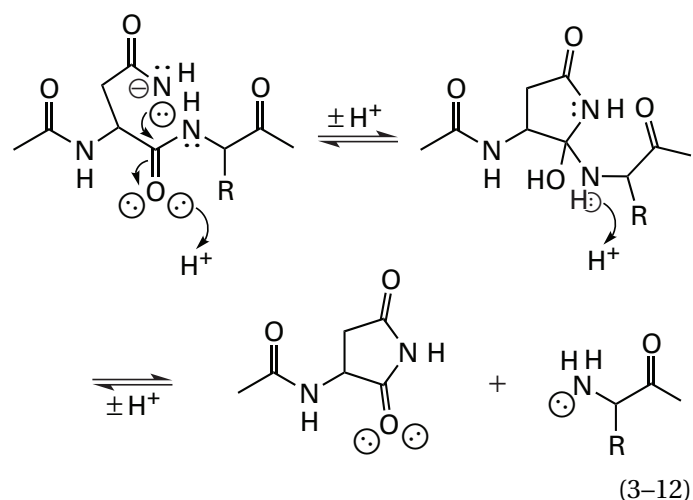
The ultimate exploitation of this class of posttranslational modifications involving five-membered rings is self-catalyzed rearrangements of an amino acid sequence. There is a group of proteins, such as vacuolar adenosinetriphosphatase,^{183,184} certain RecA proteins,¹⁸⁵ and certain DNA polymerases¹⁸⁶ that contain an internal sequence called an intein that is 400–550 aa in length. The **intein** is spliced out of these proteins coincident with the formation of a peptide bond connecting the carboxy terminus of the amino-terminal segment of the protein that precedes the intein to the amino terminus of the carboxy-terminal segment following it.



The intein can be one continuous folded polypeptide connecting the amino-terminal segment preceding it to the carboxy-terminal segment following it, or it can be the carboxy-terminal segment of one folded polypeptide that is bound noncovalently to the amino-terminal segment of a second folded polypeptide.^{187,188} In the latter instance, the intein, after it has been spliced out, is two folded polypeptides bound to each other but the other product is still one continuous, spliced polypeptide formed from the amino-terminal segment of the first folded polypeptide and the carboxy-terminal segment of the second. The intein always begins with a serine, threonine, or cysteine and ends with a histidiny asparagine, and the carboxy-terminal segment always begins with a serine, threonine, or cysteine.¹⁸⁹

An even more extensive set of similar self-catalyzed posttranslational rearrangements occurs in concanavalin A from *Canavalia ensiformis*. After the initial polypeptide is produced by the ribosome, the α -amido group of Serine 30 couples to the α -acyl group of Asparagine 281 in place of the α -amido group of Glutamate 282, releasing the amino-terminal 29 amino acids preceding Serine 30 and the carboxy-terminal nine amino acids following Asparagine 281 as two short peptides, and the polypeptide is cleaved to the carboxy-terminal sides of Asparagines 148 and 163, releasing the intervening 26 amino acids as another short peptide.¹⁹⁰ The final intact product of the splicing begins at Alanine 164 of the precursor and ends at Asparagine 148.

The two spontaneous cleavages to the carboxy-terminal sides of Asparagines 148 and 163 in concanavalin A are thought to result from an attack of the amide nitrogen of the asparagine on its own acyl carbon



in a reaction analogous to that involving the acyl oxygen of aspartate under acidic conditions (Figure 3-5). In the case of concanavalin A, this cleavage would be catalyzed by other amino acids in the protein and the product would be the imide.

The first step in intein splicing is an $N \rightarrow O$ or $N \rightarrow S$

acyl migration (Equation 3-9) of the amino-terminal segment of the protein occurring at the serine, threonine, or cysteine on the amino-terminal side of the intein (Equation 3-11).^{191,192} The amino-terminal segment is then passed by a transesterification to the oxygen or sulfur of the serine, threonine, or cysteine on the carboxy-terminal side of the upstream splice site (Equation 3-11) to form a branched intermediate in which the intein and the carboxy-terminal segment are still joined together and the amino-terminal segment is esterified to the serine, threonine, or cysteine.^{193,194} In the next step of the reaction, the peptide bond to the carboxy-terminal side of the asparagine is cleaved (Equation 3-12) to produce the free intein with an unsubstituted aspartyl imide at its carboxy terminus.¹⁸⁹ The peptide bond between the amino-terminal segment and the carboxy-terminal segment is then formed by the respective $O \rightarrow N$ or $S \rightarrow N$ acyl migration. The amino-terminal and carboxy-terminal splice sites sit next to each other in the folded protein to permit all of these rearrangements to occur in close proximity.^{188,195,196}

In the rearrangement of concanavalin A, Glutamate 282 in the asparaginyglutamate is replaced by Serine 30 to produce an asparaginyserine. The first step in this reaction is probably the cleavage of the peptide bond of the asparaginyglutamate at positions 281 and 282 to produce the aspartyl imide at the resulting carboxy terminus (Equation 3-12). The following steps in the reaction would then be, by analogy to those of intein splicing, $N \rightarrow O$ migration at Serine 30, attack of the α -amino group of Serine 30 on the aspartyl imide of Asparagine 281, and hydrolysis of the ester between the α -carboxyl group of Serine 29 and the hydroxyl group on the side chain of Serine 30.

The posttranslational modifications of the backbone of the initially synthesized polypeptide that are produced by endopeptidolytic cleavage, self-catalyzed cleavage, the formation of aspartyl imides or isoaspartyl peptide bonds, or intein splicing have usually been identified by electrophoresis of complexes between the polypeptide and dodecyl sulfate, a procedure that registers the lengths of constituent polypeptides; by electrospray mass spectrometry, a procedure that registers decreases in mass caused by loss of portions of the polypeptide; by amino-terminal sequencing, a procedure that registers newly formed amino termini; by digestion with carboxypeptidases, a procedure that identifies new carboxy-terminal sequences; and by digestion with exopeptidases, which digest the normal peptide bonds but not imides or isopeptide bonds. These analyses rely heavily on the complete amino acid sequence of the unmodified polypeptide that is the immediate product of protein synthesis. This sequence is learned from sequencing the nucleic acid encoding it. For example, even though the complete amino acid sequence of mature concanavalin A was known from direct sequencing,¹⁹⁷ the extensive rearrangements producing the final

protein went unrecognized until the complementary DNA encoding it had been sequenced.^{190,198}

The **amino terminus** of a polypeptide can be *N*-methylated,^{199,200} *N*-2-pyruvylated,²⁰¹ or *N*-acylated, either intramolecularly, as in pyroglutamate (Figure 3-16), or externally, as when it is *N*-formylated,²⁰² *N*-acetylated,²⁰³ or *N*-glucuronylated.²⁰⁴ Enzymes are available that hydrolyze pyroglutamyl groups²⁰⁵ or remove acetyl groups.²⁰⁶ In a murein lipoprotein from bacterial outer membrane²⁰⁷ and ubiquinol oxidase (cytochrome *bo*₃) from *E. coli*,²⁰⁸ each of the respective amino-terminal cysteines is *N*-acylated by a fatty acid at its α -amino group and its sulfur forms a thioether with carbon 3 of a 1,2-diacylglycerol. *n*-Tetradecanoyl amides of amino termini (Figure 3-16)²⁰⁹ were first found on protein kinases. The existence of these fatty acylated amino termini was established by isolating an amino-terminal peptide, CH₃(CH₂)₁₂COHNGly-Asn-Ala, from cAMP-dependent protein kinase and confirming its structure by chemical degradation and by mass spectrometry with fast-atom bombardment.²¹⁰ By similar procedures it was shown that recoverin was acylated at its amino terminus with a mixture of *n*-dodecanoic acid, *cis*-*n*-tetradec-5-enoic acid, and *cis,cis*-*n*-tetradeca-5,8-dienoic acid in

addition to *n*-tetradecanoic acid.²¹¹ Such chemical demonstrations of a modification at the amino terminus of a polypeptide should be distinguished from an unsupported conjecture that the amino terminus is blocked when the Edman degradation fails.

The **carboxy terminus** of a polypeptide can also be modified, for example, as the primary amide (Figure 3-16), the tyrosyl amide,²¹² or the methyl ester.²¹³ In at least one instance,²¹⁴ the primary amide at a carboxy terminus is produced from a carboxy-terminal glycine that is first monooxygenated and then decomposes with the loss of glyoxylate to leave behind its former amino group as the carboxy-terminal amide.

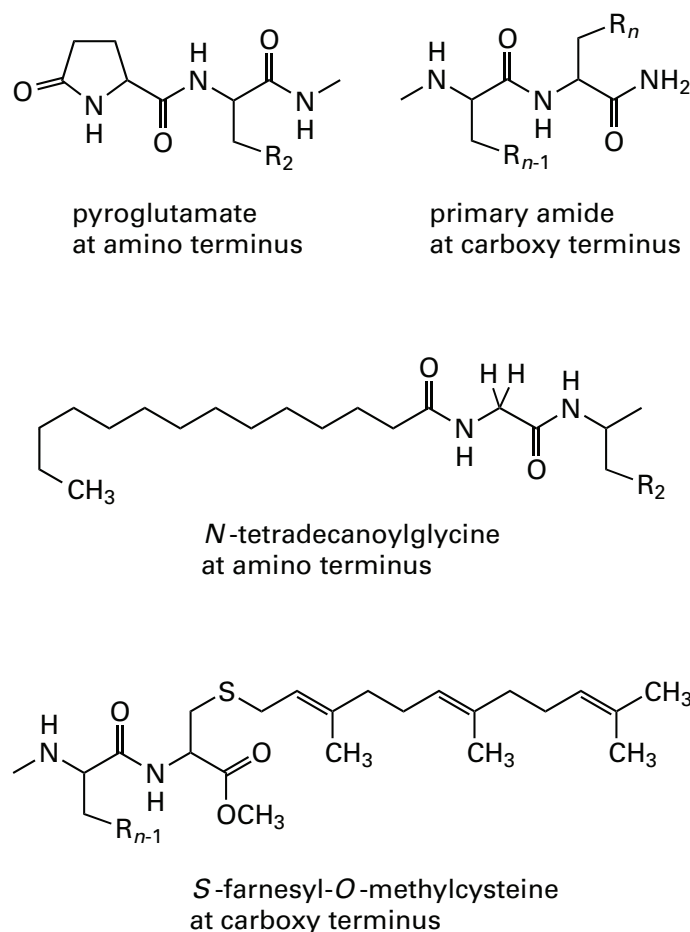
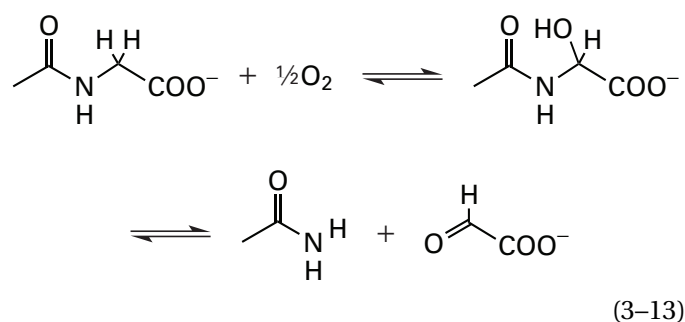


Figure 3-16: Posttranslational modifications that occur at the termini of a polypeptide.

During the posttranslational modification of several polypeptides with the carboxy-terminal sequence CXYZ (where X, Y, and Z each represent one of many possible amino acids),²¹⁵ the last three amino acids are removed,²¹⁶ and the cysteine at the new carboxy terminus is doubly modified (Figure 3-16) by **isoprenylation** and methylation. A farnesyl group²¹⁷ or a geranylgeranyl group^{218,219} is added to the sulfur of the cysteine in an allylthioether, and the new carboxy terminus is methylated to form the methyl ester.^{215,220} It is thought that these modifications make the carboxy terminus sufficiently hydrophobic to bind to biological membranes.^{221,222} There are also proteins in which the polypeptide synthesized from the messenger RNA has the carboxy-terminal sequence Cys-Cys or Cys-X-Cys, and each of the cysteines in these carboxy-terminal sequences is then geranylgeranylated.^{223,224} The geranylgeranylated proteins with the carboxy-terminal sequence Cys-X-Cys are then methylated on their terminal carboxylates,^{223,225} but those with the carboxy-terminal sequence Cys-Cys are not.²²⁵

An extensive posttranslational modification of the carboxy terminus occurs in certain proteins that are bound tightly to the extracellular surface of protozoal and animal cells.²²⁶⁻²²⁸ It has the effect of covalently connecting the carboxy terminus of the protein to a phosphatidylinositol dissolved within the bilayer of the plasma membrane. The carboxy terminus is linked directly through an amide to an ethanolamine, which is linked through a phosphate diester to the mannose of an oligosaccharide, which, in turn, is linked by a glycosidic linkage to phosphatidylinositol, a phospholipid (Figure 3-17).²²⁸⁻²³⁵ Because this

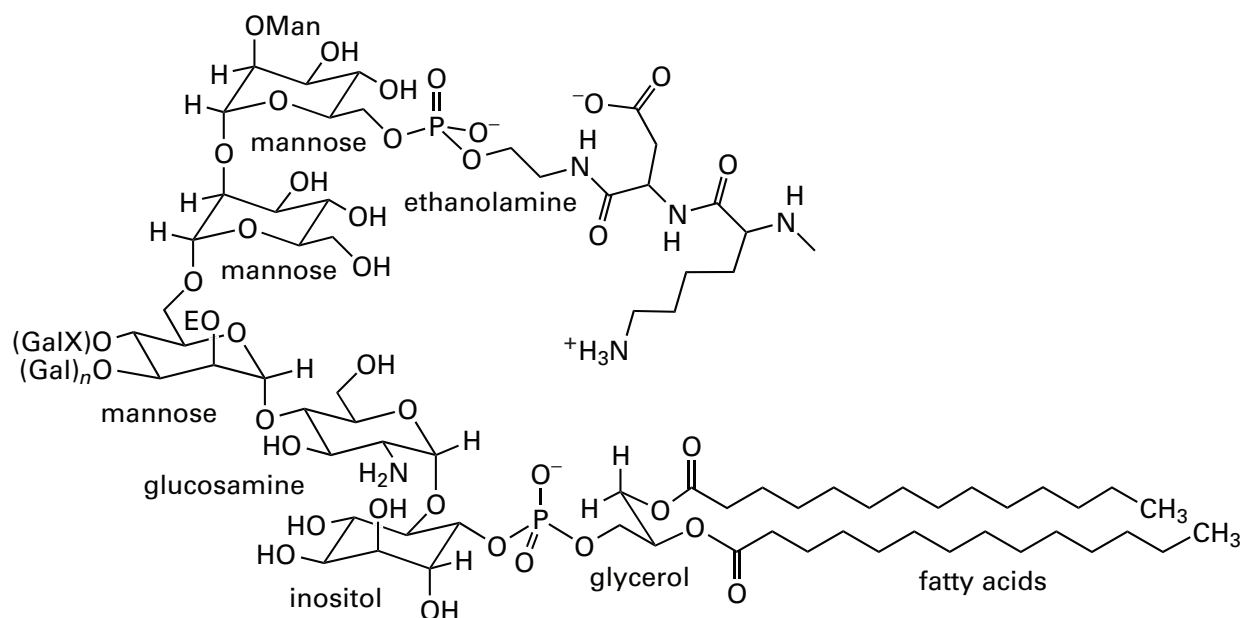


Figure 3-17: Structure of the linkage between phosphatidylinositol and the carboxy terminus of a polypeptide in a phosphatidylinositol-linked protein.²²⁸⁻²³² The carboxy-terminal amino acid sequence shown is that for the linkage at the end of the variant surface glycoprotein MITat.1.4 from *Trypanosoma brucei*.²³³ The phosphatidylinositol shown is in the ditetradecanoyl form, but saturated and unsaturated fatty acids from 12 to 22 carbons in length can be esterified at either position in place of either or both of the tetradecanoyl groups. A variant in which a tetradecanoyl group is also attached to the inositol has been reported,²³⁴ as well as one in which a ceramide replaces the diacylglycerol.²³⁰ The phosphatidylinositol is coupled to an unacetylated D-glucosamine, which is coupled in turn through a trisaccharide of D-mannoses to ethanolamine phosphate, the primary amine of which is attached in amide linkage to the carboxy terminus of the protein. A variant of the more common structure displayed here has the phosphoethanolamine attached through the 3-position of the middle mannose rather than the 6-position of the end mannose.²³⁵ Within the tetrasaccharide, the position marked (Gal)_n is either a hydrogen or an oligosaccharide of one or more galactosyl groups; the position marked Man is either a hydrogen, an α 1-mannosyl group, or a mannosyl disaccharide; the position marked (Gal X) is either a hydrogen, a β 1-galactosyl group, or a β 1-N-acetylgalactosamyl group; and the position marked E is either a hydrogen or an (O-ethanolamino)phosphoryl group.

posttranslational modification causes the protein to adhere to membranes, it is called a **glycosylphosphatidylinositol (GPI) anchor**.

In addition to the polypeptide itself and its amino and carboxy termini, **posttranslational modification of the side chain of an amino acid** can occur. When it does, the derivative remains an L- α -amino acid residue because its carboxyl group and its α -amino group are protected by the amides of the backbone. There are many posttranslationally modified amino acids that have been identified in naturally occurring polypeptides (Table 3-1, Figure 3-18).

The length of Table 3-1 gives the erroneous impression that posttranslational modifications are common. Aside from glycosylation and the phosphorylation of serines, threonines, and tyrosines, the incidence of any of these modifications is quite **limited**, often being confined to only one protein or one small family of proteins. For example, two of the earliest recognized posttranslational modifications were the 5-hydroxylysine and 4-hydroxyproline (Figure 3-18) that, with few exceptions, are formed in the posttranslational monooxygenation of only prolines and lysines that are found in segments of amino acid sequence in which every third amino acid is a glycine.⁴¹⁰ Such sequences

occur in the various collagens and proteins related to the collagens. The modifications producing covalently bound coenzymes occur only in proteins using these coenzymes to assist in catalysis of particular reactions. Some of the other posttranslational modifications, for example, the quinones of 2,5-dihydroxytyrosine (Figure 3-18),³¹² 6,7-dioxo-4-(2-tryptophanyl)tryptophan (Figure 3-18), and dehydroalanine,³⁴⁹ occur only in the active sites of particular enzymes and are designed for specific functions. 4-Carboxyglutamate (Figure 3-18)^{252,253,385,411} is found only in a few of the proteins that bind calcium strongly or that are involved in calcium metabolism.⁴¹² Thyroxine,²⁶¹⁻²⁶³ O-(3,5-diiodo-4-hydroxyphenyl)-3,5-diiodotyrosine (Figure 3-18), is found only in the protein thyroglobulin,²⁶⁶ wherein it is formed at two positions by the intramolecular condensation of two pairs of 3,5-diiodotyrosines.⁴¹³ The sole function of this large protein (2769 aa) is to produce the thyroxine, which is then liberated from the protein by its complete digestion. Diphthamide, 2-[3-carboxamido-3-(trimethylammonio)propyl]histidine (Figure 3-18), is found only in one of the elongation factors (elongation factor 2) involved in eukaryotic translation.^{267,269} The attachment of ADP-ribosyl groups to diphthamide, arginine, and asparagine side chains in one or the other of a

small group of proteins is catalyzed by bacterial toxins, and only proteins in individuals infected with these bacteria are modified in this way. There also seem to be enzymes in normal cells, however, that are capable of ADP-ribosylating a small number of proteins as part of their normal operation.^{331,414-417}

Mass spectrometry is often used to identify these posttranslational modifications on the side chains of amino acids. Electrospray mass spectrometry of a purified, intact protein is often the first indication that it contains a posttranslational modification. Because the unmodified amino acid sequence of a protein as it is produced by the ribosome is often known even before it has been purified but usually soon after, any difference between the molecular mass observed by electrospray mass spectrometry and the mass calculated from the unmodified amino acid sequence indicates a posttranslational modification. Such results were the first indications that the protein *Ner* of bacteriophage *Mu* was modified at its amino terminus with a pyruvate²⁰¹ and that bovine recoverin was modified at its amino terminus by one of several different fatty acids.²¹¹ Electrospray mass spectrometry of peptides purified from a digest of rat profilaggrin identified nine phosphopeptides by the fact that their molecular masses were 80 Da greater than those predicted from their amino acid sequences.⁴¹⁸

Normal direct probe, high-resolution mass spectrometry and mass spectrometry with electron ionization have been used to provide molecular ions and fragment ions of posttranslational modifications such as polyisoprenoids^{219,296} or 5-mercaptopuracil³⁴² that can be removed chemically from the amino acid to which they are attached. Electrospray mass spectrometry in the negative ion mode has been used in a similar way to identify the ceramide released from the GPI anchor of the arabinogalactan proteoglycan from *Pyrus communis*.²³⁰ Fast-atom bombardment feeding a conventional mass spectrometer has been used to vaporize a bispeptide containing the semicarbazide derivative of 6,7-dioxo-4-(2-tryptophanyl)tryptophan and obtain a **high-resolution mass spectrum** with a molecular ion of 940.3262 Da which was of sufficient precision to calculate a molecular formula for the modification.³⁸⁵

Fast-atom bombardment or matrix-assisted-laser-desorption ionization feeding a tandem mass spectrometer can be used to vaporize posttranslationally modified peptides purified from a digest of a protein, sort the molecular ions in the first mass spectrometer, fragment those ions, and then separate the fragments in the second mass spectrometer. The resulting pattern of fragments is often sufficient to identify the posttranslational modification. Peptides containing an α -hydroxyglycine,³²⁴ an 8 α -(*N*³-histidyl)flavin mononucleotide,⁴¹⁹ and an *N*-acetyl-*O*-phosphothreonine⁴²⁰ were analyzed in this way. Matrix-assisted-laser-desorption ionization feeding a time-of-flight mass spectrometer in either the positive-ion mode^{304,305,350} or negative-ion mode⁴²¹ has

been used to identify posttranslational modifications in peptides that have been purified from digests of proteins. The negative-ion mode is used for phosphopeptides.⁴²²

Electrospray can also be used to produce ions to feed a tandem mass spectrometer.^{161,270} Although some difficulty arises with the multiply charged ions emitted by the electrospray, they can usually be sorted out successfully in the first mass spectrometer because peptides are short enough that only a few ions are produced from each of them.¹¹¹ For example, a peptide containing covalently bound flavin from fructosyl-amino acid oxidase of *Aspergillus* was vaporized by electrospray ionization, the ionic molecule of *m/z* 659 Da was selected in the first quadrupole mass spectrometer, it was fragmented by collision-induced dissociation, and the fragment ions produced a mass spectrum in the second quadrupole mass spectrometer of the tandem. The pattern of fragments demonstrated that the flavin was covalently attached to Cysteine 342 of the protein.⁴²³ Such a system can also be used to identify the locations in the sequence of a protein at which it is phosphorylated.⁴²⁴

If the posttranslational modification cannot be identified by its mass or its pattern of fragmentation, it is usually possible to hydrolyze the polypeptide and liberate the modified amino acid. Usually the hydrolysis is performed enzymatically to avoid destruction of the modified amino acid that might occur in strong acid or strong base. Enough of the peculiar amino acid is purified to perform a proof of its structure by chemical analysis.

One way in which two or more of the amino acid side chains in a polypeptide can be modified coincidentally is during the formation of a **covalent cross-link** between them or among them. The cross-link can be intramolecular, connecting two or more amino acid side chains in the same polypeptide, or intermolecular, connecting two or more amino acid side chains in different polypeptides. There is no formal distinction between these two outcomes because the linkage is invariably made after the polypeptides have folded into their native structure and, subsequently, formed specific intermolecular complexes among themselves. This folding and intermolecular assembly is what brings the two or more amino acid side chains that will be cross-linked into atomic contact with each other. Therefore, it is irrelevant whether the amino acid side chains started out on the same polypeptide or different polypeptides or whether they are at positions within the amino acid sequence of the same polypeptide that are close to or distant from each other. The only deciding factor is that they are immediately adjacent to each other in the final structure of the mature protein.

A simple example of a covalent cross-link is an amide between a lysine side chain and a glutamate side chain. Such a cross-link is formed from a glutamine side chain and a lysine side chain, both within a

Table 3–1: Posttranslational Modifications of the Side Chains of Amino Acids in Proteins²³⁶

type of modification	derivative of side chain
phosphorylation	<i>O</i> -phosphoserine, ^{237–239} <i>O</i> -phosphothreonine, ^{238–240} <i>O</i> -phosphotyrosine, ^{239,241,242} <i>N</i> -phospholysine, ²⁴³ <i>N</i> -phosphohistidine, ^{243–245} <i>N</i> -phosphoarginine, ²⁴⁵ <i>S</i> -phosphocysteine, ²⁴⁶ <i>O</i> -phosphoaspartate, ^{247–249} <i>O</i> -phosphoglutamate ²⁵⁰
sulfation	<i>O</i> -sulfotyrosine ²⁵¹
carboxylation	4-carboxyglutamate, ^{252,253} 3-carboxyaspartate ²⁵⁴
aromatic substitution	3,5-diiodotyrosine, ^{255–257} 3-iodotyrosine, ^{255,257,258} 3,5-dibromotyrosine, ^{255,259} 3-bromotyrosine, ^{255,258} 3-bromo-5-chlorotyrosine, ^{255,260} 3-chlorotyrosine, ²⁵⁵ 3,5-dichlorotyrosine, ²⁵⁵ <i>O</i> -(3,5-diiodo-4-hydroxyphenyl)-3,5-diiodotyrosine (thyroxine, T ₄), ^{261–263} <i>O</i> -(3-iodo-4-hydroxyphenyl)-3,5-diiodotyrosine (triiodothyronine, T ₃), ^{264–266} 2-[3-carboxamido-3-(trimethylammonio)propyl]histidine (diphthamide), ^{267–269} 2-(1-mannosyl)tryptophan, ^{270,271} <i>o</i> -bromophenylalanine ²⁷²
methylation	<i>N</i> -methyllysine, ^{273–275} <i>N,N</i> -dimethyllysine, ^{274–276} <i>N,N,N</i> -trimethyllysine, ^{274,275,277} <i>N</i> ^ω -methylarginine, ^{278,279} <i>N</i> ^ω , <i>N</i> ^ω -dimethylarginine, ^{278,280} <i>N</i> ^ω , <i>N</i> ^ω -dimethylarginine, ^{281,282} <i>N</i> ^δ -methylarginine, ²⁸³ <i>N</i> ³ -methylhistidine, ²⁸⁴ <i>N</i> ¹ -methylhistidine, ¹⁵⁴ <i>O</i> -methyl-D-aspartate, ¹⁷⁸ <i>O</i> -methylglutamate, ²⁸⁵ <i>O</i> -methylisoaspartate, ²⁸⁶ <i>S</i> -methylmethionine, ²⁸⁷ <i>S</i> -methylcysteine, ¹⁵⁴ <i>N</i> -methylasparagine, ^{288,289} <i>N</i> -methylglutamine, ²⁹⁰ 2-(<i>S</i>)-methylglutamine, ^{153,291} 5-(<i>S</i>)-methylarginine ^{153,291}
alkylation	<i>N</i> -(4-amino-2-hydroxybutyl)lysine (hypusine), ^{292–294} <i>S</i> -farnesylcysteine, ^{217,295} <i>S</i> -geranylgeranyl cysteine ^{219,296}
acylation	<i>N</i> -acetyllysine, ^{297,298} <i>O</i> -palmitoylthreonine, ²⁹⁹ <i>S</i> -palmitoylcysteine, ^{300–302} <i>S</i> -stearoylcysteine ³⁰³
amidation	γ -poly(α -glutamyl)glutamate, ³⁰⁴ γ -poly(glycyl)glutamate ³⁰⁵
monooxygenation	5-hydroxylysine, ³⁰⁶ <i>N,N,N</i> -trimethyl-5-hydroxylysine, ³⁰⁷ <i>N,N,N</i> -trimethyl- <i>O</i> -phospho-5-hydroxylysine, ³⁰⁷ 4-hydroxyproline, ^{308–310} 3-hydroxyproline, ³¹¹ 2,5-dihydroxytyrosine, ³¹² β -hydroxyaspartate, ^{313–315} β -hydroxyasparagine, ³¹⁶ β -hydroxytryptophan, ^{317,318} <i>m</i> -hydroxyphenylalanine, ³¹⁹ 3-hydroxytyrosine (3,4-dihydroxyphenylalanine, DOPA), ^{320–323} α -hydroxyglycine ³²⁴
oxidation	6-deamino-6-oxolysine (allysine), ^{306,325} 6-deamino-5-hydroxy-6-oxolysine, ³⁰⁶ cysteinesulfenic acid, ³²⁶ cyteinesulfenic acid, ³²⁶ β -dethio- β -oxocysteine, ³²⁷ methioninesulfone ³²⁸
free radical	tyrosyl free radical, ³²⁹ glycyl free radical ³³⁰
ADP-ribosylation	<i>N</i> ^ω -(ADP-ribosyl)arginine, ^{331–333} <i>N</i> -(ADP-ribosyl)asparagine, ³³⁴ <i>S</i> -(ADP-ribosyl)cysteine, ³³⁵ poly(ADP-ribosyl)glutamate, ^{336,337} 1-[<i>N</i> -(ADP-ribosyl)]-2-[3-carboxamido-3-(trimethylammonio)propyl]histidine ^{338,339}
nucleotidylation	<i>O</i> -(5'-adenylyl)tyrosine, ³⁴⁰ <i>O</i> -(5'-uridylyl)tyrosine, ³⁴¹ <i>O</i> -[5'-(5-mercapto)uridylyl]tyrosine ³⁴²
hydrolysis	citrulline from arginine, ³⁴³ ornithine from arginine, ³⁴⁴ aspartate from asparagine, ^{176,345,346} glutamate from glutamine ³⁴⁶
dehydration	dehydroalanine from serine, ^{347–349} α,β -dehydrotyrosine ^{350,351}
glycosylation	<i>O</i> -poly(mannosyl)serine, ³⁵² <i>O</i> -poly(mannosyl)threonine, ³⁵² <i>O</i> -oligo[(α 1,2)galactosyl]serine, ³⁵³ <i>O</i> -[3- <i>O</i> -(β -glucosyl)- α -fucosyl]threonine, ³⁵⁴ <i>O</i> -[2- <i>O</i> -(α -glucosyl)- β -galactosyl]-5-hydroxylysine, ³⁵⁵ <i>O</i> -(β -xylosyl)serine, ^{356,357} <i>O</i> -[4- <i>O</i> -(β -galactosyl)- β -xylosyl]serine, ^{356,357} <i>S</i> -digalactosylcysteine, ³⁵⁸ <i>S</i> -triglucosylcysteine, ³⁵⁹ <i>O</i> -(glucosylarabinosyl)hydroxyproline, ³⁶⁰ <i>O</i> -(<i>N</i> -acetylglucosaminyl)serine, ^{361,362} <i>O</i> -(<i>N</i> -acetylglucosaminyl)threonine, ³⁶² <i>O</i> -poly(arabinofuranosyl)hydroxyproline, ³⁶³ <i>O</i> -[3-[D-xylosyl(α 1,3)-D-xylosyl]-D-glucosyl]serine, ³⁶⁴ <i>O</i> -poly(glucosyl)tyrosine ³⁶⁵
cross-links between side chains of two amino acids	lysine in amide linkage with aspartate, ³⁶⁶ lysine in amide linkage with glutamate, ^{367,368} cysteine in thioester with glutamate, ^{369,370} 2-(<i>S</i> -cysteinyl)histidine, ^{371,372} 3-(<i>S</i> -cysteinyl)tyrosine, ^{373,374} 3-(1-histidyl)tyrosine, ^{375,376} 3-(3-tyrosyl)tyrosine, ^{377,378} <i>O</i> -(3-tyrosyl)tyrosine, ^{379–381} 3,5-di(3-tyrosyl)tyrosine, ³⁷⁸ 3,3'-methylenebis(tyrosine), ³⁸² 3-(3-tyrosyl)- <i>O</i> -(3-tyrosyl)tyrosine, ³⁸³ 3-(<i>O</i> -tyrosyl)-5-(3-tyrosyl)tyrosine, ³⁸⁴ 6,7-dioxo-4-(2-tryptophanyl)tryptophan (tryptophan tryptophylquinone), ^{385,386} 5-hydroxy-2-(<i>N</i> -lysyl)tyrosine, ³⁸⁷ cystine
covalently bound coenzymes	heme in thioether linkages to two cysteines, ³⁸⁸ <i>S</i> -phycoerythrobilinylcysteine, ³⁸⁹ bis(<i>S</i> -cysteinyl)phycoerythrobilin, ³⁹⁰ 8a-(<i>S</i> -cysteinyl)-8a-hydroxyflavin adenine dinucleotide, ³⁹¹ 8a-(<i>S</i> -cysteinyl)flavin adenine dinucleotide, ³⁹² 8a-(<i>N</i> ³ -histidinyl)flavin adenine dinucleotide, ^{392,393} 6-(<i>S</i> -cysteinyl)flavin mononucleotide, ^{394,395} 8a-(<i>N</i> ³ -histidinyl)flavin mononucleotide, ³⁹⁶ 8a-(<i>O</i> -tyrosyl)flavin adenine dinucleotide, ^{397,398} <i>N</i> -biotinyllysine, ³⁹⁹ <i>N</i> -lipoyllysine, ⁴⁰⁰ <i>N</i> -(phosphopyridoxyl)lysine, ^{401,402} <i>N</i> -retinyllysine, ^{403,404} <i>O</i> -(4'-phosphopantetheinyl)serine ⁴⁰⁵ heme in thioether linkage to two cysteines and ether linkage to tyrosine at a <i>meso</i> position, ⁴⁰⁶ 2'-[5''-(phosphoseryl)ribosyl]-3'-dephosphocoenzyme A ^{407–409}

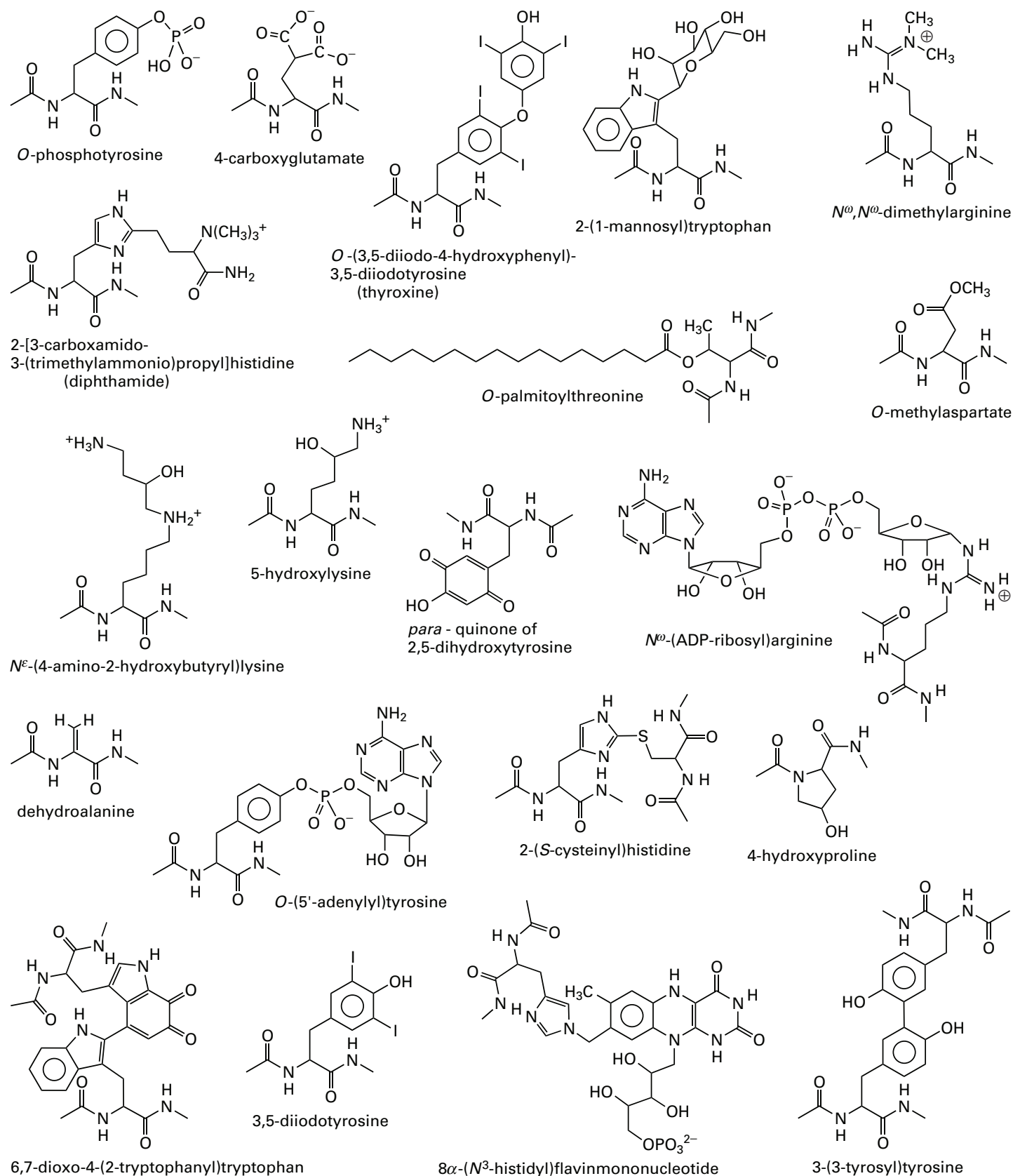
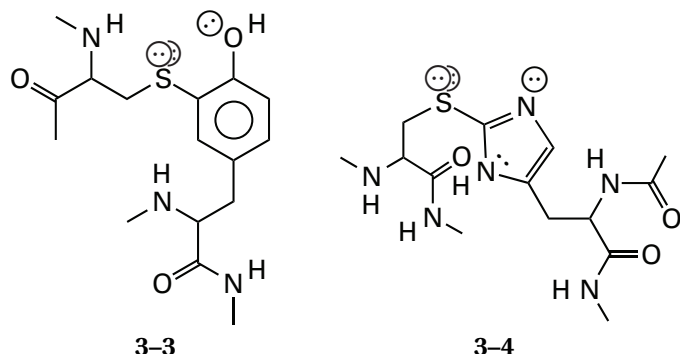


Figure 3-18: Posttranslational modifications of amino acid side chains in the interior of a polypeptide.

folded polypeptide, by the enzyme protein-glutamine γ -glutamyltransferase.⁴²⁵ Thioethers between a cysteine side chain and the aromatic ring of either a tyrosine side chain (3-3)³⁷³ or a histidine side chain (3-4)³⁷²

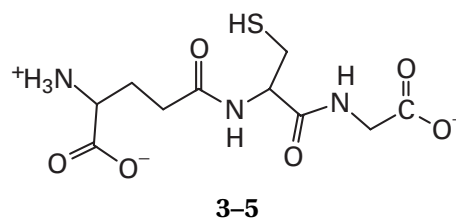


are cross-links that have been identified in galactose oxidase,³⁷⁴ and in hemocyanin,⁴²⁶ and monophenol monooxygenase,^{371,372} respectively.

A large number of cross-links form in both collagen and elastin as the direct result of the posttranslational oxidative deamination of lysine side chains in these proteins to the corresponding aliphatic 6-deamino-6-oxolysyl aldehydes in a reaction performed by the enzyme protein-lysine 6-oxidase.⁴²⁷ These aldehydes are formed in the vicinity of each other as well as in the vicinity of other lysines and of 5-hydroxylysines and 5-hydroxy-6-deamino-6-oxolysyl aldehydes, all also derived from lysine side chains. A dazzling array of aliphatic carbonyl chemistry is initiated by the formation of the reactive aldehydes in this environment, including aldol condensations, imine formations, dehydrations of β -hydroxyaldehydes, dehydrations of aliphatic alcohols, Michael additions, and oxidations (Figure 3-19). Only four out of the more than 25 cross-links that result from this uncontrolled flurry of reactions³⁰⁶ are displayed in the figure. The purpose of these cross-links is to strengthen fibers of collagen. Cross-links between tyrosine side chains, such as 3-(3-tyrosyl)tyrosine (Figure 3-18) and the other examples listed in Table 3-1, also serve to strengthen the biological fibers, films, and coatings in which they are found.

One of the most common posttranslational cross-links in a protein is the disulfide that forms when two cysteine side chains are oxidatively coupled to form one **cystine** side chain (Figure 3-20). While cysteine is unstable under the conditions necessary to hydrolyze proteins in strong acid, cystine is stable, and its appearance on the standard ion-exchange chromatogram between alanine and valine (Figure 1-3) establishes the presence of cystine side chains in a protein.

The interior of most cells has a high concentration of a small, free thiol, such as **glutathione**



that reduces back to cysteine any cystine that forms in cytoplasmic proteins in a reaction known as disulfide interchange (Figure 3-20). The net effect of disulfide interchange is to set the cystine side chains in a protein in equilibrium with the disulfide of the small thiol such that

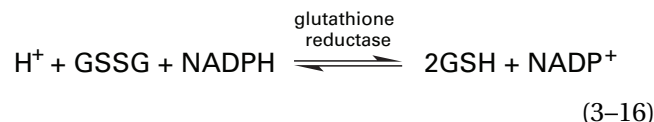
$$K_{eq} = \frac{[RS-SR][\text{prot}(SH)_2]}{[RSH]^2[\text{prot}(S-S)]} \quad (3-14)$$

where RSH is the small thiol, RS-SR its disulfide, $[\text{prot}(SH)_2]$ is the molar concentration of a particular reduced pair of adjacent cysteine side chains in the folded polypeptide, and $[\text{prot}(S-S)]$ is the concentration of cystine between these same two side chains, also in the folded polypeptide. The reaction as written is first-order in reduced protein because the oxidation of the reduced protein is intramolecular. Equation 3-14 can be rearranged:

$$\frac{K_{eq}[RSH]^2}{[RS-SR]} = \frac{[\text{prot}(SH)_2]}{[\text{prot}(S-S)]} \quad (3-15)$$

The point made by this equation is that the greater the ratio $[RSH]^2/[RS-SR]$, the less cystine will be found in proteins.

In the cytoplasm, $[RS-SR]$ is kept at a low level enzymatically. For example, when RSH is glutathione, GSH, the enzyme glutathione reductase accomplishes this:



This reaction drives the equilibrium of Equation 3-14 in the direction of the reduced protein by coupling it to the level of reduction of NADPH. The result of all these facts is that proteins confined to the cytoplasm usually⁴²⁸ do not contain cystine, while proteins removed from the cytoplasm often do contain cystine.

A protein is usually prepared for sequencing by **reducing** any cystine side chains it may contain with a small thiol such as 2-mercaptoethanol (Figure 3-20, $R = HOCH_2CH_2-$) and then **alkylating** all of its cysteines with a reagent such as iodoacetic acid

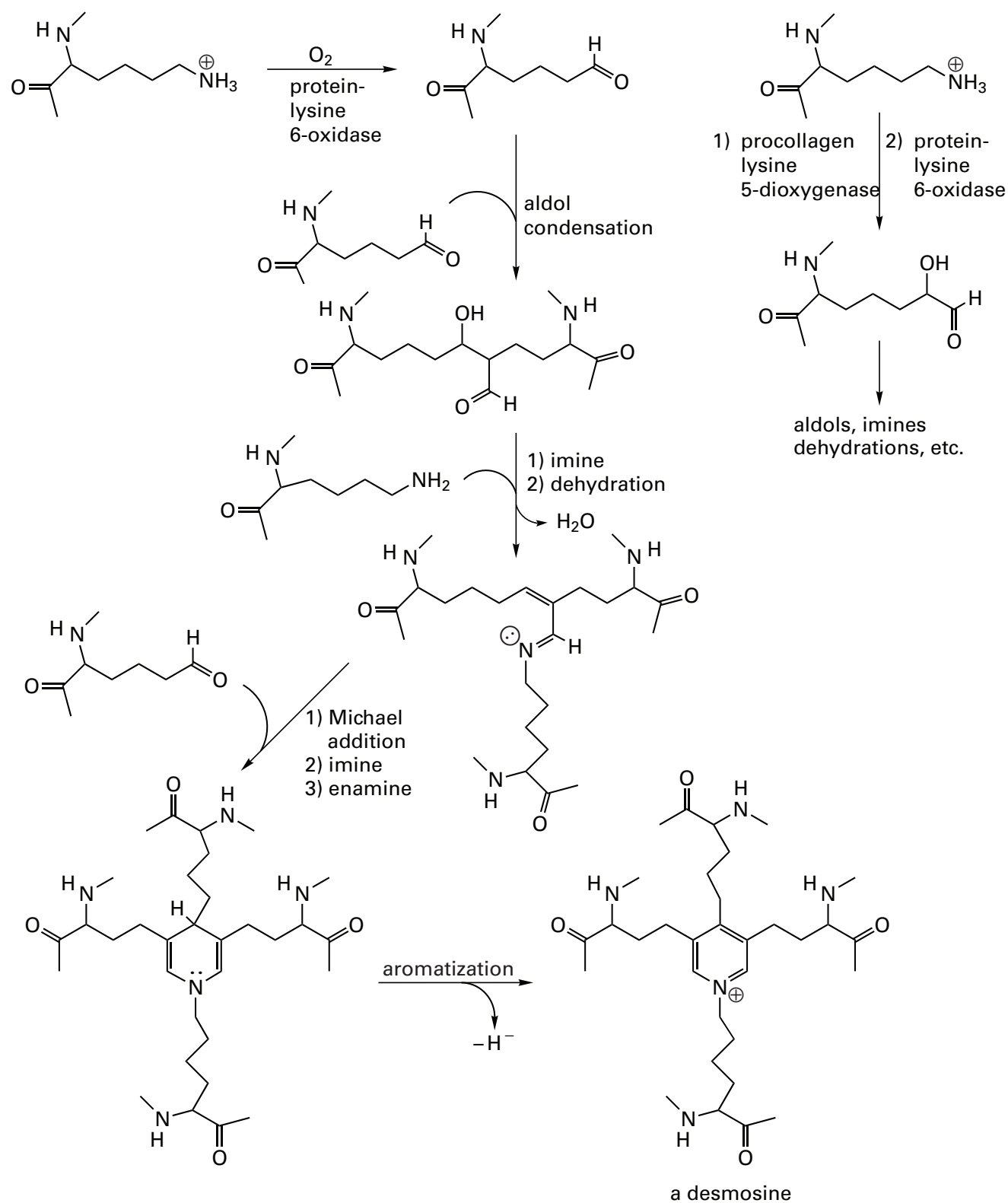


Figure 3-19: Examples of the formation of four of the more than 25 cross-links initiated by the formation of 6-deamino-6-oxolysine in collagen by protein-lysine 6-oxidase. Shown is an aldol condensation to produce the first cross-link. The β-hydroxyaldehyde can then form an imine with a lysine that dehydrates to an α,β-unsaturated imine, a product that cross-links three amino acids. The enol of another aldehyde can add to the α,β-unsaturated imine, and the initial enamine can condense to an imine with the carbonyl of the aldehyde. This forms a dihydropyridine that cross-links four amino acid residues. The pyridinium cation formed upon oxidation of the dihydropyridine is a desmosine linking the four amino acid residues. Upper right corner: 6-Deamino-5-hydroxy-6-oxolysine formed by the consecutive action of procollagen-lysine 5-dioxygenase and protein-lysine 6-oxidase produces an α-hydroxyaldehyde, which is susceptible to an even more complicated set of modifications.

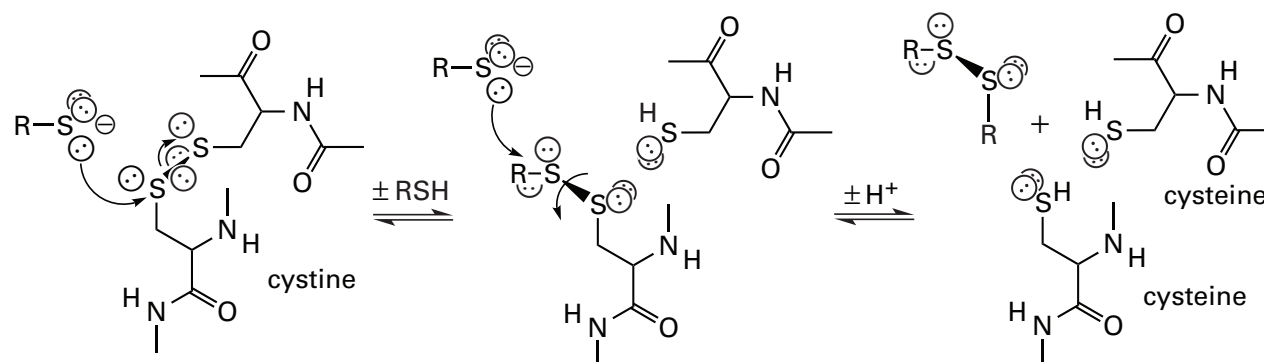
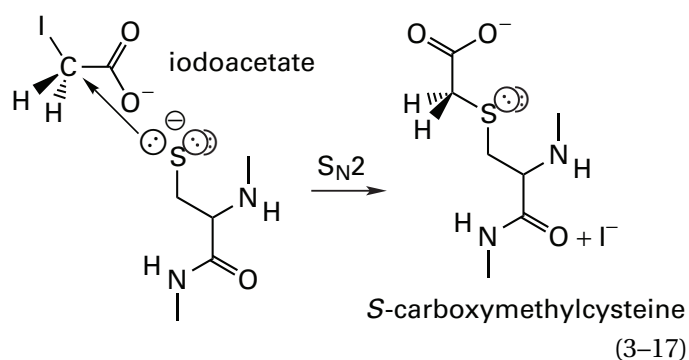
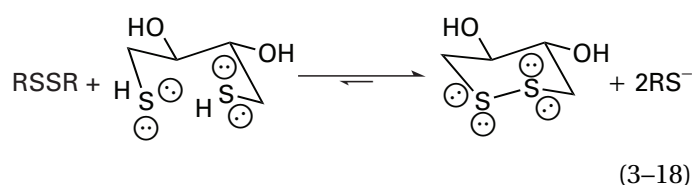


Figure 3-20: Reduction of a cystine side chain by disulfide interchange. The cystine connecting two segments of polypeptide is exposed to an external thiolate (RS^-) that displaces the cysteinyl anion by nucleophilic substitution and is in turn removed by another thiolate.



iodoacetamide, *N*-ethylmaleimide, 4-vinylpyridine, or 2-vinylpyridine to prevent their reoxidation. This creates the stable amino acid side chains *S*-(carboxymethyl) cysteine (Equation 3-17), *S*-(carboxamidomethyl) cysteine, *S*-(*N*-ethyl-2-succinimidyl)cysteine, *S*-[2-(4-pyridyl)ethyl]cysteine, or *S*-[2-(2-pyridyl)ethyl]cysteine, respectively, in place of the unstable cysteine. *S*-[2-(2-Pyridyl)ethyl]cysteine absorbs at 254 nm, and its absorption can be used to identify peptides containing cysteine.

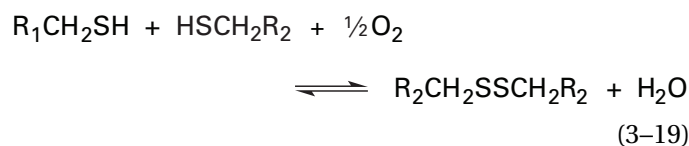
In experimental situations where cystines in a protein must be reduced, a problem arises. If the protein remains folded, its cystine side chains are generally more stable, relative to the adjacent cysteine side chains that would be produced upon its reduction, than is the disulfide of a small reducing agent such as 2-mercaptoethanol, relative to its thiol form. This results from the fact that the reaction in the protein is intramolecular in the direction of oxidation, but the reaction of the 2-mercaptoethanol is intermolecular in the direction of oxidation (Equation 3-14). The problem is solved either by using an intramolecular dithiol such as 2,3-dihydroxy-1,4-dithiobutane (**dithiothreitol**, DTT)⁴²⁹



which produces a stable disulfide as a product, or by unfolding the protein with urea or guanidinium chloride, which causes the reaction of the cysteine side chains in the protein to be formally intermolecular in the direction of oxidation, or by simultaneously applying both of these strategies.

There are a few examples in which a protein contains a cystine connecting two cysteines that are adjacent to each other in the amino acid sequence.⁴³⁰⁻⁴³³ The eight-membered ring that results, however, is unstable for steric reasons.^{432,434,435} Cystines connecting two cysteines from different folded polypeptides are also occasionally encountered,⁴³⁶ but in the majority of the extracytoplasmic proteins that contain cystine side chains as a posttranslational modification, the two cysteine side chains that are connected to each other are in the same folded polypeptide but far from each other in the amino acid sequence. For example, in ribonuclease, a protein formed from a polypeptide of 124 aa, the cystines are formed from Cysteine 96 and Cysteine 40, Cysteine 58 and Cysteine 110, Cysteine 26 and Cysteine 84, and Cysteine 65 and Cysteine 72. When the polypeptide folds to form the native structure of ribonuclease, these pairs of cysteine side chains, which were distant from each other in the unfolded polypeptide, are juxtaposed.

It is the **juxtaposition** that not only determines what cystines form but also brings the two sulfurs close enough to each other that they can react at an appreciable rate.^{437,438} They are then oxidized to cystine by molecular oxygen⁴³⁹



or by **disulfide interchange** (Figure 3-20) with cystines in protein disulfide-isomerase.⁴⁴⁰⁻⁴⁴³ Protein disulfide-isomerase⁴⁴⁴ contains a cystine so unstable that the

equilibrium constant^{445,446} for disulfide interchange between its disulfide (taking the place of RS-SR in Equation 3-14) and the disulfide in a normal extracytoplasmic protein [prot (S-S) in Equation 3-14] is about 10^{-3} . Unlike synthetic 2,3-dihydroxy-1,4-dithiobutane or the natural protein thioredoxin, both of which can cleave disulfides in native proteins, protein disulfide-isomerase forms them.

The identification of the two cysteine side chains that are connected in a particular cystine in a native protein requires that a peptide containing only those two cysteines still joined as the cystine be isolated from a digest of the protein.⁴⁴⁷⁻⁴⁵⁰ Before the protein is unfolded or digested, however, it must be treated with an alkylating agent such as *N*-ethylmaleimide under conditions capable of **capping off** all the free sulfhydryls in the preparation, which if left unalkylated would catalyze disulfide interchange (Figure 3-20) and thereby **scramble the disulfides**.⁴⁴⁸ Ideally, the peptides with intact cystine side chains used to identify the cysteines involved should be two short peptides held together by the cystine itself. For example, one of the peptides from ribonuclease isolated from a digest of the protein performed with pepsin, trypsin, and chymotrypsin was composed of the two smaller peptides NGQTNCYH and NVACK, covalently coupled by a cystine between the two cysteine side chains.⁴⁴⁷ From this result it could be concluded that, in native ribonuclease, Cysteine 65 is coupled as a cystine with Cysteine 72.

The three digestions used in the experiments just described served the purpose of producing bispeptides containing cystine that were as small as possible. This precaution avoids the confusion of having several large peptides interlaced by multiple disulfides⁴⁵¹ into one large, intractable peptide. This problem, however, is sometimes unavoidable, as in the case of thrombomodulin, in which three cystines occur within a short sequence of 14 amino acids and from which individual peptides containing each of them could not be obtained. In this case, the linkages were assigned⁴⁵² by following the rates at which the individual cysteines appeared as the cystines were slowly cleaved with the nucleophile tris(2-carboxyethyl)phosphine.⁴⁵³

Because oxidation states of cysteine (Figure 2-8) other than cystine as well as covalent modifications of cysteine⁴⁵⁴ revert to yield free cysteine upon addition of a thiol such as 2,3-dihydroxy-1,4-dithiobutane, the **indirect assignment** of a cystine based solely upon the appearance of free cysteine after the addition of a thiol cannot be trusted.^{455,456}

Procedures have been developed to assist in the **analysis of peptides containing cystine**. Sensitive methods have been described for continuously monitoring chromatograms of digests performed with endopeptidases to detect peptides containing intact cystine side chains either electrochemically after they have been reductively cleaved on the surface of an elec-

trode⁴³⁰ or colorimetrically after they have been nucleophilically cleaved with tributylphosphine.⁴⁵⁷ The bis(phenylthiohydantoin) of cystine (Figure 3-1) displays a unique relative mobility on the high-pressure liquid chromatograms used to identify the products from the steps of automated Edman degradation.^{458,459} Peptides containing cystine that have been purified from a digest of a protein can also be positively identified by mass spectrometry.^{460,461} The advantage in this instance is that the gaseous molecular ion of the bis-peptide, necessarily containing the intact cystine, is observed directly. The presence of a cystine within a peptide can be established by mass spectrometry because the mass of the peptide gradually increases by 2 Da as a result of photoreduction during successive shots from the laser during its vaporization.⁴⁶²

Suggested Reading

Carr, S.A., Biemann, K., Shoji, S., Parmelee, D.C., & Titani, K. (1982) *n*-Tetradecanoyl is the NH₂-terminal blocking group of the catalytic subunit of cyclic AMP-dependent protein kinase from bovine cardiac muscle, *Proc. Natl Acad. Sci. U.S.A.* 79, 6128-6131.

Haniu, M., Horan, T., Arakawa, T., Le, J., Katta, V., Hara, S., & Rohde, M.F. (1996) Disulfide structure and *N*-glycosylation sites of an extracellular domain of granulocyte-colony stimulating factor receptor, *Biochemistry* 35, 13040-13046.

Problem 3-12: Assume that a protein containing an intein has a cysteine at the amino terminus of the intein and a serine at the amino terminus of the carboxy-terminal segment. Write the mechanism for intein splicing involving the initial N → S migration, an S → O migration, cleavage of the peptide bond between the intein and the carboxy-terminal segment to produce the unsubstituted aspartyl imide, and the final O → N migration to produce the new peptide bond.

Problem 3-13: Draw the structure of a polypeptide with an amino-terminal cysteine residue the α-amine of which is acylated with palmitate and the sulfur of which forms a thioether with C3 of a 1,2-dipalmitoyl-3-deoxyglycerol.

Problem 3-14: A remarkable feature of the enzyme glutamate-ammonia ligase from *E. coli* is that its catalytic properties depend on the conditions of growth under which the *E. coli* from which it is purified were grown. The enzyme purified from *E. coli* grown on NH₄Cl and glucose (Type I) is less sensitive to inhibition by AMP than is the enzyme purified from *E. coli* grown on glutamate and glycerol (Type II). The Type II enzyme can be converted into Type I enzyme if it is treated with snake venom phosphodiesterase.⁴⁶³

When Type II enzyme was digested with snake venom phosphodiesterase and subsequently precipitated out of solution with trichloroacetic acid, the super-

natant solution contained material having a maximum absorbance at 260 nm.

Hydrolysis of the Type I and Type II enzymes was performed by a series of endopeptidases. Both enzymes were split into the same number of peptides, but one decapeptide from the Type II enzyme differed in chromatographic behavior from the similar decapeptide from the Type I enzyme. The single different peptide isolated from the Type II enzyme had the following composition after acid hydrolysis:

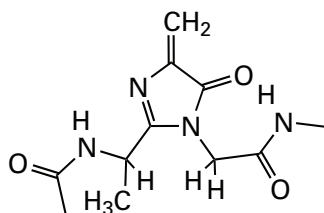
D	2
E	2
P	3
G	1
L	1
T	1
adenine	1
D-ribose	1
phosphate	1

From an acid-base titration, the following pK_a values were measured for the decapeptide isolated from Type II enzyme before and after treatment with snake venom phosphodiesterase.

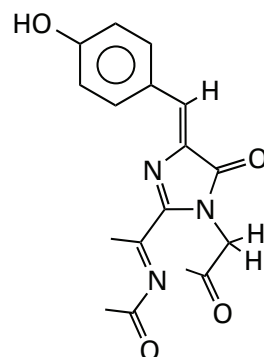
number of groups	pK_a	
	before	after
1	3.4	3.4
1	3.7	3.7
4	4–4.5	4–4.5
1	7.8	7.8
1		6.0
1		10.0

How does the covalent structure of the Type II enzyme differ from that of the Type I enzyme? Explain each result described above on the basis of the proposed structure.

Problem 3–15: Write mechanisms for the formation of the following posttranslational modification found in histidine ammonia-lyase⁴⁶⁴

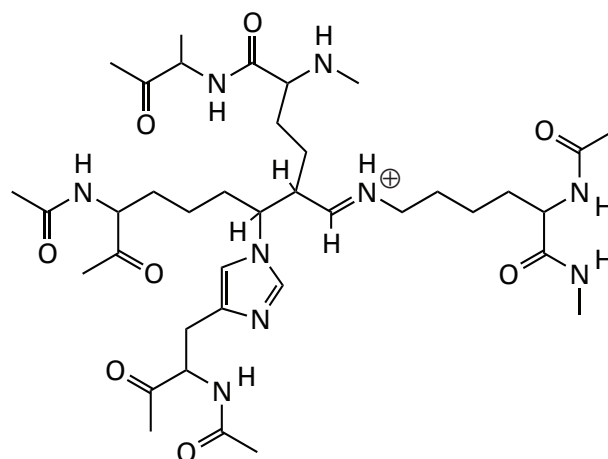


and the following posttranslational modification producing the chromophore in red fluorescent protein from *Discosoma*.⁴⁶⁵



Write resonance structures for the chromophore.

Problem 3–16: Write a step-by-step series of reactions that show how the cross-link dehydromerohistidine would form in collagen from three lysine residues and a histidine residue.



Problem 3–17: The sequence of tick anticoagulant protein is YNRLCIKPRDWIDECDSDNEGGERAYFRNGKG-GCDSFWICPEDHTGADYYSSYRDCFNACI.

A peptide was purified from a tryptic digest of the protein that produced the following results on Edman degradation.⁴⁶⁶

cycle	1	2	3	4	5	6
phenylthiohydantoins	D,G	G,W	C*,F,I	D	E,S	F
cycle	7	8	9	10	11	
phenylthiohydantoins	D,W	I,S	C*,N	E,P	E,G	

*Bis (phenylthiohydantoin) of cystine.

How are the cysteines linked to form the two cystines in the peptide? What unexpected cleavage did trypsin produce?

Oligosaccharides of Glycoproteins

Living organisms are formed from three types of covalent polymers: proteins, nucleic acids, and polysaccharides. **Polysaccharides** used biologically for structural purposes

or for the storage of carbohydrate occur as long, often branched, uniform polymers of monosaccharides (sugars). Examples of polysaccharides would be agarose (Figure 1-7), cellulose (Figure 1-2), starch,⁴⁶⁷ hyaluronic acid, chitin, and glycogen. Although the reducing ends of these polysaccharides are sometimes attached covalently to particular proteins,⁴⁶⁸ this fact is secondary to their biological roles. **Oligosaccharides** are shorter, more heterogeneous oligomers of monosaccharides. Oligosaccharides are frequently attached to recently synthesized polypeptides as posttranslational modifications of serines, threonines, or asparagines. Such posttranslational modifications produce glycoproteins.

A **glycoprotein** is any protein to which one or more oligosaccharides are covalently attached. To define the complete covalent structure of a glycoprotein, not only the amino acid sequence of the protein but also the points of attachment and the sequences of the monosaccharide in the oligosaccharides must be established. Some of the rarely occurring oligosaccharides and their sites of attachment have been listed along with the other posttranslational modifications in Table 3-1. The more commonly encountered oligosaccharides in glycoproteins from animals and plants, however, are branched oligomers attached through *N*-acetylglucosamine to asparagine side chains or through *N*-acetylgalac-

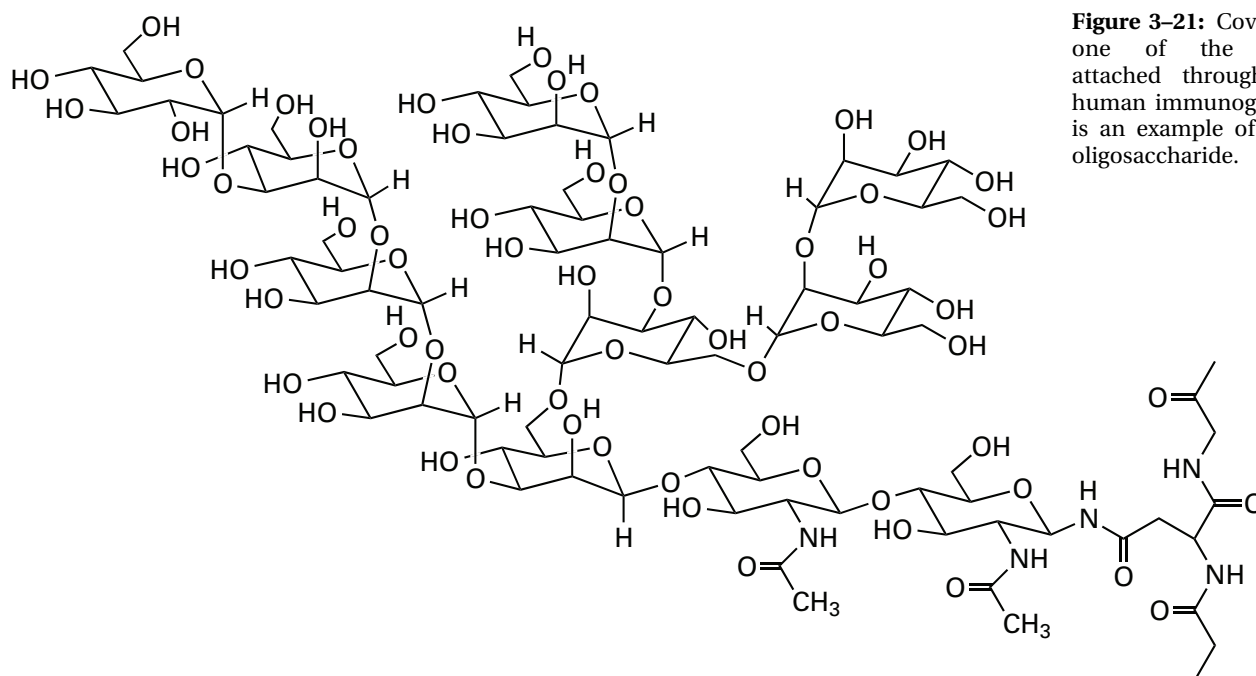


Figure 3-21: Covalent structure of one of the oligosaccharides attached through asparagine to human immunoglobulin D.⁴⁶⁹ This is an example of a high-mannose oligosaccharide.

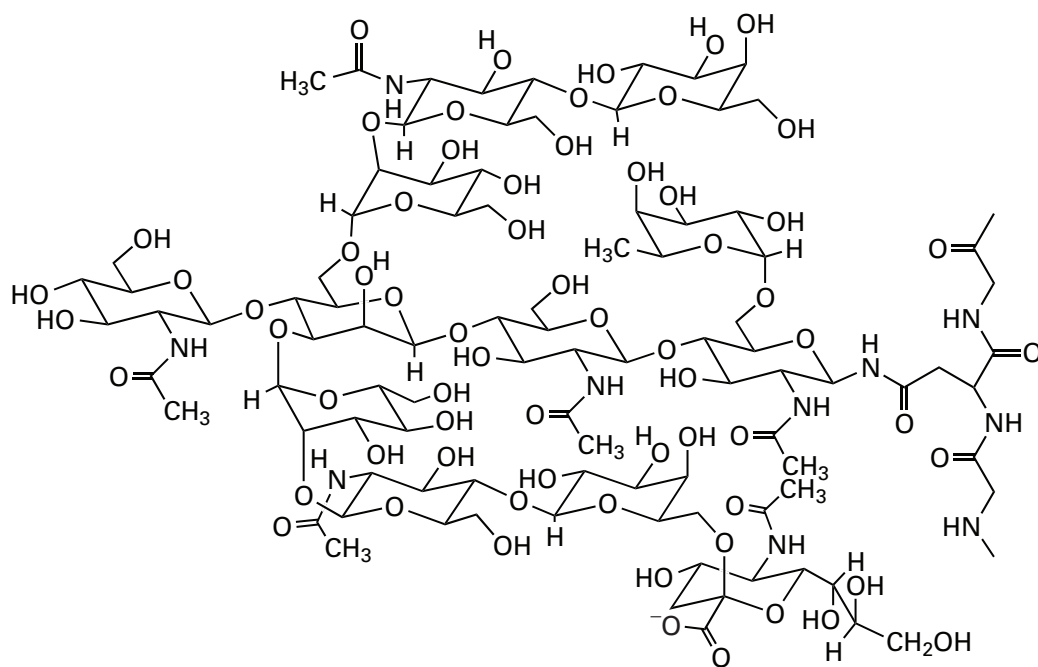


Figure 3-22: Covalent structure of one of the oligosaccharides attached through asparagine to human immunoglobulin D.⁴⁶⁹ This is an example of a complex *N*-linked oligosaccharide.

tosamine to serine and threonine side chains (Figures 3-21, 3-22, and 3-23).⁴⁶⁹⁻⁴⁷²

The monomers from which these oligosaccharides are constructed are **monosaccharides**. The eleven major monosaccharides that are found in the oligosaccharides of glycoproteins are D-mannose, D-galactose, D-glucose, N-acetyl-D-glucosamine, N-acetyl-D-galactosamine, the sialic acids, D-glucuronic acid, L-fucose, L-rhamnose, D-xylose, and L-arabinose (Figure 3-24). Several of the monosaccharides in glycoproteins can be O-sulfated,^{473,474} and mannoses and N-acetylglucosamines can be O-(2-aminoethyl)phosphonylated.⁴⁷⁵

A variety of different sialic acids are known (greater than 40) that are derivatives of either D-neuraminic acid (Figure 3-24) or the closely related D-5-deamino-5(S)-hydroxyneuraminic acid (2-keto-3-deoxy-D-glycero-D-galacto-noninic acid; 3-deoxy-D-glycero-D-galacto-nonulosonic acid).^{476,477} These two anionic monosaccharides are modified variously by N-acetylation, N-glycolylation, O-lactylation, O-sulfation, O-methylation, O-phosphorylation, and O-acetylation.⁴⁷⁸

The covalent bonds that link the monosaccharides together are those of acetals and occasionally ketals. **Glycosidic linkages** are the bonds in these acetals and ketals formed between the only carbonyl carbon on each monosaccharide, enclosed within a pyranose ring or a furanose ring as a hemiacetal, and one of the hydroxyl groups of the preceding monosaccharide in the oligomer. A glycosidic linkage is formed between the oxocarbenium cation of the pyranose or furanose and a lone pair of electrons on a nitrogen or an oxygen. An example would be

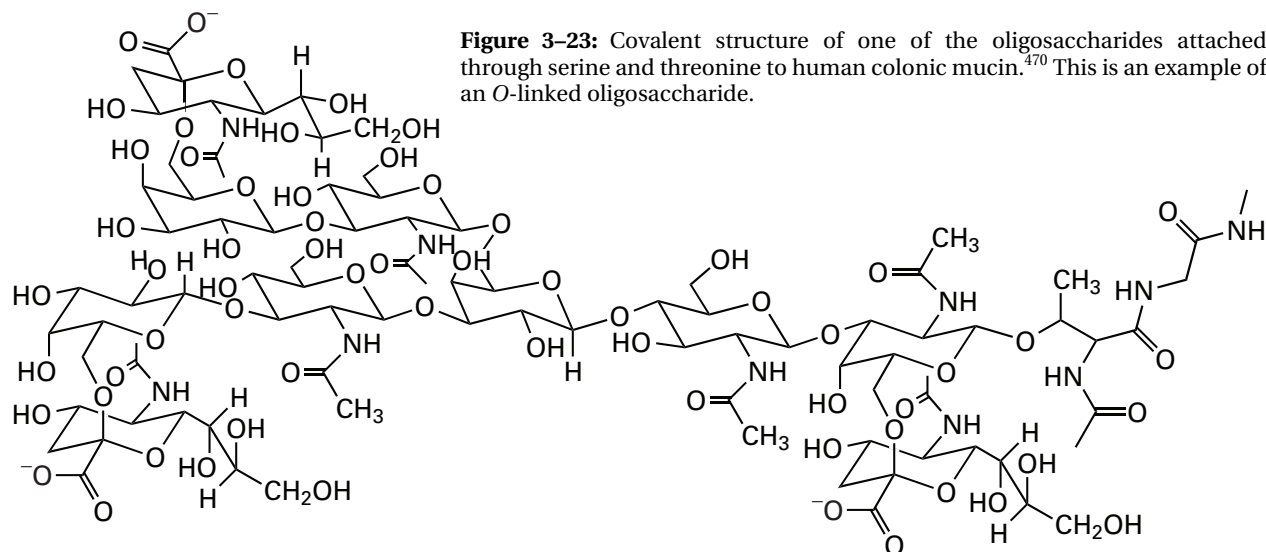
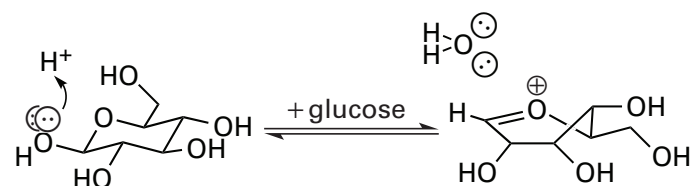
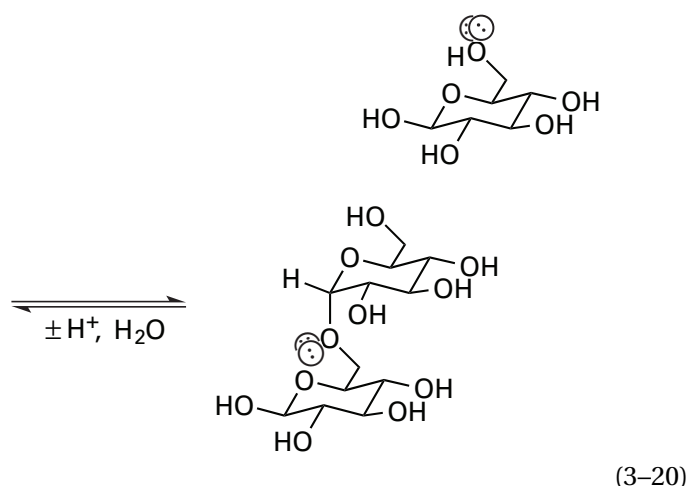


Figure 3-23: Covalent structure of one of the oligosaccharides attached through serine and threonine to human colonic mucin.⁴⁷⁰ This is an example of an O-linked oligosaccharide.



Branching of the oligosaccharide (Figures 3-21 to 3-23) occurs whenever two or more of the hydroxyl groups on one of the monosaccharides participate in glycosidic linkages.

Each of the oligosaccharides on a glycoprotein can be thought of as beginning at the monosaccharide that is attached to the polypeptide (the reducing end). The point of attachment is either an **O-glycosidic linkage**, formed between the carbonyl carbon of this initial monosaccharide and the hydroxyl group of a serine or threonine side chain, or an **N-glycosidic linkage**, formed between the carbonyl carbon of this initial monosaccharide and the amide nitrogen of an asparagine side chain. The first monosaccharide in an oligosaccharide attached to a serine or a threonine is usually N-acetylgalactosamine; the first sugar in an oligosaccharide attached to asparagine is almost always N-acetylglucosamine. Peripheral to this initial monomer, the oligomer will be found to branch at several points and end at each of several unsubstituted monosaccharides that occupy the last positions on the branches (the nonreducing ends). There are usually 2-8 monosaccharides counting from the initial monosaccharide to the end of a branch and

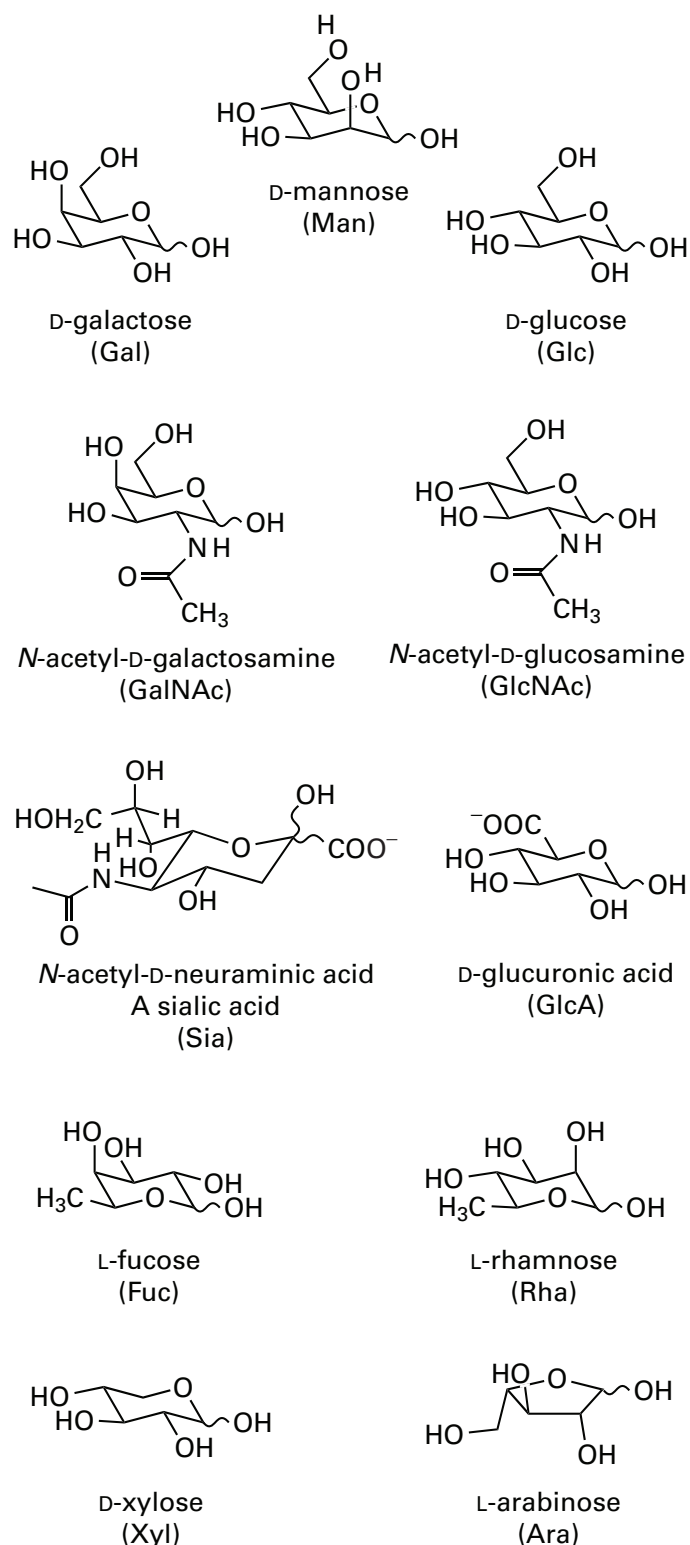


Figure 3-24: Eleven most common monosaccharides composing the oligosaccharides of glycoproteins.

1–4 branches in a typical oligosaccharide on a glycoprotein. Often a branch has only one monosaccharide.

Writing the **sequences** of the oligosaccharides on glycoproteins is complicated by the requirements that each monosaccharide must be noted, the particular

hydroxyl group in the preceding monosaccharide to which it is attached through its carbonyl carbon must be noted, and the anomeric state of its carbonyl carbon must be noted. It is usually assumed that unless otherwise stated the monosaccharides in the oligosaccharide are of the *D* stereochemistry, where, for example, L-fucose, L-rhamnose, and L-arabinose would be exceptions to be noted, and in the pyranose form, where, for example, galactofuranose and arabinofuranose would be exceptions to be noted. To confuse matters further, the sequences of oligosaccharides are usually written from right to left beginning at the right with the monosaccharide attached directly to the protein. The anomeric state of the carbonyl carbon and the hydroxyl group to which it is attached are noted to the right of the name of each monosaccharide. For example, GlcNAc(β 1,4) states that an *N*-acetylglucosamine is attached through its carbonyl carbon, carbon 1, to the 4-hydroxyl group of its immediate predecessor to the right in the written sequence by a β -anomeric acetal. In addition, any modifications of the monosaccharides, such as *O*-acetylation, *O*-sulfation, *O*-phosphorylation, de-*N*-acetylation, or *O*-methylation, must be noted. The actual structure of the oligosaccharide in Figure 3-23 can be compared to its written sequence (Table 3-2, entry 7).

There is yet a further peculiarity of oligosaccharides, that of **microheterogeneity**. Because serious biological problems do not seem to arise when unfinished oligosaccharides are produced, in contrast to the devastation that would occur if unfinished proteins and nucleic acids were produced, natural selection has not enforced uniformity on the synthesis of oligosaccharides. It may even be the case that the existing lack of uniformity has advantages for which natural selection has selected. Although a few finished glycoproteins are homogeneous, in most instances the synthesis of the oligosaccharides is an apparently haphazard stochastic process, and each oligosaccharide that ends up attached at a particular site in a glycoprotein is usually unfinished. Each, however, is unfinished in a different way. Each is missing a different set of monosaccharides. As a result, 10 or 15 different oligosaccharides may be found on the amino acid side chain at the same position in the amino acid sequence of different molecules of the same protein.

Once they have been separately identified and individually sequenced, each of these oligosaccharides can be recognized as a different, incomplete realization of only one complete sequence. This prototypical sequence is often longer than any one of the sequences of the actual oligosaccharides, but each of the sequences of the actual oligosaccharides is a piece of the prototype and every sugar in the prototype is represented in one of the actual oligosaccharides. Whether the prototype including all of the actual sequences is the most complete sequence that could have been produced or is itself only an incomplete realization of a longer sequence can never be decided unequivocally. As an example, 13 different

Table 3-2: Oligosaccharides Isolated from Human Colonic Mucin^{470,a}

(1)	Sia(α 2,6)GalNAc ^b
(2)	GlcNAc(β 1,3)GalNAc
(3)	GlcNAc(β 1,3)GalNAc Sia(α 2,6)
(4)	Gal(β 1,4)GlcNAc(β 1,3)GalNAc
(5)	Gal(β 1,4)GlcNAc(β 1,3)GalNAc Sia(α 2,6)
(6)	GlcNAc(β 1,3)Gal(β 1,4)GlcNAc(β 1,3)GalNAc
(7)	Sia(α 2,6)Gal(β 1,3)GlcNAc(β 1,3) Sia(α 2,6)Gal(β 1,3)GlcNAc(β 1,6)Gal(β 1,4)GlcNAc(β 1,3)GalNAc Sia(α 2,6)

^aOnly 7 of the 13 oligosaccharides isolated are tabulated. ^bFor abbreviations see Figure 3-27.

oligosaccharides were isolated from human colonic mucin.⁴⁷⁰ All of the other 12, of which six are presented in Table 3-2, were incomplete realizations of the largest (Table 3-2, entry 7). It is possible, however, that this largest one may be an incomplete realization of an even larger oligosaccharide that escaped detection.

Another view of microheterogeneity, in opposition to the view that it is haphazard and purposeless, is that it has a role in producing many different **glycoforms** of the same protein. This would increase the functional range of these proteins and would be advantageous in particular situations. For example, the microheterogeneity observed in the set of oligosaccharides isolated from human colonic mucin (Table 3-2) may permit the oligosaccharides on this glycoprotein to ensnare many different species of bacteria, each of which binds specifically to only one or a few oligosaccharide sequences. It is probably the case that microheterogeneity is relevant in some instances and irrelevant in others. For example, the length and amount of branching in the oligosaccharides on erythropoietin determines its biological activity,⁴⁷⁹ but the presence or absence of oligosaccharide on channel-forming intrinsic protein has no effect on its function.⁴⁸⁰ Unlike most proteins, most oligosaccharides do not assume a fixed conformation so the involvement of microheterogeneity in their biological specificity would be based mainly on differences in sequence. As the oligosaccharides, however, become more crowded⁴⁸¹ or more branched,⁴⁸² local steric effects become more numerous, and their confinement of the conformation of the oligosaccharide may contribute to differences in biological function.

From an examination of the sequences of the oligosaccharides attached to glycoproteins from animals and plants, several generalizations can be drawn. The

most common of these oligosaccharides can be divided into three classes (Table 3-3). The high-mannose oligosaccharides begin with two *N*-acetylglucosamines linked (β 1,4) to each other, the first attached to an asparagine in the protein. These oligosaccharides contain 5-9 mannoses (Figure 3-21). The complex *N*-linked oligosaccharides, because they are biosynthetically derived from the high-mannose oligosaccharides, also begin with two *N*-acetylglucosamines linked (β 1,4) to each other, followed by three branched mannose residues. Beyond this structural core, variable amounts of *N*-acetylglucosamine, galactose, fucose, various sialic acids, and occasionally *N*-acetylgalactosamine⁴⁸⁶ are attached (Figure 3-22). The *O*-linked oligosaccharides begin with an *N*-acetylgalactosamine linked to a serine or threonine on the protein and contain variable amounts of *N*-acetylglucosamine, galactose, *N*-acetylgalactosamine, fucose, and various sialic acids (Figure 3-23).

High-mannose oligosaccharides occur in all eukaryotes, but those in fungi have differences in linkage and branching patterns⁴⁸⁷ from those in plants and animals and often contain significantly more mannose.⁴⁸⁸ Most, if not all,⁴⁸⁹ of the high-mannose oligosaccharides from the proteins of plants and animals are incomplete realizations of one complete, basic structure (Table 3-3, entry 1).⁴⁷¹ This uniformity results from the fact that this unit is transferred in its entirety to the targeted asparagine side chain on the glycoprotein.⁴⁸⁴ Then, in a specific sequence of steps, catalyzed by three exomannosidases, it is shortened until all of the mannoses in (α 1,2) linkage have been removed. When only five mannoses remain, the oligosaccharide is then elongated in a highly specific sequence of steps by specific glycosyltransferases to produce complex *N*-linked oligosaccha-

Table 3–3: Representatives^{483–485} of the Three Main Classes of Oligosaccharides on Animal Glycoproteins

(1) High-Mannose Oligosaccharide
<div>Man(α1,2)Man(α1,6)</div> <div>Man(α1,2)Man(α1,3)Man(α1,6)Man(β1,4)GlcNAc(β1,4)GlcNAcβAsn^a</div> <div>Glu(α1,2)Glu(α1,3)Glu(α1,3)Man(α1,2)Man(α1,3)</div>
(2) Complex <i>N</i> -Linked Oligosaccharide
<div>Sia(α2,6)Gal(β1,4)GlcNAc(β1,6)</div> <div>Sia(α2,6)Gal(β1,4)GlcNAc(β1,2)Man(α1,6)</div> <div>GlcNAc(β1,4)Man(β1,4)GlcNAc(β1,4)GlcNAcβAsn^b</div> <div>Sia(α2,6)Gal(β1,4)GlcNAc(β1,2)Man(α1,3)</div> <div>Sia(α2,6)Gal(β1,4)GlcNAc(β1,4)</div>
(3) <i>O</i> -Linked Oligosaccharide
<div>Gal(β1,4)GlcNAc(β1,6)</div> <div>Fuc(α1,4)</div> <div>Gal(β1,3)GlcNAc(β1,3)Gal(β1,3)GalNAcSer/Thr^c</div> <div>Gal(β1,4)GlcNAc(β1,6)</div>

^aFrom Chinese hamster ovary cells. ^bFrom human plasma α₁-acid glycoprotein. ^cFrom blood group A active glycoprotein in human ovarian cyst fluid.

rides. After the last step of this elongation, other mannosidases remove two more mannoses to leave the three found in the mature complex *N*-linked oligosaccharide.

At the end of this process, most of these **complex *N*-linked oligosaccharides** are also incomplete realizations of one basic structure (Table 3–3, entry 2),^{471,485,490} but minor differences in the positions on the peripheral *N*-acetylglucosamines and galactoses at which the linkages are made have been noted^{491–493} as well as substitution of the peripheral galactoses with *N*-acetylgalactosamines.⁴⁹⁴ Short, repeating units of *N*-acetylglucosaminyl(β1,3)galactose have also been observed inserted between the peripheral galactoses and *N*-acetylglucosamines of some complex *N*-linked oligosaccharides.⁴⁹⁵ Fucoses are found attached to many of the complex *N*-linked oligosaccharides from animals in (α1,6) or (α1,3) linkage to one or the other of the *N*-acetylglucosamines in their cores⁴⁹⁶ or in (α1,3) or (α1,4) linkage to *N*-acetylglucosamines in their peripheries. Xyloses are found attached to a few of the complex *N*-linked oligosaccharides from animals⁴⁹⁷ but many of the complex *N*-linked oligosaccharides from plants⁴⁷¹ in (β1,2) linkage to the central mannose in the core.

It seems that, within the same protein, the oligosac-

charides on certain asparagine side chains will remain as high-mannose oligosaccharides exclusively, while the oligosaccharides on other asparagine side chains are completely converted to complex *N*-linked oligosaccharides.⁴⁶⁹ Occasionally, however, a **hybrid *N*-glycan** is encountered, in which one of the branches in one of these oligosaccharides is of the complex structure while the other remains of the high-mannose structure,⁴⁹⁴ presumably because the processing on the latter branch was specifically blocked.

The ***O*-linked oligosaccharides** display less uniformity than the *N*-linked. This may result from the fact that they are built up one sugar at a time rather than as intact units.⁴⁹⁸ The *O*-linked oligosaccharides drawn in Figure 3–23 and presented in Table 3–3 include some of the common structural features of this class. By far the most common monosaccharide forming the linkage to the serine or threonine is *N*-acetylgalactosamine, but oligosaccharides *O*-linked through other monosaccharides have been reported.^{499,500} The branches are constructed from the basic repeating unit, Gal(β1,3 or β1,4)GlcNAc(β1,3 or β1,4 or β1,6). Branching usually occurs at a galactose or at the initial *N*-acetylgalactosamine, rarely if ever at an *N*-acetylglucosamine. The

basic repeating unit of each branch can begin with either an *N*-acetylglucosamine (Figure 3–23) or a galactose (Table 3–3). Fucose is found in (α 1,4) and (α 1,3) linkages to penultimate *N*-acetylglucosamines in addition to (α 1,2) linkage to peripheral galactoses. The branches either end with a galactose of the repeating unit or are capped by an *N*-acetylgalactosamine in (α 1,3) or (α 1,4) linkage. Sialic acids are found in (α 2,6) or (α 2,3) linkage to galactoses or the initial *N*-acetylgalactosamine. Many variations on these patterns are observed.^{470,483,501–508} Often *O*-linked oligosaccharides are quite short. An example would be NeuNAc(α 2,3)Gal(β 1,3)[NeuNAc(α 2,6)]GalNAc.^{501,503} All these regularities seem to result from the fact that the sugars are added one at a time from the initial *N*-acetylgalactosamine outward by a limited set of glycosyltransferases. These enzymes are specific for particular sugars and attach them only to particular hydroxyl groups on particular sugars within the growing oligosaccharide.

Two of the most heavily glycosylated glycoproteins found in animals are the mucins and the proteoglycans. These two types of glycoproteins can contain up to 80% or 90% carbohydrate by mass, respectively.

The **mucins** are the glycoproteins that constitute mucus and also coat the surfaces of many types of cells.

The human intestinal mucin MUC2 is a polypeptide 5159 aa long.² Between Cysteine 1375 and Cysteine 1762 and between Cysteine 1858 and Isoleucine 4299, there are two regions of amino acid sequence that are rich in threonine (58% of the amino acids) and proline (24%) and are thought to contain the majority if not all of the sites for the *O*-linked glycosylation (Table 3–2), which occurs mainly on threonine.⁵⁰⁹ The larger of these two regions is made up almost exclusively of 101 consecutive repeats of the sequence –TTTTTVTPPTPTGTQTPTTTPI– with only a few substitutions over the entire length of 2323 aa. There are about 1100 *N*-acetylgalactosylthreonyl linkages in the entire protein,⁵⁰⁹ and if all of these are confined to the two regions rich in threonine, about 85% of the threonines in these regions carry oligosaccharides. From an examination of the repeating sequence and the fact that each oligosaccharide contains an average of four monosaccharides,⁵⁰⁹ one can gain an appreciation of how closely packed these oligosaccharides must be. Other mucins also have similar regions rich in threonine and serine, usually found in repeating sequences^{510–512} that are also heavily glycosylated.

Proteoglycans are proteins to which particular types of regular polysaccharides are attached. Proteoglycans are secreted as extracellular matrix and

Table 3–4: Polysaccharides of Proteoglycans^a

type	original repeating unit ^a	postsynthetic modification
chondroitin sulfate	(β 1,4)GlcA(β 1,3)GalNAc(β 1,4)	GalNAc-6-SO ₃ [–] GalNAc-4-SO ₃ [–]
dermatan sulfate ⁵¹³	(β 1,4)GlcA(α 1,3)GalNAc(β 1,4) (β 1,4)GlcA(β 1,3)GalNAc(β 1,4)	GalNAc-4-SO ₃ [–] GlcA \rightarrow IdoA ^b IdoA-2-SO ₃ [–]
heparin	(α 1,4)GlcA(α 1,4)GlcNAc(α 1,4)	deacetylation ^c <i>N</i> -sulfation ^c GlcNSO ₃ [–] -6-SO ₃ [–] GlcNAc-6-SO ₃ [–] glucosamine-6-SO ₃ [–] GlcA-2-SO ₃ [–] GlcNAc-3-SO ₃ [–]
heparan sulfate ⁵¹⁴	(α 1,4)GlcA(α 1,4)GlcNAc(α 1,4)	deacetylation ^c <i>N</i> -sulfation ^c GlcA \rightarrow IdoA ^d IdoA-2-SO ₃ [–] GlcNSO ₃ [–] -6-SO ₃ [–] GlcNAc-6-SO ₃ [–] glucosamine-6-SO ₃ [–]
keratan sulfate	(β 1,3)Gal(β 1,3)GlcNAc(β 1,3)	GlcNAc-6-SO ₃ [–] Gal-6-SO ₃ [–]

^aThis is the disaccharide monomer constituting the newly synthesized polymer before postsynthetic modification.

^bEpimerization of glucuronic acid at carbon 5 to form iduronic acid, the presence of which distinguishes dermatan sulfate from chondroitin sulfate. ^cDeacetylation and *N*-sulfation (–NSO₃[–]) are both incomplete so that *N*-acetylglucosamine, glucosamine, and *N*-sulfoglucosamine coexist in the same proteoglycan. ^dEpimerization of glucuronic acid at carbon 5 to form iduronic acid, the presence of which distinguishes heparan sulfate from heparin.

are the main constituents in such structures as cartilage, vascular wall, and tendon. Although they often carry the usual *N*-linked and *O*-linked oligosaccharides, by definition, they also carry at least one of a class of long polysaccharides formed from repeating disaccharides (Table 3–4). Each proteoglycan is defined by the repeating disaccharide that forms the polysaccharide that is attached to it. These polysaccharides of repeating units are heterogeneous because of a collection of postsynthetic modifications (Table 3–4) that are only partially accomplished, often concentrated within randomly spaced blocks of consecutive monosaccharides along the length of the polymer. One constant feature of the covalent structure of a proteoglycan is that each of its defining polysaccharides, except for keratan sulfate,⁵¹⁵ is *O*-linked to the protein through the oligosaccharide –GlcA(β 1,3)Gal(β 1,3)Gal(β 1,4)xylosylserine.⁵¹⁶

As with those of the mucins, the polypeptides of proteoglycans can be quite long. The polypeptide of one of the human chondroitin sulfate proteoglycans is 2293 aa in length⁵¹⁷ and has on average, 12,000 monosaccharides in its covalently attached oligosaccharides and polysaccharides.⁵¹⁸ Unlike the mucins, which contain short oligosaccharides densely packed together because they are on long strings of adjacent threonines, the proteoglycans contain long polysaccharides that can be attached to serines at isolated –Gly-Ser– or –Ser-Gly– sites scattered randomly over the sequence at intervals of about 50 aa.^{517,519} In at least one of the proteoglycans, however, there is the sequence –YS(GS)₂₄L–, to the serines of which heparin and chondroitin sulfate are attached.⁵²⁰

The **sequence of the monosaccharides** in an oligosaccharide or polysaccharide on a glycoprotein is established by chemical analysis. The starting material in this analysis is a purified preparation of the glycoprotein itself. Often, to facilitate the analysis, the oligosaccharides on the glycoprotein have been made radioactive by growing cells producing it in the presence of one or two radioactive monosaccharides, for example, [³H]mannose and [¹⁴C]glucosamine.⁵²¹ Oligosaccharides attached to a glycoprotein are isolated by digesting the protein with endopeptidases, purifying the resulting glycopeptides, and releasing the oligosaccharides from these glycopeptides by chemical or enzymatic cleavage. The **glycopeptides** produced by the digestion are usually separated on a chromatographic system, such as chromatography by reverse-phase adsorption, to separate them on the basis of only their amino acid sequence. In this way all of the oligosaccharides attached to a particular amino acid side chain in the sequence of the glycoprotein are isolated together.⁵²² From an examination of the amino acid sequences of a large number of such glycopeptides, it has been concluded that *N*-linked oligosaccharides from plants and animals are always attached to asparagines that have either a serine or a threonine two amino acids further on in the amino acid sequence (Asn-X-

Ser/Thr).^{471,523} The serines and threonines to which *O*-linked oligosaccharides are attached, however, are evidently not designated by any pattern in the surrounding sequence of amino acids⁵⁰¹ but tend to be clustered in regions of the polypeptide rich in serines, threonines, and prolines. The ultimate example of this would be the mucin MUC2.

The chemical or enzymatic cleavage used to release the several microscopically heterogeneous oligosaccharides from a particular glycopeptide depends upon the glycosidic linkage. For *N*-linked oligosaccharides, **endo-glycosidases** specific for cleavage within the common segment GlcNAc(β 1,4)GlcNAcAsn are usually used. For example, mannosyl-glycoprotein endo- β -*N*-acetylglucosaminidase (endoglycosidase H) catalyzes hydrolysis of the glycosidic linkage between two *N*-acetylglucosamines and releases the oligosaccharide missing its initial monosaccharide, while peptide-*N*⁴-(*N*-acetyl- β -glucosaminy)asparagine amidase (peptide-*N*-glycosidase F) cleaves the *N*-glycosidic linkage between an *N*-linked oligosaccharide and the asparagine on a glycoprotein or glycopeptide.⁵²⁴ Oligosaccharides in *N*-glycosidic linkage to asparagine can also be released from the glycopeptide by hydrazinolysis⁵²⁵ and reacylated with acetic anhydride. Regardless of the method by which it is released, the aldehyde at C1 of the initial *N*-acetylglucosamine in the oligosaccharide is usually reduced to the primary alcohol with Na[³H]BH₄ (Figure 3–25).⁴⁶⁹ This reduction eliminates the aldehyde, simplifies the subsequent chemistry, and makes the oligosaccharide radioactive, if it is not so already. Oligosaccharides in *O*-glycosidic linkage to serine and threonine are usually released from the glycopeptides by treatment with base, which promotes **β -elimination** (Figure 3–25). The treatment with base is performed in the presence of Na[³H]BH₄ to prevent, by reduction of the aldehyde at C1, the destruction of the oligosaccharide from its reducing end and to make the released oligosaccharide radioactive.

It is at this point that the technical consequences of microheterogeneity are experienced. Instead of one pure oligosaccharide released in a quantity equimolar to the amount of glycopeptide, a mixture of many oligosaccharides is produced, each present in a correspondingly small quantity. This mixture is first separated into neutral and anionic oligosaccharides chromatographically.⁴⁹¹ **Chromatographic systems** that separate the oligosaccharides by molecular exclusion or by anion exchange⁵²⁶ are then used to perform further separations. Chromatography by molecular exclusion provides an indication of the size of each oligosaccharide in the set. After the sialic acids have been removed from the anionic oligosaccharides by hydrolysis in mild acid and separately analyzed, the composition of each oligosaccharide is determined. This can be done by methanolysis under acidic conditions to cleave the acetals and coincidentally form methyl glycosides (Figure 3–26) that are

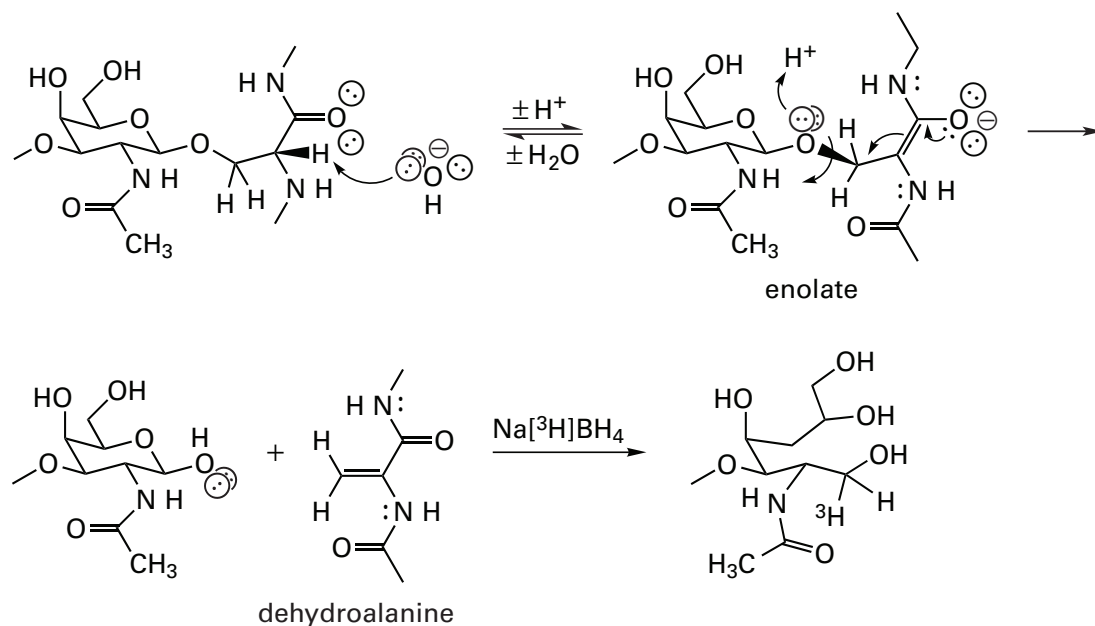


Figure 3-25: β -Elimination of an *N*-acetylgalactosamino oligosaccharide from serine by strong base. The initial step in the reaction is the removal of the proton α to the amide to produce the 1,2-diaminoenolate, which in turn ejects the alcohol to form a dehydroalanine. The reducing end of the oligosaccharide is then reductively labeled with $\text{Na}[^3\text{H}]\text{BH}_4$.

identified by their mobilities on gas chromatography. It is also possible to analyze the composition of the oligosaccharide directly by submitting it to hydrolysis in acid and separating the resulting monosaccharides and deacetylated amino sugars on chromatography by anion exchange (Figure 3-27),^{480,527} the effluent from which is monitored electrochemically.⁵²⁸

The sequence of each of the purified oligosaccharides is determined by indirection. A series of chemical and enzymatic reactions is performed on the oligosaccharide and the outcome of each of these reactions is assessed either directly or by determining the change in composition of the oligosaccharide that occurs. The results of these various reactions are gathered until in their entirety they are consistent with only one of the many possible structures for the oligosaccharide. This one structure is then considered to be the actual structure. The reactions used in this process are periodate oxidation, Smith degradation, treatment with glycosidases, and methylation. The results from these chemical analyses are often supplemented with nuclear magnetic resonance spectroscopy and mass spectrometry.

When sodium metaperiodate (NaIO_4) is dissolved in water at acidic pH (pH 3–6) it forms a mixture of acidic hydrates referred to as periodic acid (HIO_4). **Periodic acid** cleaves polyalcohols such as monosaccharides at the carbon–carbon bonds between vicinal diols and produces two carbonyls from the two hydroxyl groups (Figure 3-28). Both of the hydroxyl groups in the vicinal diol must be free for periodic acid to cleave the carbon–carbon bond between them. The disappearance of a monosaccharide during treatment with periodic acid demonstrates that, in the intact oligosaccharide, the sugar that disappeared had at least two adjacent hydroxyl groups unbonded in glycosidic linkages. It is also possible to identify the actual products of the perio-

date cleavage by mass spectrometry.²²⁸ Oxidation by periodic acid can be performed sequentially by the Smith degradation (Figure 3-28). This series of reactions takes advantage of the lability to acid of a glycosidic linkage at carbon 1 of a sugar that has been cleaved by periodic acid and the resulting aldehydes of which have been reduced with sodium borohydride. In theory this reaction should be able to cleave sugars sequentially from the ends of the branches inward, but in practice only one cycle is usually successful because the selectivity for acyclic acetals is not great.

A more informative sequence of cleavages can often be performed with **exoglycosidases**. These are enzymes that remove particular sugars from the ends of branches. They are highly specific for the sugar removed, the anomeric state of the glycosidic linkage, and sometimes the location of the hydroxyl group from which the bond has been formed. Examples of such exoglycosidases would be β -galactosidase, α -L-fucosidase, β -*N*-acetylglucosamidase, and exo- α -sialidase. An example of the specificity for the hydroxyl group would be the exo- α -2,3-sialidase from Newcastle disease virus. The release of a monosaccharide after exposure of the oligosaccharide to an exoglycosidase is evidence that that monosaccharide was at the end of a branch and attached to it by a glycosidic linkage of the designated anomeric stereochemistry. The digestions are usually performed sequentially. After each of the monosaccharides at the ends of the branches has been catalogued, each of the shortened products of the first round of digestions is then submitted to a round of digestion to identify the penultimate monosaccharides on each branch and so on until the last sugar is released. The products of each round of digestion are monitored either chromatographically⁵²⁹ or by mass spectrometry.⁵³⁰ Several specific endoglycosidases, which cleave an oligosaccharide internally at particular

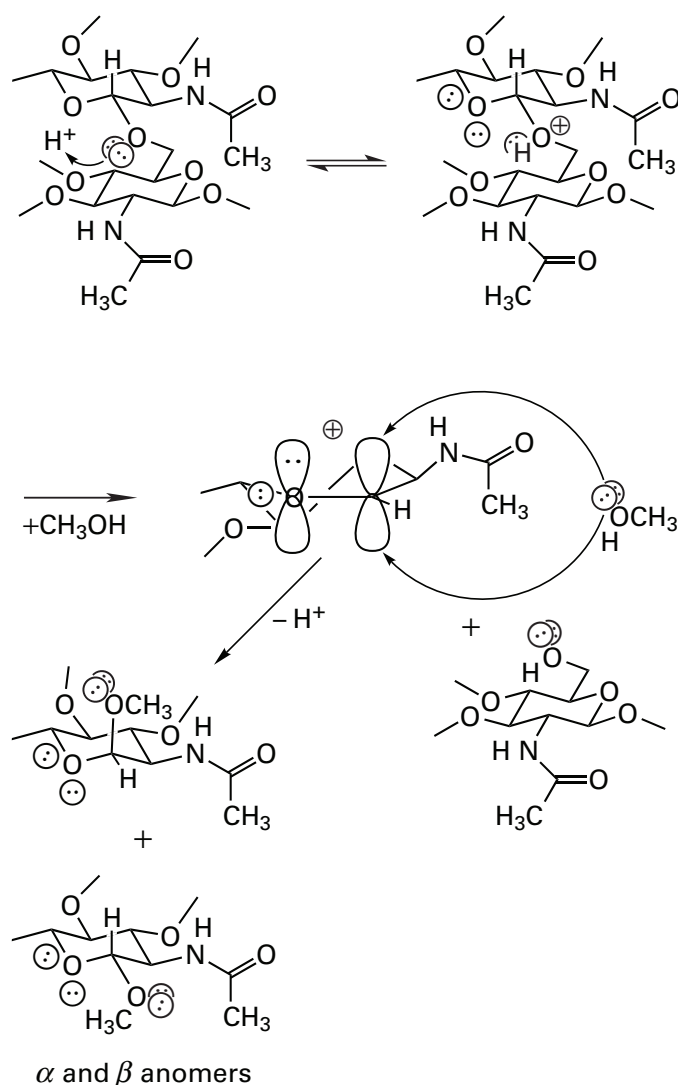


Figure 3-26: Acidic methanolysis of oligosaccharides. Protonation of the exocyclic acetal oxygen produces a leaving group, the departure of which gives the planar oxacarbenium cation. Addition of methanol to either face of the oxacarbenium cation produces a mixture of the α - and β -anomers of the methyl glycoside.

bonds with high specificity, are also available. Examples of these enzymes are endo-1,4- β -galactosidase and endo- α -sialidase. Many of the complementary DNAs isolated from the original sources of these glycosidases have been transferred to expression vectors, and the proteins are expressed in high yield by transfected bacteria. One advantage to these expression systems is that these enzymes purified from recombinant bacteria are uncontaminated by other glycosidases.

An oligosaccharide can be **chemically methylated** on all of its free hydroxyl groups. This is done by forming the sodium alkoxides of the hydroxyl groups in a solution of dimethyl sulfoxide by using the sodium salt of the dimethyl sulfoxide anion as the base. The alkoxides are then methylated with methyl iodide.⁵³¹ The methylated oligosaccharide is then hydrolyzed in acid, and the resulting monosaccharides are reduced with NaBH_4 . The

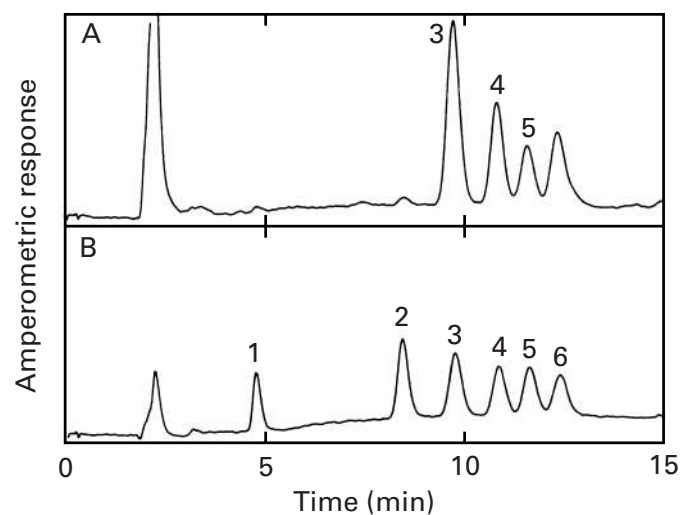


Figure 3-27: Chromatographic analysis of the hydrolysate of a glycopeptide to quantify its composition of monosaccharides.⁵²⁷ A sample (300 pmol) of a purified, homogeneous glycopeptide (with the composition $\text{Gal}_3\text{GlcNAc}_5\text{Man}_3$) was hydrolyzed for 4 h at 100°C in 4 M trifluoroacetic acid. The acid was removed by evaporation, and the hydrolysate was submitted to chromatography (panel A) on a column ($0.46\text{ cm} \times 25\text{ cm}$) of a medium for anion exchange equilibrated and eluted with 22 mM NaOH. The strong base makes the sugars sufficiently anionic to be separated by the chromatographic medium. The concentration of monosaccharide was monitored continuously with a pulsed amperometric detector (PAD). The detector responds to the current resulting from the uptake of electrons at a gold electrode (PAD response). The electrode is poised at +50 mV, which is a sufficiently positive potential to oxidize the polyols of the monosaccharides. It is this oxidation at the surface of the electrode that produces the current. Standards (25 pmol) were run (panel B) under the same conditions and separately identified as the following monosaccharides: 1, fucose; 2, galactosamine; 3, glucosamine; 4, galactose; 5, glucose; 6, mannose. From the areas of the peaks of the standards it could be calculated that the original hydrolysate contained 1.1 nmol of glucosamine, 0.79 nmol of galactose, and 0.75 nmol of mannose. The glucose observed was a contaminant. Reprinted with permission from ref 527. Copyright 1988 Academic Press, Inc.

resulting alditols are acetylated both at the hydroxyl groups produced by the hydrolysis of the glycosidic linkages and at the hydroxyl groups produced by the reduction of the aldehydes. The various methylated alditol acetates that result from this treatment are identified chromatographically. In this way, the hydroxyl groups at which the various monosaccharides were bonded in the glycosidic linkages of the original oligosaccharide can be identified because they are acetylated rather than methylated in the products.⁴⁹⁴ For example, upon methylation, the oligosaccharide drawn in Figure 3-21 yielded 1,5-diacetyl-2,3,4,6-tetramethylmannitol, 1,2,5-triacetyl-3,4,6-trimethylmannitol, 2,4-dimethyl-1,3,5,6-tetraacetylmannitol, and smaller amounts of 1,3,5-triacetyl-2,4,6-trimethylmannitol and 3,6-dimethyl-1,4,5-triacetyl-N-acetylglucosaminitol.⁴⁶⁹ The appearance of each of these products is consistent with the structure ultimately proposed for the intact oligosaccharide.

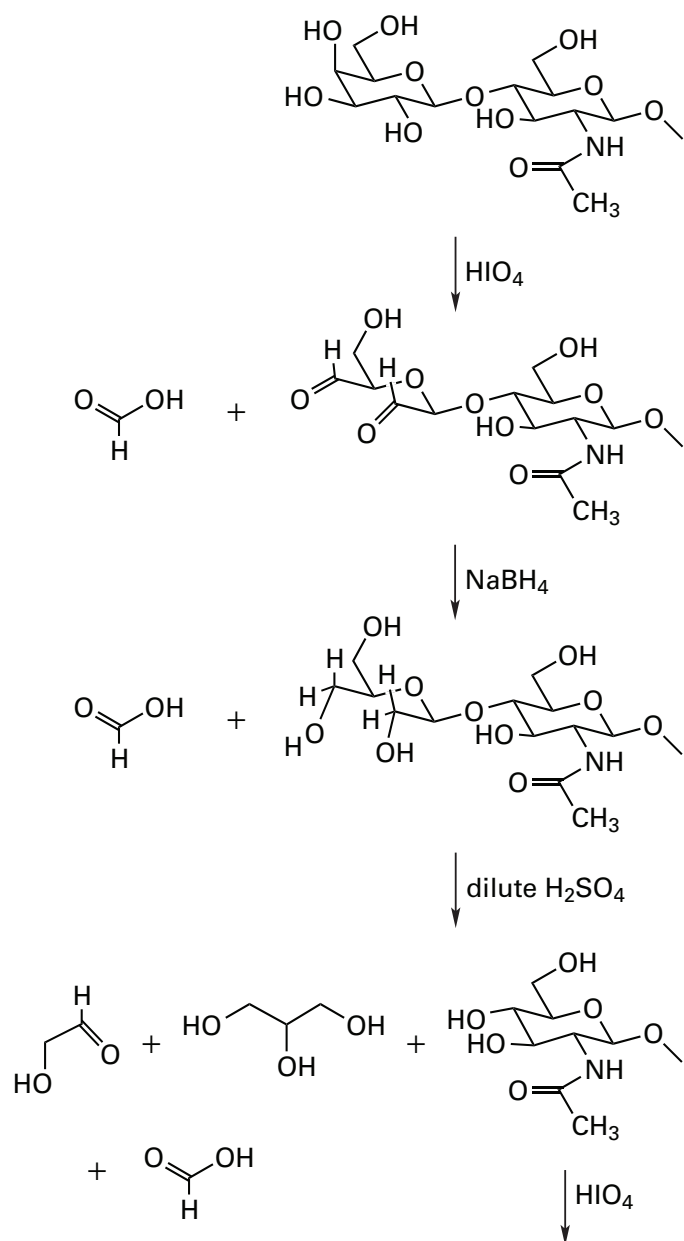


Figure 3-28: Periodate oxidation and Smith degradation. Periodate oxidation (HIO_4) cleaves the carbon-carbon bond between any two unbonded vicinal hydroxyl groups. The products are most readily understood by first putting a hydroxyl group at each position on a carbon that was involved in the carbon-carbon bond and then turning hydrates to aldehydes and orthoacids to acids. Following periodate oxidation, the aldehydes are reduced with borohydride anion, and the acetals containing the degraded monosaccharides are cleaved selectively in weak acid. This frees hydroxyl groups that were previously in glycosidic linkages and makes certain monosaccharides, which were resistant before, now susceptible to periodate oxidation.

Although it is limited by the amounts ($0.5 \mu\text{mol}$) of oligosaccharide needed, **nuclear magnetic resonance spectroscopy** has been applied to solving the sequence of oligosaccharides. The analysis has been most successful in cases where the oligosaccharide is a member of a class, such as high-mannose or complex oligosaccharides, the

structures of which are predictable and for which many well-characterized standards are available to assist in the assignments of the various resonances.^{485,489,491,507,508,532,533}

In at least one instance, however, the structures of a nested set of oligosaccharides of increasing length, from a less well-characterized class of oligosaccharides, were determined entirely by nuclear magnetic resonance spectroscopy.⁵⁰² If a standard oligosaccharide of exactly the same structure as the oligosaccharide isolated from the glycoprotein is available, the coincidence of the nuclear magnetic resonance spectrum of the standard and that of the unknown is proof of the structure of the unknown.⁵³⁴ In the nuclear magnetic resonance spectrum of an oligosaccharide, the chemical shift for the resonance of each of the various hydrogens attached to the carbons of the monosaccharides is characteristic of the monosaccharide itself, the carbon it is attached to, and whether or not the hydroxyl group on that carbon is glycosidically linked.⁵⁰⁶ Two-dimensional spectra are used to assign the set of resonances from hydrogens on the same monosaccharide.^{506,535}

Mass spectrometry has also been applied to structural studies of oligosaccharides and glycopeptides. Oligosaccharides and glycopeptides can be transferred to the gas phase as ionic molecules by fast-atom bombardment,⁵³⁶ electrospray,⁵³⁷ or matrix-assisted-laser-desorption ionization.⁵³⁸ A mass spectrometer cannot distinguish mannose, galactose, and glucose from each other, nor *N*-acetylglucosamine from *N*-acetylgalactosamine. When used without collision-induced dissociation, it can provide information only about the number of hexoses, *N*-(acetylamino)hexoses, and sialic acids present in a given glycopeptide⁵³⁹ because the molecular mass of an oligosaccharide is the same regardless of how the monosaccharides are connected and which epimers are present. Because of the unusual molecular mass of fucose, oligosaccharides containing different amounts of fucose can be distinguished by mass spectrometry.⁵⁴⁰ When mass spectrometry is combined with chemical modifications such as methylation, more information can be gathered by using mass spectra to analyze the products of the reactions.⁵⁴¹ One significant advantage of mass spectrometry is that mixtures of glycopeptides or oligosaccharides can be analyzed because each component in the mixture produces a different molecular ion.⁵⁴²

When a tandem mass spectrometer is used with an intermediate step of collision-induced dissociation, the molecular ion of the oligosaccharide can be selected by the first mass spectrometer, and the fragment ions can be registered by the second.⁵³⁷ In this way a clean sequence of fragments, each missing an additional monosaccharide, can be observed. The sequence in which the fragmentation occurs can provide information about the sequence of monosaccharides in the oligosaccharide, but because the monomeric units are usually not arranged linearly in an oligosaccharide but as branches, the order in which hexoses and *N*-acetylhexosamines are lost upon

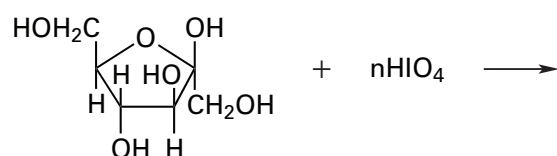
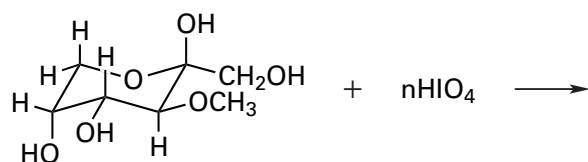
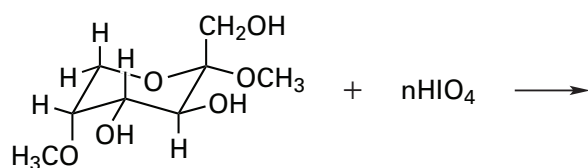
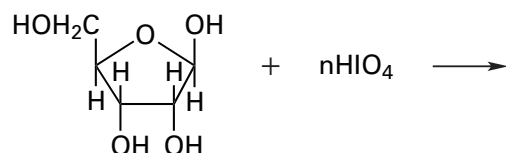
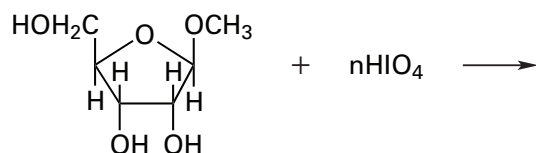
fragmentation is not sufficient to define the sequence in which they occur in the oligosaccharide. In the case of the linear, unbranched oligosaccharides of proteoglycans (Table 3-4), however, because the masses of glucuronic acid and iduronic acid differ from those of *N*-acetylgalactosamic and *N*-acetylglucosamine and because the various postsynthetic modifications change the masses of the monosaccharides, spectrometry provides significant information about sequence.⁵³⁸

Suggested Reading

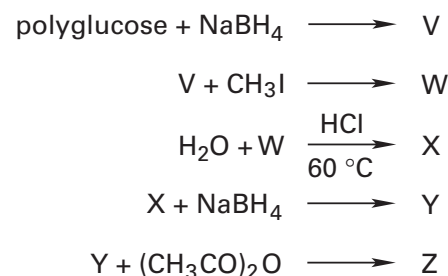
Baenziger, J.U., & Fiete, D. (1979) Structure of the complex oligosaccharides of fetuin, *J. Biol. Chem.* 254, 789-795.

van Kuik, J.A., de Waard, P., Vliegthart, J.F.G., Klein, A., Carnoy, C., Lamblin, G., & Roussel, P. (1991) Isolation and structural characterization of novel neutral oligosaccharide-alditols from respiratory-mucus glycoproteins of a patient suffering from bronchiectasis 2. Structure of twelve hepta-to-nonasaccharides, six of which possess the GlcNAc $\beta(1 \rightarrow 3)[\text{Gal } \beta(1 \rightarrow 4)\text{GlcNAc } \beta(1 \rightarrow 6)]\text{Gal } \beta(1 \rightarrow 3)\text{GalNAc-ol}$ common structural element, *Eur. J. Biochem.* 198, 169-182.

Problem 3-18: Complete the following reactions:



Problem 3-19: A polysaccharide, which is a polymer of glucose only, is treated in the following way:



Z is a mixture containing the following distribution of methylated glucitol acetates.

methylated glucitol acetate	mole percent
1,4,5-triacetyl-2,3,6-trimethylglucitol	81.9
2,3-dimethyl-1,4,5,6-tetraacetylglucitol	9.0
1,5-diacetyl-2,3,4,6-tetramethylglucitol	8.8
4-acetyl-1,2,3,5,6-pentamethylglucitol	0.2

- On average, how many monosaccharides does each molecule of the polysaccharide contain, how many branch points are there, and how many nonreducing ends are there?
- Draw structures of the linkages in the main linear polymer and the structure of a branch point.
- If the polysaccharide were treated with periodic acid, what percentage of the glucose would be destroyed?

Problem 3-20: A glycopeptide has been isolated following exhaustive pronase digestion of phytohemagglutinin from lima beans.⁵⁴³ Determine its structure from the following information. The compositions of single, intact, homogeneous glycopeptides or oligosaccharides are enclosed within parentheses.

- Composition: (mannose₄, *N*-acetylglucosamine₂, Asp)
- Exhaustive methylation, acid hydrolysis, reduction, and acetylation

methylated acetylated sugar alcohol	mol (mol of dimethyltetraacetylmannitol) ⁻¹
1,5-diacetyl-2,3,4,6-tetramethylmannitol	1.95
1,2,5-triacetyl-3,4,6-trimethylmannitol	1.10
2,4-dimethyl-1,3,5,6-tetraacetylmannitol	1.00
3,6-dimethyl-1,4,5-triacetyl- <i>N</i> -acetylglucosaminitol	1.90

- Periodate oxidation followed by mild acid hydrolysis yields a smaller glycopeptide and no free sugar of any kind. The composition of the smaller

- glycopeptide is (mannose₁, *N*-acetylglucosamine₂, Asp)
- (D) Mannosidase treatments of initial glycopeptide produced

mannosidase	mol of mannose released (mol of glycopeptide) ⁻¹
α -mannosidase (<i>Arthrobacter</i> GJM-1)	0.9
(α 1,2)-mannosidase (<i>Aspergillus niger</i>)	1.1
α -mannosidase (jack bean)	3.0
α -mannosidase (jack bean) followed by β -mannosidase (<i>A. niger</i>)	3.8
β -mannosidase (<i>A. niger</i>) alone	< 0.2

- (E) Glycopeptide core remaining after digestion with α -mannosidase (jack bean) and β -mannosidase (*A. niger*) was GlcNAc(β 1,4)GlcNAc-Asn.
- (F) The glycopeptide remaining after α -mannosidase (*Arthrobacter* GJM-1) digestion was isolated, exhaustively methylated, and hydrolyzed, and the resulting methylhexoses were reduced and acetylated. 1,5,6-Triacetyl-2,3,4-trimethylmannitol, 1,2,5-triacetyl-3,4,6-trimethylmannitol, and 1,5-diacetyl-2,3,4,6-tetramethylmannitol were obtained in approximately equal amounts.

Draw a structure for this glycopeptide that is consistent with all of these observations.

Problem 3-21: A glycopeptide has been purified from thyroglobulin⁵⁴⁴ that had been digested with pronase. From the following information, determine its complete structure. Draw the linkage to the amino acid side chain in the peptide portion. The compositions of single, intact, homogeneous glycopeptides or oligosaccharides are enclosed within parentheses.

- (A) Composition
(Asp, Gly, Val, mannose₃, acetate₂, glucosamine₂)
- (B) Pronase + α -mannosidase
Gly
Val
3 mannose
(Asp, glucosamine₂, acetate₂)
- (C) Chick oviduct extract
(mannose₃, glucosamine₂, acetate₂)
(Asp, Gly, Val)
- (D) Exhaustive methylation, acid hydrolysis, reduction, and acetylation (amounts not determined)
3,6-dimethyl-1,4,5-triacetyl-*N*-acetylglucosaminitol
1,5-diacetyl-2,3,4,6-tetramethylmannitol
2,X-dimethyl-1,Y,5,6-tetraacetylmannitol (X = 3 or 4; Y = 4 or 3)
- (E) Periodate oxidation, acid hydrolysis
(Asp, Gly, Val, mannose₁, glucosamine₂, acetate₂)

- (F) Reduction followed by acid hydrolysis of oligosaccharide from step C
3 mannose
1 glucosamine
1 glucosaminitol
2 acetate
- (G) α -Mannosidase treatment of oligosaccharide from step C
3 mannose
[di-*N*-acetylglucosaminyl(β 1,4)glucosamine]*

References

- Cladaras, C., Hadzopoulou-Cladaras, M., Nolte, R.T., Atkinson, D., & Zannis, V.I. (1986) *EMBO J.* 5, 3495–3507.
- Gum, J.R., Jr., Hicks, J.W., Toribara, N.W., Siddiki, B., & Kim, Y.S. (1994) *J. Biol. Chem.* 269, 2440–2446.
- Labeit, S., & Kolmerer, B. (1995) *Science* 270, 293–296.
- Gutte, B., & Merrifield, R.B. (1971) *J. Biol. Chem.* 246, 1922–1941.
- Nutt, R.F., Brady, S.F., Darke, P.L., Ciccarone, T.M., Colton, C.D., Nutt, E.M., Rodkey, J.A., Bennett, C.D., Waxman, L.H., Sigal, I.S., et al. (1988) *Proc. Natl. Acad. Sci. U.S.A.* 85, 7129–7133.
- Edman, P. (1953) *Acta Chem. Scand.* 7, 700–701.
- Hewick, R.M., Hunkapillar, M.W., Hood, L.E., & Dreyer, W.J. (1981) *J. Biol. Chem.* 256, 7990–7997.
- Smyth, D.G., Stein, W.H., & Moore, S. (1962) *J. Biol. Chem.* 237, 1845–1850.
- Bailey, J.M., & Shively, J.E. (1990) *Biochemistry* 29, 3145–3156.
- Doolittle, L.R., Mross, G.A., Fothergill, L.A., & Doolittle, R.F. (1977) *Anal. Biochem.* 78, 491–505.
- Tarr, G.E., Beecher, J.F., Bell, M., & McKean, D.J. (1978) *Anal. Biochem.* 84, 622–627.
- Matsudaira, P. (1987) *J. Biol. Chem.* 262, 10035–10038.
- Vandekerckhove, J., Bauw, G., Puype, M., Van Damme, J., & Van Montagu, M. (1985) *Eur. J. Biochem.* 152, 9–19.
- Moos, M., Jr., Nguyen, N.Y., & Liu, T.Y. (1988) *J. Biol. Chem.* 263, 6005–6008.
- Hirs, C.H.W., Moore, S., & Stein, W.H. (1956) *J. Biol. Chem.* 219, 623–642.
- Dixon, H.B., & Perham, R.N. (1968) *Biochem. J.* 109, 312–314.
- Butler, P.J., Harris, J.I., Hartley, B.S., & Leberman, R. (1969) *Biochem. J.* 112, 679–689.
- Houmard, J., & Drapeau, G.R. (1972) *Proc. Natl. Acad. Sci. U.S.A.* 69, 3506–3509.
- Heinrikson, R.L. (1977) *Methods Enzymol.* 47, 175–189.
- Smyth, D.G. (1967) *Methods Enzymol.* 11, 214–231.
- Masaki, T., Fujihashi, T., Nakamura, K., & Soejima, M. (1981) *Biochim. Biophys. Acta* 660, 51–55.
- Jekel, P.A., Weijer, W.J., & Beintema, J.J. (1983) *Anal. Biochem.* 134, 347–354.
- Schenkein, I., Levy, M., Franklin, E.C., & Frangione, B. (1977) *Arch. Biochem. Biophys.* 182, 64–70.

* Identified by comparison to authentic compound.

24. Drapeau, G.R. (1980) *J. Biol. Chem.* 255, 839–840.
25. Choi, K.H., Laursen, R.A., & Allen, K.N. (1999) *Biochemistry* 38, 11624–11633.
26. Sela, M., White, F.H., & Anfinsen, C.B. (1959) *Biochim. Biophys. Acta* 31, 417–426.
27. Brattin, W.J., Jr., & Smith, E.L. (1971) *J. Biol. Chem.* 246, 2400–2418.
28. Nicholas, R.A. (1984) *Biochemistry* 23, 888–898.
29. Gross, E., & Witkop, B. (1962) *J. Biol. Chem.* 237, 1856–1860.
30. Jacobson, G.R., Schaffer, M.H., Stark, G.R., & Vanaman, T.C. (1973) *J. Biol. Chem.* 248, 6583–6591.
31. Witkop, B. (1961) *Adv. Protein Chem.* 16, 221–321.
32. Burstein, Y., & Patchornik, A. (1972) *Biochemistry* 11, 4641–4650.
33. Landon, M. (1977) *Methods Enzymol.* 47, 145–149.
34. Piskiewicz, D., Landon, M., & Smith, E.L. (1970) *Biochem. Biophys. Res. Commun.* 40, 1173–1178.
35. Charbonneau, H., Tonks, N.K., Kumar, S., Diltz, C.D., Harrylock, M., Cool, D.E., Krebs, E.G., Fischer, E.H., & Walsh, K.A. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 5252–5256.
36. Hu, W., Van Driessche, G., Devreese, B., Goodhew, C.F., McGinnity, D.F., Saunders, N., Fulop, V., Pettigrew, G.W., & Van Beeumen, J.J. (1997) *Biochemistry* 36, 7958–7966.
37. Bornstein, P. (1970) *Biochemistry* 9, 2408–2421.
38. Titani, K., Koide, A., Hermann, J., Ericsson, L.H., Kumar, S., Wade, R.D., Walsh, K.A., Neurath, H., & Fischer, E.H. (1977) *Proc. Natl. Acad. Sci. U.S.A.* 74, 4762–4766.
39. Huang, I.Y., Welch, C.D., & Yoshida, A. (1980) *J. Biol. Chem.* 255, 6412–6420.
40. Koide, A., Titani, K., Ericsson, L.H., Kumar, S., Neurath, H., & Walsh, K.A. (1978) *Biochemistry* 17, 5657–5672.
41. Mahoney, W.C., & Hermodson, M.A. (1980) *J. Biol. Chem.* 255, 11199–11203.
42. Shoji, S., Parmelee, D.C., Wade, R.D., Kumar, S., Ericsson, L.H., Walsh, K.A., Neurath, H., Long, G.L., Demaille, J.G., Fischer, E.H., & Titani, K. (1981) *Proc. Natl. Acad. Sci. U.S.A.* 78, 848–851.
43. Bradshaw, R.A., Garner, W.H., & Gurd, F.R. (1969) *J. Biol. Chem.* 244, 2149–2158.
44. Petra, P.H. (1970) *Methods Enzymol.* 19, 460–503.
45. Folk, J.E. (1970) *Methods Enzymol.* 19, 504–508.
46. Hayashi, R. (1976) *Methods Enzymol.* 45, 568–587.
47. Himmelhoch, S.R. (1970) *Methods Enzymol.* 19, 508.
48. Harris, J.I., & Li, C.H. (1955) *J. Biol. Chem.* 213, 499–507.
49. Johnson, R.S., & Biemann, K. (1987) *Biochemistry* 26, 1209–1214.
50. Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F., & Whitehouse, C.M. (1989) *Science* 246, 64–71.
51. Barber, M., Bordoli, R.S., Sedgwick, R.D., & Tyler, A.N. (1981) *J. Chem. Soc., Chem. Commun.*, 325–327.
52. Surman, D.J., & Vickerman, J.C. (1981) *J. Chem. Soc., Chem. Commun.*, 324–325.
53. Karas, M., & Hillenkamp, F. (1988) *Anal. Chem.* 60, 2299–2301.
54. Shevchenko, A., Wilm, M., Vorm, O., & Mann, M. (1996) *Anal. Chem.* 68, 850–858.
55. Hirs, C.H.W., Moore, S., & Stein, W.H. (1960) *J. Biol. Chem.* 235, 633–647.
56. Canfield, R.E. (1963) *J. Biol. Chem.* 238, 2698–2707.
57. Hartley, B.S. (1964) *Nature* 201, 1284–1291.
58. Edmundson, A.B. (1965) *Nature* 205, 883–887.
59. Edelman, G.M., Cunningham, B.A., Gall, W.E., Gottlieb, P.D., Rutishauser, U., & Waxdal, M.J. (1969) *Proc. Natl. Acad. Sci. U.S.A.* 63, 78–85.
60. Titani, K., Koide, A., Ericsson, L.H., Kumar, S., Hermann, J., Wade, R.D., Walsh, K.A., Neurath, H., & Fischer, E.H. (1978) *Biochemistry* 17, 5680–5693.
61. Fowler, A.V., & Zabin, I. (1978) *J. Biol. Chem.* 253, 5521–5525.
62. Watt, K.W., Cottrell, B.A., Strong, D.D., & Doolittle, R.F. (1979) *Biochemistry* 18, 5410–5416.
63. Roberts, R.J. (1983) *Nucleic Acids Res.* 11, r135–r167.
64. Chaiken, P., Ballou, D.P., & Massey, V. (1997) *Proc. Natl. Acad. Sci. U.S.A.* 94, 7233–7238.
65. Burke, C.C., Wildung, M.R., & Croteau, R. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 13062–13067.
66. Vehar, G.A., Keyt, B., Eaton, D., Rodriguez, H., O'Brien, D.P., Rotblat, F., Oppermann, H., Keck, R., Wood, W.I., Harkins, R.N., Tuddenham, E.G.D., Lawn, R.M., & Capon, D.J. (1984) *Nature* 312, 337–342.
67. Wood, W.I., Capon, D.J., Simonsen, C.C., Eaton, D.L., Gitschier, J., Keyt, B., Seeburg, P.H., Smith, D.H., Hollingshead, P., Wion, K.L., Delwart, E., Tuddenham, E.G.D., Vehar, G., & Lawn, R.M. (1984) *Nature* 312, 330–337.
68. Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H. (1986) *Cold Spring Harbor Symp. Quant. Biol.* 51 (Pt 1), 263–273.
69. Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B., & Erlich, H.A. (1988) *Science* 239, 487–491.
70. Lundberg, K.S., Shoemaker, D.D., Adams, M.W., Short, J.M., Sorge, J.A., & Mathur, E.J. (1991) *Gene* 108, 1–6.
71. Ertl, H., Hallmann, A., Wenzl, S., & Sumper, M. (1992) *EMBO J.* 11, 2055–2062.
72. Boulter, J., Luyten, W., Evans, K., Mason, P., Ballivet, M., Goldman, D., Stengelin, S., Martin, G., Heinemann, S., & Patrick, J. (1985) *J. Neurosci.* 5, 2545–2552.
73. Kumagai, I., Pieler, T., Subramanian, A.R., & Erdmann, V.A. (1982) *J. Biol. Chem.* 257, 12924–12928.
74. Olivera, B.M., Baine, P., & Davidson, N. (1964) *Biopolymers* 2, 245–257.
75. Fisher, M.P., & Dingman, C.W. (1971) *Biochemistry* 10, 1895–1899.
76. Richards, E.G., & Lecanidou, R. (1971) *Anal. Biochem.* 40, 43–71.
77. Maxam, A.M., & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. U.S.A.* 74, 560–564.
78. Tamm, C., Hodes, M.E., & Chargaff, E. (1952) *J. Biol. Chem.* 195, 49–63.
79. Chargaff, E., Rust, P., Temperli, A., Morisawa, S., & Danon, A. (1963) *Biochim. Biophys. Acta* 76, 149–151.
80. Lawley, P.D., & Brooks, P. (1963) *Biochem. J.* 89, 127–128.
81. Maxam, A.M., & Gilbert, W. (1980) *Methods Enzymol.* 65, 499–560.
82. Temperli, A., Turler, H., Rust, P., Danon, A., & Chargaff, E. (1964) *Biochim. Biophys. Acta* 91, 462–476.
83. Hayes, D.H., & Hayes-Baron, F. (1967) *J. Chem. Soc., C*, 1528–1533.

84. Tamm, C., Shapiro, H.S., Lipshitz, R., & Chargaff, E. (1953) *J. Biol. Chem.* 203, 673–688.
85. Sanger, F., Nicklen, S., & Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* 74, 5463–5467.
86. Messing, J., Crea, R., & Seeburg, P.H. (1981) *Nucleic Acids Res.* 9, 309–321.
87. Saladino, R., Mincione, E., Crestini, C., Negri, R., Di Mauro, E., & Costanzo, G. (1996) *J. Am. Chem. Soc.* 118, 5615–5619.
88. Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B., & Hood, L.E. (1986) *Nature* 321, 674–679.
89. Prober, J.M., Trainor, G.L., Dam, R.J., Hobbs, F.W., Robertson, C.W., Zagursky, R.J., Cocuzza, A.J., Jensen, M.A., & Baumeister, K. (1987) *Science* 238, 336–341.
90. Rosenblum, B.B., Lee, L.G., Spurgeon, S.L., Khan, S.H., Menchen, S.M., Heiner, C.R., & Chen, S.M. (1997) *Nucleic Acids Res.* 25, 4500–4504.
91. Tabor, S., & Richardson, C.C. (1987) *Proc. Natl. Acad. Sci. U.S.A.* 84, 4767–4771.
92. Tabor, S., & Richardson, C.C. (1995) *Proc. Natl. Acad. Sci. U.S.A.* 92, 6339–6343.
93. Chan, W.Y., Liu, Q.R., Borjigin, J., Busch, H., Rennert, O.M., Tease, L.A., & Chan, P.K. (1989) *Biochemistry* 28, 1033–1039.
94. Eggink, G., Engel, H., Vriend, G., Terpstra, P., & Witholt, B. (1990) *J. Mol. Biol.* 212, 135–142.
95. Wei, Y., Contreras, J.A., Sheffield, P., Osterlund, T., Derewenda, U., Kneusel, R.E., Matern, U., Holm, C., & Derewenda, Z.S. (1999) *Nat. Struct. Biol.* 6, 340–345.
96. Xu, D., Ballou, D.P., & Massey, V. (2001) *Biochemistry* 40, 12369–12378.
97. Hasson, M.S., Muscate, A., McLeish, M.J., Polovnikova, L.S., Gerlt, J.A., Kenyon, G.L., Petsko, G.A., & Ringe, D. (1998) *Biochemistry* 37, 9918–9930.
98. Andersson, I. (1996) *J. Mol. Biol.* 259, 160–174.
99. Keller, B., Sauer, N., & Lamb, C.J. (1988) *EMBO J.* 7, 3625–3633.
100. Nardelli, D., Gerber-Huber, S., Van Het Schip, F.D., Gruber, M., Ab, G., & Wahli, W. (1987) *Biochemistry* 26, 6397–6402.
101. Xu, M., & Lewis, R.V. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87, 7120–7124.
102. Ann, D.K., Smith, M.K., & Carlson, D.M. (1988) *J. Biol. Chem.* 263, 10887–10893.
103. Koide, T., Foster, D., Yoshitake, S., & Davie, E.W. (1986) *Biochemistry* 25, 2220–2225.
104. Celniker, S.E., Keelan, D.J., & Lewis, E.B. (1989) *Genes Dev.* 3, 1424–1436.
105. La Spada, A.R., Wilson, E.M., Lubahn, D.B., Harding, A.E., & Fischbeck, K.H. (1991) *Nature* 352, 77–79.
106. Perutz, M.F. (1999) *Trends Biochem. Sci.* 24, 58–63.
107. Perutz, M.F., Johnson, T., Suzuki, M., & Finch, J.T. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 5355–5358.
108. Group, T.H.s.D.C.R. (1993) *Cell* 72, 971–983.
109. Gomez, J., Sanchez-Martinez, D., Stiefel, V., Rigau, J., Puigdomenech, P., & Pages, M. (1988) *Nature* 334, 262–264.
110. Haydock, P.V., & Dale, B.A. (1990) *DNA Cell Biol.* 9, 251–261.
111. Resing, K.A., Johnson, R.S., & Walsh, K.A. (1993) *Biochemistry* 32, 10036–10045.
112. Morgan, D.O., Edman, J.C., Standring, D.N., Fried, V.A., Smith, M.C., Roth, R.A., & Rutter, W.J. (1987) *Nature* 329, 301–307.
113. Warren, J.C., Murdock, G.L., Ma, Y., Goodman, S.R., & Zimmer, W.E. (1993) *Biochemistry* 32, 1401–1406.
114. Kennedy, M.C., Mende-Mueller, L., Blondin, G.A., & Beinert, H. (1992) *Proc. Natl. Acad. Sci. U.S.A.* 89, 11730–11734.
115. Seely, O., Jr., Feng, D.F., Smith, D.W., Sulzbach, D., & Doolittle, R.F. (1990) *Genomics* 8, 71–82.
116. Ferreira, G.C., & Dailey, H.A. (1993) *J. Biol. Chem.* 268, 584–590.
117. Kervinen, J., Dunbrack, R.L., Jr., Litwin, S., Martins, J., Scarrow, R.C., Volin, M., Yeung, A.T., Yoon, E., & Jaffe, E.K. (2000) *Biochemistry* 39, 9018–9029.
118. Zeghouf, M., Fontecave, M., Macherel, D., & Coves, J. (1998) *Biochemistry* 37, 6114–6123.
119. Mathis, J.R., Back, K., Starks, C., Noel, J., Poulter, C.D., & Chappell, J. (1997) *Biochemistry* 36, 8340–8348.
120. Hallis, T.M., Lei, Y., Que, N.L., & Liu, H. (1998) *Biochemistry* 37, 4935–4945.
121. Peters, R.J., Flory, J.E., Jetter, R., Ravn, M.M., Lee, H.J., Coates, R.M., & Croteau, R.B. (2000) *Biochemistry* 39, 15592–15602.
122. Stewart, J., Wilson, D.B., & Ganem, B. (1990) *J. Am. Chem. Soc.* 112, 4582–4584.
123. Butt, T.R., Jonnalagadda, S., Monia, B.P., Sternberg, E.J., Marsh, J.A., Stadel, J.M., Ecker, D.J., & Crooke, S.T. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 2540–2544.
124. van der Linden, M.P., Mottl, H., & Keck, W. (1992) *Eur. J. Biochem.* 204, 197–202.
125. Hitzeman, R.A., Leung, D.W., Perry, L.J., Kohr, W.J., Levine, H.L., & Goeddel, D.V. (1983) *Science* 219, 620–625.
126. Hinnen, A., Hicks, J.B., & Fink, G.R. (1978) *Proc. Natl. Acad. Sci. U.S.A.* 75, 1929–1933.
127. Luckow, V.A., & Summers, M.D. (1989) *Virology* 170, 31–39.
128. Medin, J.A., Hunt, L., Gathy, K., Evans, R.K., & Coleman, M.S. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87, 2760–2764.
129. Luckow, V.A. (1991) in *Recombinant DNA Technology and Applications* (Prokop, A., Bajpai, K., & Ho, C., Eds.) pp 97–152, McGraw-Hill, New York.
130. Smith, M.C., Furman, T.C., Ingolia, T.D., & Pidgeon, C. (1988) *J. Biol. Chem.* 263, 7211–7215.
131. Chattopadhyay, D., Evans, D.B., Deibel, M.R., Jr., Vosters, A.F., Eckenrode, F.M., Einspahr, H.M., Hui, J.O., Tomasselli, A.G., Zurcher-Neely, H.A., Heinrikson, R.L., et al. (1992) *J. Biol. Chem.* 267, 14227–14232.
132. Hutchison, C.A., III, Phillips, S., Edgell, M.H., Gillam, S., Jahnke, P., & Smith, M. (1978) *J. Biol. Chem.* 253, 6551–6560.
133. Zoller, M.J., & Smith, M. (1982) *Nucleic Acids Res.* 10, 6487–6500.
134. Alber, T., Sun, D.P., Wilson, K., Wozniak, J.A., Cook, S.P., & Matthews, B.W. (1987) *Nature* 330, 41–46.
135. Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.J., & Roe, B.A. (1980) *J. Mol. Biol.* 143, 161–178.
136. Wilkinson, A.J., Fersht, A.R., Blow, D.M., & Winter, G. (1983) *Biochemistry* 22, 3581–3586.

137. Sugimoto, M., Esaki, N., Tanaka, H., & Soda, K. (1989) *Anal. Biochem.* 179, 309–311.
138. Kunkel, T.A., Roberts, J.D., & Zakour, R.A. (1987) *Methods Enzymol.* 154, 367–382.
139. Taylor, J.W., Ott, J., & Eckstein, F. (1985) *Nucleic Acids Res.* 13, 8765–8785.
140. Vandeyar, M.A., Weiner, M.P., Hutton, C.J., & Batt, C.A. (1988) *Gene* 65, 129–133.
141. Weiner, M.P., Costa, G.L., Schoettlin, W., Cline, J., Mathur, E., & Bauer, J.C. (1994) *Gene* 151, 119–123.
142. Deng, W.P., & Nickoloff, J.A. (1992) *Anal. Biochem.* 200, 81–88.
143. Ho, S.N., Hunt, H.D., Horton, R.M., Pullen, J.K., & Pease, L.R. (1989) *Gene* 77, 51–59.
144. Landt, O., Grunert, H.P., & Hahn, U. (1990) *Gene* 96, 125–128.
145. Jones, D.H., & Winistorfer, S.C. (1992) *BioTechniques* 12, 528–530, 532, 534–525.
146. Reidhaar-Olson, J.F., & Sauer, R.T. (1988) *Science* 241, 53–57.
147. Climie, S., & Santi, D.V. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87, 633–637.
148. Noren, C.J., Anthony-Cahill, S.J., Griffith, M.C., & Schultz, P.G. (1989) *Science* 244, 182–188.
149. Bain, J.D., Diala, E.S., Glabe, C.G., Dix, T.A., & Chamberlin, A.R. (1989) *J. Am. Chem. Soc.* 111, 8013–8014.
150. Judice, J.K., Gamble, T.R., Murphy, E.C., de Vos, A.M., & Schultz, P.G. (1993) *Science* 261, 1578–1581.
151. Endo, Y., & Wool, I.G. (1982) *J. Biol. Chem.* 257, 9054–9060.
152. Zinoni, F., Birkmann, A., Stadtman, T.C., & Beock, A. (1986) *Proc. Natl. Acad. Sci. U.S.A.* 83, 4650–4654.
153. Ermler, U., Grabarse, W., Shima, S., Goubeaud, M., & Thauer, R.K. (1997) *Science* 278, 1457–1462.
154. Grabarse, W., Mahlert, F., Shima, S., Thauer, R.K., & Ermler, U. (2000) *J. Mol. Biol.* 303, 329–344.
155. Neurath, H. (1984) *Science* 224, 350–357.
156. Thomas, L., Leduc, R., Thorne, B.A., Smeekens, S.P., Steiner, D.F., & Thomas, G. (1991) *Proc. Natl. Acad. Sci. U.S.A.* 88, 5297–5301.
157. Lazure, C., Seidah, N.G., Pelaprat, D., & Chretien, M. (1983) *Can. J. Biochem. Cell Biol.* 61, 501–515.
158. Nakanishi, S., Inoue, A., Kita, T., Nakamura, M., Chang, A.C., Cohen, S.N., & Numa, S. (1979) *Nature* 278, 423–427.
159. Blobel, G., & Dobberstein, B. (1975) *J. Cell Biol.* 67, 852–862.
160. Yorgey, P., Lee, J., Kordel, J., Vivas, E., Warner, P., Jebaratnam, D., & Kolter, R. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 4519–4523.
161. Kelleher, N.L., Hendrickson, C.L., & Walsh, C.T. (1999) *Biochemistry* 38, 15623–15630.
162. Ekstrom, J.L., Tolbert, W.D., Xiong, H., Pegg, A.E., & Ealick, S.E. (2001) *Biochemistry* 40, 9495–9504.
163. Recsei, P.A., & Snell, E.E. (1973) *Biochemistry* 12, 365–371.
164. van Poelje, P.D., & Snell, E.E. (1990) *Biochemistry* 29, 132–139.
165. Albert, A., Dhanaraj, V., Genschel, U., Khan, G., Ramjee, M.K., Pulido, R., Sibanda, B.L., von Delft, F., Witty, M., Blundell, T.L., Smith, A.G., & Abell, C. (1998) *Nat. Struct. Biol.* 5, 289–293.
166. Recsei, P.A., Huynh, Q.K., & Snell, E.E. (1983) *Proc. Natl. Acad. Sci. U.S.A.* 80, 973–977.
167. Kapke, G., & Davis, L. (1975) *Biochemistry* 14, 4273–4276.
168. Guan, C., Cui, T., Rao, V., Liao, W., Benner, J., Lin, C.L., & Comb, D. (1996) *J. Biol. Chem.* 271, 1732–1737.
169. Fisher, K.J., Tollersrud, O.K., & Aronson, N.N., Jr. (1990) *FEBS Lett.* 269, 440–444.
170. Porter, J.A., Young, K.E., & Beachy, P.A. (1996) *Science* 274, 255–259.
171. Swallow, D.L., & Abraham, E.P. (1958) *Biochem. J.* 70, 364–373.
172. Geiger, T., & Clarke, S. (1987) *J. Biol. Chem.* 262, 785–794.
173. Haley, E.E., Corcoran, B.J., Dorer, F.E., & Buchanan, D.L. (1966) *Biochemistry* 5, 3229–3235.
174. Noguchi, S., Miyawaki, K., & Satow, Y. (1998) *J. Mol. Biol.* 278, 231–238.
175. Esposito, L., Vitagliano, L., Sica, F., Sorrentino, G., Zagari, A., & Mazzarella, L. (2000) *J. Mol. Biol.* 297, 713–732.
176. McIntire, W.E., Schey, K.L., Knapp, D.R., & Hildebrandt, J.D. (1998) *Biochemistry* 37, 14651–14658.
177. Artigues, A., Birkett, A., & Schirch, V. (1990) *J. Biol. Chem.* 265, 4853–4858.
178. McFadden, P.N., & Clarke, S. (1982) *Proc. Natl. Acad. Sci. U.S.A.* 79, 2460–2464.
179. Aswad, D.W. (1984) *J. Biol. Chem.* 259, 10714–10721.
180. Lowenson, J.D., & Clarke, S. (1992) *J. Biol. Chem.* 267, 5985–5995.
181. Johnson, B.A., Murray, E.D., Jr., Clarke, S., Glass, D.B., & Aswad, D.W. (1987) *J. Biol. Chem.* 262, 5622–5629.
182. Najbauer, J., Orpizewski, J., & Aswad, D.W. (1996) *Biochemistry* 35, 5183–5190.
183. Hirata, R., Ohsumk, Y., Nakano, A., Kawasaki, H., Suzuki, K., & Anraku, Y. (1990) *J. Biol. Chem.* 265, 6726–6733.
184. Kane, P.M., Yamashiro, C.T., Wolczyk, D.F., Neff, N., Goebel, M., & Stevens, T.H. (1990) *Science* 250, 651–657.
185. Davis, E.O., Sedgwick, S.G., & Colston, M.J. (1991) *J. Bacteriol.* 173, 5653–5662.
186. Perler, F.B., Comb, D.G., Jack, W.E., Moran, L.S., Qiang, B., Kucera, R.B., Benner, J., Slatko, B.E., Nwankwo, D.O., Hempstead, S.K., et al. (1992) *Proc. Natl. Acad. Sci. U.S.A.* 89, 5577–5581.
187. Martin, D.D., Xu, M.Q., & Evans, T.C., Jr. (2001) *Biochemistry* 40, 1393–1402.
188. Klabunde, T., Sharma, S., Telenti, A., Jacobs, W.R., Jr., & Sacchettini, J.C. (1998) *Nat. Struct. Biol.* 5, 31–36.
189. Xu, M.Q., Comb, D.G., Paulus, H., Noren, C.J., Shao, Y., & Perler, F.B. (1994) *EMBO J.* 13, 5517–5522.
190. Carrington, D.M., Auffret, A., & Hanke, D.E. (1985) *Nature* 313, 64–67.
191. Xu, M.Q., & Perler, F.B. (1996) *EMBO J.* 15, 5146–5153.
192. Shao, Y., Xu, M.Q., & Paulus, H. (1996) *Biochemistry* 35, 3810–3815.
193. Chong, S., Shao, Y., Paulus, H., Benner, J., Perler, F.B., & Xu, M.Q. (1996) *J. Biol. Chem.* 271, 22159–22168.
194. Xu, M.Q., Southworth, M.W., Mersha, F.B., Hornstra, L.J., & Perler, F.B. (1993) *Cell* 75, 1371–1377.
195. Duan, X., Gimble, F.S., & Quijcho, F.A. (1997) *Cell* 89, 555–564.

196. Ichiyanagi, K., Ishino, Y., Ariyoshi, M., Komori, K., & Morikawa, K. (2000) *J. Mol. Biol.* 300, 889–901.
197. Cunningham, B.A., Wang, J.L., Waxdal, M.J., & Edelman, G.M. (1975) *J. Biol. Chem.* 250, 1503–1512.
198. Chrispeels, M.J., Hartl, P.M., Sturm, A., & Faye, L. (1986) *J. Biol. Chem.* 261, 10021–10024.
199. Chang, C.N., Schwartz, M., & Chang, F.N. (1976) *Biochem. Biophys. Res. Commun.* 73, 233–239.
200. Stock, A., Clarke, S., Clarke, C., & Stock, J. (1987) *FEBS Lett.* 220, 8–14.
201. Rose, K., Simona, M.G., Savoy, L.A., Regamey, P.O., Green, B.N., Clore, G.M., Gronenborn, A.M., & Wingfield, P.T. (1992) *J. Biol. Chem.* 267, 19101–19106.
202. Milligan, D.L., & Koshland, D.E., Jr. (1990) *J. Biol. Chem.* 265, 4455–4460.
203. Persson, B., Flinta, C., von Heijne, G., & Jornvall, H. (1985) *Eur. J. Biochem.* 152, 523–527.
204. Lin, T.S., & Kolattukudy, P.E. (1980) *Eur. J. Biochem.* 106, 341–351.
205. Doolittle, R.F. (1972) *Methods Enzymol.* 25, 231–244.
206. Farries, T.C., Harris, A., Auffret, A.D., & Aitken, A. (1991) *Eur. J. Biochem.* 196, 679–685.
207. Hantke, K., & Braun, V. (1973) *Eur. J. Biochem.* 34, 284–296.
208. Prutsch, A., Lohaus, C., Green, B., Meyer, H.E., & Lubben, M. (2000) *Biochemistry* 39, 6554–6563.
209. Towler, D.A., Eubanks, S.R., Towery, D.S., Adams, S.P., & Glaser, L. (1987) *J. Biol. Chem.* 262, 1030–1036.
210. Carr, S.A., Biemann, K., Shoji, S., Parmelee, D.C., & Titani, K. (1982) *Proc. Natl. Acad. Sci. U.S.A.* 79, 6128–6131.
211. Dizhoor, A.M., Ericsson, L.H., Johnson, R.S., Kumar, S., Olshevskaya, E., Zozulya, S., Neubert, T.A., Stryer, L., Hurley, J.B., & Walsh, K.A. (1992) *J. Biol. Chem.* 267, 16033–16036.
212. Paturle-Lafanechere, L., Edde, B., Denoulet, P., Van Dorsselaer, A., Mazarguil, H., Le Caer, J.P., Wehland, J., & Job, D. (1991) *Biochemistry* 30, 10523–10528.
213. Xie, H., & Clarke, S. (1993) *J. Biol. Chem.* 268, 13364–13371.
214. Eipper, B.A., Perkins, S.N., Husten, E.J., Johnson, R.C., Keutmann, H.T., & Mains, R.E. (1991) *J. Biol. Chem.* 266, 7827–7833.
215. Stimmel, J.B., Deschenes, R.J., Volker, C., Stock, J., & Clarke, S. (1990) *Biochemistry* 29, 9651–9659.
216. Vorburger, K., Kitten, G.T., & Nigg, E.A. (1989) *EMBO J.* 8, 4007–4013.
217. Anderegg, R.J., Betz, R., Carr, S.A., Crabb, J.W., & Duntze, W. (1988) *J. Biol. Chem.* 263, 18236–18240.
218. Yamane, H.K., Farnsworth, C.C., Xie, H.Y., Howald, W., Fung, B.K., Clarke, S., Gelb, M.H., & Glomset, J.A. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87, 5868–5872.
219. Rilling, H.C., Breunger, E., Epstein, W.W., & Crain, P.F. (1990) *Science* 247, 318–320.
220. Ishibashi, Y., Sakagami, Y., Isogai, A., & Suzuki, A. (1984) *Biochemistry* 23, 1399–1404.
221. Hancock, J.F., Cadwallader, K., & Marshall, C.J. (1991) *EMBO J.* 10, 641–646.
222. Silvius, J.R., & l'Heureux, F. (1994) *Biochemistry* 33, 3014–3022.
223. Farnsworth, C.C., Kawata, M., Yoshida, Y., Takai, Y., Gelb, M.H., & Glomset, J.A. (1991) *Proc. Natl. Acad. Sci. U.S.A.* 88, 6196–6200.
224. Khosravi-Far, R., Lutz, R.J., Cox, A.D., Conroy, L., Bourne, J.R., Sinensky, M., Balch, W.E., Buss, J.E., & Der, C.J. (1991) *Proc. Natl. Acad. Sci. U.S.A.* 88, 6264–6268.
225. Giner, J.L., & Rando, R.R. (1994) *Biochemistry* 33, 15116–15123.
226. Ferguson, M.A., Low, M.G., & Cross, G.A. (1985) *J. Biol. Chem.* 260, 14547–14555.
227. Tse, A.G., Barclay, A.N., Watts, A., & Williams, A.F. (1985) *Science* 230, 1003–1008.
228. Ferguson, M.A., Homans, S.W., Dwek, R.A., & Rademacher, T.W. (1988) *Science* 239, 753–759.
229. Homans, S.W., Ferguson, M.A., Dwek, R.A., Rademacher, T.W., Anand, R., & Williams, A.F. (1988) *Nature* 333, 269–272.
230. Oxley, D., & Bacic, A. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 14246–14251.
231. Fankhauser, C., Homans, S.W., Thomas-Oates, J.E., McConville, M.J., Desponds, C., Conzelmann, A., & Ferguson, M.A. (1993) *J. Biol. Chem.* 268, 26365–26374.
232. Field, M.S., & Menon, A.K. (1993) in *Lipid Modifications of Proteins* (Schlesinger, M. J., Ed.) pp 83–134, CRC Press, Boca Raton, FL.
233. Ferguson, M.A., & Williams, A.F. (1988) *Annu. Rev. Biochem.* 57, 285–320.
234. Gerold, P., Striepen, B., Reitter, B., Geyer, H., Geyer, R., Reinwald, E., Risse, H.J., & Schwarz, R.T. (1996) *J. Mol. Biol.* 261, 181–194.
235. Guthrie, M.L., de Almeida, M.L., Yoshida, N., & Ferguson, M.A. (1992) *J. Biol. Chem.* 267, 6820–6828.
236. Uy, R., & Wold, F. (1977) *Science* 198, 890–896.
237. Lipmann, F. (1933) *Biochem. Z.* 262, 3–8.
238. Taborsky, G. (1974) *Adv. Protein Chem.* 28, 1–210.
239. Hunter, T. (1987) *Cell* 50, 823–829.
240. deVerdier, C. (1952) *Nature* 170, 804–805.
241. Eckhart, W., Hutchinson, M.A., & Hunter, T. (1979) *Cell* 18, 925–933.
242. Hunter, T., & Cooper, J.A. (1985) *Annu. Rev. Biochem.* 54, 897–930.
243. Chen, C.C., Bruegger, B.B., Kern, C.W., Lin, Y.C., Halpern, R.M., & Smith, R.A. (1977) *Biochemistry* 16, 4852–4855.
244. DeLuca, M., Ebner, K.E., Hultquist, D.E., Kreil, G., Peter, J.B., Moyer, R.W., & Boyer, P.D. (1963) *Biochem. Z.* 338, 512–525.
245. Smith, L.S., Kern, C.W., Halpern, R.M., & Smith, R.A. (1976) *Biochem. Biophys. Res. Commun.* 71, 459–465.
246. Pigiet, V., & Conley, R.R. (1978) *J. Biol. Chem.* 253, 1910–1920.
247. Degani, C., & Boyer, P.D. (1973) *J. Biol. Chem.* 248, 8222–8226.
248. Lewis, R.J., Brannigan, J.A., Muchova, K., Barak, I., & Wilkinson, A.J. (1999) *J. Mol. Biol.* 294, 9–15.
249. Sanders, D.A., Gillece-Castro, B.L., Stock, A.M., Burlingame, A.L., & Koshland, D.E., Jr. (1989) *J. Biol. Chem.* 264, 21770–21778.
250. Cohen-Solal, L., Cohen-Solal, M., & Glimcher, M.J. (1979) *Proc. Natl. Acad. Sci. U.S.A.* 76, 4327–4330.
251. Huttner, W.B. (1982) *Nature* 299, 273–276.
252. Nelsestuen, G.L., Zytkevich, T.H., & Howard, J.B. (1974) *J. Biol. Chem.* 249, 6347–6350.

253. Stenflo, J., Ferlund, P., Egan, W., & Roepstorff, P. (1974) *Proc. Natl. Acad. Sci. U.S.A.* 71, 2730–2733.
254. McTigue, J.J., Dhaon, M.K., Rich, D.H., & Suttie, J.W. (1984) *J. Biol. Chem.* 259, 4272–4278.
255. Welinder, B.S. (1972) *Biochim. Biophys. Acta* 279, 491–497.
256. Henze, M. (1907) *Hoppe-Seyler's Z. Physiol. Chem.* 51, 64.
257. Wolff, J., & Covelli, I. (1969) *Eur. J. Biochem.* 9, 371–377.
258. Roche, J. (1952) *Experientia* 8, 45–84.
259. Ackermann, D., & Müller, E. (1941) *Hoppe-Seyler's Z. Physiol. Chem.* 269, 146–157.
260. Hunt, S., & Breuer, S.W. (1971) *Biochim. Biophys. Acta* 252, 401–404.
261. Kendall, E.C. (1919) *J. Biol. Chem.* 39, 125–147.
262. Harington, C.R. (1944) *Proc. R. Soc. London, B* 132, 223–238.
263. McQuillan, M.T., & Trikojus, V.M. (1972) in *Glycoproteins: Their Composition, Structure, and Function* (Gottschalk, A., Ed.) pp 926–963, Elsevier, Amsterdam.
264. Gross, J., & Pitt-Rivers, R. (1953) *Biochem. J.* 53, 645–650.
265. Roche, J., Michel, R., & Tata, J. (1953) *Biochim. Biophys. Acta* 11, 543–547.
266. Mercken, L., Simons, M.J., Swillens, S., Massaer, M., & Vassart, G. (1985) *Nature* 316, 647–651.
267. Chen, J.Y., & Bodley, J.W. (1988) *J. Biol. Chem.* 263, 11692–11696.
268. Evans, D.A., & Lundy, K.M. (1992) *J. Am. Chem. Soc.* 114, 1495–1496.
269. Van Ness, B.G., Howard, J.B., & Bodley, J.W. (1980) *J. Biol. Chem.* 255, 10710–10716.
270. Hofsteenge, J., Muller, D.R., de Beer, T., Löffler, A., Richter, W.J., & Vliegenthart, J.F. (1994) *Biochemistry* 33, 13524–13530.
271. Löffler, A., Doucey, M.A., Jansson, A.M., Muller, D.R., de Beer, T., Hess, D., Meldal, M., Richter, W.J., Vliegenthart, J.F., & Hofsteenge, J. (1996) *Biochemistry* 35, 12005–12014.
272. Yoshino, K., Takao, T., Suhara, M., Kitai, T., Hori, H., Nomura, K., Yamaguchi, M., Shimonishi, Y., & Suzuki, N. (1991) *Biochemistry* 30, 6203–6209.
273. Ambler, R.P., & Rees, M.W. (1959) *Nature* 184, 56–57.
274. Paik, W.K., & Kim, S. (1971) *Science* 174, 114–119.
275. Paik, W.K., & Kim, S. (1980) *Protein Methylation*, John Wiley, New York.
276. Paik, W.K., & Kim, S. (1967) *Biochem. Biophys. Res. Commun.* 27, 479–483.
277. Hempel, K., Lange, H.W., & Birkofer, L. (1968) *Naturwissenschaften* 55, 37.
278. Ghosh, S.K., Paik, W.K., & Kim, S. (1988) *J. Biol. Chem.* 263, 19024–19033.
279. Paik, W.K., & Kim, S. (1970) *J. Biol. Chem.* 245, 88–92.
280. Baldwin, G.S., & Carnegie, P.R. (1971) *Science* 171, 579–581.
281. Karn, J., Vidali, G., Boffa, L.C., & Allfrey, V.G. (1977) *J. Biol. Chem.* 252, 7307–7322.
282. Lischwe, M.A., Cook, R.G., Ahn, Y.S., Yeoman, L.C., & Busch, H. (1985) *Biochemistry* 24, 6025–6028.
283. Zobel-Thropp, P., Gary, J.D., & Clarke, S. (1998) *J. Biol. Chem.* 273, 29283–29286.
284. Vijayasathy, C., & Rao, B.S. (1987) *Biochim. Biophys. Acta* 923, 156–165.
285. Van Der Werf, P., & Koshland, D.E., Jr. (1977) *J. Biol. Chem.* 252, 2793–2795.
286. Lowenson, J.D., & Clarke, S. (1990) *J. Biol. Chem.* 265, 3106–3110.
287. Farooqui, J.Z., Tuck, M., & Paik, W.K. (1985) *J. Biol. Chem.* 260, 537–545.
288. Swanson, R.V., & Glazer, A.N. (1990) *J. Mol. Biol.* 214, 787–796.
289. Klotz, A.V., & Glazer, A.N. (1987) *J. Biol. Chem.* 262, 17350–17355.
290. Lhoest, J., & Colson, C. (1977) *Mol. Gen. Genet.* 154, 175–180.
291. Selmer, T., Kahnt, J., Goubeaud, M., Shima, S., Grabarse, W., Ermler, U., & Thauer, R.K. (2000) *J. Biol. Chem.* 275, 3755–3760.
292. Park, M.H., Wolff, E.C., & Folk, J.E. (1993) *Biofactors* 4, 95–104.
293. Shiba, T., Mizote, H., Kaneko, T., Nakajima, T., & Kakimoto, Y. (1971) *Biochim. Biophys. Acta* 244, 523–531.
294. Wolff, E.C., Park, M.H., & Folk, J.E. (1990) *J. Biol. Chem.* 265, 4793–4799.
295. Kamiya, Y., Sakurai, A., Tamura, S., Takahashi, N., Tsuchiya, E., Abe, K., & Fukui, S. (1979) *Agric. Biol. Chem.* 43, 363–369.
296. Farnsworth, C.C., Gelb, M.H., & Glomset, J.A. (1990) *Science* 247, 320–322.
297. Gershey, E.L., Vidali, G., & Allfrey, V.G. (1968) *J. Biol. Chem.* 243, 5018–5022.
298. DeLange, R.J., Smith, E.L., Fambrough, D.M., & Bonner, J. (1968) *Proc. Natl. Acad. Sci. U.S.A.* 61, 1145–1146.
299. Stoffel, W., Hillen, H., Schreoder, W., & Deutzmann, R. (1983) *Hoppe-Seyler's Z. Physiol. Chem.* 364, 1455–1466.
300. Jing, S., & Trowbridge, I.S. (1987) *EMBO J.* 3, 2581–2585.
301. Schmidt, M.F. (1989) *Biochim. Biophys. Acta* 988, 411–426.
302. Schmidt, M., Schmidt, M.F., & Rott, R. (1988) *J. Biol. Chem.* 263, 18635–18639.
303. Bach, R., Konigsberg, W.H., & Nemerson, Y. (1988) *Biochemistry* 27, 4227–4231.
304. Redeker, V., Rossier, J., & Frankfurter, A. (1998) *Biochemistry* 37, 14838–14844.
305. Redeker, V., Levilliers, N., Schmitter, J.M., Le Caer, J.P., Rossier, J., Adoutte, A., & Bre, M.H. (1994) *Science* 266, 1688–1691.
306. Gallop, P.M., Blumenfeld, O.O., & Seifter, S. (1972) *Annu. Rev. Biochem.* 41, 617–672.
307. Nakajima, T., & Volcani, B.E. (1970) *Biochem. Biophys. Res. Commun.* 39, 28–33.
308. Udenfriend, S. (1966) *Science* 152, 1335–1340.
309. Bornstein, P. (1974) *Annu. Rev. Biochem.* 43, 567–603.
310. Berg, R.A., & Prockop, D.J. (1973) *J. Biol. Chem.* 248, 1175–1182.
311. Ogle, J.D., Arlinghaus, R.B., & Logan, M.A. (1962) *J. Biol. Chem.* 237, 3667–3673.
312. Janes, S.M., Mu, D., Wemmer, D., Smith, A.J., Kaur, S., Maltby, D., Burlingame, A.L., & Klinman, J.P. (1990) *Science* 248, 981–987.
313. McMullen, B.A., Fujikawa, K., Kisiel, W., Sasagawa, T., Howald, W.N., Kwa, E.Y., & Weinstein, B. (1983) *Biochemistry* 22, 2875–2884.

314. Fernlund, P., & Stenflo, J. (1983) *J. Biol. Chem.* 258, 12509–12512.
315. Wang, Q.P., VanDusen, W.J., Petroski, C.J., Garsky, V.M., Stern, A.M., & Friedman, P.A. (1991) *J. Biol. Chem.* 266, 14004–14010.
316. Stenflo, J., Lundwall, A., & Dahlback, B. (1987) *Proc. Natl. Acad. Sci. U.S.A.* 84, 368–372.
317. Choinowski, T., Blodig, W., Winterhalter, K.H., & Piontek, K. (1999) *J. Mol. Biol.* 286, 809–827.
318. Blodig, W., Smith, A.T., Doyle, W.A., & Piontek, K. (2001) *J. Mol. Biol.* 305, 851–861.
319. Goodwill, K.E., Sabatier, C., & Stevens, R.C. (1998) *Biochemistry* 37, 13437–13445.
320. Dorsett, L.C., Hawkins, C.J., Grice, J.A., Lavin, M.F., Merefiefield, P.M., Parry, D.L., & Ross, I.L. (1987) *Biochemistry* 26, 8078–8082.
321. Waite, J.H., & Tanzer, M.L. (1981) *Science* 212, 1038–1040.
322. Waite, J.H. (1983) *J. Biol. Chem.* 258, 2911–2915.
323. Filpula, D.R., Lee, S.M., Link, R.P., Strausberg, S.L., & Strausberg, R.L. (1990) *Biotechnol. Prog.* 6, 171–177.
324. Tajima, M., Iida, T., Yoshida, S., Komatsu, K., Namba, R., Yanagi, M., Noguchi, M., & Okamoto, H. (1990) *J. Biol. Chem.* 265, 9602–9605.
325. Stassen, F.L. (1976) *Biochim. Biophys. Acta* 438, 49–60.
326. Muchmore, C.R., Krahm, J.M., Kim, J.H., Zalkin, H., & Smith, J.L. (1998) *Protein Sci.* 7, 39–51.
327. Schmidt, B., Selmer, T., Ingendoh, A., & von Figura, K. (1995) *Cell* 82, 271–278.
328. Gouet, P., Jouve, H.M., & Dideberg, O. (1995) *J. Mol. Biol.* 249, 933–954.
329. Sjöberg, B.M., & Reichard, P. (1977) *J. Biol. Chem.* 252, 536–541.
330. Wagner, A.F., Frey, M., Neugebauer, F.A., Schafer, W., & Knappe, J. (1992) *Proc. Natl. Acad. Sci. U.S.A.* 89, 996–1000.
331. Just, I., Sehr, P., Jung, M., van Damme, J., Puype, M., Vandekerckhove, J., Moss, J., & Aktories, K. (1995) *Biochemistry* 34, 326–333.
332. Moss, J., & Vaughan, M. (1977) *J. Biol. Chem.* 252, 2455–2457.
333. Oppenheimer, N.J. (1978) *J. Biol. Chem.* 253, 4907–4910.
334. Sekine, A., Fujiwara, M., & Narumiya, S. (1989) *J. Biol. Chem.* 264, 8602–8605.
335. West, R.E., Jr., Moss, J., Vaughan, M., Liu, T., & Liu, T.Y. (1985) *J. Biol. Chem.* 260, 14428–14430.
336. Demurcia, G., & Demurcia, J.M. (1994) *Trends Biochem. Sci.* 19, 172–176.
337. Lindahl, T., Satoh, M.S., Poirier, G.G., & Klungland, A. (1995) *Trends Biochem. Sci.* 20, 405–411.
338. Pappenheimer, A.M., Jr. (1977) *Annu. Rev. Biochem.* 46, 69–94.
339. Oppenheimer, N.J., & Bodley, J.W. (1981) *J. Biol. Chem.* 256, 8579–8581.
340. Shapiro, B.M., & Stadtman, E.R. (1968) *J. Biol. Chem.* 243, 3769–3771.
341. Adler, S.P., Purich, D., & Stadtman, E.R. (1975) *J. Biol. Chem.* 250, 6264–6272.
342. Dolinger, D.L., Schramm, V.L., & Shockman, G.D. (1988) *Proc. Natl. Acad. Sci. U.S.A.* 85, 6667–6671.
343. Harding, H.W., & Rogers, G.E. (1976) *Biochim. Biophys. Acta* 427, 315–324.
344. Sletten, K., Aakesson, I., & Alvsaker, J.O. (1971) *Nat. New Biol.* 231, 118–119.
345. Midelfort, C.F., & Mehler, A.H. (1972) *Proc. Natl. Acad. Sci. U.S.A.* 69, 1816–1819.
346. Robinson, A.B., Scotchler, J.W., & McKerrow, J.H. (1973) *J. Am. Chem. Soc.* 95, 8156–8159.
347. Wickner, R.B. (1969) *J. Biol. Chem.* 244, 6550–6552.
348. Givot, I.L., Smith, T.A., & Abeles, R.H. (1969) *J. Biol. Chem.* 244, 6341–6353.
349. Langer, M., Lieber, A., & Retey, J. (1994) *Biochemistry* 33, 14034–14038.
350. Niwa, H., Inouye, S., Hirano, T., Matsuno, T., Kojima, S., Kubota, M., Ohashi, M., & Tsuji, F.I. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 13617–13622.
351. Ormo, M., Cubitt, B., Kallio, K., Gross, L.A., Tsien, R.Y., & Remington, S.J. (1996) *Science* 273, 1392–1395.
352. Nakajima, T., & Ballou, C.E. (1974) *J. Biol. Chem.* 249, 7685–7694.
353. Muir, L., & Lee, Y.C. (1969) *J. Biol. Chem.* 244, 2343–2349.
354. Hallgren, P., Lundblad, A., & Svensson, S. (1975) *J. Biol. Chem.* 250, 5312–5314.
355. Spiro, R.G. (1967) *J. Biol. Chem.* 242, 4813–4823.
356. Lindahl, V., & Róden, L. (1972) in *Glycoproteins: Their Composition, Structure, and Function*, 2nd ed. (Gottschalk, A., Ed.) pp 491–517, Elsevier, Amsterdam.
357. Lindahl, U., & Róden, L. (1966) *J. Biol. Chem.* 241, 2113–2119.
358. Lote, C.J., & Weiss, J.B. (1971) *FEBS Lett.* 16, 81–85.
359. Weiss, J.B., Lote, C.J., & Bobinski, H. (1971) *Nat. New Biol.* 234, 25–26.
360. Miller, D.H., Lamport, D.T., & Miller, M. (1972) *Science* 176, 918–920.
361. Torres, C.R., & Hart, G.W. (1984) *J. Biol. Chem.* 259, 3308–3317.
362. Hart, G.W. (1997) *Annu. Rev. Biochem.* 66, 315–335.
363. Kieliszewski, M.J., O'Neill, M., Leykam, J., & Orlando, R. (1995) *J. Biol. Chem.* 270, 2541–2549.
364. Hase, S., Nishimura, H., Kawabata, S., Iwanaga, S., & Ikenaka, T. (1990) *J. Biol. Chem.* 265, 1858–1861.
365. Rodriguez, I.R., & Whelan, W.J. (1985) *Biochem. Biophys. Res. Commun.* 132, 829–836.
366. Klostermeyer, H., Rabbal, K., & Reimerdes, E.H. (1976) *Hoppe-Seyler's Z. Physiol. Chem.* 357, 1197–1199.
367. Pisano, J.J., Finlayson, J.S., & Peyton, M.P. (1969) *Biochemistry* 8, 871–876.
368. Harding, H.W., & Rogers, G.E. (1971) *Biochemistry* 10, 624–630.
369. Sottrup-Jensen, L., Petersen, T.E., & Magnusson, S. (1980) *FEBS Lett.* 121, 275–279.
370. Thomas, M.L., Davidson, F.F., & Tack, B.F. (1983) *J. Biol. Chem.* 258, 13580–13586.
371. Klabunde, T., Eicken, C., Sacchettini, J.C., & Krebs, B. (1998) *Nat. Struct. Biol.* 5, 1084–1090.
372. Lerch, K. (1982) *J. Biol. Chem.* 257, 6414–6419.
373. Ito, N., Phillips, S.E., Stevens, C., Ogel, Z.B., McPherson, M.J., Keen, J.N., Yadav, K.D., & Knowles, P.F. (1991) *Nature* 350, 87–90.
374. Ito, N., Phillips, S.E., Yadav, K.D., & Knowles, P.F. (1994) *J. Mol. Biol.* 238, 794–814.
375. Yoshikawa, S., Shinzawa-Itoh, K., Nakashima, R., Yaono, R., Yamashita, E., Inoue, N., Yao, M., Fei, M.J.,

- Libeu, C.P., Mizushima, T., Yamaguchi, H., Tomizaki, T., & Tsukihara, T. (1998) *Science* 280, 1723–1729.
376. Proshlyakov, D.A., Pressler, M.A., DeMaso, C., Leykam, J.F., DeWitt, D.L., & Babcock, G.T. (2000) *Science* 290, 1588–1591.
377. LaBella, F., Keeley, F., Vivian, S., & Thornhill, D. (1967) *Biochem. Biophys. Res. Commun.* 26, 748–753.
378. Andersen, S.O. (1966) *Acta Physiol. Scand., Suppl.* 263, 1–81.
379. Michon, T., Chenu, M., Kellershon, N., Desmadril, M., & Gueguen, J. (1997) *Biochemistry* 36, 8504–8513.
380. Fry, S.C. (1982) *Biochem. J.* 204, 449–455.
381. Kanwar, R., & Balasubramanian, D. (2000) *Biochemistry* 39, 14976–14983.
382. Andersen, S.O. (1967) *Nature* 216, 1029–1030.
383. Fujimoto, D., Horiuchi, K., & Hirama, M. (1981) *Biochem. Biophys. Res. Commun.* 99, 637–643.
384. Nomura, K., Suzuki, N., & Matsumoto, S. (1990) *Biochemistry* 29, 4525–4534.
385. McIntire, W.S., Wemmer, D.E., Chistoserdov, A., & Lidstrom, M.E. (1991) *Science* 252, 817–824.
386. Chen, L., Durley, R., Poliks, B.J., Hamada, K., Chen, Z., Mathews, F.S., Davidson, V.L., Satow, Y., Huizinga, E., Vellieux, F.M., et al. (1992) *Biochemistry* 31, 4959–4964.
387. Wang, S.X., Mure, M., Medzihradsky, K.F., Burlingame, A.L., Brown, D.E., Dooley, D.M., Smith, A.J., Kagan, H.M., & Klinman, J.P. (1996) *Science* 273, 1078–1084.
388. Margoliash, E., & Schejter, A. (1966) *Adv. Protein Chem.* 21, 113–286.
389. Williams, V.P., & Glazer, A.N. (1978) *J. Biol. Chem.* 253, 202–211.
390. Ficner, R., Lobeck, K., Schmidt, G., & Huber, R. (1992) *J. Mol. Biol.* 228, 935–950.
391. Walker, W.H., Kenney, W.C., Edmondson, D.E., Singer, T.P., Cronin, J.R., & Hendriks, R. (1974) *Eur. J. Biochem.* 48, 439–448.
392. Edmondson, D.E., & Singer, T.P. (1976) *FEBS Lett.* 64, 255–265.
393. Singer, T.P., & Edmondson, D.E. (1974) *FEBS Lett.* 42, 1–14.
394. Mewies, M., Basran, J., Packman, L.C., Hille, R., & Scrutton, N.S. (1997) *Biochemistry* 36, 7162–7168.
395. Steenkamp, D.J., McIntire, W., & Kenney, W.C. (1978) *J. Biol. Chem.* 253, 2818–2824.
396. Willie, A., Edmondson, D.E., & Jorns, M.S. (1996) *Biochemistry* 35, 5292–5299.
397. McIntire, W., Edmondson, D.E., Singer, T.P., & Hopper, D.J. (1980) *J. Biol. Chem.* 255, 6553–6555.
398. Cunane, L.M., Chen, Z.W., Shamala, N., Mathews, F.S., Cronin, C.N., & McIntire, W.S. (2000) *J. Mol. Biol.* 295, 357–374.
399. Maloy, W.L., Bowien, B.U., Zwolinski, G.K., Kumar, K.G., Wood, H.G., Ericsson, L.H., & Walsh, K.A. (1979) *J. Biol. Chem.* 254, 11615–11622.
400. Hale, G., & Perham, R.N. (1980) *Biochem. J.* 187, 905–908.
401. Piszkiwicz, D., Landon, M., & Smith, E.L. (1970) *J. Biol. Chem.* 245, 2622–2626.
402. Tanase, S., Kojima, H., & Morino, Y. (1979) *Biochemistry* 18, 3002–3007.
403. Akhtar, M., Blosser, P.T., & Dewhurst, P.B. (1968) *Biochem. J.* 110, 693–702.
404. Bownds, D. (1967) *Nature* 216, 1178–1181.
405. Vanaman, T.C., Wakil, S.J., & Hill, R.L. (1968) *J. Biol. Chem.* 243, 6420–6431.
406. Igarashi, N., Moriyama, H., Fujiwara, T., Fukumori, Y., & Tanaka, N. (1997) *Nat. Struct. Biol.* 4, 276–284.
407. Robinson, J.B., Jr., Singh, M., & Srere, P.A. (1976) *Proc. Natl. Acad. Sci. U.S.A.* 73, 1872–1876.
408. Berg, M., Hilbi, H., & Dimroth, P. (1996) *Biochemistry* 35, 4689–4696.
409. Hoenke, S., Wild, M.R., & Dimroth, P. (2000) *Biochemistry* 39, 13223–13232.
410. Kivirikko, K.I., & Pihlajaniemi, T. (1998) *Adv. Enzymol. Relat. Areas Mol. Biol.* 72, 325–398.
411. Esmon, C.T., Sadowski, J.A., & Suttie, J.W. (1975) *J. Biol. Chem.* 250, 4744–4748.
412. Price, P.A., Poser, J.W., & Raman, N. (1976) *Proc. Natl. Acad. Sci. U.S.A.* 73, 3374–3375.
413. Ma, Y.A., Sih, C.J., & Harms, A. (1999) *J. Am. Chem. Soc.* 121, 8967–8968.
414. Moss, J., Stanley, S.J., & Watkins, P.A. (1980) *J. Biol. Chem.* 255, 5838–5840.
415. Yost, D.A., & Moss, J. (1983) *J. Biol. Chem.* 258, 4926–4929.
416. Takada, T., Iida, K., & Moss, J. (1993) *J. Biol. Chem.* 268, 17837–17843.
417. Scaife, R.M., Wilson, L., & Purich, D.L. (1992) *Biochemistry* 31, 310–316.
418. Resing, K.A., Johnson, R.S., & Walsh, K.A. (1995) *Biochemistry* 34, 9477–9487.
419. Chlumsky, L.J., Sturgess, A.W., Nieves, E., & Jorns, M.S. (1998) *Biochemistry* 37, 2089–2095.
420. Michel, H., Hunt, D.F., Shabanowitz, J., & Bennett, J. (1988) *J. Biol. Chem.* 263, 1123–1130.
421. Rudiger, M., Plessmann, U., Rudiger, A.H., & Weber, K. (1995) *FEBS Lett.* 364, 147–151.
422. Dreger, M., Otto, H., Neubauer, G., Mann, M., & Hucho, F. (1999) *Biochemistry* 38, 9426–9434.
423. Wu, X., Takahashi, M., Chen, S.G., & Monnier, V.M. (2000) *Biochemistry* 39, 1515–1521.
424. Zhang, X., Herring, C.J., Romano, P.R., Szczepanowska, J., Brzeska, H., Hinnebusch, A.G., & Qin, J. (1998) *Anal. Chem.* 70, 2050–2059.
425. Folk, J.E., & Finlayson, J.S. (1977) *Adv. Protein Chem.* 31, 1–133.
426. Gielens, C., De Geest, N., Xin, X.Q., Devreese, B., Van Beeumen, J., & Preaux, G. (1997) *Eur. J. Biochem.* 248, 879–888.
427. Williamson, P.R., & Kagan, H.M. (1986) *J. Biol. Chem.* 261, 9477–9482.
428. Toth, E.A., Worby, C., Dixon, J.E., Goedken, E.R., Marqusee, S., & Yeates, T.O. (2000) *J. Mol. Biol.* 301, 433–450.
429. Cleland, W.W. (1964) *Biochemistry* 3, 480–482.
430. Kellaris, K.V., & Ware, D.K. (1989) *Biochemistry* 28, 3469–3482.
431. Xia, Z., Dai, W., Zhang, Y., White, S.A., Boyd, G.D., & Mathews, F.S. (1996) *J. Mol. Biol.* 259, 480–501.
432. Blake, C.C., Ghosh, M., Harlos, K., Avezoux, A., & Anthony, C. (1994) *Nat. Struct. Biol.* 1, 102–105.
433. White, S., Boyd, G., Mathews, F.S., Xia, Z.X., Dai, W.W.,

- Zhang, Y.F., & Davidson, V.L. (1993) *Biochemistry* 32, 12955–12958.
434. Wang, X., Connor, M., Smith, R., Maciejewski, M.W., Howden, M.E., Nicholson, G.M., Christie, M.J., & King, G.F. (2000) *Nat. Struct. Biol.* 7, 505–513.
 435. Chandrasekaran, R., & Balasubramanian, R. (1969) *Biochim. Biophys. Acta* 188, 1–9.
 436. Strater, N., Klabunde, T., Tucker, P., Witzel, H., & Krebs, B. (1995) *Science* 268, 1489–1492.
 437. Frech, C., & Schmid, F.X. (1995) *J. Mol. Biol.* 251, 135–149.
 438. Darby, N.J., & Creighton, T.E. (1993) *J. Mol. Biol.* 232, 873–896.
 439. Lal, M., Rao, R., Fang, X.W., Schuchmann, H.P., & vonSonntag, C. (1997) *J. Am. Chem. Soc.* 119, 5735–5739.
 440. De Lorenzo, F., Goldberger, R.F., Steers, E., Jr., Givol, D., & Anfinsen, B. (1966) *J. Biol. Chem.* 241, 1562–1567.
 441. Akiyama, Y., Kamitani, S., Kusakawa, N., & Ito, K. (1992) *J. Biol. Chem.* 267, 22440–22445.
 442. Bardwell, J.C., McGovern, K., & Beckwith, J. (1991) *Cell* 67, 581–589.
 443. Hirano, N., Shibasaki, F., Sakai, R., Tanaka, T., Nishida, J., Yazaki, Y., Takenawa, T., & Hirai, H. (1995) *Eur. J. Biochem.* 234, 336–342.
 444. McCarthy, A.A., Haebel, P.W., Torronen, A., Rybin, V., Baker, E.N., & Metcalf, P. (2000) *Nat. Struct. Biol.* 7, 196–199.
 445. Zapun, A., Bardwell, J.C., & Creighton, T.E. (1993) *Biochemistry* 32, 5083–5092.
 446. Kortemme, T., Darby, N.J., & Creighton, T.E. (1996) *Biochemistry* 35, 14503–14511.
 447. Spackman, D.H., Stein, W.H., & Moore, S. (1960) *J. Biol. Chem.* 235, 648–659.
 448. Haniu, M., Horan, T., Arakawa, T., Le, J., Katta, V., Hara, S., & Rohde, M.F. (1996) *Biochemistry* 35, 13040–13046.
 449. Hoffman, R.C., Andersen, H., Walker, K., Krakover, J.D., Patel, S., Stamm, M.R., & Osborn, S.G. (1996) *Biochemistry* 35, 14849–14861.
 450. McMullen, B.A., Fujikawa, K., & Davie, E.W. (1991) *Biochemistry* 30, 2050–2056.
 451. Burman, S., Wellner, D., Chait, B., Chaudhary, T., & Breslow, E. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 429–433.
 452. White, C.E., Hunter, M.J., Meininger, D.P., Garrod, S., & Komives, E.A. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 10177–10182.
 453. Gray, W.R. (1993) *Protein Sci.* 2, 1732–1748.
 454. Kellaris, K.V., Ware, D.K., Smith, S., & Kyte, J. (1989) *Biochemistry* 28, 3469–3482.
 455. Thompson, S.A. (1992) *J. Biol. Chem.* 267, 2269–2273.
 456. Eriksson, A.E., Cousens, L.S., Weaver, L.H., & Matthews, B.W. (1991) *Proc. Natl. Acad. Sci. U.S.A.* 88, 3441–3445.
 457. Sueyoshi, T., Miyata, T., Iwanaga, S., Toyo'oka, T., & Imai, K. (1985) *J. Biochem. (Tokyo)* 97, 1811–1813.
 458. Marti, T., Rosselet, S.J., Titani, K., & Walsh, K.A. (1987) *Biochemistry* 26, 8099–8109.
 459. Hojrup, P., & Magnusson, S. (1987) *Biochem. J.* 245, 887–891.
 460. Robertson, J.G., Adams, G.W., Medzihradsky, K.F., Burlingame, A.L., & Villafranca, J.J. (1994) *Biochemistry* 33, 11563–11575.
 461. Poerio, E., Caporale, C., Carrano, L., Pucci, P., & Buonocore, V. (1991) *Eur. J. Biochem.* 199, 595–600.
 462. Solouki, T., Emmett, M.R., Guan, S., & Marshall, A.G. (1997) *Anal. Chem.* 69, 1163–1168.
 463. Shapiro, B.M., Kingdon, H.S., & Stadtman, E.R. (1967) *Proc. Natl. Acad. Sci. U.S.A.* 58, 642–649.
 464. Schwede, T.F., Retey, J., & Schulz, G.E. (1999) *Biochemistry* 38, 5355–5361.
 465. Wall, M.A., Socolich, M., & Ranganathan, R. (2000) *Nat. Struct. Biol.* 7, 1133–1138.
 466. Sardana, M., Sardana, V., Rodkey, J., Wood, T., Ng, A., Vlasuk, G.P., & Waxman, L. (1991) *J. Biol. Chem.* 266, 13560–13563.
 467. Imberty, A., Chanzy, H., Perez, S., Buleon, A., & Tran, V. (1988) *J. Mol. Biol.* 201, 365–378.
 468. Campbell, D.G., & Cohen, P. (1989) *Eur. J. Biochem.* 185, 119–125.
 469. Mellis, S.J., & Baenziger, J.U. (1983) *J. Biol. Chem.* 258, 11546–11556.
 470. Podolsky, D.K. (1985) *J. Biol. Chem.* 260, 15510–15515.
 471. Sturm, A. (1991) *Eur. J. Biochem.* 199, 169–179.
 472. Mattei, B., Bernalda, M.S., Federici, L., Roepstorff, P., Cervone, F., & Boffi, A. (2001) *Biochemistry* 40, 569–576.
 473. Spiro, R.G., & Bhoyroo, V.D. (1988) *J. Biol. Chem.* 263, 14351–14358.
 474. Roux, L., Holojda, S., Sundblad, G., Freeze, H.H., & Varki, A. (1988) *J. Biol. Chem.* 263, 8879–8889.
 475. Hard, K., Van Doorn, J.M., Thomas-Oates, J.E., Kamerling, J.P., & Van der Horst, D.J. (1993) *Biochemistry* 32, 766–775.
 476. Nadano, D., Iwasaki, M., Endo, S., Kitajima, K., Inoue, S., & Inoue, Y. (1986) *J. Biol. Chem.* 261, 11550–11557.
 477. Angata, T., Nakata, D., Matsuda, T., Kitajima, K., & Troy, F.A. (1999) *J. Biol. Chem.* 274, 22949–22956.
 478. Manzi, A.E., Dell, A., Azadi, P., & Varki, A. (1990) *J. Biol. Chem.* 265, 8094–8107.
 479. Takeuchi, M., Inoue, N., Strickland, T.W., Kubota, M., Wada, M., Shimizu, R., Hoshi, S., Kozutsumi, H., Takasaki, S., & Kobata, A. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 7819–7822.
 480. van Hoek, A.N., Wiener, M.C., Verbavatz, J.M., Brown, D., Lipniunas, P.H., Townsend, R.R., & Verkman, A.S. (1995) *Biochemistry* 34, 2212–2219.
 481. Gerken, T.A., Butenhof, K.J., & Shogren, R. (1989) *Biochemistry* 28, 5536–5543.
 482. Homans, S.W., Dwek, R.A., & Rademacher, T.W. (1987) *Biochemistry* 26, 6553–6560.
 483. Wu, A.M., Kabat, E.A., Nilsson, B., Zopf, D.A., Gruezo, F.G., & Liao, J. (1984) *J. Biol. Chem.* 259, 7178–7186.
 484. Li, E., Tabas, I., & Kornfeld, S. (1978) *J. Biol. Chem.* 253, 7762–7770.
 485. Fournet, B., Montreuil, J., Strecker, G., Dorland, L., Haverkamp, J., Vliegthart, F.G., Binette, J.P., & Schmid, K. (1978) *Biochemistry* 17, 5206–5214.
 486. Green, E.D., & Baenziger, J.U. (1988) *J. Biol. Chem.* 263, 25–35.
 487. Trimble, R.B., Atkinson, P.H., Tschopp, J.F., Townsend, R.R., & Maley, F. (1991) *J. Biol. Chem.* 266, 22807–22817.
 488. Ballou, L., Hernandez, L.M., Alvarado, E., & Ballou, C.E. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87, 3368–3372.

489. Van Kuik, J.A., Van Halbeek, H., Kamerling, J.P., & Vliegthart, J.F. (1986) *Eur. J. Biochem.* 159, 297–301.
490. Yamashita, K., Inui, K., Totani, K., Kochibe, N., Furukawa, M., & Okada, S. (1990) *Biochemistry* 29, 3030–3039.
491. Shoji, H., Takahashi, N., Nomoto, H., Ishikawa, M., Shimada, I., Arata, Y., & Hayashi, K. (1992) *Eur. J. Biochem.* 207, 631–641.
492. Pfeiffer, G., Dabrowski, U., Dabrowski, J., Stirm, S., Strube, K.H., & Geyer, R. (1992) *Eur. J. Biochem.* 205, 961–978.
493. Bendiak, B., Harris-Brandts, M., Michnick, S.W., Carver, J.P., & Cumming, D.A. (1989) *Biochemistry* 28, 6491–6499.
494. Nakata, N., Furukawa, K., Greenwalt, D.E., Sato, T., & Kobata, A. (1993) *Biochemistry* 32, 4369–4383.
495. Knibbs, R.N., Perini, F., & Goldstein, I.J. (1989) *Biochemistry* 28, 6379–6392.
496. Rudd, P.M., Downing, A.K., Cadene, M., Harvey, D.J., Wormald, M.R., Weir, I., Dwek, R.A., Rifkin, D.B., & Gleizes, P.E. (2000) *Biochemistry* 39, 1596–1603.
497. van Kuik, J.A., van Halbeek, H., Kamerling, J.P., & Vliegthart, J.F. (1985) *J. Biol. Chem.* 260, 13984–13988.
498. Strous, G.J. (1979) *Proc. Natl. Acad. Sci. U.S.A.* 76, 2694–2698.
499. Nishimura, H., Takao, T., Hase, S., Shimonishi, Y., & Iwanaga, S. (1992) *J. Biol. Chem.* 267, 17520–17525.
500. Harris, R.J., van Halbeek, H., Glushka, J., Basa, L.J., Ling, V.T., Smith, K.J., & Spellman, M.W. (1993) *Biochemistry* 32, 6539–6547.
501. Mellis, S.J., & Baenziger, J.U. (1983) *J. Biol. Chem.* 258, 11557–11563.
502. Dua, V.K., Rao, B.N., Wu, S.S., Dube, V.E., & Bush, C.A. (1986) *J. Biol. Chem.* 261, 1599–1608.
503. Thomas, D.B., & Winzler, R.J. (1969) *J. Biol. Chem.* 244, 5943–5946.
504. Slomiany, B.L., Murty, V.L., & Slomiany, A. (1980) *J. Biol. Chem.* 255, 9719–9723.
505. Adamany, A.M., Blumenfeld, O.O., Sabo, B., & McCreary, J. (1983) *J. Biol. Chem.* 258, 11537–11545.
506. Klein, A., Carnoy, C., Lamblin, G., Roussel, P., van Kuik, J.A., de Waard, P., & Vliegthart, J.F. (1991) *Eur. J. Biochem.* 198, 151–168.
507. Capon, C., Leroy, Y., Wieruszkeski, J.M., Ricart, G., Strecker, G., Montreuil, J., & Fournet, B. (1989) *Eur. J. Biochem.* 182, 139–152.
508. Hounsell, E.F., Lawson, A.M., Stoll, M.S., Kane, D.P., Cashmore, G.C., Carruthers, R.A., Feeney, J., & Feizi, T. (1989) *Eur. J. Biochem.* 186, 597–610.
509. Byrd, J.C., Nardelli, J., Siddiqui, B., & Kim, Y.S. (1988) *Cancer Res.* 48, 6678–6685.
510. Bobek, L.A., Tsai, H., Biesbrock, A.R., & Levine, M.J. (1993) *J. Biol. Chem.* 268, 20563–20569.
511. Lan, M.S., Batra, S.K., Qi, W.N., Metzgar, R.S., & Hollingsworth, M.A. (1990) *J. Biol. Chem.* 265, 15294–15299.
512. Bhargava, A.K., Woitach, J.T., Davidson, E.A., & Bhavanandan, V.P. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87, 6798–6802.
513. Maimone, M.M., & Tollefsen, D.M. (1990) *J. Biol. Chem.* 265, 18263–18271.
514. Safaiyan, F., Lindahl, U., & Salmivirta, M. (2000) *Biochemistry* 39, 10823–10830.
515. Nilsson, B., Nakazawa, K., Hassell, J.R., Newsome, D.A., & Hascall, V.C. (1983) *J. Biol. Chem.* 258, 6056–6063.
516. Gunnarsson, A., Svensson, S., & Roden, L. (1984) *Carbohydr. Res.* 133, 75–82.
517. Pluschke, G., Vanek, M., Evans, A., Dittmar, T., Schmid, P., Itin, P., Filardo, E.J., & Reisfeld, R.A. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 9710–9715.
518. Campbell, S.C., Krueger, R.C., & Schwartz, N.B. (1990) *Biochemistry* 29, 907–914.
519. Zimmermann, D.R., & Ruoslahti, E. (1989) *EMBO J.* 8, 2975–2981.
520. Avraham, S., Stevens, R.L., Gartner, M.C., Austen, K.F., Lalley, P.A., & Weis, J.H. (1988) *J. Biol. Chem.* 263, 7292–7296.
521. Kornfeld, S., Li, E., & Tabas, I. (1978) *J. Biol. Chem.* 253, 7771–7778.
522. Leonard, C.K., Spellman, M.W., Riddle, L., Harris, R.J., Thomas, J.N., & Gregory, T.J. (1990) *J. Biol. Chem.* 265, 10373–10382.
523. Spiro, R.G. (1970) *Annu. Rev. Biochem.* 39, 599–638.
524. Tai, T., Yamashita, K., Ogata-Arakawa, M., Koide, N., & Muramatsu, T. (1975) *J. Biol. Chem.* 250, 8569–8575.
525. Takasaki, S., Mizuochi, T., & Kobata, A. (1982) *Methods Enzymol.* 83, 263–268.
526. Hardy, M.R., & Townsend, R.R. (1988) *Proc. Natl. Acad. Sci. U.S.A.* 85, 3289–3293.
527. Hardy, M.R., Townsend, R.R., & Lee, Y.C. (1988) *Anal. Biochem.* 170, 54–62.
528. Lacourse, W.R., & Johnson, D.C. (1993) *Anal. Chem.* 65, 50–55.
529. Davidson, D.J., & Castellino, F.J. (1991) *Biochemistry* 30, 625–633.
530. Zhao, Y., Kent, S.B., & Chait, B.T. (1997) *Proc. Natl. Acad. Sci. U.S.A.* 94, 1629–1633.
531. Hakomori, S. (1964) *J. Biochem. (Tokyo)* 55, 205–208.
532. Nomoto, H., Takahashi, N., Nagaki, Y., Endo, S., Arata, Y., & Hayashi, K. (1986) *Eur. J. Biochem.* 157, 233–242.
533. Damm, J.B., Voshol, H., Hard, K., Kamerling, J.P., & Vliegthart, J.F. (1989) *Eur. J. Biochem.* 180, 101–110.
534. Sturm, A., Bergwerff, A.A., & Vliegthart, J.F. (1992) *Eur. J. Biochem.* 204, 313–316.
535. Kitagawa, H., Nakada, H., Kurosaka, A., Hiraiwa, N., Numata, Y., Fukui, S., Funakoshi, I., Kawasaki, T., Yamashina, I., Shimada, I., & et al. (1989) *Biochemistry* 28, 8891–8897.
536. Sasaki, H., Ochi, N., Dell, A., & Fukuda, M. (1988) *Biochemistry* 27, 8618–8626.
537. Nemeth, J.F., Hochgesang, G.P., Jr., Marnett, L.J., Caprioli, R.M., & Hochensang, G.P., Jr. (2001) *Biochemistry* 40, 3109–3116.
538. Shriver, Z., Raman, R., Venkataraman, G., Drummond, K., Turnbull, J., Toida, T., Linhardt, R., Biemann, K., & Sasisekharan, R. (2000) *Proc. Natl. Acad. Sci. U.S.A.* 97, 10359–10364.
539. Aspinall, G.O., McDonald, A.G., Pang, H., Kurjanczyk, L.A., & Penner, J.L. (1994) *Biochemistry* 33, 241–249.
540. Kitagawa, H., Nakada, H., Fukui, S., Funakoshi, I., Kawasaki, T., Yamashina, I., Tate, S., & Inagaki, F. (1991) *Biochemistry* 30, 2869–2876.

- 541. Kocharova, N.A., Knirel, Y.A., Widmalm, G., Jansson, P.E., & Moran, A.P. (2000) *Biochemistry* 39, 4755–4760.
- 542. Sasaki, H., Bothner, B., Dell, A., & Fukuda, M. (1987) *J. Biol. Chem.* 262, 12059–12076.
- 543. Misaki, A., & Goldstein, I.J. (1977) *J. Biol. Chem.* 252, 6995–6999.
- 544. Arima, T., & Spiro, R.G. (1972) *J. Biol. Chem.* 247, 1836–1848.
- 545. Drew, H.R., Wing, R.M., Takano, T., Broka, C., Tanaka, S., Itakura, K., & Dickerson, R.E. (1981) *Proc. Natl. Acad. Sci. U.S.A.* 78, 2179–2183.

Chapter 4

Crystallographic Molecular Models

To this point, it has been described how proteins are composed of long polymers of amino acids and how these polymers are posttranslationally modified by processes that alter the backbone of the polypeptide or the side chains of the amino acids or that add oligosaccharides to the polypeptides. All of the specific covalent bonds connecting all of the atoms in each of the posttranslationally altered polypeptides composing a particular protein can be defined by chemical analysis. The bond lengths and fixed bond angles of the monomers, amino acids and monosaccharides, and of the bonds coupling the monomers into polymers, amides and acetals, are known precisely. With these values, every bond length, the hybridization of every atom, and every fixed bond angle in each complete, posttranslationally modified polypeptide can be assigned unambiguously. From this information a long flexible molecular model of a particular posttranslationally modified polypeptide can be constructed with high precision.

The problem with defining the complete structure of any polymer, polypeptides included, is the rotational degrees of freedom about the large number of exocyclic, unconjugated single bonds that are present in the polymer. In a finished polypeptide there are from hundreds to tens of thousands of such single bonds. In a commercial polymer, such as polystyrene, rotation about its many single bonds causes each molecule of the polymer, even though it may be covalently identical to other molecules of the polymer in the sample, to assume a different three-dimensional structure, and if the polymer is in solution, the structure of each molecule usually changes constantly and randomly with time. The polypeptides in a protein, however, assume only one unchanging structure, or a small number of interchanging structures, uniquely determined by the amino acid sequences of those polypeptides. Each molecule of the same protein assumes the same or one of a small number of three-dimensional structures. This structure or these few structures are exclusively assumed because almost all of the exocyclic single bonds composing the backbones of the polymer and most of the exocyclic single bonds of the side chains of the amino acids are confined to particular dihedral angles. It is crystals of proteins that have provided both this insight and the opportunity to observe molecular models representing these structures.

The existence of a crystal of any protein permits certain conclusions to be drawn about that protein. As in

organic chemistry, it can be concluded that the molecules in the crystal are all covalently identical or almost identical to each other. Furthermore, if a crystal exists, the covalently identical molecules can be present only in a small number of specific three-dimensional conformations. In the case of proteins, all of the molecules in the crystal usually have the same structure or one of a small number of almost identical structures. It is now also known that the structure of a molecule of protein in a crystal is essentially identical to its only structure or one of its few structures when it is free in solution. When the crystal is submitted to X-ray crystallography, that unique structure can be observed.

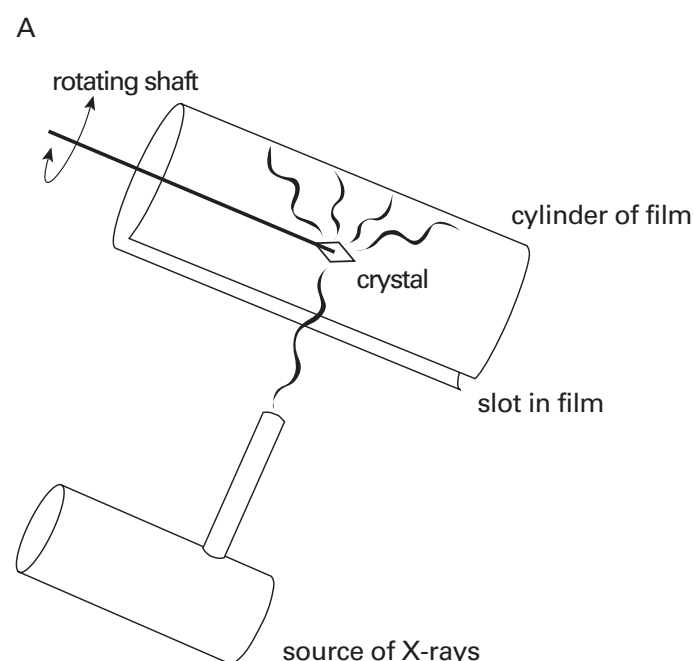
*Maps of Electron Density*¹

Suppose that one could see X-radiation. If one were to pick up a crystal of a purified protein and tumble it in his hand under a beam of X-radiation of one wavelength, it would glitter as does a jewel in a ray of sunlight. There would be, however, a peculiarity to this glitter. A jewel glitters because its facets reflect the sunlight as small individual mirrors. This means that if one follows a facet carefully as the jewel turns, one would see that it is always reflecting the sunlight and realize that the glittering sensation only arises because the eye is at rest with respect to the moving reflected beam. The glitter from a crystal of protein, however, arises because its facets produce flashes, and these flashes occur only when a facet is aligned in one precise direction relative to the direction of the incident beam of X-radiation. The reason for this is that the flashes are produced by the summation in phase of the reflections from a stack of evenly spaced, parallel mirrors. This summation in phase is **diffraction**. It is only at certain angles that the reflections sum in phase.

If one played with the crystal of protein long enough, it would become clear that there were axes running through it. Rotation about any one of these axes would produce flashes that were regularly arrayed. This regular array of flashes would be reminiscent of the array of reflections that emanates from one of the rotating mirrored spheres in a ballroom. One difference, however, would be that while each mirror on the sphere continuously reflects the spotlight when it is on the illuminated side, as can be discerned by following the reflected beams on the walls, each of the mirrors in the crystal of

protein reflects only when it passes through certain precise orientations, as could also be discerned by watching the patterns of the flashes on the walls. In addition, the mirrors in the rotating crystal, referred to as the **reflecting faces**, reflect onto the walls behind the crystal as well as in front of the crystal because the crystal is not opaque to X-radiation, as is the ballroom sphere to light, and both sides of each mirror can reflect.

The easiest way to verify this behavior is photographically. A crystal is mounted on the end of a rotating shaft the axis of which is coincident with the axis of a cylinder of photographic film. The crystal is attached to the shaft in an orientation such that one of its principal axes is parallel to the axis around which the shaft is rotating. The cylinder of film has a slot through which a beam of X-radiation perpendicular to the axis of rotation can be directed upon the crystal (Figure 4-1A).² After an appropriate exposure, the film is developed. The image observed is that of reflected flashes arrayed on lines of latitude (Figure 4-1B). Each line of latitude, referred to as a **layer line**, arises from all of the mirrors that are tilted at the same angle with respect to the beam of incident X-radiation. Because the spots on the film produced by the flashes occur along layer lines, the tilt of the mirrors relative to the axis of the crystal must be able to assume only certain values. Because the layer lines are made up of discrete spots, each the result of one flash, each mirror must reflect only when the angle between its face and the incident beam assumes unique values.



The profound insight into this curious phenomenon was the realization that the remarkable variations in the intensities of the flashes (Figure 4-1B) contained information and that, from the information they contained, the atomic structure of the molecules from which the crystal was formed could be deduced. With the promise that this is the reward, one can now ask, what are these mirrors, why do they flash, and why does each one flash with a different intensity?

A **crystal of protein** is a solution to a warehousing problem. It is a solid object formed from a huge number of the same protein molecules, neatly stacked as the boxes or barrels in a warehouse, with the vacancies between the molecules of the protein filled with water. It is, for all intents and purposes, an **infinite, three-dimensional array of identical enantiomeric objects**. It can be shown that there are only 71 ways to arrange enantiomeric objects to form an infinite array. Each crystal represents a particular one of these 71 solutions.

Each of these 71 different arrangements can be divided in its entirety into a stack of boxes, each of which is identical in its size, shape, contents, and the arrangement of its contents to every other one. These boxes are always parallelepipeds, and they are referred to as unit cells. A **unit cell** is the smallest parallelepiped of matter that, by only simple translational movements along the three axes of the crystal, can be stacked to create and fill completely the whole crystal. Keep in mind that each of these parallelepipeds is filled with molecules of protein,

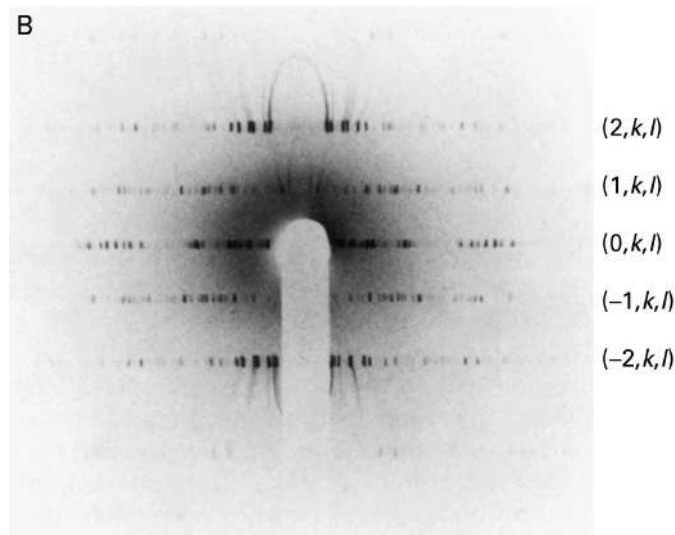


Figure 4-1: (A) Schematic drawing of a camera used to take an oscillation photograph of a crystal turning about one of its crystallographic axes. (B) A photograph from such a camera.² The crystal was aligned such that one axis of the unit cell was perpendicular to the beam of X-radiation, and the crystal was then rotated back and forth around this vertical axis back. Each of these oscillations covered the same excursion of about 20°. The axis of rotation was aligned vertically with respect to the film as it is displayed. The white shadow in the center of the photograph is of a beam stop used to protect the film from the majority of the X-radiation, which passes through and around the crystal. The beam was pointed at the circular top of the beam stop. The five layer lines are labeled as if the rotation had occurred around the **a** axis of the crystal. The middle layer line $(0,k,l)$ is the equator. Reprinted with permission from ref 2. Copyright 1968 Macmillan.

and surrounding water, that are necessarily arranged in space in certain positions and orientations. It is this arrangement of molecules that exists, not the unit cells or the planes about to be discussed.

In any crystal, three sets of planes can be constructed. Within each of these three sets, all of the constituent planes must be equidistant and parallel to each other. Every one of the planes in each of the three sets must intersect planes from both of the other two sets, and every one of the parallelepipeds that results from these intersections must be the same unit cell containing the same distribution of matter. The partition of space accomplished by three sets of planes so defined is accompanied by the creation of a network of lines, each of which is the intersection of two of these planes. This network of lines is a **lattice** (Figure 4-2) encaging a set of unit cells.

Unfortunately, each crystal can be divided into several different lattices, each of which satisfies the definition. In addition, any translational movement, no matter how small, of any one of these lattices produces another equally satisfactory lattice. Usually crystallographers follow certain conventions in choosing the fundamental lattice that will be used. These conventions are designed in part to reveal any underlying symmetry that

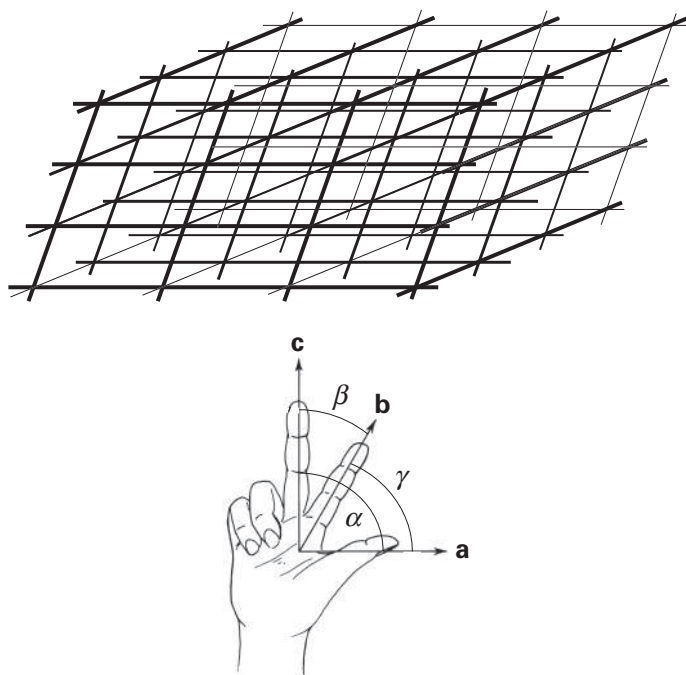


Figure 4-2: A triclinic lattice. This lattice is the most general because none of the sides of the unit cells (a , b , or c) is the same length and none of the three angles (α , β , or γ) is 90° or 120° . If $\alpha = \beta = 90^\circ$, the lattice would be **monoclinic**. If $\alpha = \beta = \gamma = 90^\circ$, the lattice would be **orthorhombic**. If $a = b$ and $\alpha = \beta = \gamma = 90^\circ$, the lattice would be **tetragonal**. If $a = b = c$ and $\alpha = \beta = \gamma \neq 90^\circ$, the lattice would be **rhombohedral**. If $a = b$, $\alpha = \beta = 90^\circ$, and $\gamma = 120^\circ$, the lattice would be **hexagonal**. If $a = b = c$ and $\alpha = \beta = \gamma = 90^\circ$, the lattice would be **cubic**. The axes are defined by the right-hand rule. The hand is reprinted with permission from ref 2. Copyright 1968 Macmillan.

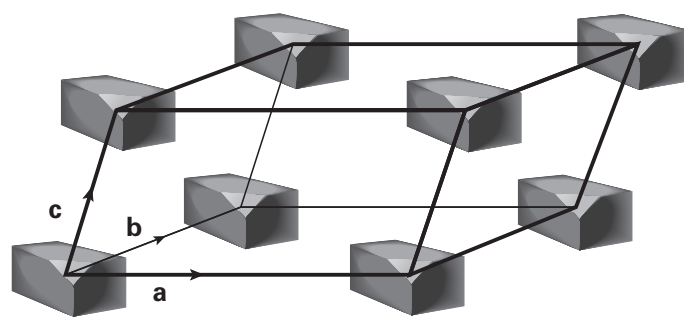


Figure 4-3: Fundamental unit cell in a triclinic lattice showing the relationship between the distribution of matter and the boundaries of the fundamental unit cell. Even if the fundamental unit cell has not been chosen to enclose one of the repeating objects, it contains a total of one complete object.

may be present within the crystal and in part to simplify the extensive calculations that are involved in producing a map of electron density. One lattice is chosen, however, during a procedure known as indexing, and this choice defines the **fundamental unit cell** (Figure 4-3). The three **axes** of the fundamental unit cell are conventionally designated a , b , and c by the right-hand rule (Figure 4-2). The length of the fundamental unit cell along each axis, in nanometers, is designated a , b , or c , respectively.

There are other ways to divide the space occupied by the crystal into different sets of unit cells by using other **sets of parallel planes**. This can be most easily seen by starting with a two-dimensional lattice (Figure 4-4). Each set of parallel lines in the figure is constructed so that its members pass through the origins of the fundamental unit cells, and the origin of each fundamental unit cell is contained in one of the lines of each set. This two-dimensional lattice can then be thought of as one of the lattice planes in a three-dimensional monoclinic crystal. In this case, the view presented in the figure would be down the c axis, and each line would be the intersection of a plane perpendicular to the page. Each set of planes parallel to each other and perpendicular to the two-dimensional lattice would create a new array of unit cells (Figure 4-5), and within a particular crystal, every one of these arrays would contain the same number of unit cells. Most of these arrays do not produce lattices because their unit cells are not formed from three intersecting sets of parallel planes, but they are arrays of genuine unit cells nevertheless. By extension it is clear that there is an infinite number of ways to divide a lattice into a set of unit cells that are bounded by at least one set of parallel planes (Figure 4-5).

Each of these sets of parallel planes is identified by giving it an **index**, (h, k, l) . The index is referred to the axes of the fundamental unit cell. From an examination of Figure 4-5, it can be seen that the parallel planes perpendicular to the page that define a given set of unit cells always intersect the axes of the fundamental unit cell at intervals that are the quotient of the length of the fundamental unit cell along that axis (a , b , or c , respectively) and an integer. As the tilt of the planes relative to that axis

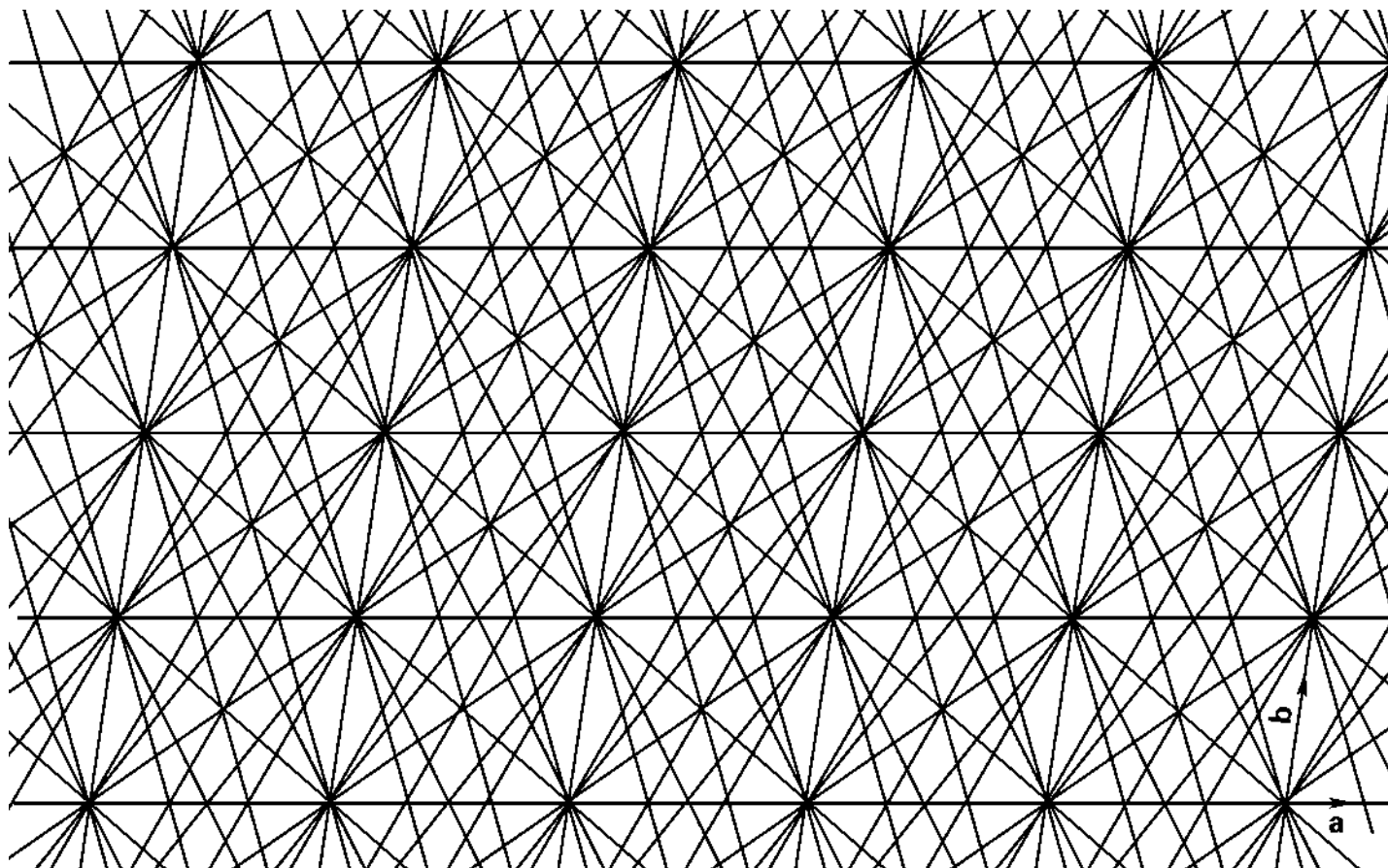


Figure 4-4: Sets of parallel lines, defining sets of unit cells, that pass through a two-dimensional lattice with axes **a** and **b**. The sets of parallel lines with indices (1,1), (2,1), and (3,1) and (-1,1), (-2,1), and (-3,1) are presented.

increases, so does the magnitude of this integer, monotonically and continuously. A given set of parallel planes (Figure 4-6) is assigned three integers, h , k , and l . The magnitude of the integer h is the number of segments into which the planes divide the length of the fundamental unit cell along its **a** axis. The magnitude of the integer k is the number of segments into which the planes divide the length of the fundamental unit cell along the **b** axis; and the magnitude of the integer l , along the **c** axis. When the set of planes is parallel to one of the axes of the fundamental unit cell, as all of the planes in Figure 4-5 are to the **c** axis, it is assigned 0 for the respective index.

The **signs of the integers** assigned to each reflection are determined by the relative progressions of the planes along the three axes. If, as one progresses from one plane to the next along the **a** axis in a positive direction, the intersections of the successive planes with the **b** axis progresses also in a positive direction, as they do in Figure 4-6, then the signs of the integers h and k are the same. If, however, as one progresses from one plane to the next along the **a** axis in a positive direction, the intersections of the successive planes with the **b** axis progress in a negative direction, as they do in Figure 4-5, then the signs of the integers h and k are opposite each other. The same holds for the relationship between the signs of the integers h and l .

Each plane in a set of parallel planes has two **faces**, and either can reflect X-radiation. The two reflections, one from each of the two sides of that set of reflecting planes, are a **Friedel pair**.² In the indices (h,k,l) assigned respectively to the two reflections of the Friedel pair, the signs of the integers h , k and l are opposite. For example, the two reflections with indices (3,-2,4) and (-3,2,-4), respectively, are from the opposite faces of the same set of parallel planes and are a Friedel pair.

The reflections from the faces of the sets of reflecting planes are produced by the electrons in the crystal and they are emitted by diffraction.

Electrons **scatter X-radiation**, and molecules are clouds of electrons confined within atomic and molecular orbitals. The molecule or molecules of protein and the molecules of water distributed through any unit cell in a crystal are clouds of electrons, and they will scatter X-radiation. Electrons scatter X-radiation by being excited to vibrate by the oscillating electric field of the incident beam and then radiating X-radiation of the same wavelength in all directions.

Cut a crystal of protein across its entire width with any plane parallel to the set of planes of a given index (Figure 4-5). Examine one of the two smooth, flat faces produced by that random transection of the crystal. That face contains a particular amount of electron density

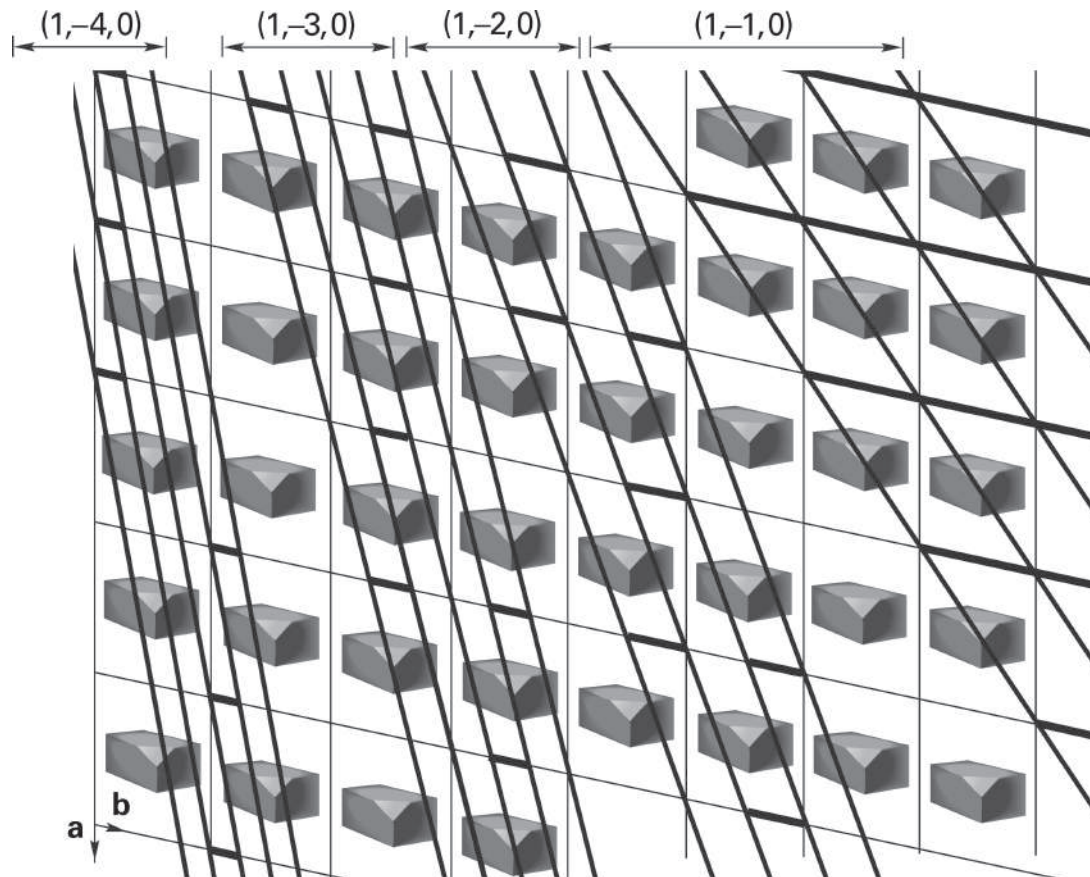


Figure 4-5: Sets of unit cells created by sets of parallel planes. Assume this to be a monoclinic lattice viewed down the c axis and each line the intersection of a perpendicular plane $(h,k,0)$ with the page. Each new set of parallel planes produces a new set of unit cells. The index of each set is given at the top. Each type of unit cell cuts the repeating object into different segments, but in each unit cell in each set of parallel planes, the segments, if put together, form one complete object.

from those atoms within each unit cell that were transected by the plane. Each unit cell defined by the set of planes parallel to the transection contributes exactly the same amount of electron density to the face because it is sliced at exactly the same angle and at exactly the same level. All of the electron density in the entire face will scatter X-radiation. The electron density in the face scatters the X-radiation just as the silver on the smooth, flat surface of a mirror scatters light, and the face is therefore **a mirror for X-radiation**. All of the quanta of X-radiation reflected by that mirror at a certain angle will be reflected

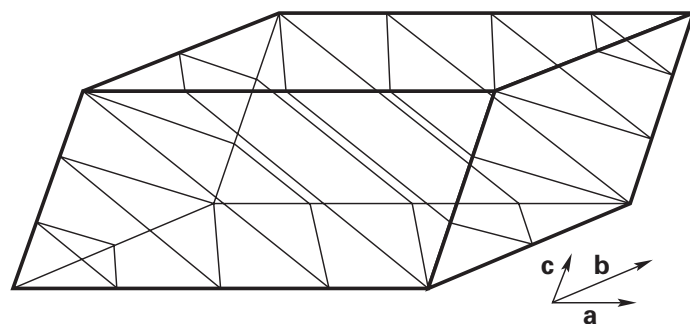


Figure 4-6: Assignment of an index to a set of planes creating the reflecting faces. The index h , k , or l relative to a given axis, a , b , or c , respectively, is the number of segments into which the respective axis is intersected by the set of planes over the length of the fundamental unit cell. The index of this set of parallel planes is $(4,2,3)$.

in phase as is the case with any planar mirror (Figure 4-7). As a result, the scattering elements can be anywhere in the reflecting face and the regularly arrayed, repeating pattern of electron density can be translated along axes parallel to the plane, without affecting either the amplitude or the phase of the reflection. It is this insensitivity of reflected electromagnetic radiation to translation that creates the requirement that a unit cell be only a translational repeating unit. The **amplitude of the reflection** produced by this mirror will be proportional to the quantity of electron density it contains, which is equal to the amount of electron density contributed by each unit cell times the number of unit cells it transects.

Consider the two planes parallel to the one just described at a distance the width of one unit cell above it and at a distance the width of one unit cell below it. Each of these two planes creates a reflecting face with an

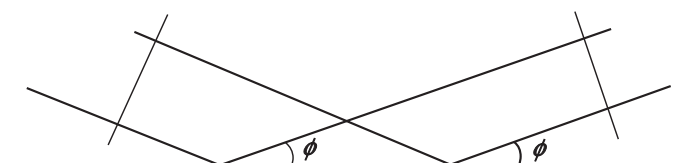


Figure 4-7: Incident electromagnetic radiation at an angle ϕ to a plane of reflection emerges from the reflection in phase regardless of the locations of the points on the plane at which reflection occurs.

identical orientation and an identical repeating pattern of electron density to the one just described. The three reflecting faces considered so far, however, are undistinguished members of a set of reflecting faces evenly spaced throughout the entire crystal that each contain an identical repeating pattern of electron density, that are each the distance of a unit cell above and below their neighbors, and that together include identical transections through all of the unit cells of the same index in the crystal. Each of the members in this set of faces will produce a reflection. If the crystal is being rotated in a beam of X-radiation, when the angle of the incident beam of X-radiation assumes one of a set of particular values, θ_{hkl} , with respect to the set of planes that produced this set of reflecting faces, the reflections from all of the reflecting faces in the set will add in phase to produce a burst or flash of X-radiation by **diffraction** (Figure 4-1B).

The values of θ_{hkl} at which this diffracted reflection occurs is defined by Bragg's law²

$$\theta_{hkl} = \sin^{-1} \left(\frac{n\lambda}{2d_{hkl}} \right) \quad (4-1)$$

where n is any integer, λ is the wavelength of the incident X-radiation, and d_{hkl} is the perpendicular distance, or **Bragg spacing**, between the reflecting faces, which is the width of the unit cells between the planes. Only diffracted reflections are emitted by the crystal because when the incident angle of the X-radiation on a set of faces is not equal to one of these values θ_{hkl} , there are so many reflections from that set of faces that are out of phase with each other that all of them cancel completely. From Equation 4-1 it follows that X-radiation is diffracted by every set of reflecting faces for which the spacing between the planes is larger than $\lambda/2$. The distance $\lambda/2$ is the diffraction limit. Sets of faces with spacings less than the **diffraction limit** do not diffract the X-radiation, and therefore their reflections cannot be observed.

Now consider a plane transecting the crystal parallel to one of the reflecting faces just described, so that it has the same index but at a distance ds above it (Figure 4-8). Consider the reflecting face created by this plane that faces the same direction as the reflecting faces just described. This new reflecting face is a member of a second set of reflecting faces each a distance d_{hkl} apart and each a distance ds above a member of the first set. This second set of reflecting surfaces will diffract the X-radiation at the same incident angle that the first set did because its spacing and angular disposition is the same. But the amplitude of the diffraction from the second set will be different from the amplitude of the diffraction from the first set because even though all of the unit cells are also sliced by the second set, a reflecting face in the second set transects a different region of the unit cell and therefore contains a different amount of electron density from each unit cell than did a reflecting

face in the first set. The phase of the diffraction from the second set will also be different from the phase of the diffraction from the first set of reflecting faces because the second set is displaced a distance ds from the first.

This process of slicing the crystal with sets of reflecting faces each displaced from the set before it by a distance ds can be repeated until the entire unit cell, and hence the entire crystal, has been sliced (Figure 4-8). Each one of these different sets of reflecting faces will diffract at the same angle, θ_{hkl} , because they all have the spacing of the planes of the given index. The single amplitude and single phase of the total diffracted reflection produced by the complete set of all of these reflecting faces will be the sum of the individual amplitudes and individual phases of all of the component sets. The diffracted reflection from the complete set will be observed at the angle θ_{hkl} to the incident beam of X-radiation, and its amplitude and phase will necessarily contain information concerning the distribution of electron density within the crystal. Each complete set of faces of a given index passing through the lattice will produce its own diffracted reflection. Each of the spots on the film in Figure 4-1 is the diffracted reflection from the complete set of reflecting faces of a particular index.

The phase of the reflection from the set of faces hkl is designated α_{hkl} . The **phase of the reflection** is the distance between a crest of the emitted wave and a point of reference common to all of the emitted X-radiation. The phase is expressed in units of wavelength so that, were its value 1, there would be an integral number of wavelengths between the point of reference and the crest. Because the wave is periodic, the phase is expressed as a dimensionless fraction between 0 and 1. Because the

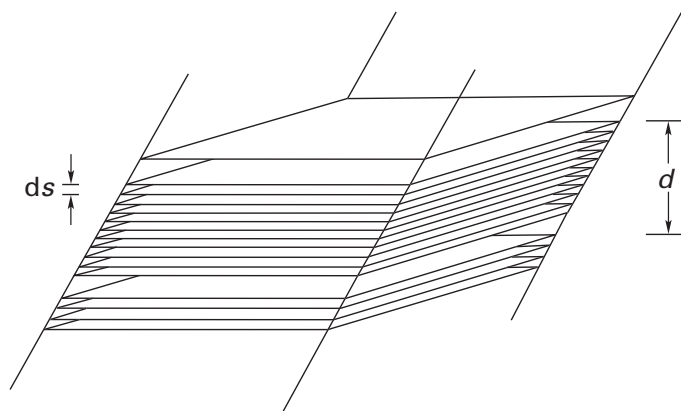


Figure 4-8: Reflecting faces within three consecutive unit cells of height d each a distance ds apart from its neighbors. The first 11 reflecting faces in the unit cell in the middle are shown. The top three faces of that same set of 11 in the bottom unit cell are also shown as well as the bottom face of the same set in the top unit cell. Each reflecting face extends over the whole crystal but only its intersection with the respective unit cell is shown in the figure. Each reflecting face in the set represented here is parallel to two of the axes of the unit cell. Each reflecting face in the stack within each unit cell has a different two-dimensional distribution of electron density because each plane producing a reflecting face cuts a different section through the unit cell.

wavelength of the X-radiation used is usually less than 0.2 nm, the differences in phase among the different reflections are less than 0.2 nm, and because coherent sources are unavailable, they are immeasurable.

Because the choice of the lattice used to define the fundamental unit cell was arbitrary, the choice of the **boundaries of the fundamental unit cell** is arbitrary. Because every plane passing through the unit cell parallel to its boundaries (Figure 4-8) is no better than any other, the plane chosen as the upper boundary of the fundamental unit cell can be anywhere so long as the plane below it chosen for the lower boundary is the one at the level where the pattern repeats. Crystals are seldom cooperative and the molecules packed within them almost never fit as intact entities into a neat box. But this is irrelevant because the solution to a crystallographic calculation gives the distribution of electrons in the fundamental unit cell. The distribution of electrons in any number of adjacent fundamental unit cells can be constructed by simply stacking fundamental unit cells next to each other. If a large enough pile of fundamental unit cells is made, a complete molecule will be found somewhere in the pile.

Each of the thousands of diffracted reflections emerging from the rotating crystal must be assigned an index. For example, each reflection in Figure 4-1B originated from one of the complete sets of reflecting faces, and the index of that set must be assigned to that reflection. The problem of **indexing** is a game of mirrors. As with all games, it is captivating and takes on a life of its own. The crystallographer plays the game in **reciprocal space**; and, as such, learning to live in reciprocal space is a rite of passage. But it is not necessary to live in reciprocal space unless one is a crystallographer; it is enough to know that this can be done with certainty. The concept of reciprocal space simplifies this process of assignment. A familiarity with reciprocal space also permits an engineer to design an X-ray camera that can display the reflections on a sheet of photographic film in the order of their index number. A precession photograph (Figure 4-9) is the product of such a camera.

The intensities of all of the reflections the values of h , k , and l of which fall within chosen limits are measured and indexed. From the measured intensity of each reflection, the amplitude of the structure factor of that reflection can be calculated. The **structure factor** of a reflection is a vector. The phase of that vector is the phase of the reflection, and the square of the amplitude of that vector is a number directly proportional to the measured intensity of the reflection. The constant of this proportionality can be calculated from a consideration of the geometry and the dimensions of the instrument used to measure the reflection. The **amplitude of the structure factor** is the quotient of the square root of the measured intensity of the reflection and this constant of proportionality. In this way, the amplitude of the structure factor is the amplitude of the reflection that has

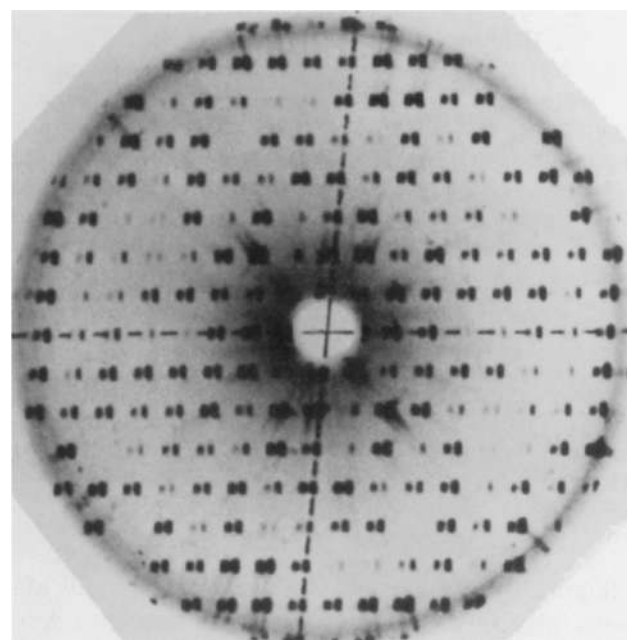


Figure 4-9: Precession photograph of the diffraction of X-radiation by a crystal of egg-white lysozyme from *Gallus gallus* and its isomorphous replacement both in the same triclinic lattice.¹² This is a section through the full three-dimensional pattern of reflections, taken with a Buerger precession camera. Only reflections with an index h of 0 were arrayed by the camera in this particular section. Two photographs are superimposed slightly out of horizontal register to show changes in intensities produced by isomorphous introduction of heavy atoms into the crystal. Left spot of each pair, native lysozyme; right spot, crystal after HgBr_4 has diffused in. This is a photograph of an array of the $0kl$ set of reflections mechanically arranged by the camera with the l axis horizontal and k axis nearly vertical. The index of each reflection can be assigned by inspection from its position in the array. One can consider this photograph as an array of all of the reflections in an equatorial layer line from a rotation photograph about the a axis, such as the one in Figure 4-1B, laid out by the precession camera systematically upon the field. The photograph contains all reflections needed to compute a projection of the structure down the a axis to Bragg spacings of 0.4 nm. Reprinted with permission from ref 12. Copyright 1964 Academic Press.

been corrected for the geometric and instrumental details involved in the measurement. Consequently, it is a property only of the unit cell itself and not of the method by which the measurement was made. Together the amplitudes of the properly indexed structure factors produce a data set. A **data set** is a three-dimensional matrix centered on an origin (0,0,0) in which is entered the amplitudes of all the structure factors calculated from the intensities of the respective reflections that have been measured from a given crystal. Each amplitude is entered at the location in the matrix that has an index identical to the index of its reflection. The amplitude entered at position hkl in this matrix is designated as F_{hkl} and is a positive number.

The rectangular sheet of photographic film that was once used to record the amplitudes of the diffracted reflections (Figure 4-1B) has been replaced by a **charge-coupled device**, which is a rectangular plate that can

measure simultaneously the magnitude of the flux of X-radiation through each pixel on its surface. If the density of pixels is constant, the larger the surface area of the detector, the greater the number of reflections that can be measured simultaneously, and the less will be the damage to the crystal caused by the X-radiation.³

The greater the flux of X-radiation from the source, the shorter will be the interval needed to collect a complete data set. The sources of X-radiation with the greatest fluxes are the large **synchrotrons** located at national laboratories around the world.⁴ The additional advantage to X-radiation from synchrotrons is that these sources produce a broad spectrum of wavelengths so that any narrow range of wavelength within this spectrum can be chosen for the experiment. If the crystal diffracts effectively, it is now possible to collect hundreds of thousands^{3,5} of unique reflections* to Bragg spacings of less than 0.1 nm,^{5,6} which approaches the diffraction limit of the available wavelengths.

The size and shape of the fundamental unit cell can be defined from the angles at which the reflections emerge from the crystal and hence the spacings of the reflections over the surface of the detector (Figure 4-1B). The size and shape of the fundamental unit cell and the indexed data set itself are the only directly measurable quantities available to the crystallographer, and they are ultimately the information used to calculate the distribution of electrons in the crystal of a protein. The unobservable phases, the values of which are inescapably required for the calculation, must be ascertained indirectly by comparing several data sets, each obtained from an altered form of the original crystal.

At this point it is possible to explain the **pattern of reflections in the oscillation photograph** in Figure 4-1B. The central layer line, which intersects at its midpoint the axis of the collimated beam of X-radiation in the camera, is referred to as the equator. Assume that the axis of the crystal chosen for rotation is the **a** axis and that the axis of the beam of X-radiation is perpendicular to the axis about which the crystal is rotated. Define the angle ν , which is the angle between the axis of rotation and a diffracted reflection (Figure 4-10). The complete sets of reflecting faces with an index h of 0, referred to as the $(0,k,l)$ sets of faces, contain only faces parallel to the axis of rotation. As the rotation of the crystal brings each set of these $(0,k,l)$ faces into the proper angle θ_{0kl} with respect to the beam of X-radiation, diffracted reflection occurs. Because each set of these faces is parallel to the axis about which the crystal is rotated and because the angle of reflection equals the angle of incidence, each flash emerges at an angle $\nu = 90^\circ$. The reflections from the $(0,k,l)$ sets of faces are the layer line of reflections on the equator. There is, however, no easily explained pattern in which the reflections with particular values for k

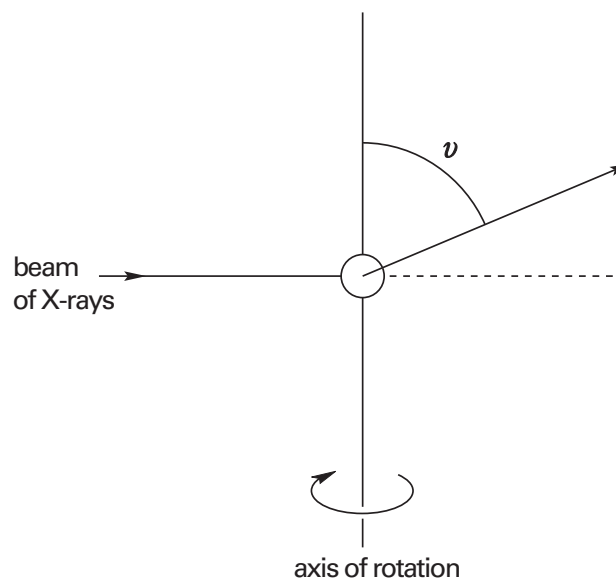


Figure 4-10: Definition of ν , the angle between the axis of rotation and the diffracted reflection.

and l are distributed along the equator. Now consider all the sets of faces with an index h of 1, referred to as the $(1,k,l)$ sets of faces. Each of the faces in these sets makes an angle with the axis of rotation such that all values of ν for these sets are the same and all the reflections lie on the first layer line. The same argument can be made for the other layer lines. Therefore, in a crystal rotated about the **a** axis, all of the reflections with the same first index lie on the same layer line, and each successive layer line out from the equator is of successively higher or successively lower first index. In each layer line, however, the pattern in which the reflections with successive second or successive third indices occur is complex. Each of these reflections, however, can be assigned a full index unambiguously by the crystallographer.

Any piece of matter, at a given instant ($<10^{-12}$ s), has a particular distribution of molecules, and this distribution of molecules causes the matter to have a **distribution of electron density**, $\rho(x,y,z)$. If the matter is a gas or a liquid, the rapid redistribution of the molecules causes $\rho(x,y,z)$ to change over its full extent with time. The collection and measurement of the intensities of the reflections necessary to perform a determination of molecular structure usually takes hours, and the distribution of electron density of a liquid or a gas averaged over the period of the measurement is absolutely uniform. The liquid regions of aqueous solvent within a crystal of protein, which account for 40–75% of its volume,⁷ are, as a result, featureless. Any portion of the protein the position of which fluctuates over dimensions greater than those of an atom, for example a flexible segment of polypeptide, is also featureless. To the extent that the matter in a crystal is a solid, its molecules remain fixed in space, and solids have well-defined distributions of electron density. The fixed portions of the molecules of protein in the crystal remain in place, except for thermal vibrations,

* Unique reflections count only those not duplicated by the symmetry of the crystal and count only one of the Friedel pairs.

and the experimentally measurable distribution of electron density in the regions of the crystal that contain the fixed portion of the protein is the average over these small vibrational displacements. Within these limits, it is featured. The distribution of featured electron density in a crystal is a periodic function, by definition, and it is this periodicity that leads to the reflections.

It can be shown² that for a crystal of protein

$$\rho(x, y, z) = \frac{1}{V} \sum_{h=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} F_{hkl} \exp[-2\pi i(hx + ky + lz - \alpha_{hkl})] \quad (4-2)$$

where V is the volume of the fundamental unit cell, F_{hkl} are all of the amplitudes, properly indexed, and α_{hkl} are all of the properly indexed phases of the structure factors of the reflections. The coordinates x , y , and z in Equation 4-2 are referred to the major axes **a**, **b**, and **c**, respectively, of the fundamental unit cell (Figure 4-11). This usually produces a **coordinate system** that is not orthogonal. The lengths a , b , and c of the fundamental unit cell along the three major axes **a**, **b**, and **c** are expressed in absolute length (nanometers). The three distances x , y , and z that are the coordinates of a point within the fundamental unit cell are measured along the three major axes (Figure 4-11), but the units in which these three distances are expressed in Equation 4-2 are relative distances along these axes, where $x = Xa^{-1}$, $y = Yb^{-1}$, and $z = Zc^{-1}$, and X , Y , and Z are the absolute distances (nanometers) along each respective axis. The integers h , k , and l ; the coordinates of a point in the unit cell, x , y , and z ; and the phase of each structure factor, α_{hkl} , are all dimensionless numbers. The 2π by which they are multiplied in Equation 4-2 converts these dimensionless numbers to units of radians. This is the purpose that 2π serves in all of the remaining equations of this section.

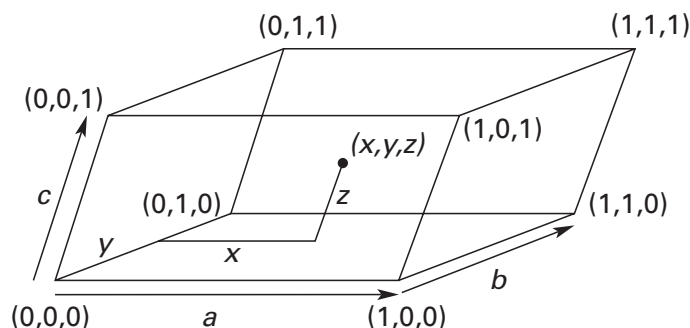


Figure 4-11: Assignment of coordinates x , y , and z to a point within a fundamental unit cell. The coordinate axes for the distances x , y , and z are the crystallographic axes **a**, **b**, and **c**, respectively. The distances in each direction are measured along these axes regardless of the angles between them. The numbers in parentheses are the values for x , y , and z , respectively, at the respective corners of the unit cell.

The **imaginary portion** of Equation 4-2 is somewhat disconcerting because the intention of the equation is to calculate a real electron density. This conundrum is solved by noting that

$$\exp(iw) = \cos w + i \sin w \quad (4-3)$$

and that, because a complete data set is a complete set of Friedel pairs, all terms in $i \sin w$ cancel in pairs.² As a result

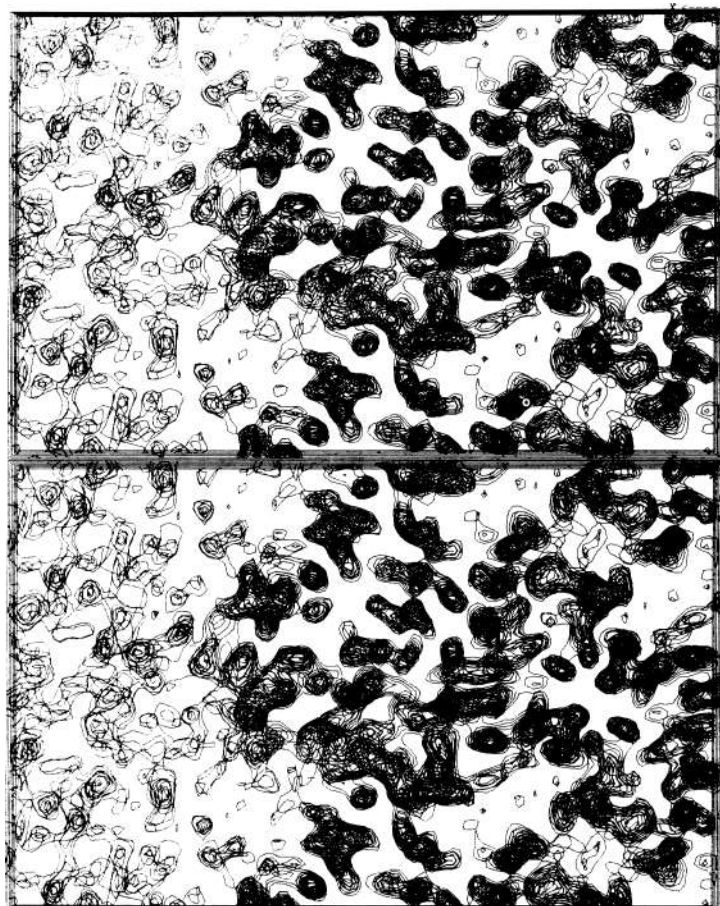
$$\rho(x, y, z) = \frac{1}{V} \sum_h \sum_k \sum_l F_{hkl} \cos[2\pi(hx + ky + lz - \alpha_{hkl})] \quad (4-4)$$

The value of examining Equation 4-2 is that it demonstrates that the electron density at any point in the fundamental unit cell can be calculated explicitly by inserting the amplitudes of the structure factors, properly indexed; the coordinates of that point; and the respective phases. The way that the **calculation of a map of electron density** is performed is to divide the fundamental unit cell into a large number of points with coordinates x_p , y_q , z_r . Insertion of all available values for F_{hkl} , α_{hkl} , h , k , l , and the coordinates of the point x_p , y_q , z_r into Equation 4-2 produces the value of $\rho(x,y,z)$ at the point x_p , y_q , z_r . All points in three-dimensional space with values of $\rho(x_p, y_q, z_r)$ within certain narrow ranges are connected by contoured surfaces, or all points in sections through the fundamental unit cell with values of $\rho(x_p, y_q, z_r)$ within certain narrow ranges are connected by contour lines, and this procedure produces a map of electron density (Figure 4-12).⁸

The summations of Equation 4-2 or Equation 4-4 are theoretically infinite, but any finite number of structure factors will give an approximate solution. In any case, the data set can never contain reflections from beyond the diffraction limit. The effect of summing over only a **finite number of structure factors** in Equation 4-2 is to blur $\rho(x,y,z)$. The fewer the structure factors used, the more blurred $\rho(x,y,z)$ becomes. Usually 10,000–500,000 independent reflections are measured and indexed to calculate a map of electron density. The minimum Bragg spacing of this data set usually will be 0.1–0.3 nm, and the individual values of h , k , and l will lie between about –40 and 40.

The decision about which structure factors to include in the data set is based on the ability of the instruments to measure reflections with large angles of θ_{hkl} , the ability of the crystal to diffract from sets of faces with small Bragg spacings, and other technical limitations. Once the decision has been made, the data set is collected so that it includes the amplitudes of the structure factors from as many as possible of the reflections arising from sets of faces the Bragg spacings of which is greater than a minimum value. The universal choice of

Figure 4-12: Stereoview of the map of electron density (Bragg spacing ≥ 0.25 nm) calculated from crystals of deoxyribonuclease I from *Bos taurus*.⁸ The experimental phases were estimated by use of isomorphous derivatives obtained with solutions of TbCl_3 , K_2PtCl_6 , $\text{Pb}(\text{NO}_3)_2$, and the three binary combinations of these three salts. Portions of six contiguous sections (0.08 nm section⁻¹) through the fundamental unit cell (monoclinic; $a = 13.2$ nm, $b = 5.5$ nm, and $c = 3.8$ nm; $\beta = 91.4^\circ$) perpendicular to the b axis were stacked together. The contours represent three levels of electron density and were drawn through points on each section the electron densities of which were 10%, 20%, and 30%, respectively, of the maximum electron density observed. Reprinted with permission from ref 8. Copyright 1984 IRL Press.



“resolution” as the word used to report this **minimum magnitude of the Bragg spacings** between the faces, the reflections from which have been included in the data set, is unfortunate. This choice suggests that the property being noted is the same property as the resolution defined in optics, which it is not. The minimum magnitude of the Bragg spacings does place an upper limit on the quality of the final map of electron density, but it is not the only factor determining its quality. The difficulty is not collecting and indexing reflections, which ultimately determines the minimum Bragg spacing, but establishing accurate phases for each reflection.⁷

In practice, at a given minimum value for the Bragg spacing, it is the quality of the phases that defines the quality of the map of electron density.* It has been shown⁹ that if all of the correct amplitudes are used but all of the phases are set at the same arbitrary value, the map of electron density calculated is meaningless. If, however, all of the correct phases are used and all of the amplitudes are set at the same arbitrary value, a fairly accurate map of electron density can be calculated.† It is less misleading to refer to the data set as one “to Bragg spacings of” rather than as one “at a resolution of”.

The use of Equation 4-2 or 4-4 requires that the phase, α_{hkb} of each structure factor in a data set be estimated. One way such an estimate can be accomplished for crystals of protein is by **multiple isomorphous replacement**.¹⁰ Suppose there are two crystals of protein, alike in almost every way and hence isomorphous. The only difference between them is that, at one or a few specific locations on each of the molecules of protein in one of the crystals, an atom or several atoms that have a large number of electrons has been attached; the other crystal is of the unadorned protein. One of the requirements placed on the bound atoms is that they have high electron density, in other words, a large number of electrons in a small volume. For this reason, a **heavy atom** such as xenon, iodine, mercury, gold, or uranium is usually chosen. Another requirement is that these heavy atoms occupy specific points in the fundamental unit cell. This requirement is fulfilled if they are bound to particular amino acids or clusters of amino acids on the surface of each molecule of protein. For example, the PtCl_4^{2-} used to phase the structure factors from a crystal of phosphocarryer protein III^{glc} happened to be chelated by two histidines that by chance were adjacent to each other on the surface of the protein.¹¹ The reflections from the two crystals, the one containing the fixed heavy atoms and the one without them, will have the same values for all θ_{hkl} and hence the same geometric display of reflections,

* The quality of the phases is quantified by the **figure of merit**, which ranges from zero (completely unreliable) to 1.0 (perfect.)

† Even if, however, the phases were estimated perfectly but only amplitudes and phases for reflections to Bragg spacings of 0.5 nm were used in the calculation, the resulting map of electron density would not contain sufficient information to reveal the structure of a protein. A minimum Bragg spacing of 0.3–0.35 nm is required.

but the amplitudes of the reflections will differ (Figure 4–9)¹² as well as the phases.

The structure factor of a reflection from a set of faces with the index hkl can be represented as a vector \mathbf{F}_{hkl} . The length of the vector is the amplitude of the structure factor, F_{hkl} , and its direction is defined by the phase of the reflection, $2\pi\alpha_{hkl}$ radians. Because the computations are performed in complex space (Equation 4–2), complex coordinates are chosen to represent this vector:

$$\mathbf{F}_{hkl} = F_{hkl}(\cos 2\pi\alpha_{hkl} + i \sin 2\pi\alpha_{hkl}) = F_{hkl} \exp(2\pi i \alpha_{hkl}) \quad (4-5)$$

The real component of the vector \mathbf{F}_{hkl} is $F_{hkl} \cos 2\pi\alpha_{hkl}$, and the imaginary component is $F_{hkl} \sin 2\pi\alpha_{hkl}$. The amplitude of the vector is

$$F_{hkl} = F_{hkl}(\cos^2 2\pi\alpha_{hkl} + \sin^2 2\pi\alpha_{hkl}) \quad (4-6)$$

Equation 4–2 states that the electron density is the Fourier transform of the structure factors. It follows that the structure factors must be the Fourier transforms of the electron density. As a result, the amplitude and phase of a given structure factor from a crystal can be calculated if the distribution of atoms in a fundamental unit cell of that crystal is known

$$\mathbf{F}_{hkl} = \sum_j f_j \exp[2\pi i(hx_j + ky_j + lz_j)] \quad (4-7)$$

where f_j is the scattering factor for atom j and (x_j, y_j, z_j) is its position in the unit cell. The scattering factor is determined by the number of electrons in atom j and their distributions over their respective orbitals. Because the sizes of the orbitals are of the order of the wavelength of the X-radiation, the numerical value of the scattering factor for a given atom is a function of the angle θ_{hkl} of the reflection. As θ_{hkl} increases, the scattering produced by the electrons around an atom decreases as a result of interference. Values for scattering factors have been tabulated for all atoms and systematic values of θ .

It can be seen that, since Equation 4–7 is a summation

$$\mathbf{F}_{hkl, H+P} = \mathbf{F}_{hkl, H} + \mathbf{F}_{hkl, P} \quad (4-8)$$

where $\mathbf{F}_{hkl, H+P}$ is the structure factor from the crystal containing the heavy atom, $\mathbf{F}_{hkl, P}$ is the same structure factor from the unadorned crystal, and $\mathbf{F}_{hkl, H}$ would be the structure factor from a crystal in which only the heavy atoms were present at the same locations they occupy in the existing isomorph. This summation can be presented geometrically (Figure 4–13A).^{7,13} If one knows where the heavy atoms are located in the fundamental unit cell, the

vectors $\mathbf{F}_{hkl, H}$, both amplitudes and phases, can be calculated with Equation 4–7. Discovering the locations of the heavy atoms in a given isomorph is an art, the description of which is dramatic but not germane to this discussion. Their locations are eventually determined, and these locations are used to calculate each of the values of $\mathbf{F}_{hkl, H}$.

Unless the heavy atom chosen displays strong anomalous dispersion, at least two isomorphous crystals, each substituted with a heavy atom in a different way are required for a unique determination of the phases. The data that are available are $F_{hkl, P}$, $F_{hkl, (H+P)g}$, and $\mathbf{F}_{hkl, Hg}$, where the index g refers to each of the several isomorphous replacements from the crystals of which reflections have been measured. From Equations 4–5 and 4–8, these data provide a set of **simultaneous vector equations** equal in number to the number of isomorphous replacements for each structure factor. In theory, any two of these vector equations can be solved for the phase, $2\pi\alpha_{hkl, P}$, of structure factor $\mathbf{F}_{hkl, P}$; in practice, as many as are available are used.

There is a **geometric solution** to this set of simultaneous vector equations. The amplitude of the vector $\mathbf{F}_{hkl, P}$ is known from the data set, but not its phase, $2\pi\alpha_{hkl, P}$. Therefore, what is known about $\mathbf{F}_{hkl, P}$ defines a circle of radius $F_{hkl, P}$ with its center at point P (Figure 4–13B). Both the amplitude and the phase of a given $\mathbf{F}_{hkl, H}$ are known, and this vector can be placed so that its head is at point P. Its tail defines the position, point D, of the tail of vector $\mathbf{F}_{hkl, H+P}$ from the isomorphous derivative in the vector sum (Figure 4–13A). The phase of vector $\mathbf{F}_{hkl, H+P}$ is unknown but its amplitude, which is known, defines a second circle with its center at point D. Because the vector sum must balance (Equation 4–8), the two points where the two circles intersect must represent two possibilities for the one actual vector sum. In theory, the correct one of the two possibilities can be determined by going through the same steps with the data from a second isomorphous replacement because the phase of $\mathbf{F}_{hkl, P}$ must be the same in both, and only one of the two possibilities for $\mathbf{F}_{hkl, P}$, namely, the one defined by the actual vector sum, should be the same in both.

A particularly gratifying example of this way of choosing the correct point defining the head of vector $\mathbf{F}_{hkl, P}$ was the definition of the phase of structure factor $\mathbf{F}_{9,1,-2}$ for a crystal of hemoglobin by use of six different isomorphous replacements (Figure 4–13C). All seven circles intersect at approximately the same point and define the phase $\alpha_{9,1,-2}$. This is, of course, the best example from the thousands of structure factors for hemoglobin; and, in practice, the circles almost never intersect in the same spot or even near the same spot. The phase of each $\mathbf{F}_{hkl, P}$ must be estimated by taking a statistical average of all of the points of intersection.⁷ The uncertainty in this average value for each phase is then used to weight the contribution of the respective structure factor to the summations of Equation 4–2 or 4–4.

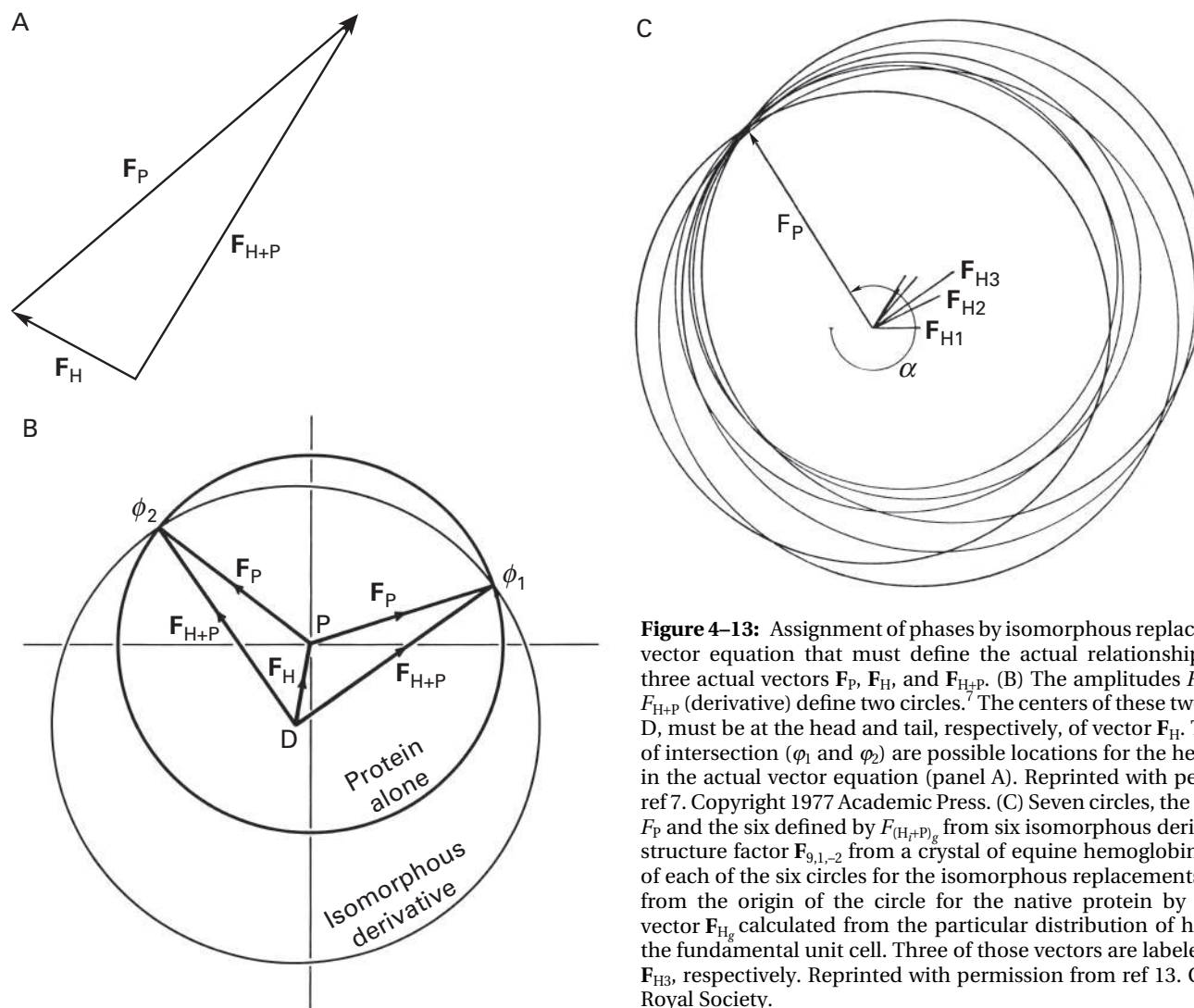


Figure 4-13: Assignment of phases by isomorphous replacement. (A) The vector equation that must define the actual relationship between the three actual vectors F_P , F_H , and F_{H+P} . (B) The amplitudes F_P (parent) and F_{H+P} (derivative) define two circles.⁷ The centers of these two circles, P and D, must be at the head and tail, respectively, of vector F_H . The two points of intersection (ϕ_1 and ϕ_2) are possible locations for the head of vector F_P in the actual vector equation (panel A). Reprinted with permission from ref 7. Copyright 1977 Academic Press. (C) Seven circles, the one defined by F_P and the six defined by $F_{(H+P)_g}$ from six isomorphous derivatives, for the structure factor $F_{9,1,-2}$ from a crystal of equine hemoglobin.¹³ The origins of each of the six circles for the isomorphous replacements are displaced from the origin of the circle for the native protein by the respective vector F_{H_g} calculated from the particular distribution of heavy metals in the fundamental unit cell. Three of those vectors are labeled F_{H1} , F_{H2} , and F_{H3} , respectively. Reprinted with permission from ref 13. Copyright 1961 Royal Society.

Some examples of isomorphous replacements that have been used in crystallographic investigations should make these considerations less abstract. Each isomorphous replacement is a different crystal, usually obtained by soaking a crystal of the unmodified protein in a solution of a compound containing the heavy atom. These compounds can be simple ions, such as Sm^{3+} , WO_4^{2-} , $\text{Pt}(\text{CN})_4^{2-}$, $\text{Au}(\text{CN})_2^-$, Hg^{2+} , or $\text{Pt}(\text{NH}_3)_2^{4+}$. Such ions are chelated at certain specific locations in the unit cells by functional groups on the surface of the protein in the crystal. Xenon gas at high pressure also produces isomorphous replacements¹⁴ by associating with hydrophobic pockets within the protein.¹⁵ Some organomercuric compounds, such as ethyl mercurithiosalicylate or diphenylmercury, are bound at specific locations on the protein while other organomercuric compounds, such as ethylmercury chloride, mersalyl, *o*-mercuriphenol, or *p*-mercuribenzoate, react covalently with the thiols of cysteines on the protein.¹⁶ As many as three mercuric ions, Hg^{2+} , can be noncovalently associated with the lone pairs of electrons of a cysteine.¹⁷ More complicated organomercury compounds such as 5-mercuride-oxyuridine

monophosphate,¹⁸ 3-acetoxymethyl-4-aminobenzene-sulfonamide,¹⁹ and ethylmercuriphosphate²⁰ have been designed as analogues of ligands specific to the protein. An oligonucleotide containing 5-iodouracil can be used to attach an iodine atom to a protein that normally binds the unsubstituted oligonucleotide.²¹

At least two and perhaps as many as six (Figure 4-13C) or seven isomorphous replacements are made. The isomorphous replacements used to obtain the phases for maltose binding protein were made by soaking crystals in K_2PtCl_4 , $\text{Pb}(\text{NO}_3)_2$, sodium mersalyl, dysprosium iodide, and glucosyl(α 1,4)-6-iodo-6-deoxyglucose.²² The last compound is a specific ligand for the binding protein. In the case of trimethylamine-*N*-oxide reductase (cytochrome *c*), six separate isomorphous replacements were prepared with $\text{Ta}_6\text{Br}_{14}$, $(\text{NH}_4)_2\text{OsCl}_6$, $(\text{NH}_4)\text{IrCl}_6$, K_2PtCl_6 , sodium ethylmercurithiosalicylate, and sodium bis(*N*-methylhydantoinato)gold, respectively.²³ In the case of chloramphenicol *O*-acetyltransferase, $\text{Sm}(\text{NO}_3)_3$, $\text{KAu}(\text{CN})_2$, K_2PtCl_4 , *p*-mercuribenzoate, and *p*-iodochloramphenicol, the last of which is a good substrate for the

enzyme, provided useful isomorphous replacements.²⁴ In the case of apoferritin, however, only two isomorphous replacements, made with *p*-mercuribenzoate and $K_2UO_2F_5$, were used to determine the phases to Bragg spacings of 0.28 nm.²⁵

Once the positions of two or more separate sets of heavy metal atoms are known within the fundamental unit cell, the reagents can be used in pairs to generate additional unique isomorphous replacements. The advantage is that because the positions in each of the original isomorphous replacements are already available, the positions in the combined isomorphous replacement can be readily established. Isomorphous replacements were made from crystals of alcohol dehydrogenase with $K_2Pt(CN)_4$ and $KAu(CN)_2$, and the positions of the platinum and gold, respectively, in the resulting fundamental unit cells were determined. In combination, these two anions produced a third isomorphous replacement.²⁶ From crystals of deoxyribonuclease I, it was possible to make three isomorphous replacements, one each with $TbCl_3$, K_2PtCl_4 , and $Pb(NO_3)_2$, which could then be used in the three possible combinations to generate three additional, unique isomorphous replacements.⁸

Today, however, phases are usually estimated by taking advantage of the **anomalous dispersion** of the heavy atoms in only one isomorphous derivative.^{27–29} The real and imaginary components of the scattering factors f_j (Equation 4–7) for atoms such as copper,²⁹ selenium,³⁰ holmium,³¹ terbium,³² tantalum,³³ uranium,³⁴ platinum,³⁵ and bromine³⁶ change with the wavelength of the X-radiation in the vicinity of their respective absorption edges. The changes are dramatic enough that if data sets are gathered at three or four different wavelengths properly chosen with respect to the absorption edge of the heavy atom, those data sets can be equivalent, in terms of the differences produced in the intensities of the reflections, to sets of reflections measured from three or four isomorphous replacements. The advantage of this procedure is that the same crystal containing the heavy atoms is used for all of the measurements, so that the errors associated with combining data from different crystals are avoided.

The appropriate heavy atoms are usually incorporated into the crystal by soaking. In the case of basic blue copper protein, however, only the copper ion already within the native protein was used as the heavy atom, and data sets gathered at four different wavelengths were sufficient to establish experimental phases to Bragg spacings of 0.25 nm with no isomorphous replacement at all.²⁹ It is also possible to take advantage of the anomalous dispersion of one isomorphous derivative in combination with the normal diffraction from several others to establish experimental phases.³⁷ A common way of introducing a heavy atom susceptible to anomalous dispersion³⁰ is to express the protein to be crystallized in a bacterium auxotrophic for methionine growing on a

medium containing selenomethionine rather than methionine. The selenium atoms end up at each position in the sequence of the protein normally occupied by methionine and are positioned at precise locations in the unit cell by the tertiary structure of the protein.

There is an additional component of a crystal of protein that is formally equivalent to a set of heavy atoms in an isomorphous derivative. This component is the featureless aqueous solvent that surrounds the protein. The fact that it should be featureless allows it to be used, much as an electron-rich atom is used, to improve the phases by **solvent flattening**.³⁸ A map of electron density is prepared with the available estimates of the phase for each structure factor gathered from isomorphous replacement and anomalous dispersion. If the map is clear enough that the boundary between protein and solvent can be defined (Figure 4–12), all of the region of the fundamental unit cell occupied by solvent is forced to have the same uniform electron density even though in this original map it was not uniform in density (notice the noise in the regions of the map occupied by solvent in Figure 4–12). From this geometric solid of uniform electron density, a set of structure factors equivalent to an additional $F_{hkl,H}$ for Equation 4–8 could be calculated with Equation 4–7 and used as an additional constraint on the phases (Figure 4–13), but solvent flattening is more successful if used iteratively.

The updated map of electron density with the vaguely defined features of the protein and the solvent that has been purposely flattened is used in its entirety to calculate a set of phases. These calculated phases are used in combination with the available estimates of the phase from isomorphous replacement to arrive at a set of improved phases. These improved phases and the observed amplitudes are used to calculate a new map of electron density. The regions of solvent in the new map are defined, the electron density in these regions is again forced to be uniform, and the process is repeated. As the iterations progress, the solvent in each new map becomes flatter and the protein more detailed. In theory³⁸ and in practice,³⁹ the method can provide adequate phases in the absence of measurements of anomalous dispersion when only one isomorphous heavy atom derivative is available. Usually, however, solvent flattening is used to improve the phases that have been gathered with multiple isomorphous replacements or by anomalous dispersion.

Because the quality of the final map of electron density (Figure 4–12) depends so heavily on the quality of the phases, the uninvolved observer can evaluate the results only if she is informed. It is important to learn how many isomorphous replacements were made, how many wavelengths were used for anomalous dispersion, and which data sets were used to calculate the phases. It is also essential to see at least a portion of the calculated, **unrefined map of electron density** (Figure 4–12) to get a feeling for its quality.⁴⁰ It must be emphasized that,

unless a map of electron density is already available for a closely related protein, the calculation of the initial map of electron density from the phases derived from isomorphous replacement is an unavoidable step in crystallography, and the quality of this map can affect significantly the remainder of the process. The work involved in obtaining this initial map is extensive, and many crystallographic experiments are designed specifically to avoid this work.

When the map of electron density within the fundamental unit cell or from several neighboring fundamental unit cells is examined, the electron density that corresponds to the intact molecule of protein can be discerned. Since a large fraction of the crystal is liquid water, which is featureless, the protein, which is fixed and highly featured, stands out (Figure 4-12). The compact globule of electron density eventually assigned to an individual molecule of protein usually has an overall size and shape consistent with its amino acid sequence, its frictional coefficient, and other molecular parameters. Within this globular solid, features can be seen in relatively sharp detail, but only seldom at atomic resolution.

Suggested Reading

Stout, G.H., & Jensen, L.H. (1989) *X-ray Structure Determination, A Practical Approach*, 2nd ed., Wiley, New York.

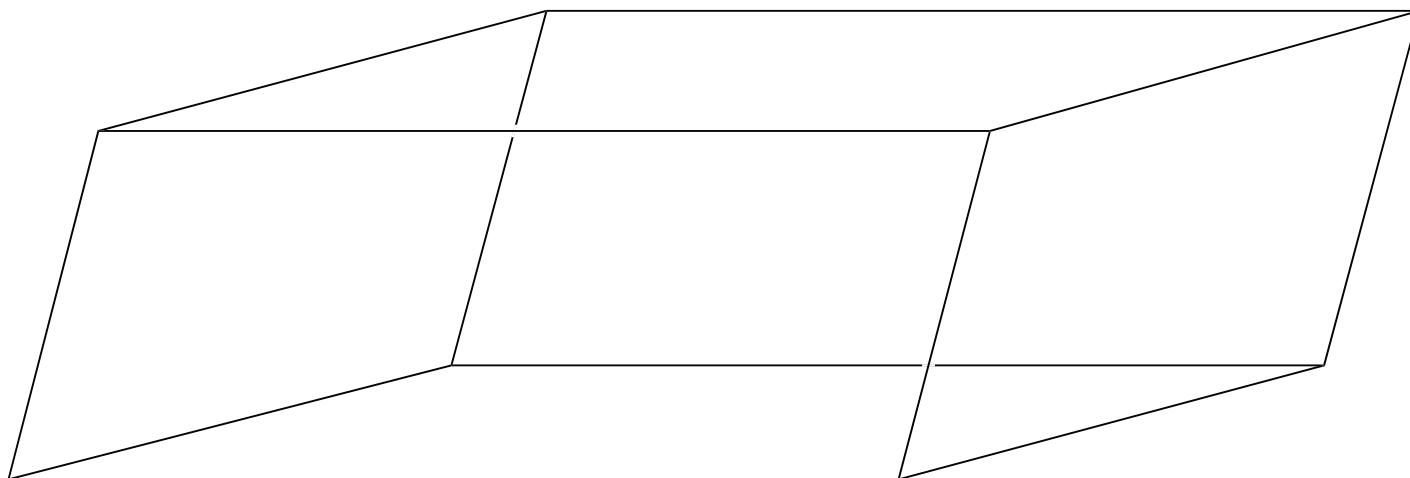
Problem 4-1: Below there is a generic unit cell. Make several xerographic copies of it. Draw a right-handed set of axes labeled **a**, **b**, and **c** next to each of your copies of the unit cell. Draw a diagram of the (4,2,3) set of reflecting planes passing through the first unit cell as in Figure 4-6. Draw a diagram of the (4,-2,3) set of reflecting planes passing through the second unit cell. Draw a diagram of the (4,2,-3) set of reflecting planes passing through the third unit cell. Label each of your diagrams with the index number.

Problem 4-2: The amplitude of a particular structure factor from a crystal of protein, F_p , is 22.2. The amplitude of the structure factor with the same index from a crystal of the first isomorphous replacement, F_1 , is 24.2. The structure factor with the same index calculated from the established positions of the heavy metal ions in the unit cell of the first isomorphous replacement has an amplitude F_{H1} of 5.4 and a phase of 110° . The amplitude of the structure factor with the same index from a crystal of the second isomorphous replacement F_2 is 21.0. The structure factor with the same index calculated from the established positions of the heavy metal ions in the unit cell of the second isomorphous replacement has an amplitude F_{H2} of 8.9 and a phase of 65° . Estimate graphically the phase for the structure factor of this index from the crystal of protein alone.

Problem 4-3: Pig heart citrate (*si*) synthase crystallizes from solution at pH 7.4. The crystals are tetragonal. The dimensions of the fundamental unit cell are $a = b = 7.74$ nm and $c = 19.64$ nm.⁴¹ A crystal was submitted to diffraction with X-radiation generated from a rotating anode of copper. The $K\alpha$ emission of the copper ($\lambda = 0.154$ nm) was selected for the source of the X-radiation. On graph paper draw a view of the fundamental unit cell looking down the c axis with the set of (2,-4,0) faces intersecting it. At what angle θ to the incident beam of X-radiation will the reflection from the (2,-4,0) set of faces emerge from the crystal?

The Molecular Model

An **irregular tube of electron density** can be observed to meander through and account for the globule of featured electron density assigned to the intact molecule of protein in the map of electron density. Sections of one such continuous tube can be seen embedded in the flat slice



of electron density presented in Figure 4-12. This tube is the polypeptide of the protein (2-8) that has folded to assume the native structure of the molecule. It is into this tube that a molecular model of the known covalent structure of the polypeptide must be fit.

Once the covalent sequences of the polypeptides, the covalent sequences and points of attachment of any covalently bound oligosaccharides, and the identity and points of attachment of any other posttranslational modifications have been established and even before a map of electron density is available, it is possible to construct a **molecular model** of the fully modified and glycosylated polypeptide known to constitute a molecule of protein. Such a model would incorporate bond lengths and bond angles that have been measured with high precision during crystallographic studies of small molecules. These small molecules used as standards are molecules the covalent structures of which are identical to segments of polypeptide, the side chains of the amino acids, segments of oligosaccharide, or the monosaccharides in the oligosaccharide. As with any molecular model of such a size and complexity, the one of a polypeptide would be a flexible, protean object that assumes a new shape each time rotation around one of its acyclic single bonds occurs.

It is this long, flexible model that must be fit, amino acid by amino acid, into the map of electron density. Until recently, the process of fitting the model into the map was always performed visually by the crystallographer.^{40,42} It is now possible,⁴³ however, for a computer to fit the model into the map automatically. Nevertheless, the success of this automated process for a particular map of electron density still must be carefully evaluated by the crystallographer,⁴⁴ and the fit must be altered accordingly by manual adjustments. To determine whether or not the molecular model has been correctly fit into the map of electron density, there are no automated rules that are as reliable as the judgment and accumulated knowledge of the crystallographer. If careful human evaluation of each fit is not performed routinely, there is a risk that the frequency at which incorrect crystallographic molecular models are published will increase as more and more crystallographic molecular models are produced in an automated fashion.

One criterion that the molecular model of the polypeptide has been successfully fit into the map of electron density is the correspondence between the sequence in which amino acids of different sizes (Figure 4-14) are known to occur along the amino acid sequence of the polypeptide and the sequence in which protrusions of different size occur at regular intervals along the tube of electron density (Figure 4-15). The 20 different **amino acids** are, in order of increasing electron density (Figure 4-14), glycine, alanine, serine, proline, cysteine, valine, threonine, aspartate, asparagine, leucine, isoleucine, glutamate, glutamine, methionine, lysine, histidine, phenylalanine, arginine, tyrosine, and trypto-

phan. In terms of electron density, many of them are indistinguishable, for example, valine and threonine or aspartate, asparagine, leucine, and isoleucine, and only a few of them, for example, tryptophan, tyrosine, and phenylalanine, are of sufficient size and peculiar enough shape to be identified unambiguously with the protrusions jutting out from the continuous tube in the map of electron density (Figure 4-15).^{*} Together, however, the sequence in which the amino acids are arranged in a given protein and their relative sizes usually provide sufficient reassurance that the molecular model of the polypeptide has been fit into the map correctly.

An additional reassurance that the polypeptide has been properly fit into the map can be obtained from anomalous dispersion. If the protein has been expressed so that it contains selenomethionine instead of methionine, the electron density at the locations in the map that are occupied by the selenium atoms will vary in intensity when the wavelength of X-radiation used to produce the reflections is varied near the absorption edge of the selenium. These variations in intensity can be used to locate the positions at which the methionines must end up after the molecular model has been fit properly into the tube of electron density.⁴⁵ Although the anomalous dispersion from sulfur itself is weak, under appropriate circumstances the positions of both the cysteines and the methionines in the map of electron density of a molecule of protein expressed normally can be located by the anomalous dispersion of their sulfurs.⁴⁶

The reassurance provided by the agreement between the known amino acid sequence of the polypeptide and the sequence of the sizes of the protrusions along the continuous tube of electron density or the positions of atoms capable of anomalous dispersion is not inconsequential. It is rarely the case that the tube of electron density representing the polypeptide in the map of electron density is continuous over its entire length. Portions of the polypeptide that are so flexible that they vibrate too widely will not contribute to the diffraction and will produce no structured electron density. Often segments of the polypeptide can assume several different conformations while in the crystal. The movement among these conformations within a particular molecule of the protein can be rapid or a particular conformation can be statically occupied. Regardless of the rate at which the conformations interconvert, if at any given instant these segments from different molecules of the protein in the crystal assume different conformations, this disorder will prevent them from contributing to the diffraction and hence to the structured electron density. Occasionally

^{*} Although the side chain of cysteine has the same number of electrons as those of threonine and valine and the side chain of methionine the same number as those of glutamate, glutamine, and lysine (Figure 4-14), the sulfurs in cysteine and methionine, because they have 10 core electrons, produce strong localized features of electron density (Figure 4-15) that permit them to be distinguished from the others.

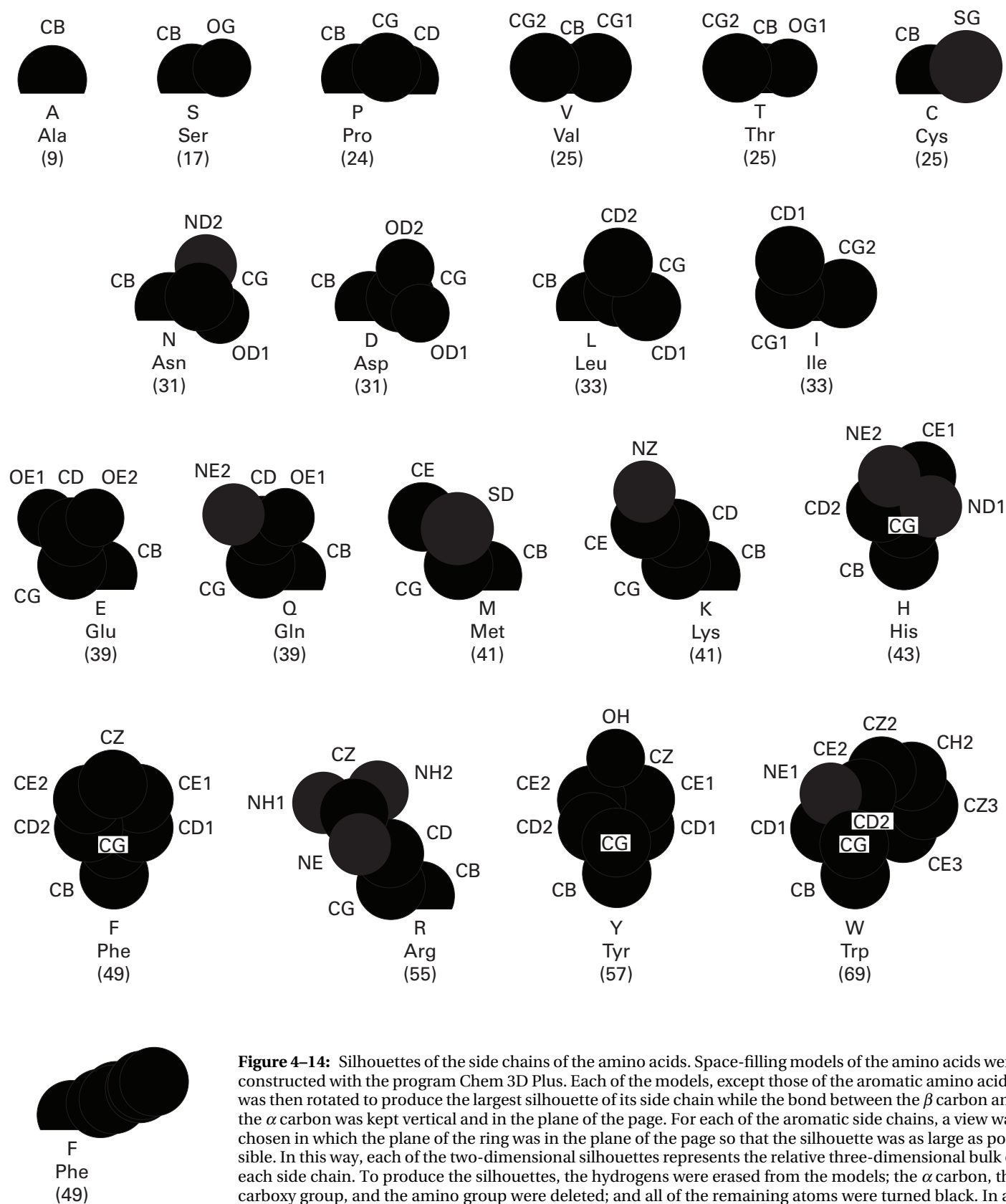


Figure 4-14: Silhouettes of the side chains of the amino acids. Space-filling models of the amino acids were constructed with the program Chem 3D Plus. Each of the models, except those of the aromatic amino acids, was then rotated to produce the largest silhouette of its side chain while the bond between the β carbon and the α carbon was kept vertical and in the plane of the page. For each of the aromatic side chains, a view was chosen in which the plane of the ring was in the plane of the page so that the silhouette was as large as possible. In this way, each of the two-dimensional silhouettes represents the relative three-dimensional bulk of each side chain. To produce the silhouettes, the hydrogens were erased from the models; the α carbon, the carboxy group, and the amino group were deleted; and all of the remaining atoms were turned black. In all of the silhouettes, except those of the aromatic side chains, the α carbon would occupy the position of the label. In each of the silhouettes of the side chains of the aromatic amino acids, the β carbon is directly above the label. The number of electrons in each side chain is indicated in parentheses. The standard crystallographic code for each atom in each side chain is indicated. A silhouette of phenylalanine viewed edge on is also included.

such unstructured electron density fills a large void in the map representing a significant portion (100–200 aa) of the molecule of protein.⁴⁷ Usually, however, it is short segments of the tube that are blurred or missing, breaking its continuity. Another source of ambiguity is when one segment of the polypeptide crosses another segment too closely to follow each tube confidently through the intersection. Missing segments of electron density and ambiguous crossings have led to serious errors in the fitting of polypeptides into maps of electron density,^{26,40,48,49} and these errors are often corrected by paying close attention to the patterns of the protrusions along the tube and correlating them with the sequence.⁵⁰

Such **errors** in tracing the polypeptide occur frequently enough that any crystallographic molecular model should be considered provisional until it has been shown to agree with other independent observations. This is an important point because crystallography has often assumed the mantle of infallibility. Even incorrect crystallographic molecular models look completely convincing. It should be pointed out, however, that they are convincing not because of the way the polypeptide is folded but because they are constructed with covalent bonds of the correct lengths and angles. These latter features are not indicative of the reliability of the crystallographic molecular model because they were incorporated into it automatically.

After the polypeptide has been fit into the map of electron density, certain regular patterns, collectively referred to as **secondary structure**, can be seen in the arrangement of the polyamide backbone. The regular patterns that are seen in the crystallographic molecular models of proteins are **α helices**, **β structures**, and **β turns**. α Helices and β structures were first observed in hypothetical models built by Pauling and his collaborators (Figure 4–16).^{51–53} After the models had been built, it was found that certain of their dimensions were consistent with molecular dimensions that had been observed in patterns of X-ray diffraction from oriented fibers of protein such as hair and silk,^{52,54} but it was not until much later that these structures were actually observed in maps of electron density of proteins. Several strands of β structure (Figure 4–15) are often joined together to form pleated sheets (Figure 4–16). Such sheets can be formed from strands all running parallel to each other or alternating in their orientation and thus each antiparallel to its neighbors or from a mixture of these two arrangements. β Turns (Figure 4–16D) were first noticed by Venkatachalam in the crystallographic molecular models of a cyclic hexapeptide, a short tetrapeptide, and the protein lysozyme.⁵⁵

Aside from the lengths of the covalent bonds of the polyamide backbone, the main structural element responsible for these secondary structures is the **hydrogen bond**, which is a noncovalent interaction that forms between the dipole of the nitrogen–hydrogen bond of one amide and one or two of the lone pairs of electrons

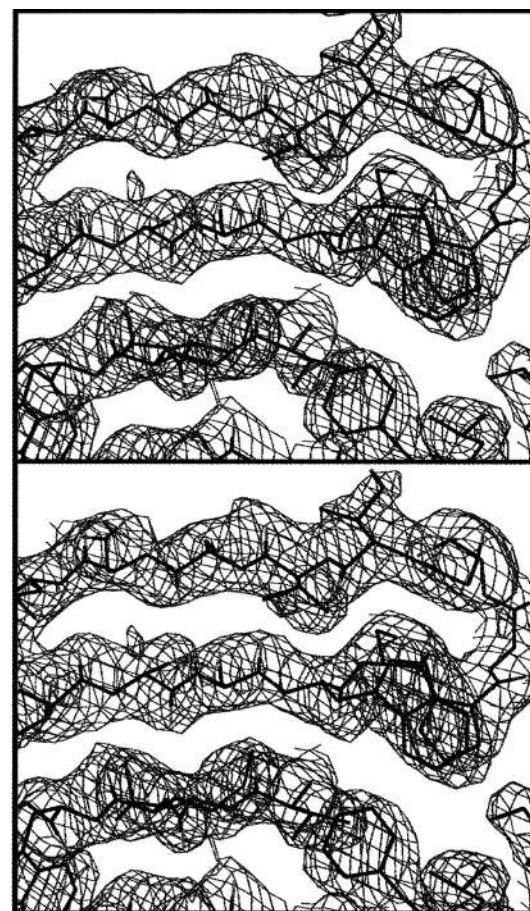


Figure 4–15: Fitting the molecular model of the polypeptide into the map of electron density (Bragg spacing ≥ 0.21 nm) for galactose oxidase.¹³⁷ The unrefined map of electron density shown was calculated with phases that were estimated by use of isomorphous derivatives obtained with K_2PtCl_4 , H_2IrCl_6 , and $Pb(NO_3)_2$ and then improved by iterative solvent flattening. The figure shows skeletal models of the segments –VLMW– and –TSSWDPSTGIVSDR– from the sequence of the protein fit into the three vertical tubes of electron density. These are three strands of an antiparallel β sheet. The large protrusions for the two tryptophans, the methionine, and the phenylalanine, which intrudes into the image from the left, are significant features along the course of the tube representing the polypeptide that confirm the fit. The isoleucine, valines, aspartate, and leucine are less dramatic protrusions. The contortion of the polypeptide at the proline is yet another indication that the sequence has been matched properly with the map of electron density.

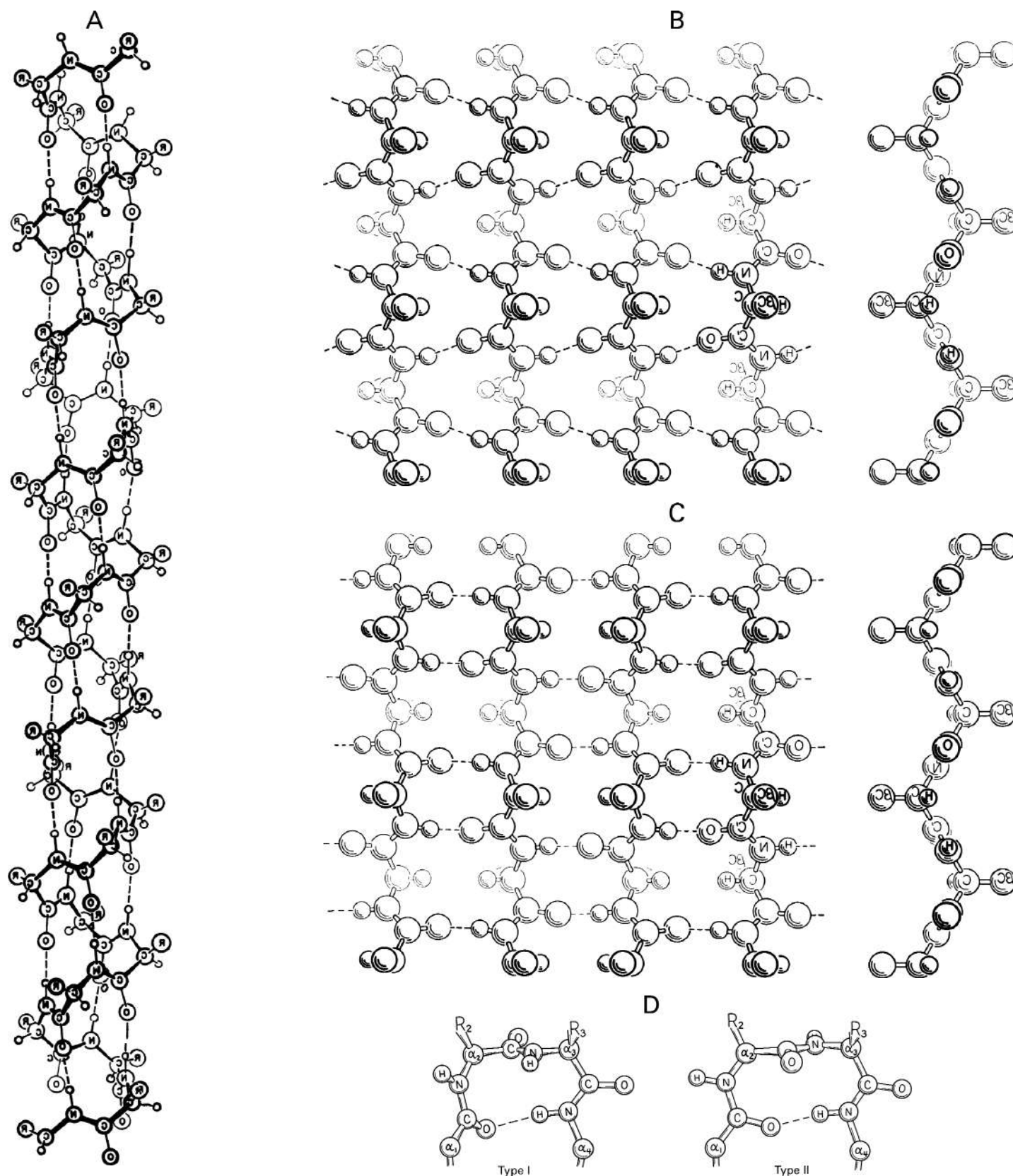


Figure 4-16: Four types of secondary structure found in molecules of protein: (A) α helix,⁵¹ (B) parallel β sheet,⁵² (C) antiparallel β sheet,⁵² and (D) two types of β turn (type I and type II).⁵³ The polyamide backbone can be traced by the pattern $\dots N, C\alpha, CO, N, C\alpha, CO, N, C\alpha, CO \dots$ and the side chains are the groups protruding (marked R, O , or H , respectively). Side views of the β sheets are shown to the right of each overhead view to demonstrate the pleats. Reprinted with permission from refs 51–53. Copyright 1951 National Academy of Sciences and 1981 Academic Press.

on the acyl oxygen of another amide of the backbone of the polypeptide. These hydrogen bonds are indicated in Figure 4-16 by dashed lines. A hydrogen bond connects the acyl oxygen contributed to the polyamide backbone by each amino acid in a sequence coiled into an α helix with the amido nitrogen-hydrogen contributed to the backbone by the amino acid four positions farther along (Figure 4-16A). In pleated sheets of parallel β structure (Figure 4-16B), the amido nitrogen-hydrogen and the acyl oxygen contributed by an amino acid in one of the polypeptides are connected by hydrogen bonds to the acyl oxygen and amido nitrogen-hydrogen, respectively, of amino acids two positions apart from each other in the sequence of a neighboring polypeptide to form a ring containing 12 atoms. In pleated sheets of antiparallel β structure (Figure 4-16C), hydrogen-bonded rings of 14 atoms and 10 atoms alternate along a ladderlike structure. The only structural element that defines the conformation of a β turn (Figure 4-16D) is a hydrogen bond between the acyl oxygen of the first amino acid in the turn and the amido nitrogen-hydrogen of the fourth and last amino acid in the turn.

Secondary structures enforce particular geometries on the conformation of the polypeptide. The β turn causes the polypeptide to double back on itself, often to form a hairpin the two tines of which are cross-connected in antiparallel β structure. An α helix has a right-handed pitch,* and the absolute stereochemistry of the L-amino acids causes each side chain, the R groups in Figure 4-16A, to cant toward its amino terminus. The side chains protrude from the helical core at intervals of about 100° . β Structure is pleated when viewed from the side (Figure 4-16B,C) owing to unavoidable steric requirements resulting from the angles of the covalent bonds along the polypeptide. In pleated sheets of β structure, the side chains of the amino acids in the sequence of each strand alternately protrude to one side and then the other of the surface in which the strands of polypeptide lie.

When the molecular model of the polypeptide has been fit into the tube of electron density, the final structure of its conformation represents, within the accuracy of the map of electron density, a skeleton of the actual molecule of protein. This **crystallographic molecular model** is the product of fitting atomically accurate molecular models of known covalent structures into a map of electron density.† The resulting arrangement of

the segments of secondary structure in three dimensions produces a representation of the tertiary structure of the folded polypeptide. The **tertiary structure** of a protein is the complete conformation into which its polypeptide is folded in its native form.

An example of a crystallographic molecular model is the one constructed for the protein penicillopepsin (Figure 4-17).⁵⁶ To obtain a full understanding of this molecular model, it must be viewed stereoscopically. The five panels of Figure 4-17 show drawings of the same view of the model. In Figure 4-17A, all of the atoms in the crystallographic molecular model are displayed in **skeletal representation**; in Figure 4-17B, the side chains of the amino acids have been removed to focus attention on only the polyamide **backbone** of the polypeptide and its hydrogen bonds, which are indicated by dashed lines; and in Figure 4-17C, only the α carbons of each amino acid are displayed, each connected to its two immediate neighbors in the amino acid sequence by line segments to create an **α -carbon diagram**. In all of the panels, the amino terminus is on the upper right at about 10 o'clock and the carboxy terminus is to the back at about 8 o'clock. You should follow the polypeptide [... N, C α , CO, N, C α , CO, N, C α , CO ...] through the whole drawing in Figure 4-17B. Note the α helices, β structures, and β turns. Compare what you see to the drawings presented by Pauling (Figure 4-16A-C). Note that α helices are rigid tubes while β structures are sinuous and flexible. Note the pleats in the β sheets. Distinguish between sheets of β structure formed from three or more strands and ribbons of β structure formed from only two strands. Now follow the polypeptide through the crystallographic molecular model in Figure 4-17A. Note the disposition of the side chains along secondary structures, and try to identify some of the amino acids.

The tertiary structure observed in a crystallographic molecular model is often presented diagrammatically (Figure 4-17D) in a **cartoon** where flat arrows are used to represent strands of polypeptide in β structure, with the head of the arrow at the carboxy terminus of the strand to provide the direction in which the chain is oriented, and cylinders are used to represent α helices. The tertiary structure of penicillopepsin, which you have explored in detail in Figure 4-17A, is represented, in the same orientation, by the diagram in Figure 4-17D. Follow the polypeptide through Figure 4-17, panels B and D, simultaneously.

The first three of the representations of the structure of a protein molecule that have been presented so far are skeletons of the crystallographic molecular model. The advantage of the skeletons is that the whole molecule can be examined simultaneously even in its interior. As with all molecules, flesh resides upon the bones in the form of the electron clouds that produced the map of electron density in the first place. It is possible to construct a model of a molecule of protein from space-filling units of the kind developed by Pauling and

* Put the four fingers of your right hand together, bent inward and horizontal, and put your thumb up. As you slide your fingers around a right-handed helix in the direction in which they are pointed, the helix rises in the direction in which your thumb is pointed. As you slide the fingers of your left hand around a left-handed helix, the helix rises in the direction of the thumb.

† To view a crystallographic molecular model on your own computer, find the file of the coordinates for the model in which you are interested at <http://betastaging.rcsb.org/pdb/Welcome.do> and download the file as text. Open the file with the program SwissPdbViewer, which can be obtained free of charge from us.expasy.org/spdbv/.

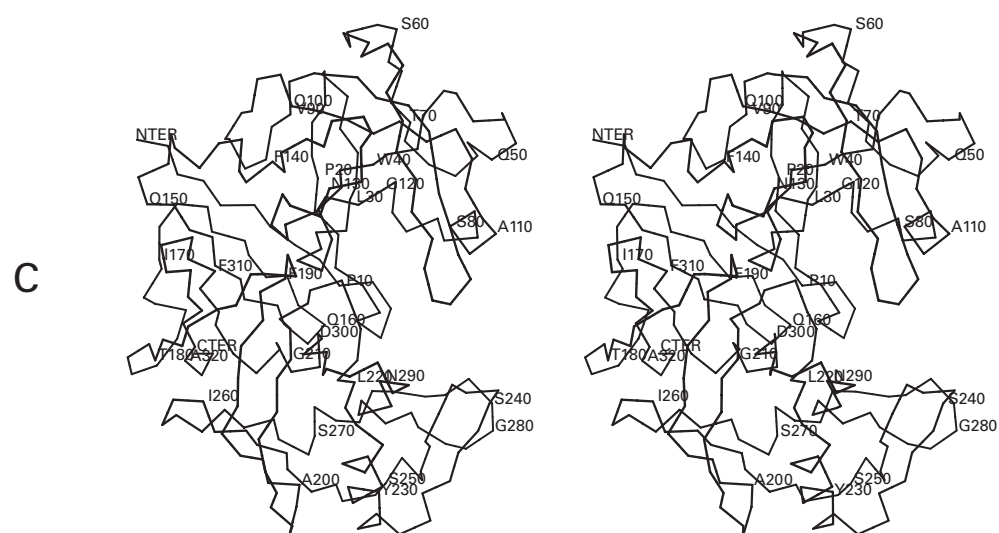
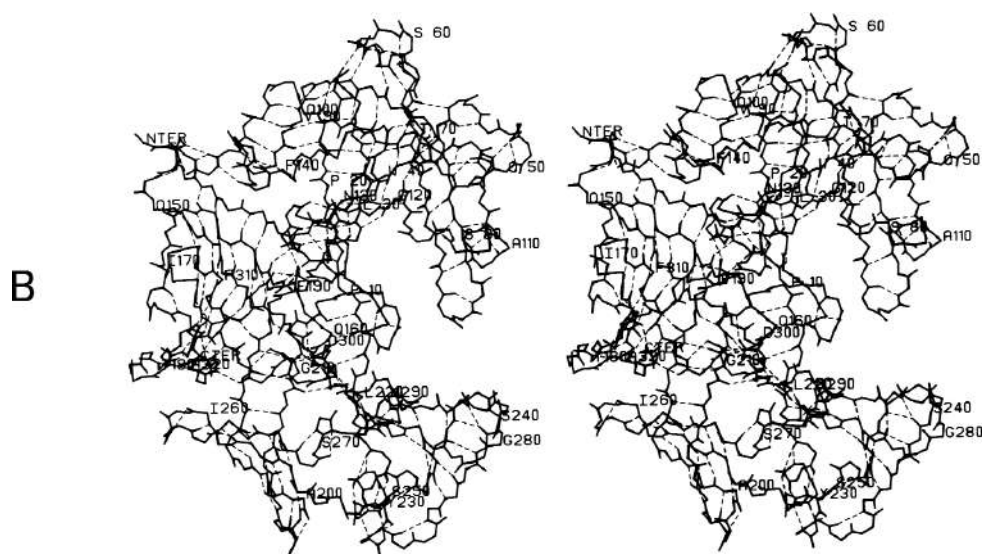
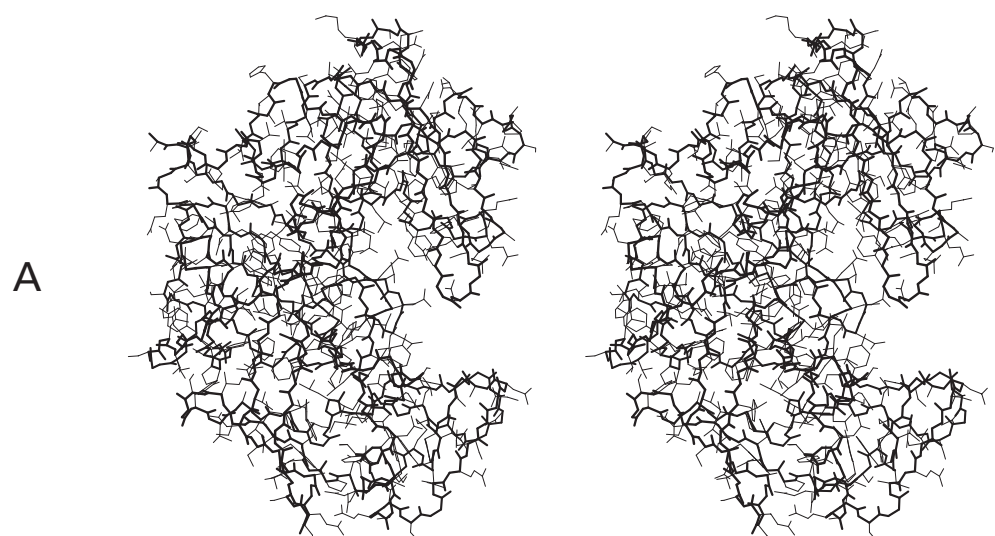


Figure 4-17: Crystallographic molecular model of penicillopepsin, from the mold *Penicillium anthinellum*.⁵⁶ In the first skeletal drawing (A), both the peptide backbone (heavy line segments) and the side chains (light line segments) of the amino acids are displayed, and no potential hydrogen bonds are indicated. This drawing was produced with MolScript.¹³⁹ In the second skeletal drawing (B), the side chains are left out and the crystallographer has assigned subjectively the locations of hydrogen bonds (dashed lines). Every tenth amino acid is identified and numbered to assist you in tracing the chain. Reprinted with permission from ref 56. Copyright 1983 Academic Press. In the α -carbon diagram (C), the positions of the α carbons of the amino acids in the crystallographic molecular model are designated by points and the points are joined by line segments. This α -carbon diagram often gives a clearer picture of the patterns of secondary and tertiary structure. This drawing was produced with MolScript.¹³⁹ In the cartoon (D), the skeletal drawing of panel B is represented diagrammatically. In a space-filling representation (E), each atom in the crystallographic molecular model is represented by a sphere with its van der Waals radius. This drawing was produced with MolScript.¹³⁹ As in the stereo image of Figure 3-9, black spheres are carbon atoms; gray, nitrogen; and white, oxygen.

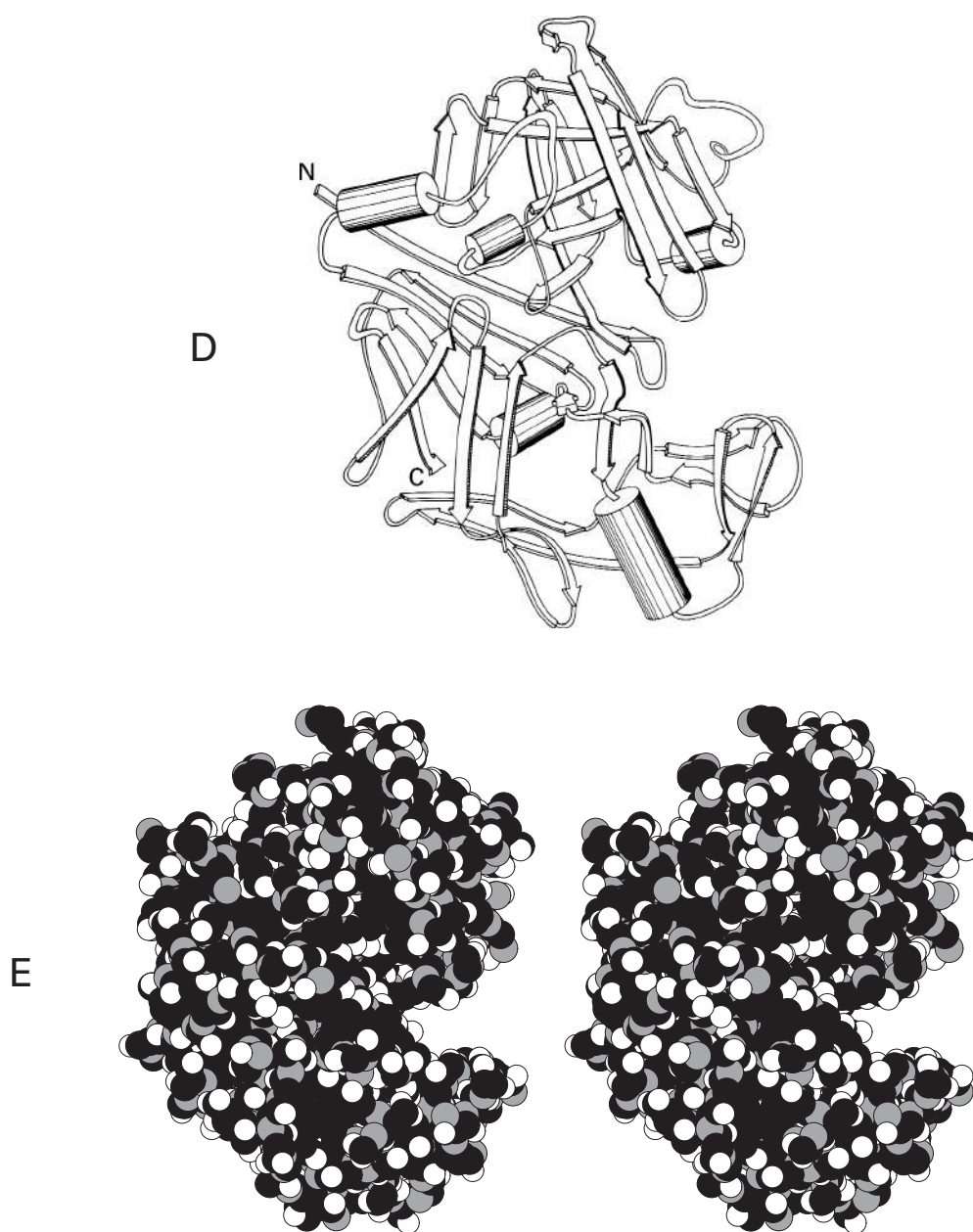
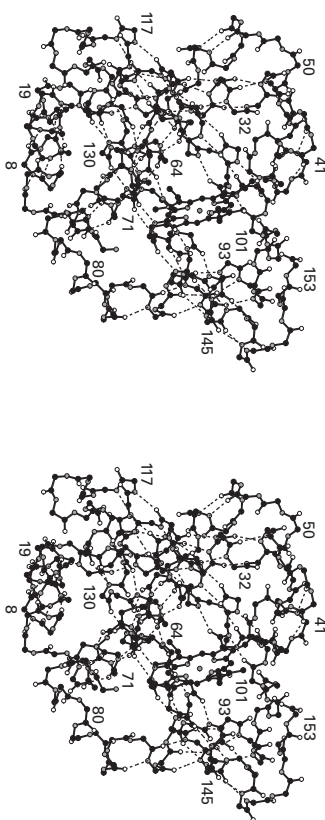


Figure 4-18: Ball and stick drawing of the polyanide backbone of myoglobin. The positions of the atoms in the backbone are those of the atoms in the crystallographic molecular model (Bragg spacing ≥ 0.17 nm) of myoglobin from *Physiter catodon* built from a data set collected from a crystal at 40 K.⁵⁹ The color code of the atoms in this figure, as in Figures 3-9 and 4-17, and for most of the ball and stick drawings in the book is black for carbon, gray for nitrogen, and white for oxygen. The radii of the balls for the atoms are equal to the van der Waals radii of the respective atoms divided by 6.4, so the sizes of the balls differ with the type of atom. Consequently, had there been a sulfur in the model, it would have been indicated by a gray ball larger than the gray balls of the nitrogens. Hydrogen bonds are indicated by dashed lines. Only the hydrogen bonds within the eight α helices designated by the crystallographers are drawn. The heme within the protein is drawn in its entirety; the iron at its center is drawn as a gray sphere. The polyanide backbone begins with a nitrogen and ends with a carboxylate. This drawing was produced with MolScript.¹³⁹



Corey (CPK models) after the coordinates of the individual atoms in three dimensions have been gathered from the skeletal model. A three-dimensional photograph of such a model of the protein lysozyme can be seen in Volume 243 of the *Journal of Biological Chemistry*.⁵⁷ This photograph and a similar drawing found in Volume 32 of *Biochemistry*⁵⁸ produce a reliable mental image of the molecular structure of a properly folded polypeptide. A

stereoisomer of a **space-filling representation** of penicillopepsin is presented in Figure 4-17E in the same orientation as the other drawings of the crystallographic molecular model of the protein.

The space-filling representation in Figure 4-17E emphasizes the tight packing of the atoms of the protein in its tertiary structure. Penicillopepsin is an example of a **globular protein** is a protein in which the entire polypeptide is folded into a compact structure the three dimensions of which are of the same order of magnitude. There are also many proteins in which **globular units** are held together by flexible segments of polypeptide as well as **fibrous proteins** in which the folded polypeptide forms a structure severely elongated in one of its dimensions.

Penicillopepsin (Figure 4-17) contains mostly antiparallel β structure where adjoining strands are often connected at one end by β turns. Myoglobin (Figure 4-18)⁵⁹ is an example of a protein that is almost entirely α helical. Most proteins are mixtures of these two major types of secondary structure, β turns, and a certain amount of **random meander**.

If one assumes for the moment that a particular crystallographic molecular model has been constructed correctly and represents, to the level of precision of the initial map of electron density, the molecules of protein as they are packed in the crystal, what relationship does this structure have to the molecules of protein when they are in solution in the cytoplasm?

The existence of the diffracted reflections requires that every fundamental unit cell in the crystal contain the same distribution of structured matter. If the fundamental unit cell contains only one molecule of protein, then every molecule of protein in the crystal must have exactly the same structure. It has always been observed, however, that in fundamental unit cells containing several molecules of the same protein, the several maps of electron density for the several molecules are quite similar and differ from each other only at the surfaces of the molecules where **differences in crystal packing** have caused flexible side chains or short, flexible loops of backbone to assume somewhat different orientations. For example, most of the polypeptide in the four asymmetrically arrayed molecules of adenylosuccinate synthase in the fundamental unit cell coincides to within 0.03 nm when the four separate crystallographic molecular models are superposed. This coincidence is well within the limits of the accuracy of the model, but four **loops** of 5-9 amino acids on the surface of the protein deviate in their positions among the four different crystallographic molecular models by 0.15-0.5 nm because of differences in crystal packing.⁶⁰ From the fact that only differences such as these are usually observed, it follows that all of the structured regions of the molecules of protein within a given crystal usually have essentially the same conformation.

Over the inflexible portion of a globular protein or within each of the globular regions of a protein in which

globular structures are held together by flexible unstructured segments of polypeptide, there is little doubt that the one structure present in the crystal is the same as the unique structure, or is the same as one of a limited number of unique structures, assumed by the protein in free solution and therefore its **native structure**. First, unlike the usual anhydrous crystals of small molecules, a crystal of protein is 40–70% water.⁷ This water usually surrounds each molecule of protein almost entirely, and the contacts between molecules of protein in the crystal are adventitious and not extensive.⁶¹ Consequently, the molecule of protein is still dissolved in the same **aqueous solution** from which it crystallized. Second, there are many instances in which the same protein has been crystallized under two or more different conditions and was found to be incorporated into the two or more **different fundamental unit cells** with completely different orientations, yet the respective maps of electron density were almost indistinguishable from each other and could be superposed.⁷ For example, the polypeptides in the two crystallographic molecular models of subtilisin from *Bacillus alcalophilus* produced from the two nonisomorphous crystal types coincided to within less than 0.1 nm except at two short surface loops.⁶² T4 Lysozyme has been crystallized in 25 different nonisomorphous forms and crystallographic molecular models have been prepared from all of them. This molecule contains two independently folded globular portions connected by a flexible segment of polypeptide, and the angle between these two portions can vary by up to 45° over the different crystals, but within each of the two portions the conformation into which the polypeptide is folded is always the same.⁶³ Third, crystals of a protein usually retain its **enzymatic activity**,⁷ albeit sometimes at a lower rate, and this also indicates that the structure of the protein has not changed during its crystallization. In fact, when crystals of protein are suspended in an organic solvent that is sufficiently immiscible with water that the crystals retain all their water of crystallization and their interior remains a separate aqueous phase, the protein is unaffected and the crystals retain their enzymatic activity.⁶⁴ Fourth, Raman spectroscopy can be performed on solids as well as liquids, and when the **Raman spectrum** of ribonuclease in solution was compared to its Raman spectrum in the crystal, the two were virtually identical in the region of the amide III vibrations, a region that would be sensitive to any changes in the structure of the polypeptide chain that might have occurred during crystallization.⁶⁵ Finally, molecular models from a number of proteins in solution have been obtained by **nuclear magnetic resonance spectroscopy**, and at their level of accuracy, they are indistinguishable from the crystallographic molecular models of the same proteins.^{66,67}

There are proteins that have been shown to be able to assume two stable structures in rapid equilibrium with each other in solution, and in some cases, two different crystals can be made, each exclusively

incorporating one of these respective structures. The crystallization and elucidation of the structures of deoxyhemoglobin and oxyhemoglobin provide an example.⁶⁸ When such crystals are exposed to a ligand that binds to the protein they contain and coincidentally elicits the change in the structure of the protein, the crystals will often shatter⁶⁹ as the protein assumes the new structure, which is incompatible with the former crystal lattice. In addition to presenting another observation consistent with the conclusions that the molecules of protein in the crystal retain the potentialities that they assume in solution, this observation suggests why some crystals are not enzymatically active. If expression of enzymatic activity requires that the protein change its shape slightly and reversibly each time it catalyzes the reaction and that change in shape is sterically hindered by the lattice of the crystal, the protein would not be able to display activity.

Crystals of citrate (*si*) synthase provide an example of such a situation.⁴¹ This protein can be crystallized under different sets of conditions that yield two different types of crystals containing **two different conformations** of the protein. From a careful examination of the maps of electron density for these two conformations, it became clear that each time the enzyme in free solution converts acetyl-S-CoA and oxaloacetate into citrate and coenzyme A, it passes back and forth between these two conformations. Neither crystal is enzymatically active, but upon dissolving either, full activity is restored. The conclusion drawn was that the packing of the molecules of protein in the crystal sterically prevented the movement between the two conformations necessary for enzymatic activity, not that either crystallographic molecular model was unrepresentative of the enzyme.

The most compelling argument for the identity of the structure seen in the crystal and the structure assumed by the protein in solution is that the structure seen in the crystal makes sense. Over the more than three decades that crystallographic molecular models of high accuracy have been available for examination, what has been seen has consistently provided reasonable explanations for the behavior of the respective proteins in solution. These explanations have stimulated experiments to test those explanations that have usually yielded informative results. Often an experiment will rule out a hypothesis based on an examination of the structure, but the more informed reexamination of the structure that then occurs usually turns up the original error of judgment. The fact that a crystallographic molecular model makes sense is an unambiguous verification that it represents the actual structure of the molecule of protein even when it is in the crystal, let alone in solution.

Suggested Reading

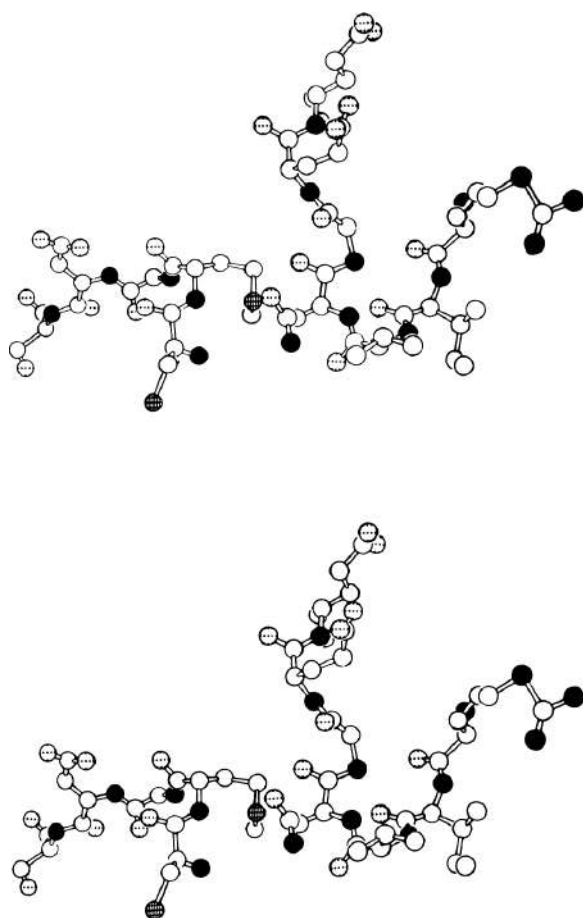
Wyckoff, H.W., Tsernoglou, D., Hanson, A.W., Know, J.R., Lee, B., & Richards, F.M. (1970) The three-dimensional structure of

ribonuclease S: Interpretation of an electron density map at a nominal resolution of 2 Å, *J. Biol. Chem.* 245, 305–328.

Brändén, C., and Jones, T.A. (1990) Between objectivity and subjectivity, *Nature*, 343, 687–689.

Problem 4–4: The structure below was drawn from a crystallographic molecular model of a particular protein.⁷⁰ It depicts only a small portion of the entire molecule. Trace the polypeptide backbone through the structure.

- How many lengths of polymer enter the figure?
- Identify as many of the amino acids as you can along the polymer, and write out its sequence or sequences.
- Identify the symbols for the individual atoms. Which atoms are not depicted? Why?



Refinement

The result of fitting the molecular model of a protein into its map of electron density, either manually or automatically by computer, is an initial crystallographic molecular model. At this stage, the accuracy of the

crystallographic molecular model is sufficient to define the patterns in which secondary structures are arranged. Because individual atoms, however, usually do not appear in the initial map of electron density, the initial crystallographic molecular model usually does not have sufficient accuracy to establish atomic details. These are of importance in their own right as well as being essential to understanding most of the biological functions of proteins. If the data set has been gathered to narrow enough Bragg spacing (0.3–0.25 nm or less), the accuracy of a crystallographic molecular model can be improved significantly by the process of refinement. The **refinement** of a crystallographic molecular model is the systematic adjustment of the positions of its atoms and the uncertainties of those positions and the addition to the model of portions of its covalent structure unobserved initially as well as molecules of solutes and water so that the amplitudes of the set of structure factors calculated from the model reproduce the observed amplitudes of the data set as closely as possible.

Although the fold of the polypeptide chain and the general positions of the side chains of the individual amino acids usually do not change significantly upon refinement, the atomic details of both the polypeptide and the side chains almost always change dramatically. If it is the case that the dramatic changes occurring during refinement actually do bring the molecular model closer to reality, then the atomic details observed in initial, unrefined molecular models are best ignored until the refinement has validated their existence.

The first step in a refinement is to calculate the amplitudes of the structure factors that the initial molecular model itself would produce, so that these amplitudes can be compared to the observed amplitudes of the data set. Once the initial molecular model of the polypeptide has been fit into the map of electron density to the satisfaction of the crystallographer, the coordinates within the fundamental unit cell of each of its atoms other than the hydrogens can be determined by direct measurements of the model. Often, if the initial map of electron density is of high enough quality, some individual molecules of water and solutes can be observed and included in the initial model and their coordinates also measured. There are always large regions of bulk solution in which individual molecules of water and solutes are never delineated because these regions are fluid in the crystal and thus unstructured. These regions of bulk solvent are included in the initial model as geometric solids of the appropriate uniform electron density.

A set of theoretical structure factors can be calculated by Fourier transformation (Equation 4–7) from the coordinates of the atoms, the shape of the geometric solid occupied by the solvent, and the scattering functions for each atom and for the solvent. The amplitudes of this set of calculated structure factors are referred to as the **calculated amplitudes**, and the set of these ampli-

tudes is designated F_c . The set of simultaneously **calculated phases** of those structure factors is designated as α_c . The amplitudes of the original experimental data set or any subset thereof are referred to as the **observed amplitudes** and designated F_o . The set of phases estimated by isomorphous replacement are euphemistically referred to as the **observed phases**, and the set containing their values is designated as α_o . All of these designations, F_c , α_c , F_o , and α_o , refer to three-dimensional matrices, each containing 5000–100,000 elements, all individually indexed as either F_{hkl} or α_{hkl} .

The **only directly observed quantities** are the observed amplitudes F_o , and they are the only parameters against which the success of the construction of any molecular model can be judged. If the molecular model were an exact representation of the molecules of protein, small solutes, and water within the crystal and there were no systematic errors in the observed data set, the calculated amplitudes F_c would be identical to the observed amplitudes F_o . It is traditional to quantify the degree of this correspondence with a **crystallographic R -factor**:

$$R \equiv \frac{\sum_h \sum_k \sum_l |F_{o,hkl} - F_{c,hkl}|}{\sum_h \sum_k \sum_l F_{o,hkl}} \quad (4-9)$$

where $F_{o,hkl}$ and $F_{c,hkl}$ are the observed and calculated amplitudes, respectively, of the structure factor hkl . The summation is performed over all available pairs of corresponding observed and calculated amplitudes or some subset of the available pairs. Once the Bragg spacings included in the data set are less than about 0.5 nm, so that the electron density of the solvent can be properly reproduced, the differences between the initial molecular model and the real structure usually become more significant the smaller the Bragg spacing, and the value of the R -factor has a tendency to increase in magnitude as the data set is expanded to include the amplitudes of structure factors of smaller and smaller Bragg spacing.⁷¹ Therefore the minimum Bragg spacings of the reflections included in the data set must be known to assess the significance of the value of the R -factor.

The value of the R -factor is often presented as a measure of the **validity** of a particular crystallographic molecular model. Such claims should be ignored. An incorrect crystallographic molecular model can give a reasonable R -factor.⁴⁰ For example, an incorrect crystallographic molecular model (Bragg spacing ≥ 0.2 nm)⁴⁹ for the ferredoxin from *Azotobacter vinelandii* had an R -factor of 0.24, while the later, presumably correct, crystallographic molecular model (Bragg spacing ≥ 0.2 nm)⁷² had an R -factor of 0.21. An incorrect crystallographic molecular model (Bragg spacing ≥ 0.3 nm)⁴⁸ for the *ras* protein had an R -factor of 0.29, while the later, presumably correct, crystallographic molecular model (Bragg spacing ≥ 0.26 nm)⁷³ had an R -factor of 0.23. It is not the

value of the R -factor that should be used as a validation of the model but the agreement of the model with independent chemical observations. In the case of the ferredoxin from *A. vinelandii*, it was disagreements between the earlier crystallographic molecular model and several direct chemical observations of the protein that prompted a reevaluation.⁷² Both of these examples were situations in which the chains were incorrectly traced in the original maps of electron density, and this produced very large errors in the molecular model.⁴⁰ Smaller errors may often go undetected.

Usually, the initial molecular model yields an R -factor of 0.30–0.60. This means that the amplitudes calculated from the model differ on the average by 30–60% from the observed amplitudes. At first glance this seems alarming because a completely random acentric structure would give an R -factor of 0.59. It is not so disturbing, however, because it is obvious from direct observation (Figure 4–12) that a unique structure has been defined by the map of electron density. Nevertheless, such a large value of the R -factor indicates that the initial molecular model does not duplicate the structure of the actual molecule very accurately and suggests that there is **room for improvement**. The improvements made in the structure after the initial model has been constructed are the refinements. The goal of refinement is to produce a molecular model the calculated structure factors of which have amplitudes as close as possible to the respective observed amplitudes. To accomplish this goal, the positions of each of the atoms in the model are adjusted in such a way that the R -factor decreases in magnitude. Only when it is realized that models of molecules of protein have 500–10,000 atoms that are not hydrogen and that the movement of any one of these atoms in the model affects the amplitudes of all of the structure factors in the set F_c is the task of refinement placed in a proper perspective.

The most easily understood way to perform a refinement proceeds by calculating **difference maps of electron density**. When two sets of crystallographic amplitudes are available for the same structure or for two structures so similar that the same set of phases, α_{hkl} , can be used for both, a difference map of electron density, $\Delta\rho(x,y,z)$, can be calculated:

$$\Delta\rho(x,y,z) = \frac{1}{V} \sum_h \sum_k \sum_l (F_{hkl} - F'_{hkl}) \exp[-2\pi i(hx + ky + lz - \alpha_{hkl})] \quad (4-10)$$

where F_{hkl} and F'_{hkl} refer to the entries in the two available sets of amplitudes. Equation 4–10 produces a map that has positive electron density wherever $\rho(x,y,z)$ is greater than $\rho'(x,y,z)$ and negative electron density wherever $\rho(x,y,z)$ is less than $\rho'(x,y,z)$, where $\rho(x,y,z)$ and $\rho'(x,y,z)$ are the two maps of electron density that would be cal-

culated directly from the respective amplitudes and phases.

Difference maps of electron density have many more uses than in refinement; but, in this particular instance, F_{hkl} are the entries in the set F_o and F'_{hkl} are the entries in the set F_c , and α_{hkl} are almost always those in the set of calculated phases. The intention of such a difference map of electron density is to indicate where the molecular model differs from the actual molecule. Where there is positive electron density in the map, the actual molecules in the crystal have matter in that location that is not present in the molecular model. Where there is negative electron density, there is matter at a certain location in the crystallographic molecular model that is not present in the actual molecules in the crystal. Adjustments can then be made in the model at the locations where significant differences occur. For example, the phenyl ring of a phenylalanine in the molecular model, sitting in negative electron density, can be moved over to occupy positive electron density. Unfortunately, as adjustments are made at one point in the molecular model, unavoidable shifts occur elsewhere, and the new R -factor calculated with the adjusted model usually does not change dramatically and often increases. A new difference map must be calculated to locate the new problems and the process repeated. This approach is an example of **manual tuning**, and it is slow and ultimately unsatisfactory.

There are several approaches to refinement that are designed to discover the optimal shifts of all of the atoms simultaneously by computation rather than manual tuning. These techniques are all based on the **minimization of a particular multivariate function** by solving large sets of simultaneous differential equations through matrix methods. Suppose that there is a multivariate function θ where

$$\theta = f(x_1, x_2, x_3, \dots, x_n) \quad (4-11)$$

Suppose also that the values for the variables x_j have been assigned initial magnitudes and that one wishes to discover the individual shifts, Δx_j , in the magnitudes of the assigned values of each of the variables that will produce a minimum numerical value for θ . It can be shown⁷⁴ that

$$\mathbf{A}_{i,j} \times \mathbf{h}_j = \mathbf{k}_i \quad (4-12)$$

where $\mathbf{A}_{i,j}$ is the square matrix ($n \times n$) the elements of which are

$$a_{ij} = \frac{\partial^2 \theta}{\partial x_i \partial x_j} \quad (4-13)$$

\mathbf{h} is the vector the elements of which are the individual shifts, Δx_j , in each x_j required to minimize θ , and \mathbf{k} is the

vector the elements of which are $\partial \theta / \partial x_i$. Equation 4-12 is solved⁷⁵ for \mathbf{h} , and its solution defines the shifts, Δx_j , in each variable Δx_j , required to produce a minimum value for θ .

Suppose the variables x_j are the positions of the atoms j in the molecular model of a protein, and

$$\theta = \sum_h \sum_k \sum_l w_{hkl} (F_{o,hkl} - F_{c,hkl})^2 \quad (4-14)$$

where $F_{o,hkl}$ is the observed amplitude of a particular structure factor hkl , $F_{c,hkl}$ is the calculated amplitude of the same structure factor hkl , and w_{hkl} is a weight assigned to a given structure factor. The magnitude of w_{hkl} varies with the certainty of the value of each observed amplitude. The values of F_o are fixed quantities, but the values of F_c are direct functions of the positions of the atoms j of the molecular model in the fundamental unit cell. These positions are the variables x_j . The solution of Equation 4-12, the vector \mathbf{h} , would then be a list of the shifts in the positions of each atom j of the molecular model that would produce a minimum value of θ and, presumably, a minimum value of the R -factor. This differs from manual tuning in that the effects of all shifts are considered simultaneously.

This conceptually simple but computationally complex approach suffers from the drawback that the shifts of the atoms j are unconstrained. In other words, every atom j in the molecular model would be allowed to shift independently regardless of the shifts imposed upon its neighbors. Consequently, atoms j would drift away from the neighbors to which they are connected through covalent bonds to produce unrealistic structures. When the data set extends to small enough Bragg spacing, this is not a problem because each atom j is represented by a sphere of electron density in the map which confines it to the vicinity of its proper location. But because the data set usually does not extend to such small Bragg spacings, some other means must be used to confine the individual atoms j of the molecular model. As one is dealing with the covalent structure of a polypeptide rather than a distribution of unconnected atoms j , the bond lengths and bond angles of the covalent bonds connecting the atoms j can be used to correlate their motions. For example, when any one of the carbon atoms j in the phenyl ring of a phenylalanine is shifted, all the others must also be shifted accordingly because they are all covalently attached to each other. To accomplish this, **constraints** are added to the minimization to force the motions of the bonded atoms j to be correlated.⁷⁶ The definition of θ is changed so that

$$\theta = \sum_h \sum_k \sum_l w_{hkl} (F_{o,hkl} - F_{c,hkl})^2 + \sum_q w_q (d_{s,q}^2 - d_{c,q}^2)^2 \quad (4-15)$$

where $d_{s,q}$ is the ideal, standard distance between any two atoms j that are rigidly connected by the covalent structure, for example, one of the ortho carbons and the para carbon of a phenyl ring, $d_{c,q}$ is the distance between them in the final, refined molecular model, and w_q is a weight the magnitude of which is chosen on the basis of how constrained the particular distance must be. If the two atoms j the positions of which are x_i and x_j , respectively, are directly attached to each other, w_q is large. If there are three or four covalent bonds between them, w_q is small. By adding the second term in Equation 4–15, bond distances and any rigid bond angles, such as those in a phenyl ring, are retained during the minimization.

The choice of which bond angles and bond distances to constrain is a subjective one that has a significant effect on the final crystallographic molecular model. Phenyl, indolyl, or imidazolyl rings are obvious, but exocyclic bond angles less so. It is usually unwise to constrain these bond angles in any structure other than the routine polypeptide, for example, in a covalently bound enzymatic inhibitor.⁷⁷ When accurate values for bond angles and bond lengths for an Fe_2S_2 cluster were available from a model compound, these were used as constraints early in the refinement of the crystallographic molecular model of the ferredoxin from *Anabaena* but then were removed from the process later on to incorporate the actual differences between the structures of the cluster in the protein and in the model compound.⁷⁸ On the contrary, it was concluded that the orientations of the ligands from the protein to the two irons in the nuclear cluster in the crystallographic molecular model of ribonucleoside-diphosphate reductase from *Escherichia coli* were inconsistent with spectral studies when no constraints on those orientations were applied but that constraining its structure to a conformation consistent with the spectral observations produced as satisfactory a refinement.⁷⁹ A **compromise** must be made between including enough constraints to hold the atoms j together and in reasonable orientation and including so many constraints that ideality replaces reality.

Alternatively, it has been proposed⁸⁰ that θ can be written as

$$\theta = \sum_h \sum_k \sum_l w_{hkl} (F_{o,hkl} - F_{c,hkl})^2 + w_e E_p \quad (4-16)$$

where E_p is a theoretically calculated value of the **potential energy** for the molecular model and w_e is a weight given to this term. The weight w_e is arbitrarily adjusted to make it more or less important during the refinement. In this approach, covalent bonds between atoms j remain because their distortion would produce a major increase in E_p . This approach has an advantage over the consideration of only interatomic distances (Equation 4–15) because any shift in an atom j in the molecular model causing it to overlap another atom j automatically causes

E_p to increase dramatically. The disadvantage of using E_p is that once overlaps are eliminated and covalent bonds are retained, the refinement is influenced by a large number of noncovalent forces imposed by the theoretical function and these may or may not be realistic. These biases influence the shifts of the atoms j dictated by \mathbf{h} .

Even in a rigid, anhydrous crystal of a small molecule, the atoms j and functional groups retain rotational and vibrational motion, which displaces them continuously and rapidly from their mean positions. The **vibrational and rotational motion** of the atoms j of a macromolecule of protein in a hydrous crystal are much more dramatic. There are vibrational motions involving segments of the polypeptide as well as those of the individual atoms j , and the water surrounding the molecule of protein does not sterically hinder the rotational or vibrational motion of its functional groups so dramatically as the immediate neighbors hinder the rotational motions of functional groups in an anhydrous crystal. Often the vibrational and rotational motion that occurs within the molecule of protein in the crystal is sufficient to blur the electron density of a side chain or a segment of the polypeptide so extensively that it is never present even in the refined map of electron density. Every atom j for which electron density is observed, however, is subject to vibrational if not rotational motion, and the extent of the resulting displacements of each atom j differs depending on the rigidity of its bonding and the rigidity of its surroundings. During the refinement of the crystallographic molecular model, it is possible to estimate the magnitudes of the actual displacements from its mean position experienced by each atom j in the molecule within the crystal.

The scattering factors f_j inserted into Equation 4–7 are affected by the displacement of each atom j from its mean position. It was noted earlier that, because of interference, as θ_{hkl} increases, the scattering from a given atom j decreases. Vibrational motion and rotational motion, because they also increase interference, also cause the scattering produced by the electrons around an atom j to decrease. It has been shown¹ that for the scattering factor of atom j in a molecule within a crystal

$$f_j = f_{0,j} \exp[-8\pi^2 \bar{u}_j^2 (\sin^2 \theta_{hkl}) \lambda^{-2}] \quad (4-17)$$

where $f_{0,j}$ is the scattering factor for atom j at rest, obtained from the usual table listing scattering factors as a function of scattering angle, and \bar{u}_j^2 is the **mean square amplitude of the displacement** in all directions of atom j from its mean position, regardless of the reason for that displacement. This mean square amplitude of the displacement incorporates not only the vibrational and rotational motion experienced by the atom j but also static disorder that may occur within the crystal lattice and that consequently affects the position of the atom j when it is averaged over the whole crystal. It is customary to define a **B value** (temperature factor) for atom j

$$B_j \equiv 8\pi^2 \bar{u}_j^2 \quad (4-18)$$

to simplify Equation 4-17 and obscure \bar{u}_j^2 . The units of the B value are nanometers².

The procedure for refining the position of a given atom j in a crystallographic molecular model (Equation 4-15) is usually expanded to include a refinement of both its position and its B value.⁸¹ By use of Equation 4-17, the individual B_j are incorporated as variables into Equation 4-7 in addition to the variables of mean position x_j , y_j , and z_j . The distances constraining the atoms j in Equation 4-15 are replaced with variances of interatomic distances. In this way the B values quantifying the displacements of the atoms j can be explicitly estimated during the refinement as well as the mean positions of the atoms j . It is also possible to resolve \bar{u}_j^2 into its three components along the **a**, **b**, and **c** axes and refine these three **anisotropic thermal parameters** for each atom j .⁸¹ Every atom j in a molecule of protein in a crystal undergoes its own particular motion with which are associated displacements from its mean position. It is important to remember that B values or anisotropic thermal parameters obtained crystallographically are only estimates of the actual magnitudes of these displacements.

The B value for a particular atom j in the crystallographic molecular model of a protein, as an estimate of its mean displacement from its mean position, provides an indication of its confinement. Usually, the atoms j in the interior of the molecule are confined by the surrounding atoms j and have low B values while those on the exterior, exposed to the water, have high B values. The flexibility of a segment of polypeptide is indicated by the set of B values for the atoms j of which it is composed. The atoms j in the most flexible or statically disordered segments of polypeptide or side chains, however, do not have B values assigned to them because they do not contribute to the diffraction and hence do not display any structured electron density in the map.

It has been pointed out that if only a portion of the available amplitudes is used for refinement with Equation 4-15 or another like it, the R -factor calculated with the portion not used is a more unbiased measure of the validity of the final model than the R -factor calculated with all of the amplitudes.⁸² The observed amplitudes and calculated amplitudes are divided at random into a test set containing 10% of them and a working set containing 90% of them. The working sets are the amplitudes $F_{o,hkl}$ and $F_{c,hkl}$ used in Equation 4-15, while the test sets are the amplitudes $F_{o,t}$ and $F_{c,t}$ used at each cycle of refinement to calculate a **free R -factor** (R_{free}) with Equation 4-9. In this way, the set of calculated amplitudes, $F_{c,t}$, used to calculate R_{free} are not the same calculated amplitudes, $F_{c,hkl}$, guiding the refinement, and R_{free} becomes a more independent measurement of the success of the refinement than the complete R -factor.

The use of a free R -factor during refinement should

eliminate errors in fitting the molecular model into the map of electron density. A standard R -factor, by design (compare Equations 4-9 and 4-15), must always decrease as the refinement progresses. If, however, the molecular model has been fit into the map of electron density incorrectly, for example, if the polypeptide has been incorrectly traced, then the free R -factor should remain constant or increase as the refinement progresses and the standard R -factor automatically decreases. As a result, the free R -factor has become an important indicator of the validity of a crystallographic molecular model.

In the simplest, but most time-consuming, format for refining a crystallographic molecular model (Figure 4-19),⁸³ Equation 4-15 is used to calculate shifts in the atoms j that cause θ to assume a minimum value. These shifts are used to create a new set of atomic positions x_j and another cycle of refinement is performed with these new positions as initial values for the x_j . In each of the first few of these cycles the value of R , and ideally that of R_{free} , drops (Figure 4-19), but the decreases become more modest at each cycle until no further progress can be made. The reason for this is that the process of refinement has become trapped within a **local minimum** of the function θ because refinement performed in this way will only progress as long as θ is decreasing. Manual tuning, however, is a way to escape from the local minimum in which the process is trapped. At some point

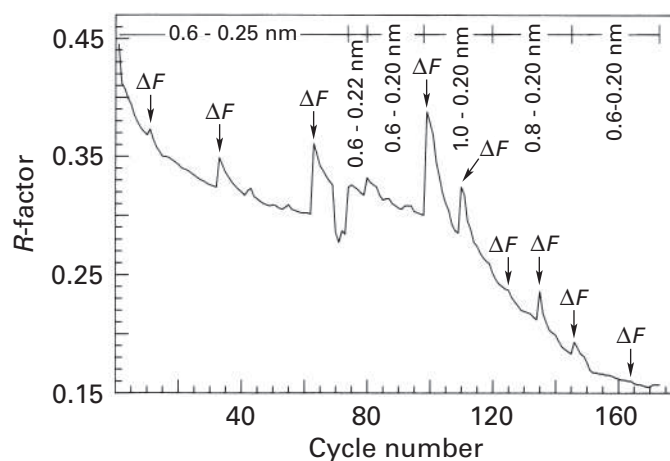


Figure 4-19: Progress of a refinement of the crystallographic molecular model of deoxyribonuclease I.⁸³ An initial molecular model was constructed by fitting a molecular model of the polypeptide into the initial map of electron density. The R -factor of the initial map was 0.45. The R -factor is presented as a function of the number of cycles of least-squares refinement performed. At the cycles indicated by ΔF , difference maps were constructed from F_o and F_c by use of the calculated phases of the molecular model to that point. These difference maps were used to rebuild the model manually and establish a new trajectory for the refinement. At cycle 98, molecules of water were added to the molecular model at locations identified in the maps of difference electron density. The range of Bragg spacings included in the data sets at each cycle is indicated at the top of the figure. The final R -factor of the refined molecular model was 0.16. Reprinted with permission from ref 83. Copyright 1986 Academic Press.

(designated ΔF in Figure 4–19), the decision is made to calculate a difference map of electron density. Adjustments of the current molecular model are made by manual tuning, and this allows the minimization to enter realistically a new trajectory. After this trajectory reaches a new local minimum, more tuning is performed. This strategy, however, is never followed today.

The manual adjustments performed at various times during this simplified strategy for refinement require a significant amount of time. Whenever the refinement reaches a plateau and no further progress is evident (Figure 4–19), the molecular model must be examined in detail and manually adjusted with the assistance of difference maps of electron density before a new trajectory can be initiated. It has been found that one way to avoid such time-consuming manual adjustments during the refinement is to combine molecular dynamics and refinement.⁸⁴

In a **molecular dynamics** simulation, atoms j are positioned in space, for example, by the fitting of the molecular model into the initial map of electron density and perhaps an initial round of refinement. A global potential energy function E_p , incorporating the individual potential energy functions of the covalent bonds and the nonbonded interactions, is calculated. The atoms j are then given kinetic energies appropriate to a certain temperature and allowed to move for a short interval (less than 1 fs) within this global potential energy function according to classical laws of motion. The new positions in turn create a new global potential energy function and the atoms j are allowed to move again in response to the new component potential energy functions, and so forth.

When molecular dynamics is used in crystallographic refinement, the global potential energy function, $E_{p,i}$ for each step i in the usual molecular dynamics calculation is augmented by an effective potential energy

$$E_{p,f} = w_x \sum_h \sum_k \sum_l w_{hkl} (F_{o,hkl} - F_{c,hkl})^2 \quad (4-19)$$

where w_x is a weighting factor chosen so that $E_{p,f}$ has the same magnitude as E_p and F_c is the set of the amplitudes of the structure factors calculated for the instantaneous distribution of atoms j after each step in the molecular dynamics calculation. This effective potential energy, $E_{p,f}$, constrains the atoms j during the molecular dynamic trajectory to the vicinity they occupied in the original molecular model, but if a high enough kinetic energy is applied, the atoms j can move as much as 0.3 nm from their initial positions.^{85,86} This is what allows the structure to break out of local minima of the function θ .

The process proceeds in several steps referred to as **simulated annealing**.⁸⁷ Initially a high kinetic energy is applied to the atoms j (high temperature), and then the kinetic energy is decreased to finish within a minimum of

potential energy. It is while the kinetic energy is high that local minima of potential energy, and hence local minima of the function θ , can be passed through. It has been shown that in this way the R -factor can be minimized with much less need for manual adjustment of the molecular model.⁸⁸ Because rather high simulated temperatures are used, however, unexpectedly large movements of segments of the model can occur,⁸⁹ so the necessity to examine difference maps of electron density and perform manual adjustments remains. Nevertheless, with proper precautions, refinement by molecular dynamics converges on the same structure as refinement performed entirely by least-squares minimization and manual adjustment.^{85,86,88}

The use of simulated annealing by molecular dynamics for the purpose of pushing the refinement out of local minima includes coincidentally a large number of hidden constraints in the **potential functions** used for covalent bonds and nonbonding interactions that are necessary to cause the atoms j to move in each step. These potential functions were not constructed for solutes in aqueous solution, which is what a molecule of protein in a crystal is. As a result, they introduce **significant, uncontrolled biases** into the final molecular model.

These biases are most clearly manifested in the final positions of the charged side chains. The choice of a charge number for a side chain has a dramatic effect on its location in the final crystallographic molecular model in which simulated annealing is used for refinement.⁸⁹ This, however, should not be the case for charged groups in aqueous solution of moderate ionic strength. A compilation of the frequency of hydrogen bonds between oppositely charged donors and acceptors of hydrogen bonds in crystallographic molecular models found them to be no more frequent than hydrogen bonds between a neutral donor and a neutral acceptor or between a charged donor or acceptor and a neutral acceptor or donor, respectively.⁹⁰ This compilation was gathered from crystallographic molecular models refined by manual adjustments rather than by simulated annealing. Another compilation,⁹¹ gathered after refinement by simulated annealing had become widespread, found that hydrogen bonds between oppositely charged donors and acceptors were almost 5 times more frequent. Since the actual frequency cannot change, this latter result suggests that refinement by simulated annealing does consistently introduce artifacts into crystallographic molecular models. Because the potential functions for the attraction between these oppositely charged groups in simulations performed by simulated annealing are unrealistically strong, it would not be surprising if this procedure produced such interactions artifactually. Another indication of the unreliability of assignments of hydrogen bonds between oppositely charged donors and acceptors is that their identity usually changes significantly, and often dramatically, from an earlier version to

a later version of a crystallographic molecular model even though both versions were built from refined maps of electron density calculated from data sets gathered from the same crystal.

There are now at least five widely used procedures for refining crystallographic molecular models. It is reassuring that even though the final models prepared with each method differ detectably,⁹² when two of them are used to refine the same model with the same data set, the two refinements usually converge to a common structure.⁹³ Often two different refinement procedures are purposely used to reassure the investigators that a peculiar aspect of the molecular model is real.^{77,94}

As a refinement progresses, there is a noticeable improvement in the shape and continuity of the tube of refined electron density representing the polypeptide.^{35,95} Segments of the polypeptide in the initial molecular model often move during the refinement, sometimes as much as 1 nm, to assume their positions in the final molecular model,^{96,97} especially in regions where the initial map of electron density was vague. Elements of secondary structure missing in the initial map of electron density can appear and elements of secondary structure in the initial map can occasionally disappear upon refinement, and positions assigned to specific amino acids in the sequence of the protein within secondary structures can shift dramatically.⁹⁷

Locations where the published **amino acid sequence** is in error become obvious,⁹⁸ and it is sometimes possible visually to read the amino acid sequence of a segment of the protein as yet unsequenced.⁹⁹ If the map of initial electron density has been calculated from a data set gathered to narrow enough Bragg spacing (<0.16 nm), the electron density for individual amino acids in the initial map can be sharp enough that they can be tentatively identified and their side chains incorporated into the original molecular model even though the sequence of the protein is unavailable.¹⁰⁰ As the refinement progresses, mistakes in these initial assignments become obvious and can be corrected.

The electron density for the carbohydrate attached to a glycoprotein, if it is not disordered in the crystal, becomes progressively more detailed. The electron density for coenzymes, which are almost always held rigidly within the protein, also becomes easier to interpret. Posttranslational modifications, sometimes unexpected,¹⁰¹ as for example 3-(S-cysteinyl)tyrosine and β -hydroxytryptophan (Table 3-1), and sometimes hoped for, as for example the ester intermediate in the self-catalyzed pyruvylation of aspartate 1-decarboxylase (Equation 3-9),¹⁰² begin to appear in the difference maps of electron density. Previously unaccounted-for molecules of **water** (oxygen atoms j) and anions and cations from the crystallization solution that are bound at specific locations on the surface or in the interior of the molecules of protein begin to appear in the difference maps and they become sharp and repro-

ducible features. Any of these features not incorporated into the initial molecular model appear in a difference map of electron density because they are fixed at certain locations in the real fundamental unit cell by their specific covalent bonds and noncovalent interactions with the molecules of protein but are as yet missing in the model.

When the identity, location, and structure of each of these fixed molecules or portions of the covalent structure of the polypeptide that were not included in the initial molecular model, because they did not appear in the initial map of electron density, become sufficiently unambiguous, they are incorporated into the molecular model at that cycle of the refinement. Their inclusion causes a significant decrease in the *R*-factor because they are as real a feature of the actual crystallographic unit cell as the individual amino acids in the polypeptide, and they contribute accordingly to F_o . For example, in the refinement for deoxyribonuclease I (Figure 4-19), the inclusion of the water molecules observed in difference maps at cycle 98 caused the molecular model to be much more realistic and permitted the refinement to produce a significantly lower minimum of the *R*-factor than it had before they were included. This reasonable consequence suggests that the refinement is registering reality, but all of the changes taking place during the refinement are adequate evidence that a crystallographic molecular model is always provisional.

There are now more than 30 crystallographic molecular models of proteins that have been fit into maps of electron density calculated from data sets with minimum Bragg spacings so narrow (≤ 0.1 nm) that the individual atoms j , and in one instance even bonding electron density,¹⁰³ are clearly observed in the initial map.^{5,6,104-106} In these instances, few if any constraints were required during refinement. Nevertheless, almost all of even the most recently constructed crystallographic molecular models of proteins have not had the benefit of such accurate maps of electron density.⁴⁴ Consequently, regardless of how the refinement is performed, ideal bond lengths and bond angles are almost always enforced upon the crystallographic molecular model because if they were not, the refinement could not be performed at all. Therefore, if a refinement were performed entirely by the computer, the final molecular model would be confined by all of these implicit and often unsubstantiated constraints. To verify that the process of refinement has not biased the final structure, careful inspections of difference maps of electron density are always required to identify locations where the actual structure of the protein deviates from these simple expectations.

This inspection is routinely done by using **omit maps of difference electron density** (Figure 4-20). A segment of amino acids, a coenzyme, or a posttranslational modification in the final refined crystallographic molecular model is omitted, and the truncated model that

results is used to calculate $F_{c,omit}$ and $\alpha_{c,omit}$. The observed data set F_o and $F_{c,omit}$ and $\alpha_{c,omit}$ are used to calculate (Equation 4–10) a difference map of electron density. In this difference map the omitted segment appears as positive electron density. This positive electron density has the advantage that its details are defined only by the observed data set because nothing is present at this location in the truncated molecular model. The atoms j in the refined molecular model in this region are adjusted, if necessary, to fit within this difference electron density and added back to the molecular model. Then another segment of the updated molecular model is omitted and so forth over the whole structure. In this way, an attempt is made to incorporate into the final molecular model the ways in which the actual structure of the protein deviates from the ideal structure dictated by ideal bond lengths and bond angles and empirical functions of potential energy used during the refinement. It should be stressed at this point that the goal of all refinement is to produce a crystallographic molecular model that reproduces as accurately as possible the actual structure of the molecule of protein, including all of its perversities,¹⁰⁷ rather than some ideal structure consistent with a set of theoretical potential energy functions.

There is one interesting and enlightening aspect of the process of producing an omit map. After a segment of the molecular model has been omitted, it is necessary to perform additional cycles of refinement on the molecular model missing that segment before calculation of the $F_{c,omit}$ and $\alpha_{c,omit}$ used to produce the omit map of difference electron density.^{108–110} The reason for this requirement is that the positions of all of the atoms j in the initial refined model, not just the atoms j omitted themselves, contains information about the positions of the atoms j omitted. This information would be transmitted to the calculated amplitudes, $F_{c,omit}$, but even more critically to the calculated phases, $\alpha_{c,omit}$, that would be used to calculate the difference map of electron density were it not purged by repositioning the remaining atoms j in the truncated molecular model by additional cycles of refinement. These additional cycles must be performed on the model in which the omitted atoms j are missing and hence unable to bias the calculations performed during the cycles of refinement.

Now that the molecular model of the polypeptide is fit by computer into the map of electron density and the manual adjustments that were once performed manually by the crystallographer during refinement have been replaced by molecular dynamics, the intimate human involvement in these procedures that once occurred no longer occurs. There is no computational algorithm that approaches the acuity of an experienced human intellect. Consequently, at some point in the complete process the product must be examined carefully by the crystallographer if errors are to be avoided. It is the omit maps that present the most obvious opportunity for this **human intervention**.

The omit maps calculated for successive segments of the molecular model (Figure 4–20) must be examined carefully by the crystallographer to ensure that they do actually represent that segment of the model. If the polypeptide has been incorrectly traced and the wrong



Figure 4–20: Omit map of electron density.¹³⁸ The initial crystallographic molecular model for the amino-terminal domain of a variant glycoprotein from *Trypanosoma brucei* was built from a map of electron density calculated from observed amplitudes (Bragg spacing ≥ 0.29 nm) and phases from multiple isomorphous replacement. The molecular model was refined with the assistance of simulated annealing. From this refined molecular model, the first 31 amino acids of the polypeptide were omitted. The truncated molecular model was submitted to 40 additional cycles of refinement before phases, $\alpha_{c,omit}$, and amplitudes, $F_{c,omit}$, were calculated from it. These phases and amplitudes along with the observed amplitudes were used to calculate an omit map of difference electron density. The portion of that map including the segment of the polypeptide from Glutamine 12 to Glutamine 24 is presented. A molecular model of this segment of polypeptide in its final conformation is positioned within the omit map.

segment of the molecular model has been fit into a particular segment of electron density, that error will be obvious in an omit map of that segment of electron density.¹¹¹ The fit of the molecular model into the omit map of electron density must be adjusted manually by the crystallographer, not because a computer could not do so but because she must be convinced that the fit justifies the final conformation imposed upon the molecular model. Only in this way, with properly calculated omit maps and properly adjusted conformations, can the ideal structure resulting from the theoretical biases of the automated fitting and refinement be replaced by the real structure dictated by the observed amplitudes. For example, a hydrogen bond produced solely by the constraints of the refinement for which there is no evidence within the observed amplitudes will not appear in a properly calculated omit map and must be removed from the crystallographic molecular model. The conformation of a side chain produced by the constraints of the refinement may differ significantly from the conformation observed in an omit map and must be adjusted accordingly.

In addition to the polypeptide, the process of refinement adjusts the conformations of coenzymes and oligosaccharides in the crystallographic molecular model. **Coenzymes** can be either covalently bonded to the polypeptide as additional examples of posttranslational modifications (Table 3-1) or enclosed within it so tightly that they form an integral structural component. In either case, the coenzyme never leaves the protein and is incorporated with the protein into a crystal. At this point, these molecules will be considered to be merely small clouds of electrons that have interesting shapes. Usually the existence and covalent structure of these coenzymes is known before the protein is crystallized.

The electron density contributed by coenzymes known to be associated with a protein is always clearly featured because these molecules are enclosed within the protein and precisely aligned for functional purposes. The shapes of most coenzymes are unique, and they can usually be placed unambiguously into one of the envelopes of electron density unfilled by the polypeptide, but the decision as to when during the refinement they are included in the molecular model depends on the situation. If the initial, unrefined map is detailed enough and the coenzyme is large enough and of a shape peculiar enough, it can be inserted into its electron density at the same time the polypeptide is fit into its map of electron density. For example, the envelopes of electron density representing the four bacteriochlorophylls *b* in the initial map of electron density for the photosynthetic reaction center calculated from the phases estimated by isomorphous replacement were clear enough that molecular models of bacteriochlorophyll *b*, with its characteristic queue, could be inserted into several of them (Figure 4-21)¹¹² even before the polypeptide could be fit into its electron density. In fact,

the envelopes of electron density for the bacteriochlorophylls *b* could be distinguished from the envelopes for the almost identical bacteriopheophytins by the bulge of electron density due to the magnesium ions present in the former but missing from the latter. Usually, however, a coenzyme is added to the model at a step in the refinement when its electron density in the difference map becomes detailed enough to insert it unambiguously, but adjustments, often major ones,¹¹³ are made in its position and configuration as the map of electron density becomes more detailed during the cycles of refinement. The precise orientation of a coenzyme in the model is assigned in the end with an omit map (Figure 4-22).¹¹⁴

The crystallographic molecular model for myoglobin (Figure 4-18) displays the characteristic, intimate association between a coenzyme and the polypeptide that enfolds it. In this case, the heme is embraced by the α helices arranged to compose the entire structure, the purposes of which are to isolate the heme from the solution, to prevent two hemes from colliding, and to permit the heme to dissolve in water in addition to providing a fifth ligand to the iron.

Often **ligands** that are known to be specifically bound by the protein are included during the crystallization and are bound by the protein in the crystal. Significant changes can occur in the position and orientation of a ligand during refinement. For example, the molecular model of methotrexate inserted into the initial map of electron density of dihydrofolate reductase had to be adjusted significantly during refinement.¹¹⁵ The final position and orientation of the ligand are assigned in the final crystallographic molecular model by fitting them into features of electron density in omit maps.¹¹⁶

The positions in the amino acid sequence of a glycoprotein at which the **oligosaccharides** are attached are often known, so the locations of these serines, threonines, or asparagines in the map of electron density can be identified as soon as the polypeptide has been fit into it. Oligosaccharides are located on the outer surface of a protein and usually protrude into the aqueous phase surrounding it (Figure 4-23).⁸³ Under these circumstances, they are fully solvated, flexible, and structureless. This absence of a fixed structure is carried into the crystal, and the region within the fundamental unit cell occupied by the oligosaccharide is often featureless. Attempts to assign an atomic structure to such regions are probably irrelevant to an understanding of the behavior of an oligosaccharide in a biological situation where it will have no defined structure anyway. Sometimes, however, the carbohydrate is surrounded sufficiently by protein to assume a defined conformation and produce structured electron density. If the initial map of electron density is calculated from a data set gathered to narrow enough Bragg spacing and the oligosaccharide is sufficiently confined by the structure of the protein, a molecular model built from its previously determined sequence of monosaccharides can be unambiguously fit into that initial

Figure 4-21: Electron density assigned to a bacteriochlorophyll *b*,¹¹² one of the coenzymes in the reaction center from *Rhodospseudomonas viridis*. A skeletal model of the known atomic structure of the coenzyme has been placed within an envelope of electron density located in the initial, unrefined map. Note the bulge of electron density in the center of the coenzyme that results from the magnesium ion with its 10 core electrons. Reprinted with permission from ref 112. Copyright 1984 Academic Press.

map.¹¹⁷ With initial maps of lower quality, the conformation of the oligosaccharide in the crystallographic molecular model is adjusted during the refinement (Figure 4-24).^{118,119}

As the refinement progresses, isolated spherical features of electron density larger than those that can be assigned to molecules of water appear and become more prominent in the map of electron density. These are negative **ions** such as chloride or bromide or positive ions such as Na^+ , K^+ , Mg^{2+} , Ca^{2+} , Fe^{2+} , Co^{2+} , Mn^{2+} , Zn^{2+} , or Cu^{2+} . Often these features, such as the chloride ions in human collagenase 3,¹²⁰ can be assigned by their scattering strength and the ionic character of the protein surrounding them. Such an assignment is strengthened by showing that if the designated ion is incorporated into the molecular model during the refinement, its refined *B* value ends up in the same range as the refined *B* values for the other atoms *j* in the model.¹²¹ If an incorrect assignment had been made, the refinement would have adjusted the *B* value to an unrealistic magnitude to compensate for the incorrect scattering factor incorporated into the calculations (Equation 4-17).

It is also possible to use their anomalous dispersion to identify ions.¹²² For example, when the wavelength of the X-radiation was decreased from 0.154 to 0.1377 nm, passing through the absorption edges for Ni (0.1488 nm) and Cu (0.1381 nm), the magnitude of the electron density for a metal ion in the map calculated from structure factors from a crystal of nitrite reductase (NO-forming) did not decrease significantly, but when it was further decreased to 0.104 nm, below the absorption edge of Zn (0.1284 nm), it decreased significantly.¹²³ This effect demonstrated that the sphere of electron density in the map at this location represented a Zn^{2+} ion. A difference map of electron density calculated from a data set gathered at a wavelength of 0.0870 nm and a data set gathered at 0.1488 nm, where iron has significant anomalous dispersion compared to other transition metal ions,

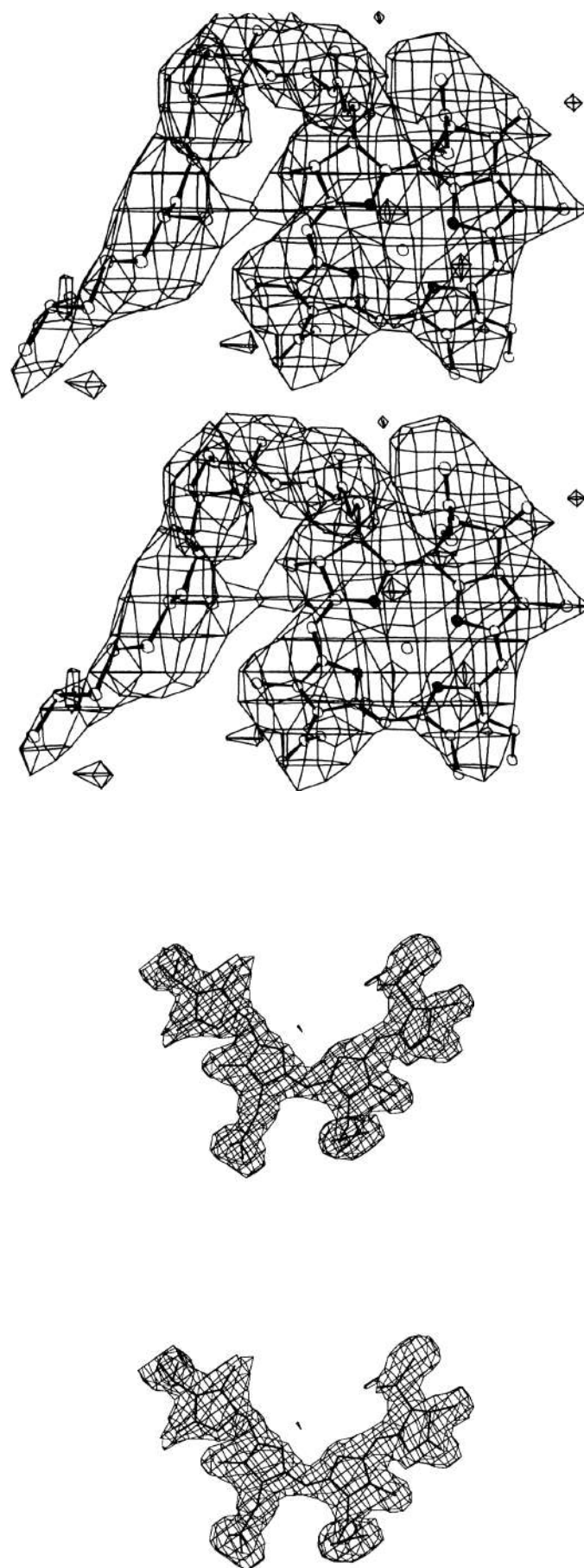
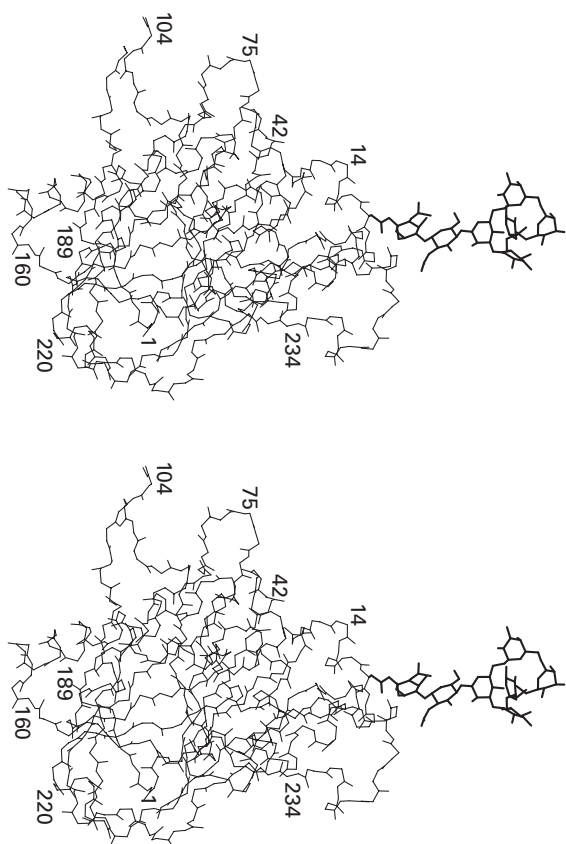


Figure 4-22: Omit map of electron density for one of the phycoerythrobilins in B-phycoerythrin from *Porphyridium sordidum*.¹¹⁴ A crystallographic molecular model constructed from an initial map of electron density (Bragg spacing ≥ 0.22 nm) was refined against the observed amplitudes with the assistance of simulated annealing. The phycoerythrobilin covalently bound by Cysteines 50 and 61 of the β subunit of the protein was then omitted from the molecular model, and an omit map of electron density was calculated. The final conformation chosen for the coenzyme is positioned within the electron density.

Figure 4-23: Skeletal drawing of the glycoprotein bovine deoxyribonuclease I. The skeletal drawing of the polyanide backbone (260 aa) is drawn with thin lines; and the oligosaccharide and the side chain of the asparagine (at the top of the structure), with thick lines. The positions of the atoms are from the crystallographic molecular model of the protein (Bragg spacing ≥ 0.20 nm).⁸³ Structured electron density could be observed in the vicinity of Asparagine 18 in the crystallographic model, but it was only large enough to contain the first two *N*-acetylglucosamines and one of the mannoses of the high-mannose oligosaccharide (GlcNAc₂Man₅; incomplete realization of the first entry in Table 3-3) known to be attached at this asparagine. The other four mannoses of the molecular model were arbitrarily positioned. The core of the structure is a laminate of two pleated β sheets. The laminate is flanked above and below by α helices and random meander. There was no electron density for the two amino acids preceding Cysteine 104 because they were either vibrating too widely or statically disordered in the crystal. The locations of convenient positions in the amino acid sequence of the protein are numbered to assist you in tracing the chain of the polypeptide. This drawing was produced with MolScript.¹³⁹



showed strong electron density at the eight positions occupied by the metal ions in the two tetranuclear clusters in hybrid-cluster protein from *Desulfovibrio vulgaris*, an observation demonstrating that all of those metal ions are irons.¹²⁴

Although they can be readily observed in maps of scattering density calculated from neutron diffraction,

atoms *j* of **hydrogen** are almost never observed in X-ray crystallographic studies of proteins because, unlike atoms *j* of carbon, nitrogen, oxygen, and sulfur, atoms *j* of hydrogen have no inner-shell electrons. Because they are in smaller orbitals, inner-shell electrons have the highest electron density and produce most of the features observed in the usual maps of electron density. In general, if hydrogens are present in a crystallographic molecular model, it is because the crystallographer knows they are there even though they were not observed. When, however, crystallographic molecular models obtained from data sets gathered to Bragg spacings of less than 0.1 nm are submitted to extensive refinement, spherical features of positive electron density appear in difference maps of electron density at positions that are occupied by hydrogens in the real molecule of protein.^{5,6,105,106} These features arise because the molecular model has no hydrogens but the molecule of protein does. Of particular interest are those features of difference electron density that can be assigned to the hydrogens in hydrogen bonds.^{125,126}

Another peculiar feature that becomes apparent as a refinement progresses is that a few of the amino acids display **alternative conformations** in the map of electron density. At the position of such an amino acid in an omit map of difference electron density, a feature having the shape of the superposition of two different rotational isomers of the amino acid is found (Figure 4-25).¹²⁷ This feature of electron density arises because in the crystal the actual amino acid spends part of its time in one conformation and part of its time in the other so the electron density in the map, which is averaged over the period in which the measurement was made, represents both conformations simultaneously. Maps of electron density calculated from data sets gathered to narrow Bragg spacing, because of their higher quality, reveal a greater frequency of alternative conformations and more subtle examples of alternative conformations, such as the *exo* and *endo* conformations of particular prolines¹²⁸ or the alternative conformations of a cystine,¹²⁹ and the features of electron density defining the alternative conformations are much sharper and less ambiguous.¹²⁶ In maps of electron density of low quality, however, most of the alternative conformations are never distinguished even upon refinement and their existence simply increases the *B* values for the functional groups that assume them.

Inherent in the process of refinement is the ability to produce a crystallographic molecular model by **molecular replacement**. Many proteins for which data sets have been gathered for the first time are closely related to other proteins for which a crystallographic molecular model is already available. Such a close relationship can be established by aligning the amino acid sequences of the two proteins. For example, the amino acid sequence of the aspartic endopeptidase from Rous sarcoma virus ($n_{aa} = 124$) can be aligned with the amino acid sequence of the aspartyl endopeptidase from

human immunodeficiency virus, type 1 ($n_{aa} = 99$), so that there are 30 identical positions and four gaps, all in the shorter protein.¹³⁰ It necessarily follows that the structures of these two proteins are superposable. The three long gaps in the shorter amino acid sequence (10, 5, and 6 amino acids, respectively) can be assumed to represent loops on the surface of the larger protein missing from the smaller. A crystallographic molecular model, produced by multiple isomorphous replacement, was available for the larger of the two proteins, that from Rous sarcoma virus.¹³¹ Crystals of the protein from the human immunodeficiency virus were produced, and a data set was collected from them. The side chains of the amino acids in the crystallographic molecular model of the protein from Rous sarcoma virus were replaced with the corresponding side chains in the aligned amino acid sequence of the protein from the human immunodeficiency virus. The loops corresponding to the gaps in the alignments were removed from the model to produce a preliminary molecular model for the protein from the human immunodeficiency virus.

This model was computationally aligned in the fundamental unit cell defined by the data set collected from crystals of the protein from human immunodeficiency virus. This preliminary model was then submitted to refinement to produce a final structure with an R -factor of 0.18.¹³² As this example illustrates, the purpose of using molecular replacement is to avoid the experimental difficulties of obtaining phases. Because there are so many proteins for which crystallographic molecular models are already available (<http://betastaging.rcsb.org/pdb/Welcome.do>), the likelihood that the protein in a new crystal is related closely enough to one for which a model has already been made is fairly high. Consequently, many of the newly reported maps of electron density have been calculated by molecular replacement.

How much more reliable is the refined crystallographic molecular model than the initial model built into the map of electron density produced by the phases determined experimentally? It is true that the R -factor is much smaller, but this is not surprising because the decrease occurred automatically. A significant decrease in the free R -factor is more reassuring. These decreases state that the refined molecular model, although it is usually not remarkably different from the original molecular model, has an electron density that produces structure factors the amplitudes of which are much closer to the observed amplitudes, which are the only directly observed quantities.

Often the success of the refinement is touted by showing that the map of electron density calculated from F_o and α_c of the final molecular model has features that very closely resemble phenyl rings or other equally characteristic side chains, but this is illusory. A map of electron density constructed from F_c and α_c would have to

have features precisely resembling these side chains because F_c and α_c are calculated from the molecular model itself, which always has ideal bond angles and bond lengths for the entire polypeptide, and the minimization has automatically caused F_c to be as close as

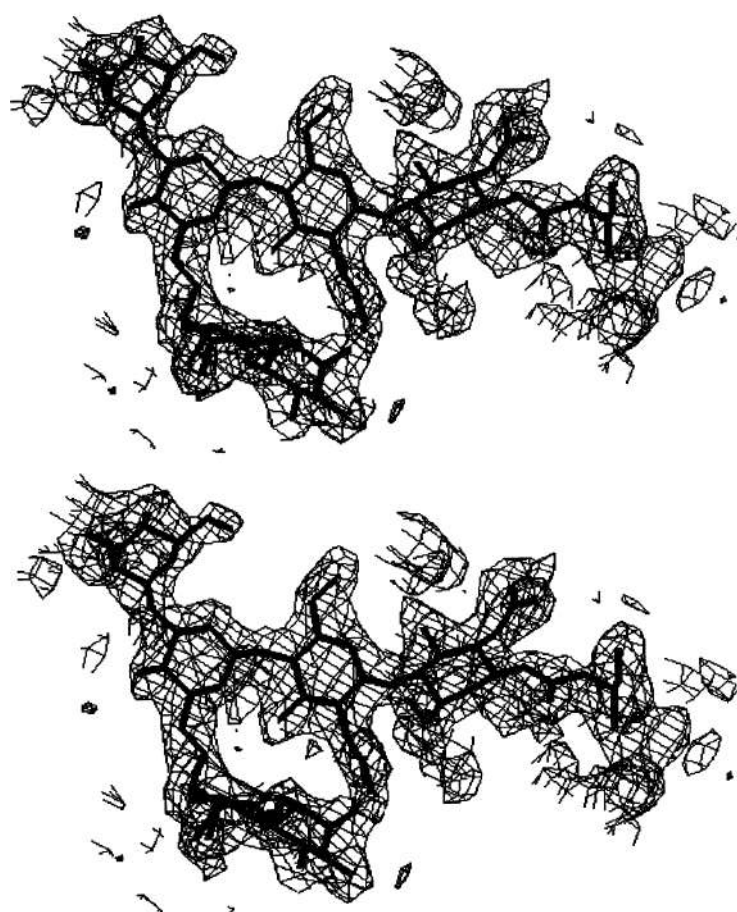
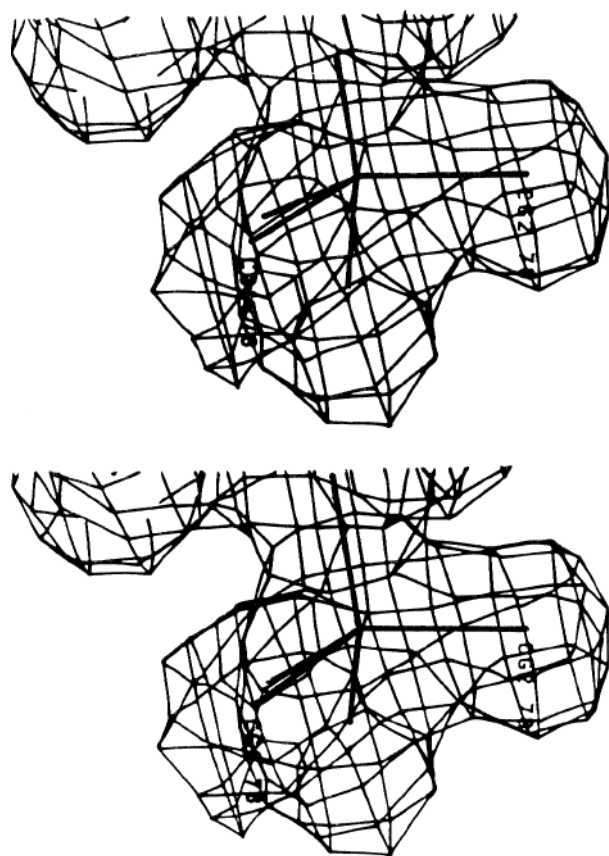


Figure 4-24: Fitting of an oligosaccharide into its assigned electron density.¹¹⁸ The oligosaccharide $\text{Man}(\alpha 1,3)\text{Man}(\alpha 1,6)[\text{Man}(\beta 1,4)\text{GlcNAc}(\beta 1,4)\text{GlcNAc}]$ was inserted into the final, refined map of electron density (Bragg spacing ≥ 0.22 nm) for the glucan 1,4- α -glucosidase from the fungus *Aspergillus awamori* to fill the electron density adjacent to Asparagine 395. The oligosaccharide was included in the refinement from the beginning. Asparagine 395 is at the bottom of the structure.

Figure 4-25: Alternate conformations of Valine 178 in ribonuclease T₁ from *Aspergillus oryzae*.¹²⁷ As the refinement of the map of electron density (Bragg spacing ≥ 0.15 nm) progressed, an unaccounted-for peak of electron density appeared that was continuous with the electron density for Valine 178 and was too close to it to have represented a molecule of water. All of the atoms of Valine 178 were then omitted from the molecular model, it was submitted to several cycles of refinement, and the omit map of electron density displayed in the figure was calculated. It contained electron density from two rotamers of Valine 178. Judging from the intensities of the two different features of electron density unique to each rotamer (those to the top and to the right), the two conformations are about equally populated. Skeletal models of the two alternative conformations of the valine are superposed within the envelope of difference electron density.



possible to F_0 . Furthermore, as has already been noted, the phases used in the calculation have more impact on the final map of electron density than the amplitudes. Therefore, it is not surprising that the details explicitly put into the model by the crystallographer should reappear when the map of electron density is reconstructed from F_0 and α_c or from any combinations of F_c , F_0 , and α_c .

A similar difficulty in evaluating the success with which a particular refinement has reproduced the real structure of the molecule of protein is the nature of the constraints applied. To progress efficiently, a refinement usually must be forced to retain bond lengths and bond angles or forced to be at the minimum of a function for potential energy; however, every constraint enforced upon the refinement is automatically incorporated into the final structure regardless of its reality. This fact is easily verified by examining crystallographic molecular models refined by different methods. The clear **imprint of the constraints** chosen for each method remains in each of the molecular models refined by that method.⁹² If one of the constraints in Equation 4-15 is that every peptide bond shall be planar, the planarity of the peptide bonds in the final structure cannot be cited as a measure of the success of the refinement. In fact, it is probably more an admission of its failure to detect the normal deviations from planarity.¹³³ If one of the constraints inadvertently introduced by Equation 4-16 is that the conformation along no carbon-carbon bond can eclipse vicinal methyls or methylenes, the absence of such eclipsed conformations cannot be cited as a measure of the success of the refinement.

These are not minor criticisms. The structure of a map of electron density usually improves so breathtakingly upon refinement that those changes that were enforced by the crystallographer must be clearly separated from those changes that arise only from the real molecular structure in the crystal. It is only these unconstrained and often unexpected features of the refined map of electron density which clearly state that it is an improvement.^{107,134}

The crystallographic molecular model, even after extensive refinement, should never be confused with the actual structure of the molecule itself. The crystallographic molecular model is no more than the coordinates of the centers of all of the atoms j in the model that was built by the computer from line segments, that was fit into the map of electron density, and that was then refined against the available amplitudes. It is only as accurate as the phases that were finally decided upon, which still incorporate some of the errors in the experimental phases, the arbitrary constraints imposed during refinement, and the inherent biases in the formula for calculating its structure factors, which assumes spherical atoms j and harmonic rectilinear atomic motions. The fact that most crystallographic molecular models change, often dramatically, as further refinement is performed or as additional data are collected from the same crystals to narrower Bragg spacing,¹³⁵ even though the real structure of the protein has not changed, is the most obvious demonstration that the model is not the structure of the protein but is a work of art representing the structure of the protein. It is unfortunate that the words "structure of the protein" are so widely used when referring to the crystallographic molecular model of the pro-

tein. This habit only adds to the misleading impression of infallibility associated with crystallography.

Suggested Reading

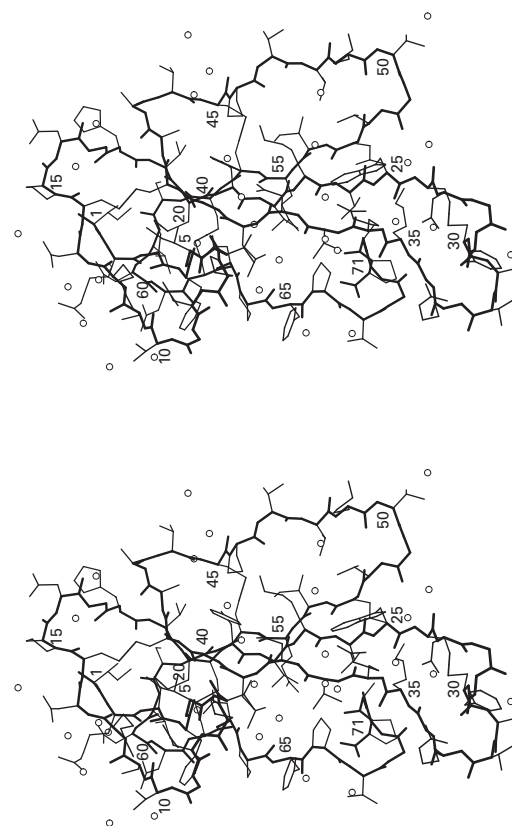
Oefner, C., & Suck, D. (1986) Crystallographic refinement and structure of DNase I at 2-Å resolution, *J. Mol. Biol.* 192, 605–632.

Problem 4–5: The figure to the right¹³⁶ is a stereo view of the crystallographic molecular model of a protein containing 71 amino acids. This drawing was produced with MolScript.¹³⁹ By examining the molecular model in stereo, you will be able to ascertain almost all of the amino acid sequence of the protein. You will not be able to distinguish threonine from valine, glutamate from glutamine, or asparagine from aspartate. Make an educated guess for threonine and valine, and just choose at random for asparagine and aspartate and glutamine and glutamate. If you can't make out an amino acid, put an X in its position in the sequence.

- (A) Write out the amino acid sequence of the protein in one-letter code. Number every tenth amino acid in your sequence to keep everything in register.
- (B) Which pairs of amino acids in the protein are cystines? Identify each pair by the sequence positions of the two cysteines that form the cystine.
- (C) What do the isolated atoms *j* scattered around in the crystallographic molecular model represent?
- (D) How did the crystallographer distinguish between threonine and valine, between glutamate and glutamine, and among aspartate, asparagine, and valine?

References

1. Stout, G.H., & Jensen, L.H. (1989) *X-ray Structure Determination, A Practical Guide*: 2nd ed, Wiley, New York.
2. Stout, G.H., & Jensen, L.H. (1968) *X-ray Structure Determination; a Practical Guide*, Macmillan, New York.
3. Andersson, I. (1996) *J. Mol. Biol.* 259, 160–174.
4. Wilson, K.S. (1998) *Nat. Struct. Biol.* 5 Suppl, 627–630.
5. Kuhn, P., Knapp, M., Soltis, S.M., Ganshaw, G., Thoene, M., & Bott, R. (1998) *Biochemistry* 37, 13446–13452.
6. Dauter, Z., Wilson, K.S., Sieker, L.C., Meyer, J., & Moulis, J.M. (1997) *Biochemistry* 36, 16065–16073.
7. Matthews, B.W. (1977) in *The Proteins*: 3rd ed. (Neurath, H., & Hill, R. L., Eds.) Vol. III, pp 404–590, Academic Press, New York.
8. Suck, D., Oefner, C., & Kabsch, W. (1984) *EMBO J.* 3, 2423–2430.
9. Fraser, R.D.B., & MacRae, T.P. (1969) in *Physical Principles and Techniques in Protein Chemistry Part A* (Leach, S. J., Ed.) pp 59–100, Academic Press, New York.



10. Bokhoven, C., Schoone, J.C., & Bijvoet, J.M. (1951) *Acta Crystallogr.* 4, 275–280.
11. Worthylake, D., Meadow, N.D., Roseman, S., Liao, D.I., Herzberg, O., & Remington, S.J. (1991) *Proc. Natl. Acad. Sci. U.S.A.* 88, 10382–10386.
12. Dickerson, R.E. (1964) in *The Proteins*: 2nd ed. (Neurath, H., Ed.) Vol. II, pp 603–778, Academic Press, New York.
13. Cullis, A.F., Muirhead, H., Perutz, M.F., Rossmann, M.G., & North, A.C.T. (1961) *Proc. R. Soc. London, A* 265, 15–38.
14. Weston, S.A., Camble, R., Colls, J., Rosenbrock, G., Taylor, I., Egerton, M., Tucker, A.D., Tunnicliffe, A., Mistry, A., Mancina, F., de la Fortelle, E., Irwin, J., Bricogne, G., & Pauptit, R.A. (1998) *Nat. Struct. Biol.* 5, 213–221.
15. Tilton, R.F., Jr., Kuntz, I.D., Jr., & Petsko, G.A. (1984) *Biochemistry* 23, 2849–2857.
16. Sugio, S., Petsko, G.A., Manning, J.M., Soda, K., & Ringe, D. (1995) *Biochemistry* 34, 9661–9669.
17. Weaver, L.H., Grutter, M.G., & Matthews, B.W. (1995) *J. Mol. Biol.* 245, 54–68.
18. Ollis, D.L., Brick, P., Hamlin, R., Xuong, N.G., & Steitz, T.A. (1985) *Nature* 313, 762–766.
19. Liljas, A., Kannan, K.K., Bergstaen, P.C., Waara, I., Fridborg, K., Strandberg, B., Carlsson, U., Jearup, L., Leovgren, S., & Petef, M. (1972) *Nat. New Biol.* 235, 131–137.
20. Blake, C.C., & Evans, P.R. (1974) *J. Mol. Biol.* 84, 585–601.

21. Kissinger, C.R., Liu, B.S., Martin-Blanco, E., Kornberg, T.B., & Pabo, C.O. (1990) *Cell* 63, 579–590.
22. Spurlino, J.C., Lu, G.Y., & Quijcho, F.A. (1991) *J. Biol. Chem.* 266, 5202–5219.
23. Schneider, F., Lowe, J., Huber, R., Schindelin, H., Kisker, C., & Knablein, J. (1996) *J. Mol. Biol.* 263, 53–69.
24. Leslie, A.G. (1990) *J. Mol. Biol.* 213, 167–186.
25. Banyard, S.H., Stammers, D.K., & Harrison, P.M. (1978) *Nature* 271, 282–284.
26. Breandaen, C.I., Eklund, H., Nordstroom, B., Boiwe, T., Seoderlund, G., Zeppezauer, E., Ohlsson, I., & Akeson, A. (1973) *Proc. Natl. Acad. Sci. U.S.A.* 70, 2439–2442.
27. Blow, D.M. (1958) *Proc. R. Soc. London, A* 247, 302–336.
28. Bijvoet, J.M. (1954) *Nature* 173, 888–891.
29. Guss, J.M., Merritt, E.A., Phizackerley, R.P., Hedman, B., Murata, M., Hodgson, K.O., & Freeman, H.C. (1988) *Science* 241, 806–811.
30. Yang, W., Hendrickson, W.A., Crouch, R.J., & Satow, Y. (1990) *Science* 249, 1398–1405.
31. Weis, W.I., Kahn, R., Fourme, R., Drickamer, K., & Hendrickson, W.A. (1991) *Science* 254, 1608–1615.
32. Kahn, R., Fourme, R., Bosshard, R., Chiadmi, M., Risler, J.L., Dideberg, O., & Wery, J.P. (1985) *FEBS Lett.* 179, 133–137.
33. Cramer, P., Bushnell, D.A., Fu, J., Gnatt, A.L., Maier-Davis, B., Thompson, N.E., Burgess, R.R., Edwards, A.M., David, P.R., & Kornberg, R.D. (2000) *Science* 288, 640–649.
34. Shapiro, L., Fannon, A.M., Kwong, P.D., Thompson, A., Lehmann, M.S., Grubel, G., Legrand, J.F., Als-Nielsen, J., Colman, D.R., & Hendrickson, W.A. (1995) *Nature* 374, 327–337.
35. Ryu, S.E., Kwong, P.D., Truneh, A., Porter, T.G., Arthos, J., Rosenberg, M., Dai, X.P., Xuong, N.H., Axel, R., Sweet, R.W., et al. (1990) *Nature* 348, 419–426.
36. Geiger, J.H., Hahn, S., Lee, S., & Sigler, P.B. (1996) *Science* 272, 830–836.
37. Xu, R.X., Hassell, A.M., Vanderwall, D., Lambert, M.H., Holmes, W.D., Luther, M.A., Rocque, W.J., Milburn, M.V., Zhao, Y., Ke, H., & Nolte, R.T. (2000) *Science* 288, 1822–1825.
38. Wang, B.C. (1985) *Methods Enzymol.* 115, 90–112.
39. Rypniewski, W.R., Breiter, D.R., Benning, M.M., Wesenberg, G., Oh, B.H., Markley, J.L., Rayment, I., & Holden, H.M. (1991) *Biochemistry* 30, 4126–4131.
40. Brändén, C., & Jones, T.A. (1990) *Nature* 343, 687–689.
41. Remington, S., Wiegand, G., & Huber, R. (1982) *J. Mol. Biol.* 158, 111–152.
42. Wyckoff, H.W., Tsernoglou, D., Hanson, A.W., Knox, J.R., Lee, B., & Richards, F.M. (1970) *J. Biol. Chem.* 245, 305–328.
43. Perrakis, A., Morris, R., & Lamzin, V.S. (1999) *Nat. Struct. Biol.* 6, 458–463.
44. Kleywegt, G.J., & Jones, T.A. (2002) *Structure* 10, 465–472.
45. Lawrence, C.M., Rodwell, V.W., & Stauffacher, C.V. (1995) *Science* 268, 1758–1762.
46. Story, R.M., Weber, I.T., & Steitz, T.A. (1992) *Nature* 355, 318–325.
47. Gruez, A., Pignol, D., Zeghouf, M., Coves, J., Fontecave, M., Ferrer, J.L., & Fontecilla-Camps, J.C. (2000) *J. Mol. Biol.* 299, 199–212.
48. de Vos, A.M., Tong, L., Milburn, M.V., Matias, P.M., Jancarik, J., Noguchi, S., Nishimura, S., Miura, K., Ohtsuka, E., & Kim, S.H. (1988) *Science* 239, 888–893.
49. Ghosh, D., O'Donnell, S., Furey, W., Jr., Robbins, A.H., & Stout, C.D. (1982) *J. Mol. Biol.* 158, 73–109.
50. Eklund, H., Nordstroom, B., Zeppezauer, E., Seoderlund, G., Ohlsson, I., Boiwe, T., Seoderberg, B.O., Tapia, O., Breandaen, C.I., & Akeson, A. (1976) *J. Mol. Biol.* 102, 27–59.
51. Pauling, L., Corey, R.B., & Branson, H.R. (1951) *Proc. Natl. Acad. Sci. U.S.A.* 37, 205–211.
52. Pauling, L., & Corey, R.B. (1951) *Proc. Natl. Acad. Sci. U.S.A.* 37, 729–740.
53. Richardson, J.S. (1981) *Adv. Protein Chem.* 34, 167–339.
54. Perutz, M.F. (1951) *Nature* 167, 1053–1054.
55. Venkatachalam, C.M. (1968) *Biopolymers* 6, 1425–1436.
56. James, M.N., & Sielecki, A.R. (1983) *J. Mol. Biol.* 163, 299–361.
57. Harte, R.A., & Rupley, J.A. (1968) *J. Biol. Chem.* 243, 1663–1669.
58. Sakon, J., Liao, H.H., Kanikula, A.M., Benning, M.M., Rayment, I., & Holden, H.M. (1993) *Biochemistry* 32, 11977–11984.
59. Teng, T.Y., Srajer, V., & Moffat, K. (1994) *Nat. Struct. Biol.* 1, 701–705.
60. Silva, M.M., Poland, B.W., Hoffman, C.R., Fromm, H.J., & Honzatko, R.B. (1995) *J. Mol. Biol.* 254, 431–446.
61. Crosio, M.P., Janin, J., & Jullien, M. (1992) *J. Mol. Biol.* 228, 243–251.
62. Sobek, H., Hecht, H.J., Aehle, W., & Schomburg, D. (1992) *J. Mol. Biol.* 228, 108–117.
63. Zhang, X.J., Wozniak, J.A., & Matthews, B.W. (1995) *J. Mol. Biol.* 250, 527–552.
64. Zaks, A., & Klivanov, A.M. (1988) *J. Biol. Chem.* 263, 3194–3201.
65. Yu, N., & Jo, B.H. (1973) *J. Am. Chem. Soc.* 95, 5033–5037.
66. Redfield, C., & Dobson, C.M. (1990) *Biochemistry* 29, 7201–7214.
67. Bycroft, M., Sheppard, R.N., Lau, F.T., & Fersht, A.R. (1990) *Biochemistry* 29, 7425–7432.
68. Shaanan, B. (1983) *J. Mol. Biol.* 171, 31–59.
69. Haurowitz, F. (1938) *Hoppe-Seyler's Z. Physiol. Chem.* 254, 266–274.
70. Freer, S.T., Kraut, J., Robertus, J.D., Wright, H.T., & Nguyen Huu, X. (1970) *Biochemistry* 9, 1997–2009.
71. Luzzati, P.V. (1952) *Acta Crystallogr.* 5, 802–810.
72. Stout, C.D. (1989) *J. Mol. Biol.* 205, 545–555.
73. Pai, E.F., Kabsch, W., Krengel, U., Holmes, K.C., John, J., & Wittinghofer, A. (1989) *Nature* 341, 209–214.
74. Waser, J. (1963) *Acta Crystallogr. A* 16, 1091–1094.
75. Hetenes, M.R., & Stiefel, E. (1952) *J. Natl. Bur. Stand.* 49, 409–436.
76. Konnert, J.H. (1976) *Acta Crystallogr.* A32, 614–617.
77. Takahashi, L.H., Radhakrishnan, R., Rosenfield, R.E., Jr., Meyer, E.F., Jr., & Trainor, D.A. (1989) *J. Am. Chem. Soc.* 111, 3368–3374.
78. Jacobson, B.L., Chae, Y.K., Markley, J.L., Rayment, I., & Holden, H.M. (1993) *Biochemistry* 32, 6788–6793.

79. Yang, Y.-S., Baldwin, J., Ley, B.A., Bollinger, J.M., Jr., & Solomon, E.I. (2000) *J. Am. Chem. Soc.* 122, 8495–8510.
80. Jack, A., & Levitt, M. (1978) *Acta Crystallogr.* A34, 931–935.
81. Konnert, J.H., & Hendrickson, W.A. (1980) *Acta Crystallogr.* A36, 344–350.
82. Brunger, A.T. (1992) *Nature* 355, 472–475.
83. Oefner, C., & Suck, D. (1986) *J. Mol. Biol.* 192, 605–632.
84. Bruenger, A., Karplus, M., & Petsko, G.A. (1989) *Acta Crystallogr.* A45, 50–61.
85. Bruenger, A.T., Kuriyan, J., & Karplus, M. (1987) *Science* 235, 458–460.
86. Johnson, L.N., Acharya, K.R., Jordan, M.D., & McLaughlin, P.J. (1990) *J. Mol. Biol.* 211, 645–661.
87. Kirkpatrick, S., Gelatt, C.D., & Vecchi, M.P. (1983) *Science* 220, 671–680.
88. Brunger, A.T. (1988) *J. Mol. Biol.* 203, 803–816.
89. Weiss, W.I., Brunger, A.T., Skehel, J.J., & Wiley, D.C. (1990) *J. Mol. Biol.* 212, 737–761.
90. Kyte, J. (1995) *Structure in Protein Chemistry*, 1st ed., p 234, Garland Publishing, New York.
91. Stickle, D.F., Presta, L.G., Dill, K.A., & Rose, G.D. (1992) *J. Mol. Biol.* 226, 1143–1159.
92. Laskowski, R.A., Moss, D.S., & Thornton, J.M. (1993) *J. Mol. Biol.* 231, 1049–1067.
93. Ogata, C.M., Gordon, P.F., de Vos, A.M., & Kim, S.H. (1992) *J. Mol. Biol.* 228, 893–908.
94. Blanchard, H., & James, M.N. (1994) *J. Mol. Biol.* 241, 574–587.
95. Ji, X., Zhang, P., Armstrong, R.N., & Gilliland, G.L. (1992) *Biochemistry* 31, 10169–10184.
96. Gros, P., Betzel, C., Dauter, Z., Wilson, K.S., & Hol, W.G. (1989) *J. Mol. Biol.* 210, 347–367.
97. Wilmanns, M., Priestle, J.P., Niermann, T., & Jansonius, J.N. (1992) *J. Mol. Biol.* 223, 477–507.
98. Breiter, D.R., Meyer, T.E., Rayment, I., & Holden, H.M. (1991) *J. Biol. Chem.* 266, 18660–18667.
99. Li, H.M., Wang, D.C., Zeng, Z.H., Jin, L., & Hu, R.Q. (1996) *J. Mol. Biol.* 261, 415–431.
100. Kumar, P.R., Eswaramoorthy, S., Vithayathil, P.J., & Viswamitra, M.A. (2000) *J. Mol. Biol.* 295, 581–593.
101. Ursby, T., Adinolfi, B.S., Al-Karadaghi, S., De Vendittis, E., & Bocchini, V. (1999) *J. Mol. Biol.* 286, 189–205.
102. Albert, A., Dhanaraj, V., Genschel, U., Khan, G., Ramjee, M.K., Pulido, R., Sibanda, B.L., von Delft, F., Witty, M., Blundell, T.L., Smith, A.G., & Abell, C. (1998) *Nat. Struct. Biol.* 5, 289–293.
103. Jelsch, C., Teeter, M.M., Lamzin, V., Pichon-Pesme, V., Blessing, R.H., & Lecomte, C. (2000) *Proc. Natl. Acad. Sci. U.S.A.* 97, 3171–3176.
104. Teeter, M.M., Roe, S.M., & Heo, N.H. (1993) *J. Mol. Biol.* 230, 292–311.
105. Anderson, D.H., Weiss, M.S., & Eisenberg, D. (1997) *J. Mol. Biol.* 273, 479–500.
106. Housset, D., Habersetzer-Rochat, C., Astier, J.P., & Fontecilla-Camps, J.C. (1994) *J. Mol. Biol.* 238, 88–103.
107. Schreuder, H.A., Prick, P.A., Wierenga, R.K., Vriend, G., Wilson, K.S., Hol, W.G., & Drenth, J. (1989) *J. Mol. Biol.* 208, 679–696.
108. Lauble, H., Kennedy, M.C., Beinert, H., & Stout, C.D. (1994) *J. Mol. Biol.* 237, 437–451.
109. Baca, M., Borgstahl, G.E., Boissinot, M., Burke, P.M., Williams, D.R., Slater, K.A., & Getzoff, E.D. (1994) *Biochemistry* 33, 14369–14377.
110. Mosimann, S.C., Newton, D.L., Youle, R.J., & James, M.N. (1996) *J. Mol. Biol.* 260, 540–552.
111. Bewley, M.C., Marohnic, C.C., & Barber, M.J. (2001) *Biochemistry* 40, 13574–13582.
112. Deisenhofer, J., Epp, O., Miki, K., Huber, R., & Michel, H. (1984) *J. Mol. Biol.* 180, 385–398.
113. Bruns, C.M., & Karplus, P.A. (1995) *J. Mol. Biol.* 247, 125–145.
114. Ficner, R., Lobeck, K., Schmidt, G., & Huber, R. (1992) *J. Mol. Biol.* 228, 935–950.
115. Bolin, J.T., Filman, D.J., Matthews, D.A., Hamlin, R.C., & Kraut, J. (1982) *J. Biol. Chem.* 257, 13650–13662.
116. Thompson, T.B., Garrett, J.B., Taylor, E.A., Meganathan, R., Gerlt, J.A., & Rayment, I. (2000) *Biochemistry* 39, 10662–10676.
117. Ida, K., Norioka, S., Yamamoto, M., Kumasaka, T., Yamashita, E., Newbigin, E., Clarke, A.E., Sakiyama, F., & Sato, M. (2001) *J. Mol. Biol.* 314, 103–112.
118. Aleshin, A., Golubev, A., Firsov, L.M., & Honzatko, R.B. (1992) *J. Biol. Chem.* 267, 19291–19298.
119. Shaanan, B., Lis, H., & Sharon, N. (1991) *Science* 254, 862–866.
120. Gomis-Ruth, F.X., Gohlke, U., Betz, M., Knauper, V., Murphy, G., Lopez-Otin, C., & Bode, W. (1996) *J. Mol. Biol.* 264, 556–566.
121. Burkhard, P., Tai, C.H., Jansonius, J.N., & Cook, P.F. (2000) *J. Mol. Biol.* 303, 279–286.
122. Brautigam, C.A., Sun, S., Piccirilli, J.A., & Steitz, T.A. (1999) *Biochemistry* 38, 696–704.
123. Murphy, M.E., Turley, S., Kukimoto, M., Nishiyama, M., Horinouchi, S., Sasaki, H., Tanokura, M., & Adman, E.T. (1995) *Biochemistry* 34, 12107–12117.
124. Cooper, S.J., Garner, C.D., Hagen, W.R., Lindley, P.F., & Bailey, S. (2000) *Biochemistry* 39, 15044–15054.
125. Sanders, D.A., Moothoo, D.N., Raftery, J., Howard, A.J., Helliwell, J.R., & Naismith, J.H. (2001) *J. Mol. Biol.* 310, 875–884.
126. Esposito, L., Vitagliano, L., Sica, F., Sorrentino, G., Zagari, A., & Mazzarella, L. (2000) *J. Mol. Biol.* 297, 713–732.
127. Martinez-Oyanedel, J., Choe, H.W., Heinemann, U., & Saenger, W. (1991) *J. Mol. Biol.* 222, 335–352.
128. Wilson, M.A., & Brunger, A.T. (2000) *J. Mol. Biol.* 301, 1237–1256.
129. Czapinska, H., Otlewski, J., Krzywda, S., Sheldrick, G.M., & Jaskolski, M. (2000) *J. Mol. Biol.* 295, 1237–1249.
130. Weber, I.T., Miller, M., Jaskaolski, M., Leis, J., Skalka, A.M., & Wlodawer, A. (1989) *Science* 243, 928–931.
131. Miller, M., Jaskaolski, M., Rao, J.K., Leis, J., & Wlodawer, A. (1989) *Nature* 337, 576–579.
132. Wlodawer, A., Miller, M., Jaskaolski, M., Sathyanarayana, B.K., Baldwin, E., Weber, I.T., Selk, L.M., Clawson, L., Schneider, J., & Kent, S.B. (1989) *Science* 245, 616–621.
133. MacArthur, M.W., & Thornton, J.M. (1996) *J. Mol. Biol.* 264, 1180–1195.
134. Artymiuk, P.J., & Blake, C.C. (1981) *J. Mol. Biol.* 152, 737–762.

135. Czjzek, M., Payan, F., Guerlesquin, F., Bruschi, M., & Haser, R. (1994) *J. Mol. Biol.* 243, 653–667.
136. Betzel, C., Lange, G., Pal, G.P., Wilson, K.S., Maelicke, A., & Saenger, W. (1991) *J. Biol. Chem.* 266, 21530–21536.
137. Ito, N., Phillips, S.E., Yadav, K.D., & Knowles, P.F. (1994) *J. Mol. Biol.* 238, 794–814.
138. Freymann, D., Down, J., Carrington, M., Roditi, I., Turner, M., & Wiley, D. (1990) *J. Mol. Biol.* 216, 141–160.
139. Kraulis, P.J. (1991) *J. Appl. Crystallogr.* 24, 946–950.

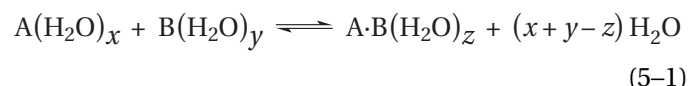
Chapter 5

Noncovalent Forces

Crystallographic studies have demonstrated that a molecule of protein, dissolved in aqueous solution, is composed of polypeptides, each of which is folded into a structure that is the same as or closely similar to the structure of all of the other polypeptides of the same amino acid sequence. A polypeptide, as it emerges from the ribosome, however, is a fluid polymer of undefined structure. Each newly synthesized polypeptide then folds spontaneously to assume its unique secondary and tertiary structure.

The folding of polypeptides to form the native structure of a protein, the association of folded polypeptides to form multimeric proteins, and the binding of substrates, coenzymes, or other molecules to proteins usually proceed without the formation of covalent bonds and are consequently controlled by noncovalent forces. It appears that four noncovalent forces are involved in these chemical reactions: ionic interactions, hydrogen bonds, the hydrophobic effect, and van der Waals forces. In the refined crystallographic molecular model of a folded protein, the consequences of these noncovalent forces are evident. The chemical and physical properties of these interactions, as they occur in aqueous solution, must be understood before those consequences can be appreciated. Therefore, a discussion of these interactions must precede a detailed description of the atomic details of refined crystallographic molecular models. None of the four categories of noncovalent forces—ionic interactions, hydrogen bonds, the hydrophobic effect, and van der Waals forces—can be completely separated from all of the others. Van der Waals forces must play a part in each of the other three phenomena, hydrogen bonds can be considered to be special cases of ionic interactions, almost all ionic interactions in biochemical situations involve hydrogen bonds, and the hydrophobic effect is to a large degree the reflection of hydrogen bonding in the solvent. It is informative, however, to discuss each of these categories separately to focus on their unique properties.

Each of these types of interactions can be considered to be a special case of a noncovalent association between two molecules, A and B, or between two segments, A and B, of the same molecule. For the situations under discussion, our attention will be directed to such a reaction as it would occur in aqueous solution. A general chemical equation for these associations is



The species $A(H_2O)_x$ and $B(H_2O)_y$ are the separated solutes dissolved in water and surrounded on all sides by water. Presumably, there are a certain number of water molecules, x and y , respectively, that are significantly affected by the presence of A or B. The effects of the solute on the surrounding molecules of water and the effects of the surrounding molecules of water on the solute are referred to as **solvation** or **hydration**. Around a particular molecule of solute at a particular instant, a particular number of water molecules are affected significantly by the presence of the solute. This number fluctuates with time, and the coefficients x , y , and z are intended to represent averages over a range of possible configurations for the hydration. When A and B associate to form the noncovalent complex, that complex will also be surrounded by water, and there will be a number of water molecules, z , that are significantly influenced by the complex. As A·B always has a smaller surface area than the sum of the surface areas of A and B, z should be less than $x + y$, and $(x + y - z)$ molecules of water will return to the bulk phase of the water when the complex is formed.

The change in standard free energy for the overall reaction can be expressed as

$$\Delta G^\circ = \Delta G^\circ_{A \cdot B} + \Delta G^\circ_{\text{hyd}(A \cdot B)} + \Delta G^\circ_{r H_2O} - \Delta G^\circ_{\text{hyd}(A)} - \Delta G^\circ_{\text{hyd}(B)} \quad (5-2)$$

where $\Delta G^\circ_{\text{hyd}(i)}$ refers to the standard free energy of hydration between each of the solutes and its surrounding waters of hydration, $\Delta G^\circ_{A \cdot B}$ refers to the direct standard free energy of interaction between A and B, and $\Delta G^\circ_{r H_2O}$ is the change in standard free energy experienced by the $x + y - z$ molecules of water as they leave shells of hydration and return to the bulk aqueous phase. It will become apparent that all of the terms on the right-hand side of Equation 5-2 except sometimes the first are remarkably influenced by the fact that this reaction occurs in water as a solvent. No noncovalent interaction has the same outcome in any other solvent. For example, hydrogen bonds and ionic interactions are stable in

almost any other solvent but are dissociated by water, while the hydrophobic effect is observed only when the solvent is water. To appreciate fully these influences of water on the outcome of noncovalent associations, the properties of liquid water itself must be understood.

Water

The properties of liquid water, when considered in their entirety, are unlike those of any other liquid. For example, the **surface tension** of water at 20 °C is 73 dyne cm⁻¹, while those of most other liquids are between 20 and 40 dyne cm⁻¹. The **relative permittivity**,* ϵ_r , of water at 20 °C is 80.2, while the relative permittivities of other liquids, with few exceptions, are less than 30. The high **melting point** and **boiling point** of water, for a molecule of its size and composition, are well-publicized anomalies. Not only are the numerical values of the physical constants anomalous, but the qualitative behaviors of the thermodynamic properties of the liquid, when it is exposed to variations of physical forces such as pressure, temperature, electric field, and electromagnetic energy, are unique. The details of these peculiarities provide an intuitive picture of the structure of liquid water that can serve as a basis for understanding the behavior of solutes such as polypeptides in this solvent. Unfortunately, there is no adequate molecular model for the structure of liquid water, and an informed intuitive picture is the closest approach to reality currently available.

A water molecule in the dilute, ideal **vapor** is an oxygen atom bonded covalently to two hydrogen atoms. Quantum mechanical calculations^{1,2} of the isolated molecule in the vacuum seem³ to support the conventional **orbital picture** of an oxygen hybridized sp^3 with two covalent bonds to two hydrogens and two σ lone pairs of electrons; these four substituents are oriented tetrahedrally around the oxygen. The HOH bond angle⁴ is 104.5°, distorted from 109.5° by the electron repulsion of the lone pairs or by a rehybridization, driven by energy of promotion, that gives the oxygen–hydrogen σ bonds more p character. The oxygen–hydrogen bond lengths are 0.096 nm.

In more concentrated vapor, **dimers of water** form (Figure 5-1).^{5,6} From results of molecular beam microwave spectroscopy, the mean structure of the dimer can be calculated.^{5,6} The two oxygens are separated by a distance of 0.298 nm. One of the four hydrogens lies on the line of centers between the two oxygens, and it is covalently bonded to one of them, which is referred to as the proton donor. The other oxygen, which is referred to as the proton acceptor, has two of the four hydrogens covalently bonded to it. The plane defined by

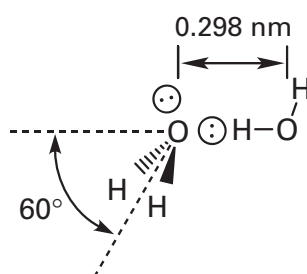


Figure 5-1: Selected dimensions of the dimer of two molecules of water in the gas phase. The distances and angles were obtained by microwave spectroscopy.^{5,6}

the two hydrogens and the oxygen of the proton acceptor is inclined at an angle of 60° to the line of centers between the two oxygen atoms. This means that the four substituents around the acceptor—the two hydrogens, the shared hydrogen, and the lone pair of electrons—are tetrahedrally arrayed in the dimer. This arrangement suggests that the oxygen–hydrogen bond on the donor points directly at one of the two σ lone pairs of electrons on the acceptor oxygen. The axis of the sp^3 orbital in which that σ lone pair resides should be congruent with the line of centers. The interaction between the hydrogen–oxygen σ bond on the donor molecule of water and the σ lone pair of electrons on the acceptor is an unhindered, intermolecular example of a **hydrogen bond**. The formation of dimers and higher oligomers in steam contributes significantly to its nonideal behavior at higher concentrations of water in the gas phase.

The **ice** that is in equilibrium with liquid water at atmospheric pressure and 0 °C is known as ice Ih. Ice Ih is a **tetrahedral diamond lattice** of oxygen atoms (Figure 5-2A),⁴ each 0.276 nm from its nearest neighbor.⁷ The oxygens are held in the lattice by hydrogen bonds to each of their four nearest neighbors. Between any oxygen atom and each of its four nearest neighbors in the lattice is one hydrogen atom. At any instant, each hydrogen is covalently bound to one of the two oxygens between which it is found, and every oxygen has only two hydrogens covalently bound to it. These two requirements create a situation in which only a predictable number of arrangements for these hydrogens can occur, and this number of arrangements can explain almost exactly the observed residual entropy of ice Ih at 0 K.⁴ There is a significant amount of **empty space** in ice Ih (Figure 5-2B), and this is one of the properties permitting it to be less dense than the liquid water with which it can be in equilibrium.

The structure and properties of ice Ih and water vapor have been exhaustively investigated and unambiguously established. At atmospheric pressure, **liquid water** lies between these two extremes on the phase diagram, and its properties can be compared with them. From the transitions between solid and liquid and between liquid and vapor, insight into the structure of the liquid can be gained.

When ice melts, the reaction involves a **standard enthalpy of fusion**, and when the liquid vaporizes, the reaction involves a **standard enthalpy of vaporization**. The enthalpy of water at atmospheric pressure can be

* The relative permittivity or **dielectric constant** of a substance is its permittivity relative to the permittivity of vacuum.

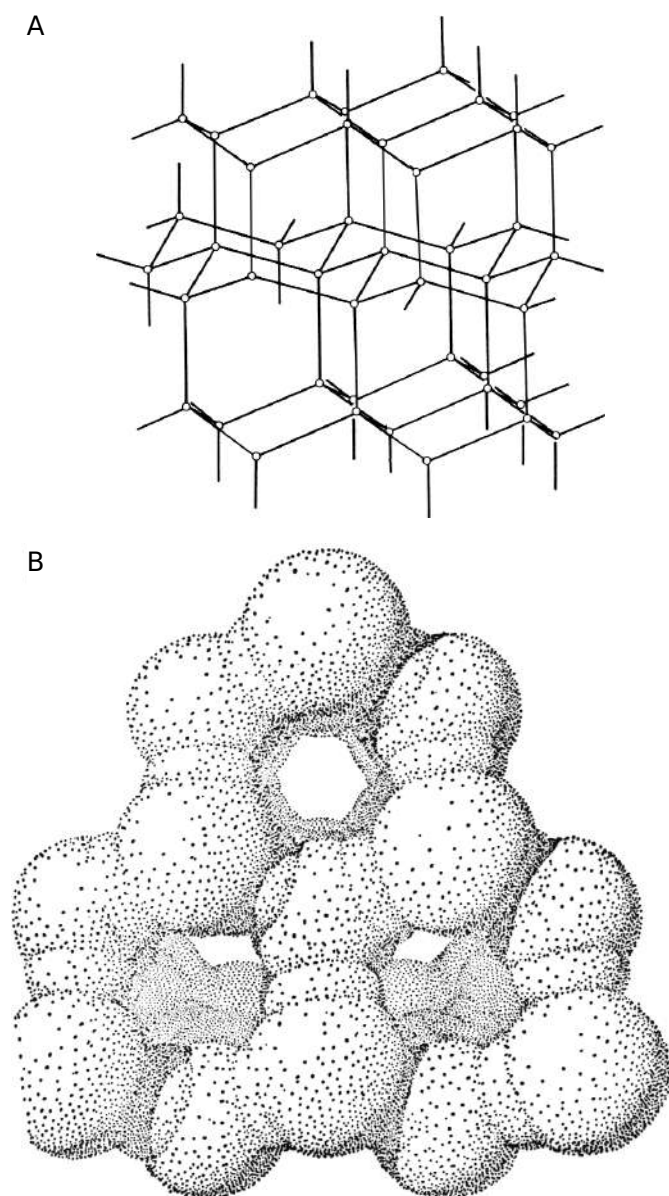


Figure 5-2: Structure of ice Ih. (A) Tetrahedral lattice.⁴ In a tetrahedral lattice, each atom or molecule occupies a position indicated by the small open circles. It forms equivalent connections with its four nearest neighbors, and because every connection is equivalent and every nearest neighbor is the same, the nearest neighbors are arrayed tetrahedrally and at equal distance. (B) Representation of space-filling models of molecules of water arrayed on the tetrahedral lattice of ice Ih. The spheres are atoms of oxygen, and the hydrogens, one for each connection of the lattice, are sandwiched between the oxygens. Reprinted with permission from ref 4. Copyright 1969 Clarendon Press.

plotted as a function of absolute temperature (Figure 5-3).⁴ On this plot, the discontinuities at the melting point and boiling point are the standard enthalpy of fusion and the standard enthalpy of vaporization, respectively. At 25 °C, the standard enthalpy of fusion is 8.0 kJ mol⁻¹, and the standard enthalpy of vaporization is 44 kJ mol⁻¹. This standard enthalpy of vaporization is more than twice that of a liquid such as chloromethane

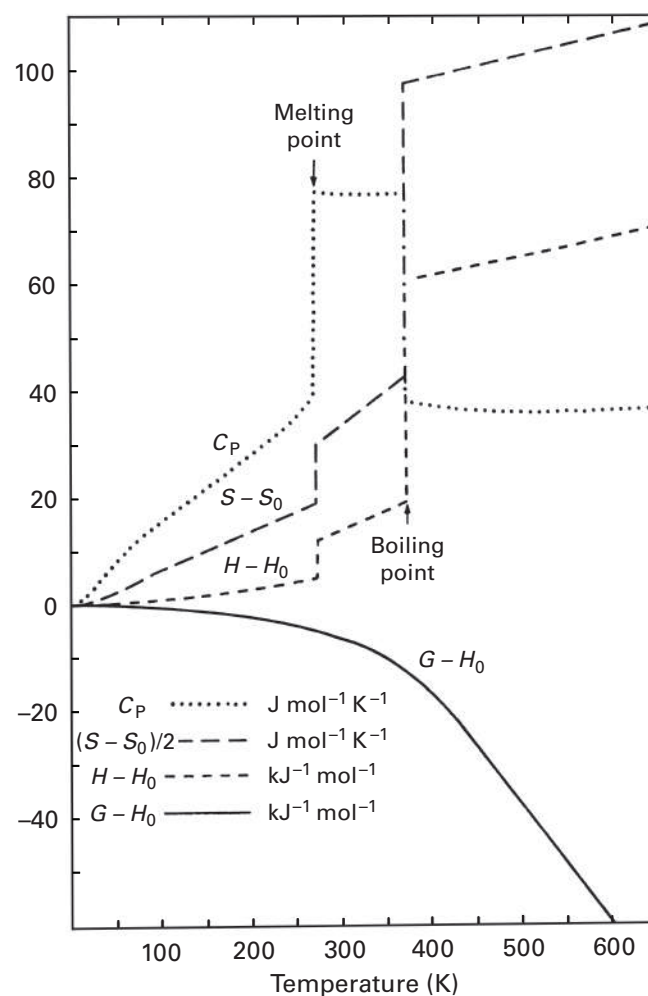


Figure 5-3: Enthalpy ($H-H_0$), entropy ($S-S_0$), free energy ($G-H_0$), and isopiestic heat capacity (C_p) as a function of temperature for water at unit atmosphere (101.3 kPa) pressure.⁴ The quantities H_0 and S_0 are the absolute enthalpy and absolute entropy, respectively, at 0 K. Enthalpy and free energy are in units of kilojoules mole⁻¹, and heat capacity is in units of joule mole⁻¹ kelvin⁻¹. The values of entropy, also in joule mole⁻¹ kelvin⁻¹, are arbitrarily divided by 2 to put them on the same scale. Adapted with permission from ref 4. Copyright 1969 Clarendon Press.

($\Delta H_{\text{vap}} = 19 \text{ kJ mol}^{-1}$ at 25 °C), which is a polar solvent ($\epsilon_r = 13$) containing molecules incapable of forming hydrogen bonds but much larger than water (50.5 g mol⁻¹). This comparison illustrates the fact that the standard enthalpy of vaporization of water is anomalously large. In the sum of the two reactions, fusion and vaporization, all of the hydrogen bonds in ice Ih are lost. The fact that most of the standard enthalpy change occurs upon vaporization and the fact that the heat of vaporization is anomalously large suggest that liquid water retains most of the hydrogen bonds present in ice Ih.

The high isochoric **heat capacity**, C_V , of liquid water (Figure 5-3) also indicates that it is highly structured. Calculations of the isochoric heat capacities of both ice Ih and water vapor, from the known vibrational,

translational, and rotational energy levels of these two substances, agree quite closely with observed values.⁴ The observed values of the isochoric heat capacity of liquid water, however, are almost twice that calculated from its estimated vibrational, translational, and rotational energy levels (Figure 5-4).⁴ This excess or **configurational heat capacity** can be explained by postulating that much of the hydrogen-bonded structure of ice remains in the liquid and its gradual deterioration as the temperature is raised is responsible for the anomalous absorption of heat. The high and relatively constant value for the heat capacity throughout the range of temperature between 0 and 100 °C suggests that the hydrogen-bonded network in the liquid is gradually and constantly deteriorating as the temperature is rising.

Another indication of the extensive hydrogen-bonded structure in liquid water is its high static relative permittivity ($\epsilon_r = 88$ at 0 °C), which is almost equivalent to that of ice Ih ($\epsilon_r = 99$ at 0 °C). The large value for the relative permittivity of ice Ih is usually explained semiquantitatively⁴ as a result of the high correlation among the orientations of the individual dipole moments of the water molecules caused by their rigid arrangement in the hydrogen-bonded lattice. When an electric field is applied, the dipole moments reorient cooperatively, producing the large relative permittivity. The fact that the

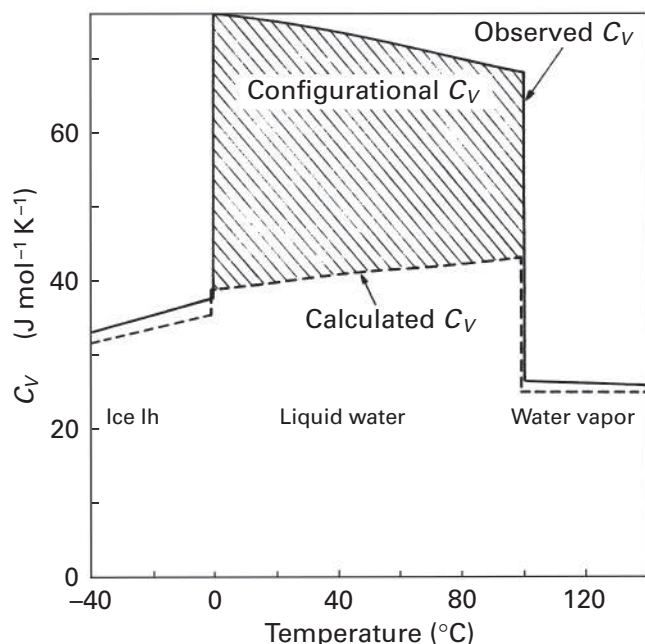


Figure 5-4: Separation of the observed isochoric heat capacity C_V of water (solid line) into calculated (dashed line) and configurational (shaded difference) components.⁴ The heat capacity of ice Ih was calculated from the two vibrational absorption bands of lowest energy ($\nu = 840 \text{ cm}^{-1}$ and $\nu = 230 \text{ cm}^{-1}$); the heat capacity of water vapor was calculated from the vibrational, rotational, and translational energies of the water molecules; and the heat capacity of the liquid was calculated on the assumption that each molecule in the liquid has three hindered degrees of translation and three hindered librations. Adapted with permission from ref 4. Copyright 1969 Clarendon Press.

liquid has almost the same relative permittivity as the solid indicates that much of the lattice remains.

The **molar volume** of ice ($19.6 \text{ cm}^3 \text{ mol}^{-1}$) is somewhat greater than that of liquid water ($18.0 \text{ cm}^3 \text{ mol}^{-1}$) at 0 °C and much greater than the molar volume that would be expected if spheres the radius of molecules of water (0.14 nm) were randomly packed in an unstructured, disordered array ($10 \text{ cm}^3 \text{ mol}^{-1}$).⁴ The large molar volume of ice Ih is due to the vacant space created by the fact that oxygens are held in a tetrahedral array by the hydrogen-bonded network (Figure 5-2B). When ice melts, the molecules of water are allowed to occupy some of the vacant space in the hydrogen-bonded lattice and the density increases. A related fact is that the molar volume of liquid water increases as the temperature is decreased below 4 °C, presumably because the expansion caused by the strengthening of the hydrogen-bonded lattice is greater than the usual contraction experienced by most liquids resulting from the decrease in thermal energy. It is only above 4 °C that the latter effect becomes dominant. The contraction of water upon melting and the expansion of the liquid upon cooling below 4 °C are almost unprecedented. Diamond, silicon, and germanium are tetrahedral solids that also float upon their melts, as ice floats upon water. Aside from these peculiar features, the molar volumes of ice and liquid water at 0 °C are both large and not that different from each other. Consequently, much of the vacant space created by the hydrogen-bonded lattice in the solid remains in the liquid.

The fact that much of the vacant space remains in liquid water also explains the unique decrease in **isothermal compressibility** that occurs in liquid water as temperature is raised from 0 to 50 °C (Figure 5-5).⁴ The isothermal compressibility of a liquid, κ_T , is the fractional decrease in volume produced by the application of pressure at constant temperature:

$$\kappa_T = -\frac{1}{V} \left(\frac{\partial V}{\partial P} \right)_T \quad (5-3)$$

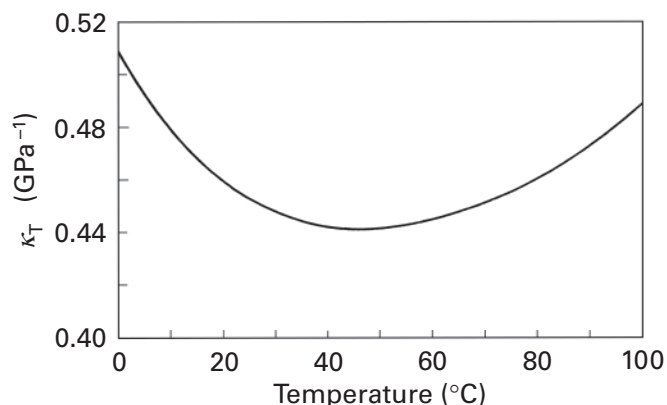


Figure 5-5: Isothermal compressibility κ_T (gigapascals⁻¹) for liquid water at unit atmosphere (101.3 kPa) pressure, presented as a function of temperature.⁴ Reprinted with permission from ref 4. Copyright 1969 Clarendon Press.

in units of reciprocal pressure (pascal^{-1}). In almost every other liquid, isothermal compressibility increases monotonically with temperature. In liquid water at low temperatures, most of the structured vacant space of ice Ih remains when the transition from solid to liquid occurs, and this structured vacant space is gradually replaced with randomly distributed, unstructured vacant space, similar to that in other liquids as the temperature is raised. The high compressibility at low temperatures results from the ability of the lattice to decrease its volume upon the application of pressure at the expense of the significant vacant space among the oxygen atoms.

The idea that liquid water at lower temperatures retains a structure similar to that of ice Ih is also supported by the small **cubic expansion coefficient** of liquid water. Upon heating at atmospheric pressure between temperatures of 20 and 30 °C, other liquids expand about 4 times more rapidly than does water (Figure 5–6).⁸ As pressure is applied, however, the cubic expansion coefficient for water increases while the coefficients of thermal expansion for other liquids decrease. At high pressures, both water and other liquids have about the same cubic expansion coefficient. If liquid water at atmospheric pressure is extensively hydrogen-bonded with an expanded structure similar to that of ice Ih (Figure 5–2B), then as the temperature is raised, the decrease in structured empty volume due to the deterioration of this hydrogen-bonded network could almost cancel the increase in unstructured volume due to increased thermal motion. As pressure is applied, however, it causes the hydrogen-bonded network to deteriorate or restructure and the liquid to have a more normal cubic expansion coefficient. In this view, the ability of pressure to

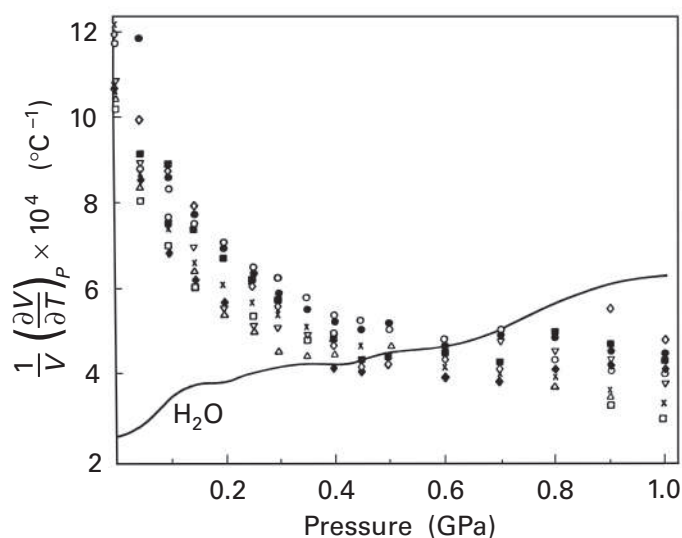


Figure 5–6: Cubic expansion coefficient for several liquids at 25 °C as a function of applied pressure.⁸ The liquids are (x) PCl_3 , (o) CH_3OH , (v) CS_2 , (•) $\text{C}_2\text{H}_5\text{Cl}$, (■) $\text{C}_2\text{H}_4\text{I}$, (Δ) $\text{C}_2\text{H}_5\text{OH}$, (▽) $\text{C}_3\text{H}_9\text{OH}$, (◆) isobutyl alcohol, and (□) $n\text{-C}_5\text{H}_{11}\text{OH}$. The solid curve is the cubic expansion coefficient for liquid water at 25 °C as a function of pressure. Reprinted with permission from ref 8. Copyright 1970 American Association for the Advancement of Science.

change the structure of the liquid is due to the fact that liquid water is in an extensively hydrogen-bonded form at normal pressures but not at higher pressures.

Such a transition between an ordered and a less ordered state caused by an increase in pressure would also explain why the application of pressure decreases the **viscosity** of liquid water rather than increasing it as it does the viscosities of other liquids.⁸ The viscosity of water is anomalously large in the first place ($\eta = 1.00 \text{ mPa s}$ at 20 °C) compared to the viscosity of liquids such as acetonitrile ($\eta = 0.36 \text{ mPa s}$ at 20 °C), pentane ($\eta = 0.24 \text{ mPa s}$ at 20 °C), and carbon disulfide ($\eta = 0.36 \text{ mPa s}$ at 20 °C).

Additional evidence for the retention of a significant fraction of the hydrogen-bonded lattice in liquid water is provided by **scattering of X-radiation**. When a beam of X-radiation is passed through a liquid, it is scattered by the electrons of the molecules in the liquid. The intensity of the scattered X-radiation varies as a function of the angle between the incident beam and the direction at which the scattered radiation emerges from the solution. This angular dependence of the intensity can be used to calculate a **radial molecular correlation function**, $G_M(r)$. This function is an approximation⁹ of the variation of electron density as a function of the radial distance from any one molecule in the liquid. The actual variation of electron density is distinguished from its approximation by designating it as $g(r)$. The function $G_M(r)$ registers any local variations in the electron density of the liquid, relative to the mean electron density of the liquid, that are maintained around any one of the molecules. Because it is a relative quantity, the value of $G_M(r)$ is unity when the electron density is equal to the mean electron density. Any variations in density that are observed are assumed to be permanent features of the structure of the liquid. As $G_M(r)$ or $g(r)$ is proportional to the electron density as a function of radial distance from a central molecule, they can be used to calculate the total number of molecules of solvent in a spherical shell between r_1 and r_2 by the integral⁹

$$n_s = \frac{p_0}{\gamma} \int_{r_1}^{r_2} 4\pi r^2 g(r) dr \quad (5-4)$$

where n_s is the number of molecules of solvent in that shell, p_0 is the bulk electron density of the liquid, and γ is the number of electrons in each molecule of solvent (10 in the case of water).

The radial molecular correlation function for liquid water has been determined over a range of temperatures from 4 to 200 °C (Figure 5–7).^{7,9,10} At fairly long distances from any one molecule of water ($>0.8 \text{ nm}$), the function becomes unity and does not vary noticeably. Therefore, beyond 0.8 nm from any given molecule the liquid is, on the average, homogeneous. A significant peak of density occurs, however, at 0.28 nm. Integration of this peak¹¹ for

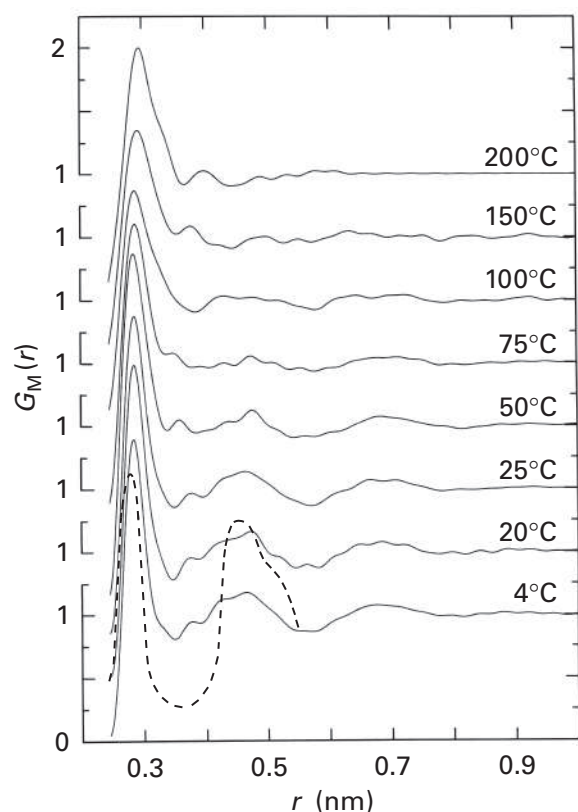


Figure 5-7: Molecular correlation functions for liquid water^{7,9} and ice Ih.¹⁰ The molecular correlation functions for liquid water at several temperatures (solid lines) were calculated from the angular dependence of the intensity of the scattered X-rays from samples of pure water through which a collimated beam of X-rays was passed. A molecular correlation function for liquid molecules arranged on the lattice of ice Ih (dashed line) was calculated from the length of the hydrogen bonds in ice Ih (0.276 nm) and the fact that the oxygen atoms lie upon a tetrahedral diamond lattice. The calculation was performed with the assumption that the distributions of electron density around the maxima defined by the lattice could be approximated by error functions. The width of the first error function was made the same as the width of the first maximum in liquid water at 4 °C, and the widths of the two subsequent error functions were made proportional to the square of their distances from the origin.⁷ Adapted with permission from ref 7, originally from ref 9. Copyright 1971 American Institute of Physics.

the curve at 4 °C, upon the assumption that it is a Gaussian function, indicates that it is produced by about four nearest neighbors. In ice there are four nearest neighbors to each water molecule and they are held at a distance of 0.276 nm. It can be assumed that these are retained in the liquid. That the peak is centered at a distance so close to the hydrogen-bonded distance in ice has been interpreted to mean that each water molecule in the liquid has about four **hydrogen-bonded nearest neighbors**.

A radial molecular correlation function can be calculated¹¹ for liquid molecules of water confined to the tetrahedral lattice of ice Ih (Figure 5-7). In ice Ih, there are four nearest neighbors at 0.276 nm, 12 next neighbors at 0.45 nm, and 12 farther neighbors at 0.52 nm

(Figure 5-2). A distribution of liquid molecules of water confined to the diamond lattice of ice Ih would produce a radial molecular correlation function with a distinct minimum between the first four neighbors and the next group of 24 (Figure 5-7).¹⁰

When the radial molecular correlation functions of liquid water at 4 °C and liquid molecules of water confined to the lattice of ice Ih are compared, several differences are noted. Although still a prominent feature, the maximum in ice Ih centered at around 0.5 nm is considerably broadened in the liquid. This indicates that the hydrogen-bonded network has become considerably more elastic in water than in ice, permitting the second and third groups of neighbors to approach the molecule at the origin much more closely, rather than being held at a distance by a rigid lattice. There also seems to be too much electron density in the actual liquid between the first maximum and the second maximum.¹⁰ This has been interpreted to mean that molecules are able to break out of the lattice and become **interstitial molecules of water**, transiently occupying the vacant spaces (Figure 5-2B).¹²

So far the discussion has emphasized similarities between ice Ih at 0 °C and liquid water at low temperatures. There are, of course, remarkable differences. The most obvious is the fact that ice is a solid and water is a liquid. Even though ice Ih is a solid, however, it, like liquid water, is able to flow. In order for condensed matter to flow, layers of molecules in that matter must be able to slide past layers of other molecules above and below them. In the case of water or ice Ih the manifestation of this ability requires extensive and simultaneous disruption of continuous layers of hydrogen bonds in the liquid or solid as it flows. This capacity to flow is far more evident in water than in ice. It is quantified by values for the viscosity of the liquid and the solid. Liquid water at 0 °C has a viscosity of 1.8 mPa s, and ice Ih at 0 °C has a viscosity of about 10^{16} mPa s.^{13,14} The difference between liquid water and ice Ih is so large because, to flow, hydrogen bonds must be broken simultaneously over significant regions. There are, however, measurements that quantify the behavior of individual molecules of water.

When individual molecules in water change their relative positions, hydrogen bonds must be broken and re-formed elsewhere. The capacity to change positions is reflected in the process of self-diffusion, a measure of the rate at which the average molecule of water diffuses through a condensed phase of water molecules. The **self-diffusion coefficient** for ice Ih is about 10^{-10} cm² s⁻¹ at 0 °C, and for liquid water it is 1.4×10^{-5} cm² s⁻¹ at 5 °C.⁴ This difference of 10^5 demonstrates that water molecules can exchange their hydrogen-bonded neighbors far more rapidly in the liquid than in the solid. To the extent that this exchange involves breaking and making of hydrogen bonds, the hydrogen bonds in the liquid are weaker than those in the solid.

An even more easily understood measurement of the rate at which a molecule of water can detach itself from the hydrogen-bonded network in the liquid in order to reorient itself is the **dielectric relaxation** of liquid water. The relative permittivity of a chemical substance is a function of the frequency of the alternating electric field used to measure it. Tabulated relative permittivities are usually static relative permittivities that are measured with an alternating electric field with a frequency of alternation so low that the measured values may be confidently extrapolated to zero frequency. The low frequency of alternation allows the molecules in the substance more than ample time to align themselves, as far as they are able, with the electric field while the measurement is made. If the frequency of the applied field, however, is gradually increased, at some point the molecules in the substance are unable to invert their alignments at rates sufficient to keep up with the alternations of the applied field. Their inability to keep up results from intermolecular forces that hinder their rotation. The dielectric relaxation time is the time that an applied field must be in operation before $\exp(-1)$ of the increase in relative permittivity due to the rotation of the molecules aligning themselves with the field has occurred. The dielectric relaxation time of ice Ih at 0 °C is 2×10^{-5} s, that of liquid water at 0 °C is 2×10^{-11} s, and that of a water molecule in a dilute solution of water in benzene is 1×10^{-12} s.⁴ A similar value for the rotational correlation time of a water molecule in ice Ih (1.5×10^{-5} s at -6 °C)¹⁵ has been measured by nuclear magnetic resonance. Although a water molecule in liquid water is constrained so that it rotates 20 times more slowly than it does in a condensed phase lacking hydrogen bonds, it rotates 10^6 times faster in liquid water than in ice. This again demonstrates that the hydrogen bonds between water molecules in liquid water are weaker than those in ice.

This weakening of the hydrogen bonds in the liquid is also reflected in a shift that occurs in the frequency of the maximum **infrared absorption** of the oxygen-hydrogen stretching vibration of water when it melts.⁴ The frequency at which a covalent bond absorbs infrared electromagnetic energy is correlated with its bond energy. The greater the bond energy, the higher the frequency of the light required to excite its vibration. In the case of the stretching frequency of the oxygen-hydrogen bond in water, the stronger the hydrogen bond in which it participates, the weaker will be the covalent oxygen-hydrogen bond itself and the lower the frequency of its absorption. The **stretching frequency** of the oxygen-hydrogen bond in ice Ih is 3220 cm^{-1} , in liquid water it is 3490 cm^{-1} , and in the dilute vapor it is 3700 cm^{-1} . In the dilute vapor, no hydrogen bond weakens the oxygen-hydrogen bond. In ice Ih, a strong, fixed hydrogen bond weakens the oxygen-hydrogen bond significantly. In liquid water, less than half the decrease in frequency between the vapor and ice Ih occurs, presumably because the hydrogen bonds formed when the vapor is

converted into the liquid are weaker than the hydrogen bonds formed when the vapor is converted into the solid.

The weakening of the hydrogen bonds upon melting that is indicated by both the increases in self-diffusion and dielectric relaxation and the increase in the stretching frequency of the oxygen-hydrogen bond requires that the dissociation constant for the hydrogen bond in liquid water be significantly larger than that for ice Ih. This increase in dissociation constant upon melting may be large enough to produce a significant population of unbonded molecules of water in the liquid, presumably the interstitial water the existence of which is implied in the radial molecular correlation function.

There are infrared spectra of liquid water which suggest that there are two distinct species of oxygen-hydrogen bonds in the liquid,¹⁶ and these two species could represent intact and broken hydrogen bonds.¹⁷ It is possible to fit the temperature dependence of both these infrared spectra and the heat capacity of the liquid with a model of the liquid that assumes that there are only two types of oxygen-hydrogen bonds present, those participating in intact hydrogen bonds and those the hydrogen bonds of which are broken.¹⁷ From such a fit, the standard free energy of formation of a hydrogen bond in liquid water is estimated to be -2.0 kJ mol^{-1} at 25 °C; and the fraction of broken hydrogen bonds, 0.30 at 25 °C. From this fraction it would follow that at a given instant about 3–4% of the molecules of water would be either attached to the lattice by only one hydrogen bond or completely free of the lattice. There are, however, results suggesting both that these infrared spectra do not result from only two populations of oxygen-hydrogen bonds⁸ and that no simple two-state model can explain both the cubic expansion coefficient and the temperature coefficient of the isothermal compression of liquid water simultaneously.⁷ Consequently, the question of the molar concentration of intact hydrogen bonds in liquid water remains open.

The mental picture of liquid water that forms intuitively as its peculiarities are described is presently more adequate than any sophisticated physical model of its structure. The impression that is formed from a consideration of these properties is that liquid water retains most of the hydrogen bonds that are present in ice Ih but that these hydrogen bonds are more elastic, weaker, and break and re-form much more rapidly than those in ice Ih.

Suggested Reading

Eisenberg, D., & Kauzmann, W. (1969) *The Structure and Properties of Water*, Clarendon Press, Oxford, England.

Problem 5-1: The isopiestic heat capacity of a substance, C_p , is defined as the amount of heat required to raise one mole of the substance one degree in temperature at constant pressure. The units of this quantity are joules degree⁻¹ mole⁻¹.

A substance has a certain intrinsic enthalpy at 0 K, and this intrinsic enthalpy increases as the temperature increases and the substance absorbs heat:

$$H_T - H_0 = \int_0^T C_p dT + \Delta H_{pc}$$

where H_T is the intrinsic enthalpy at $T = T$, H_0 is the intrinsic enthalpy at $T = 0$ K, and ΔH_{pc} is the sum of the enthalpy changes for all phase transitions between 0 K and T . The heat capacity of H_2O is the following function of temperature (Figure 5–3):

$$C_p = (0.172 \text{ J K}^{-2} \text{ mol}^{-1})T \quad [T = 0-60\text{K}]$$

$$C_p = 2.47 \text{ J K}^{-1} \text{ mol}^{-1} + (0.129 \text{ J K}^{-2} \text{ mol}^{-1})T \quad [T = 60-273\text{K}]$$

$$C_p = 77.5 \text{ J K}^{-1} \text{ mol}^{-1} \quad [T = 273-373\text{K}]$$

and

$$\Delta H_{\text{fus}} = 6.0 \text{ kJ mol}^{-1} \quad \text{at } 0^\circ \text{C}$$

$$\Delta H_{\text{vap}} = 40.7 \text{ kJ mol}^{-1} \quad \text{at } 100^\circ \text{C}$$

- (A) Use these experimental data to draw a graph of $H_T - H_0$ as a function of temperature from 0 to 375 K.

The intrinsic entropy of a substance is related to the heat capacity by the following equation:

$$S_T - S_0 = \int_0^T \frac{C_p}{T} dT + \Delta S_{pc}$$

where S_T is the intrinsic entropy at $T = T$, S_0 is the intrinsic entropy at $T = 0$, and ΔS_{pc} is the sum of the entropy changes for all phase transitions.

When phase transitions occur, $\Delta G^\circ = 0$ at the transition temperature. Use $\Delta H_{\text{fus}}^\circ$ and $\Delta H_{\text{vap}}^\circ$ to calculate $\Delta S_{\text{fus}}^\circ$ and $\Delta S_{\text{vap}}^\circ$.

- (B) Draw a graph of $S_T - S_0$ as a function of T .
- (C) If $G_T - G_0 = H_T - H_0 - S_T T + S_0 T$, are changes in G , as the temperature changes, greater than or less than changes in H in the case of H_2O ?

Standard States and Units of Concentration

Whenever standard entropy or its representative, standard free energy, are calculated from experimental

observations, a decision must be made on the **standard state** to be used. Unlike those of the standard enthalpy and those of the heat capacity, the numerical values of both standard entropy and standard free energy depend significantly on this choice of standard states.^{18,19} When dealing with reactants and products dissolved in solution, such as molecules of proteins and their ligands in water or alkanes in hexadecane, the choice of standard state, other than the obvious conventions of standard temperature and pressure, is a choice of the units in which the concentrations of the reactants and products are to be expressed. The desire in choosing the units for the concentrations is to eliminate any contributions to the entropy arising simply from the act of dispersing the solutes in the solvent and inescapably from the volumes of the solutes and the solvent. These contributions are the **entropy of mixing**. The reason for eliminating entropy of mixing is that the entropy that remains is the entropy of only the reaction itself and changes in the entropy of solvation that accompany the reaction.

It can be assumed, as seems reasonable, that the thermodynamic activity of benzene should be the same whether it is dissolved in octane, decane, dodecane, tetradecane, or hexadecane. It has been shown experimentally²⁰ that this assumption is valid only if the thermodynamic activity of benzene is expressed in units of **corrected volume fraction** as defined by the equation¹⁸

$$a_{A,j} = \gamma_{v,A,j} \phi_{A,j} \exp\left[1 - \left(V_{A,j}/V_j\right)\right] \quad (5-5)$$

where $a_{A,j}$ is the thermodynamic activity of solute A (benzene in the experiment) when it is dissolved in solvent j (octane, decane, dodecane, tetradecane, or hexadecane in the experiment), $\gamma_{v,A,j}$ is the activity coefficient necessary to convert real behavior into ideal behavior, $V_{A,j}$ is the volume of a mole of solute A when it is dissolved in a solution with solvent j , V_j is the volume of a mole of solvent j in the solution, and $\phi_{A,j}$ is the **volume fraction** of solute A in the solution with solvent j :

$$\phi_{A,j} = \frac{n_A V_{A,j}}{n_A V_{A,j} + n_j V_j} = [A] V_{A,j} \quad (5-6)$$

where n_A and n_j are the moles of solute A and solvent j , respectively, in the solution and $[A]$ is its molar concentration.* Most measurements of activity are performed in such a way that the activity coefficient $\gamma_{v,A,j}$ is insignificantly different from 1 or becomes 1 by extrapolation. That the thermodynamic activity of a solute should be

* One must remember that molarity is defined as moles liter⁻¹ and the volume of a mole of a substance is defined as centimeters³ mole⁻¹.

defined by Equation 5-5 was predicted theoretically^{21,22} before it was verified experimentally. Expressing activities of solutes by using Equation 5-5 can be thought of as correcting the concentration of the solute in units of mole fraction for the differences in the volumes of solute and solvent because when the volumes of a mole of solvent and a mole of solute are the same and $\gamma_{v,A,j}$ is 1, Equation 5-5 becomes

$$a_{A,j} = \frac{n_A}{n_A + n_j} = x_{A,j} \quad (5-7)$$

where $x_{A,j}$ is the mole fraction of solute A in solvent j . Equation 5-7 is **Raoult's law**.

The difficulty with the corrected volume fraction is deciding what volume to use for the volume of a mole of solute in the solution. When the solute is a liquid dissolved miscibly in a nonpolar liquid, the molar volume of the solute, V_m , which is the volume of a mole of the pure liquid solute, is a reasonable choice. If, however, the solute is a gas or a solid at the temperature of the measurement, its volume in the solution may be quite different from its volume at the temperature or pressure required to liquify it. In water, even a solute the pure phase of which is a liquid at the temperature of the measurement may have a volume in the solution that is significantly different from its molar volume. In the present discussion, the partial molar volume of the solute at infinite dilution has been chosen as an approximate estimate of the volume of a mole of the solute in the solution. Unlike the partial molar volume, however, which is only an estimate of the volume of a mole of the solute in the solution, the actual volume of a mole of the solute in the solution does not vary with the concentration of that solute.

The **partial molar volume** (centimeters³ mole⁻¹) of solute A or solvent j is defined as the increase in the volume of the solution, ∂V , that occurs when an infinitesimally small number of moles, ∂n , of solute A or solvent j is added to the solution

$$\bar{V} = \left(\frac{\partial V}{\partial n} \right)_{T,P} \quad (5-8)$$

at the concentrations of solute and solvent at which the measurement is made. If the solvent and solute are both hydrocarbons, it is usually assumed that the partial molar volumes of solvent and solute are

$$\bar{V} = \frac{M}{\rho} = V_m \quad (5-9)$$

where M is the molar mass (grams mole⁻¹) and ρ is the density (grams milliliter⁻¹) of the pure phase of the sol-

vent or the solute and V_m is the molar volume.²⁰ Equation 5-9 is also used to estimate the partial molar volumes of other solvents, including water, when solutions are dilute.

The partial molar volumes of most solutes when they are dissolved in water are significantly less than those defined by Equation 5-9.^{23,24} If the solute is a hydrocarbon, it occupies a significant portion of the empty space already present in the water (Figure 5-2). Direct measurement of partial molar volumes of hydrocarbons in water have rarely been performed, usually because such solutes are poorly soluble in water. In the absence of such measurements, the algorithms of Traube^{23,24} are used to estimate partial molar volumes of hydrocarbons in water and those of most other solutes as well.

Traube concluded from direct measurement that the partial molar volume of any neutral solute (centimeter³ mole⁻¹) when it is dissolved in water is the sum of the partial molar volumes of its atoms and functional groups plus the covolume, which is a universal correction. The partial molar volumes of the atoms and functional groups at 25 °C are for hydrogen, 3.1 cm³ mole⁻¹; carbon, 10.0 cm³ mole⁻¹; nitrogen, 1.5 cm³ mole⁻¹; the oxygen in an ether, 5.5 cm³ mole⁻¹; a hydroxyl group (-OH), 5.4 cm³ mole⁻¹; the oxygen in an amide, thioester, ketone, or aldehyde (=O) 5.5 cm³ mole⁻¹; an acyl group (-COO-), 15.9 cm³ mole⁻¹; phosphorus, 17.1 cm³ mole⁻¹; sulfur, 15.6 cm³ mole⁻¹; chlorine, 13.3 cm³ mole⁻¹; bromine, 17.8 cm³ mole⁻¹; and iodine, 21.6 cm³ mole⁻¹. From the sum of the partial molar volumes of its atoms and functional groups, 8.2 cm³ mole⁻¹ must be subtracted for each monocyclic ring, either saturated or unsaturated, and 26.6 cm³ mole⁻¹ for each bicyclic aromatic ring, such as a naphthyl group. To the final sum for the constituents of a particular molecule, a covolume of 12.5 cm³ mole⁻¹ must be added.

When ions are dissolved in water, their charge constricts the solvent in their vicinity. These electrostrictions vary between -10 and -30 cm³ mole⁻¹ for the addition of salts of monovalent ions or monovalent zwitterions.²⁴ How **electrostriction** is to be treated in estimating the volumes of ions to be used in calculating their corrected volume fractions is unclear. The volumes of their ionic solids, however, are also poor estimates of their volumes in solution.

When a reaction takes place in solution, its **equilibrium constant** can be defined by use of units of corrected volume fraction (Equation 5-5), which have the convenient advantage that they are dimensionless. For example, the equilibrium constant for the association



occurring in solvent j when activity coefficients can be ignored, would be

$$K_{\text{eq}} = \frac{a_{\text{C},j}}{a_{\text{A},j} a_{\text{B},j}} = \frac{[\text{C}]}{[\text{A}][\text{B}]} \left(\frac{\bar{V}_{\text{C},j}}{\bar{V}_{\text{A},j} \bar{V}_{\text{B},j}} \right) \exp \left(\frac{\bar{V}_{\text{A},j} + \bar{V}_{\text{B},j} - \bar{V}_{\text{C},j} - \bar{V}_j}{\bar{V}_j} \right) \quad (5-11)$$

If the reaction proceeds with no change in volume

$$\bar{V}_{\text{A},j} + \bar{V}_{\text{B},j} = \bar{V}_{\text{C},j} \quad (5-12)$$

and

$$K_{\text{eq}} = \frac{a_{\text{C},j}}{a_{\text{A},j} a_{\text{B},j}} = \frac{[\text{C}]}{[\text{A}][\text{B}]} \left(\frac{\bar{V}_{\text{A},j} + \bar{V}_{\text{B},j}}{\bar{V}_{\text{A},j} \bar{V}_{\text{B},j}} \right) \exp(-1) \quad (5-13)$$

Because the terms to the right of the quotient of the molar concentrations are not significant functions of the concentration of reactant and product, the quotient of the molar concentrations is a constant, namely, the equilibrium constant that is usually measured when units are molarity. But if entropy of mixing is to be eliminated, the equilibrium constant in units of corrected volume fraction (Equation 5-11) should be used for the calculation of the standard free energy of the reaction:

$$\Delta G^\circ = -RT \ln K_{\text{eq}} \quad (5-14)$$

If the two molecules that are associating are identical—for example, the association is the formation of a hydrogen bond between two molecules of phenol—the quotient of partial molar volumes in Equation 5-13 becomes $2\bar{V}_{\text{A}}^{-1}$, where A refers to one of the two identical molecules; if a small molecule, such as a ligand, associates with a macromolecule, such as a protein, the quotient of the partial molar volumes becomes \bar{V}_{A}^{-1} , where A refers to the ligand. These two situations, where both molecules are the same and where one is much larger than the other, respectively, are the limits on the quotient:

$$\frac{1}{\bar{V}_{\text{A},j}} < \frac{\bar{V}_{\text{A},j} + \bar{V}_{\text{B},j}}{\bar{V}_{\text{A},j} \bar{V}_{\text{B},j}} \leq \frac{2}{\bar{V}_{\text{A},j}} \quad (5-15)$$

where solute A is the smaller of the two molecules.

Another reaction in which the choice of standard state and units of concentration significantly affects the actual value of the standard free energy is the transfer of solute A from water to solvent *j*:



An aqueous phase and a phase of another solvent, for example hexadecane, are placed in contact with each other directly or indirectly. The solute of interest, for example benzene, is added to the system at low concentration, and its partition between the two phases is allowed to reach equilibrium.²⁰ The concentration of the solute in each phase is measured, and a partition coefficient, $K_{\text{p},\text{A}}$, is calculated. Although the concentration of solute A in each of the two phases is initially tabulated in units convenient to the method of measurement,²⁵ for example grams of solute (grams of solvent)⁻¹, units for the concentrations used to calculate the standard free energies of transfer, and hence the definition of standard state, has, as always, a significant effect on the magnitude of the standard free energy of transfer. If it is assumed that corrected volume fractions are the proper units and also that activity coefficients can be ignored, the **partition coefficient** for the transfer of solute A from water to solvent *j* is

$$K_{\text{p},\text{A}} = \frac{a_{\text{A},j}}{a_{\text{A},\text{H}_2\text{O}}} = \frac{\phi_{\text{A},j}}{\phi_{\text{A},\text{H}_2\text{O}}} \exp \left[\left(\frac{\bar{V}_{\text{A},\text{H}_2\text{O}}}{\bar{V}_{\text{H}_2\text{O}}} \right) - \left(\frac{\bar{V}_{\text{A},j}}{\bar{V}_j} \right) \right] \quad (5-17)$$

and the **standard free energy of transfer** is

$$\Delta G^\circ_{\text{A},\text{H}_2\text{O} \rightarrow j} = \lim_{a_{\text{A}} \rightarrow 0} -RT \ln \left(\frac{K_{\text{p},\text{A}} \bar{V}_{\text{A},\text{H}_2\text{O}}}{\bar{V}_{\text{A},j}} \right) = \lim_{a_{\text{A}} \rightarrow 0} -RT \ln \left[\frac{[\text{A}]_j}{[\text{A}]_{\text{H}_2\text{O}}} \exp \left(\frac{\bar{V}_{\text{A},\text{H}_2\text{O}}}{\bar{V}_{\text{H}_2\text{O}}} - \frac{\bar{V}_{\text{A},j}}{\bar{V}_j} \right) \right] \quad (5-18)$$

The limit defines the standard state as the solutions at infinite dilution, a condition at which both activity coefficients are 1 and can definitely be ignored.

The goal of the rather complex definitions of activities and hence standard states that has just been described is to arrive at a value for the standard free energy of transfer that is only the difference between the **standard free energy of solvation** for solute A by solvent *j* and the standard free energy of solvation for solute A by water.¹⁸ The use of Equation 5-5 for thermodynamic activities eliminates the contributions of the entropies of mixing to the standard free energy of transfer. The quotient $\bar{V}_{\text{A},\text{H}_2\text{O}}/\bar{V}_{\text{A},j}$ in Equation 5-18 corrects for the work performed when the volume of the system increases as solute is transferred from water to solvent *j*. In order to focus only on standard free energies of solvation, the conditions must be such that solute A is surrounded entirely by either molecules of water or molecules of solvent *j*, and no molecule of

either solute A or the molecules of water or solvent that surrounds it is affected in its behavior by the presence of another molecule of solute A. This is the reason for the limit in Equation 5-18, which defines the standard state of infinite dilution. In this way, the only contributions to the difference in standard free energy of solvation are the specific interactions between the molecule of solute A and the solvent j or the water.

The choice of units of concentration and standard state is also critical in calculating the transfer of a solute from the gas phase to a solution. The usual choice of standard state for the solution in such a reaction is the solute at infinite dilution in the solvent so that the solute is fully solvated and no interactions occur among the molecules of solute. The usual choice of standard state for the gas is the real gas extrapolated to zero pressure in order to eliminate the nonideal behavior of the real gas represented by its virial coefficients. Because of the proportionality between molarity and pressure, the practical units of concentration for a gas are usually pressure, but the thermodynamic activity of the gas should be defined as its molarity.¹⁸

To avoid both the standard entropy of mixing and changes in volume at constant pressure during the transfer of solute from the gas phase to a solution, the volume occupied by a mole of the solute in the gas phase would have to be equal to the volume occupied by a mole of the solvated solute in the solution.^{18,26} Consider a large volume of solution at standard state in contact with a large volume of the gaseous solute. Only when the pressure of the gas is such that 1 mol of gaseous solute has the same volume as the partial molar volume of the solute in the solution will the volume of the system not change when 1 mol of solute is transferred from the gas to the solution at constant pressure. Only under these circumstances is the transfer both isochoric and isobaric. As a result, the standard entropy of mixing is 0, no work is performed by the system, and the transfer occurs at constant pressure. To achieve this condition, the gas must be compressed mathematically to a volume equal to the partial molar volume of the solute in the solution. The standard free energy change for the compression of the gaseous solute to a volume equal to its partial molar volume in the solution is

$$\Delta G^\circ_{\text{compression}} = -RT \ln([A]_g \bar{V}_{A,j}) \quad (5-19)$$

where $[A]_g$ is its molar concentration in the gas phase

$$[A]_g = \frac{p_A}{2.479 \text{ kPa L mol}^{-1}} \quad (5-20)$$

where p_A is the partial pressure (pascals) of the gaseous solute A at 25 °C. When all of these considerations are combined with Equation 5-5,¹⁸ the equation for the stan-

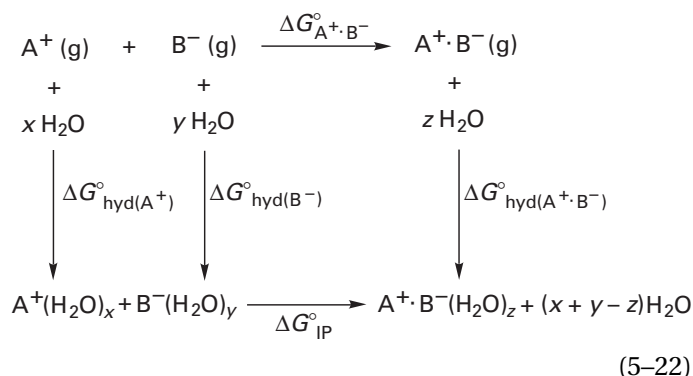
dard free energy of transfer of solute A from the gas phase to a solution in solvent j , $\Delta G^\circ_{A,g \rightarrow j}$, becomes

$$\Delta G^\circ_{A,g \rightarrow j} = -RT \ln \left[\frac{\phi_{A,j}}{[A]_g \bar{V}_{A,j}} \exp \left(1 - \frac{\bar{V}_{A,j}}{\bar{V}_j} \right) \right] = -RT \ln \left[\frac{[A]_j}{[A]_g} \exp \left(1 - \frac{\bar{V}_{A,j}}{\bar{V}_j} \right) \right] \quad (5-21)$$

Again, the intention of Equation 5-21 is to apply the appropriate corrections so that the standard free energy of transfer is only the standard free energy of solvation for solute A by solvent j .

Ionic Interactions

The possibility that a positively charged cation might interact favorably with a negatively charged anion and bring two molecules or two segments of the same polypeptide together has a lasting appeal. Such an association seems plausible because, as everyone knows, unlike charges attract each other. When a positive ion in a solution encounters a negative ion and a complex between these two ions is formed, it is referred to as an **ion pair**. In terms of Equation 5-1, a hydrated ion pair forms whenever a hydrated anion associates with a hydrated cation. In this reaction, the various changes of standard free energy identified in Equation 5-2 can be separately considered by writing the following thermodynamic cycle:



It is easiest to consider the **changes of standard enthalpy** first because they constitute the main contribution to the changes in standard free energy and they have been measured least ambiguously. The standard enthalpy change when a positive ion and a negative ion associate in the gas phase, $\Delta H^\circ_{A^+,B^-}$, is governed by electrostatics. To the extent that the Born-Haber cycle is able to provide accurate estimations of crystal lattice energies

only from simple electrostatic theory,²⁷ the **standard enthalpy of formation of an ion pair in the gas phase** from the two separated ions, A^+ and B^- , if electron repulsion is ignored, should be

$$\Delta H^\circ_{A^+, B^-} = (z_{A^+})(z_{B^-})e_a^2 \left(\frac{1}{a_{A^+} + a_{B^-}} \right) N_A \quad (5-23)$$

where z_{A^+} and z_{B^-} are the charge numbers of the ions, e_a is the elementary charge (1.602×10^{-19} C), and a_{A^+} and a_{B^-} are the radii of the two ions. Values for the ionic radii in crystalline lattices, based on crystallographic studies of salts,²⁸ are usually used for a_{A^+} and a_{B^-} .^{7,29} The standard enthalpy of formation defined by Equation 5-23 can be presented for monovalent ions ($z_{A^+} = -z_{B^-} = 1$) as a function of the sum of the two ionic radii (Figure 5-8).

When an ion is transferred from the gas phase to water, there is a large release of heat.* This large negative change in standard enthalpy is referred to as the **standard enthalpy of hydration**, $H^\circ_{\text{hyd(l)}}$. Measurements of these standard enthalpies of hydration have been tabulated⁷ for a number of monovalent, divalent, and trivalent spherical ions. The values for the spherical monovalent cations and anions can be presented as a function of their ionic radii (Figure 5-8).

The large negative **standard enthalpies of hydration for ions** are commonly explained to be the result of the ability of the fixed charge on an ion to gather around itself a layer of tightly held molecules of water that are oriented either with the positive ends of their dipoles, their hydrogens, toward an anion or the negative ends of their dipoles, their lone pairs, directed toward a cation (Figure 5-9). This explanation is probably incorrect. From measurements of the standard enthalpy of formation for complexes between monovalent cations in the gas phase and 1-7 molecules of water, it has been concluded³⁰ that about four molecules of water are sufficient to hydrate a cation such as NH_4^+ , H_3O^+ , H_2COH^+ , Li^+ , or Na^+ . This result suggests that the innermost shell of the **layer of hydration** around an ion is not large. Furthermore, when the standard enthalpy changes for the formation of 1:1 complexes between a molecule of water and various cations and anions in polar nonaqueous solvents were determined, the values observed were quite small ($0 > \Delta H^\circ > -13$ kJ mol⁻¹).³¹ These two results suggest that the large standard enthalpies of hydration observed for ions arise far more from the influence exerted by the ion over a significant region of the water surrounding it than from the specific, intimate noncovalent contacts between the ion and its immediate neighbors.

In electrostatics the **self-charging energy** is the

* This large release of heat when any ion is transferred from the gas phase to water should not be confused with the small releases or absorptions of heat that occur when the ions in the solid crystals of a salt are dissolved in water.

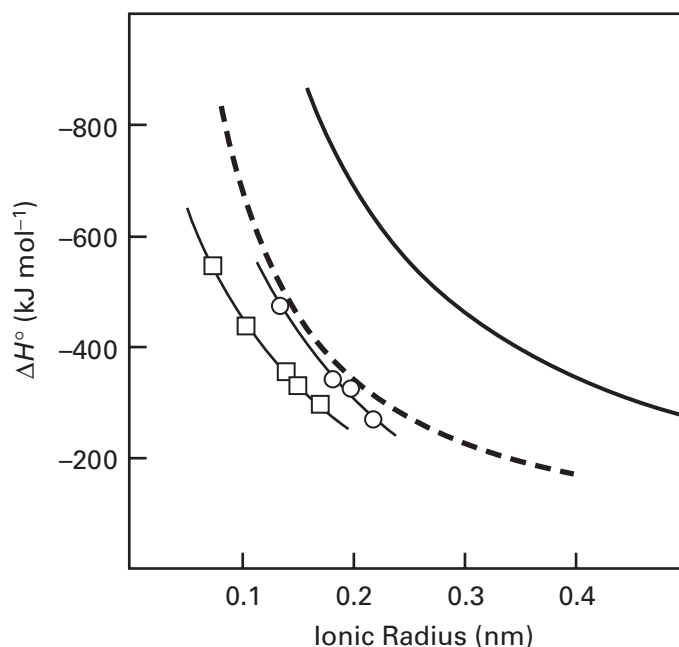


Figure 5-8: Electrostatic enthalpies and standard enthalpies of hydration. The standard enthalpy change for bringing together a monovalent cation and a monovalent anion in a vacuum is presented as a function of the sum of the two respective ionic radii (solid dark line), as calculated from Equation 5-23. The standard enthalpies of hydration⁷ for monovalent cations (\square) are presented as a function of their ionic radii. The ions are, in order of increasing radius, Li^+ , Na^+ , K^+ , Rb^+ , and Cs^+ . The line connecting the points is drawn by hand. The standard enthalpies of hydration⁷ for monovalent anions (\circ) are presented as a function of their ionic radii. The ions are, in order of increasing ionic radius, F^- , Cl^- , Br^- , and I^- . The line connecting the points is drawn by hand. The standard enthalpy change for the hydration of a monovalent ion of either charge, based on the assumption that the standard enthalpy of hydration is due only to the difference in self-charging energies in the vacuum and in water (Equation 5-25), is presented as a function of ionic radius (dark dashed line). All enthalpies are presented in kilojoules mole⁻¹ for a standard temperature of 25 °C.



Figure 5-9: Schematic drawing of molecules of water with the negative ends of their dipoles directed toward a cation and the positive ends of their dipoles directed toward an anion.

energy required to charge a sphere of a given radius a in a medium of relative permittivity ϵ_r .³² The self-charging energy, E_{sc} , for placing the charge $z_j e_a$ on an ion j of radius a_j would be

$$E_{\text{sc}} = \frac{z_j^2 e_a^2}{2a_j \epsilon_r} N_A \quad (5-24)$$

The standard enthalpy change $\Delta H^\circ_{\text{sc}}$ associated with the electrostatic energy required to move an ion from the vacuum ($\epsilon_r = 1$) to water ($\epsilon_r = 78$) at 25 °C would be the difference in the two self-charging energies:

$$\Delta H^\circ_{\text{sc}} = \frac{z_j^2 e_a^2}{2a_j} \left(\frac{1}{\epsilon_{r,\text{H}_2\text{O}}} - 1 \right) N_A \quad (5-25)$$

This standard enthalpy change for a monovalent ion can be presented as a function of ionic radius (Figure 5–8). It can be seen that, for a sphere of unit elementary charge, the value of the difference in self-charging energy between the vacuum and a medium the relative permittivity of which is equal to that of water is close to the experimentally observed standard enthalpy of hydration for a spherical anion (○) of the same radius.

The observed standard enthalpies of hydration for the cations, however, are less than the values expected from differences in self-charging energies. The large differences between $\Delta H^\circ_{\text{sc}}$ and $\Delta H^\circ_{\text{hyd(A}^+)}$ have been explained as being due either to increases in the “effective radius”³³ of the cations or to decreases in the “effective dielectric constant around the ion.”³⁴ Which of these views is more realistic is unknown. Nevertheless, self-charging can explain the shapes and slopes of the curves connecting the experimental values for standard enthalpies of hydration and the majority if not all of the absolute value of each. Because Equation 5–25 accounts for the majority of the standard enthalpy of hydration for simple ions and because it is the large bulk relative permittivity of water that causes the result to be so large, it follows that it is the bulk dielectric of the water, rather than local interactions, that is responsible for the large standard enthalpies of hydration.

The **bulk relative permittivity** of any solvent is a measure of the macroscopic response of that solvent to a fixed electrostatic charge. Consequently, in the calculation of Equation 5–25, a solvent is treated as a uniform continuum and no account is taken of the properties of its individual molecules. The high relative permittivity of water, however, is actually a result of the cooperative behavior of the molecules of water in the liquid over a significant volume. Because the high relative permittivity of water arises from the correlation of the individual dipoles of the molecules of water, the necessity to rely on that relative permittivity to explain the large enthalpy of hydration means that an ion influences the structure of the water over a significant distance, not just in its immediate vicinity.

The **standard enthalpy of hydration for a monovalent ion pair**, $\Delta H^\circ_{\text{hyd(A}^+\text{B}^-)}$, can be estimated from electrostatic theory just as standard enthalpies of hydration were estimated. It should be equal to the difference between the sum of the standard enthalpies of charging

the two ions separately and the standard enthalpy of bringing them to within a certain distance of each other:

$$\Delta H^\circ_{\text{hyd(A}^+\text{B}^-)} = \frac{e_a^2}{2} \left(\frac{1}{a_{\text{A}^+}} + \frac{1}{a_{\text{B}^-}} - \frac{2}{d_{\text{A}^+\text{B}^-}} \right) \left(\frac{1}{\epsilon_{r,\text{H}_2\text{O}}} - 1 \right) N_A \quad (5-26)$$

where $d_{\text{A}^+\text{B}^-}$ is the distance between the monovalent ions in the ion pair; when the two ions are as close together as possible, $d_{\text{A}^+\text{B}^-}$ equals $a_{\text{A}^+} + a_{\text{B}^-}$. It can be shown that

$$\frac{0.828}{a_0} < \left(\frac{1}{a_{\text{A}^+}} + \frac{1}{a_{\text{B}^-}} - \frac{2}{a_{\text{A}^+} + a_{\text{B}^-}} \right) \leq \frac{1}{a_0} \quad (5-27)$$

where a_0 is the radius of the smaller of the two monovalent ions. It follows that the standard enthalpy of hydration for a monovalent ion pair is slightly less than the standard enthalpy of hydration for the smaller of the two ions alone.³²

The overall change in standard enthalpy for the formation of an ion pair in aqueous solution should be

$$\Delta H^\circ_{\text{IP}} = \Delta H^\circ_{\text{A}^+\text{B}^-} + \Delta H^\circ_{\text{hyd(A}^+\text{B}^-)} - \Delta H^\circ_{\text{hyd(A}^+)} - \Delta H^\circ_{\text{hyd(B}^-)} \quad (5-28)$$

From an examination of Figure 5–8, it becomes clear that for monovalent ions this change in standard enthalpy is a small difference between several large numbers, and its value could be either positive or negative. This conclusion, that the standard enthalpy change has a small value, is supported by estimating the electrostatic energy involved in bringing two monovalent ions of opposite charge together in a medium with a uniform relative permittivity equal to that of water:

$$\Delta H^\circ_{\text{IP}} \cong \frac{e_a^2}{\epsilon_{r,\text{H}_2\text{O}}} \left(\frac{1}{a_{\text{A}^+} + a_{\text{B}^-}} \right) N_A \quad (5-29)$$

For $a_{\text{A}^+} \geq 0.1 \text{ nm}$ and $a_{\text{B}^-} \geq 0.1 \text{ nm}$, $-9 \text{ kJ mol}^{-1} < \Delta H^\circ_{\text{IP}} < 0 \text{ kJ mol}^{-1}$.

These changes in standard enthalpy do demonstrate quite clearly why an ion pair sequestered in the middle of a folded polypeptide would be unstable relative to the separated hydrated ions in solution. The only reason an ion pair is almost stable in aqueous solution is that there is considerable standard enthalpy of hydration for the ion pair itself, $\Delta H^\circ_{\text{hyd(A}^+\text{B}^-)}$. In the center of a pro-

tein, this standard enthalpy of hydration would not be exerted and the ion pair would be much less stable. There will never be sufficient electrostatic energy in the ion pair alone to overcome the large negative standard enthalpies of hydration that are lost when the separated ions are removed from water during the folding of the protein. This fact can be verified by examining Figure 5–8. The total standard enthalpy of hydration lost is the sum of the two values for the individual enthalpies of hydration. The standard enthalpy of association gained is that for the sum of the two ionic radii. The former is always of a greater magnitude than the latter.

The **standard entropies of hydration**,³⁵ in marked contrast to the standard enthalpies of hydration, are small. Values of the entropies of hydration for a series of small monovalent ions of either charge lie between -67 and $+21 \text{ J K}^{-1} \text{ mol}^{-1}$ when the two standard states are chosen as the molten salt at one mole fraction in the ion and the ideal solution at one mole fraction in the ion.³⁵ At 298 K, these standard entropies of hydration would cause the standard free energies of hydration to differ from enthalpies of hydration by less than 4%, certainly less than the error in the estimation of enthalpies of hydration from experimental data.⁷

The small standard entropies of hydration seem at first glance to be inconsistent with the formation of a region of oriented water around an ion, which is the explanation given for the large standard enthalpies of hydration. The apparent inconsistency is usually explained by noting that the region of oriented molecules of water surrounding either an anion or a cation cannot merge flawlessly with the hydrogen-bonded lattice of the bulk water. Therefore, there must be an outer spherical shell of disorder between the inner sphere of order and the order of the lattice beyond the influence of the ion. The negative standard entropy change of forming the sphere of oriented water should be canceled by the positive standard entropy change of forming this outer shell of the disordered transition.⁷

Explanations of the large enthalpies of hydration and the small entropies of hydration both predict that an ion will affect the **structure of the water** well beyond the few molecules in its immediate vicinity. Direct evidence for such an extended region of oriented water comes from measurements of the repulsion of hydration.^{36,37} When two identical surfaces that have dense arrays of both negative and positive ions spread over them—both, however, in exactly equal concentration so that each of the two surfaces is electrostatically neutral—are brought together in water, a repulsive force between the surfaces is evident. This repulsive force becomes significant when the two surfaces come within about 2 nm of each other and increases in magnitude exponentially as the distance is decreased. It has been proposed that this repulsive force is the resistance of the layers of hydration around the ions on each surface to their interpenetration. That this **repulsion of the layers of hydration** extends out from each sur-

face by at least 1 nm indicates that each ion orders the waters around it over a significant distance.*

In the case of molecules of protein, the ion pairs that have received the most attention are those that would form between the carboxylate ion of an aspartate or a glutamate and the ammonium ion of a lysine or the guanidinium ion of an arginine. The association constant in water³⁸ for the ion pair between an acetate ion and an ammonium ion is around 0.5 M^{-1} , and that for the ion pair between an acetate ion and a guanidinium ion is somewhat less than 0.5 M^{-1} . Consequently, the concentration of either an ammonium or a guanidinium cation would have to be greater than 2 M for half of the acetate anion in the solution to be complexed with it. These weak interactions have free energies of formation of around -3 kJ mol^{-1} when the concentrations are expressed in units of corrected volume fraction. They are probably the result of hydrogen bonding between the anion and cation rather than ionic interactions, because the complex is stronger for ammonium than guanidinium and there is no evidence that small monovalent cations and anions that lack donors and acceptors of hydrogen bonds associate to form ion pairs in water.

Several additional observations demonstrate that ion pairs between an ammonium ion and a carboxylate ion are unstable relative to the separated ions. The dielectric increment is the change in the relative permittivity of a solution with the concentration of an added solute. The dielectric increments for a series of zwitterionic amino acids containing an ammonium and a carboxylate, namely, glycine, 3-aminopropionate, 4-aminobutyrate, 5-aminopentanoate, and 6-aminohexanoate, have been measured. The values display a monotonic increase with the number of methylenes between the positively charged ammonium ion and the negatively charged carboxylate ion. The values observed are in agreement with theoretical calculations of their magnitude from a simple model in which the distance between the elementary positive charge and the elementary negative charge is determined only by random, unbiased rotation around the carbon-carbon bonds connecting them.²⁴ Were the formation of an ion pair between an ammonium cation and a carboxylate anion a favorable interaction in aqueous solution, this regularity could not have occurred. In glycine, an intramolecular ion pair cannot form. In 3-aminopropionate and 4-aminobutanoate, excellent intramolecular ion pairs, forming rings five and six atoms in size, should form even more readily than a similar intermolecular ion pair. If these intramolecular ion pairs were able to form, how-

* The distance of this repulsive force (1 nm) is about the distance (0.7 nm) calculated for the decrease in the electrostatic field around a univalent ion in water to a potential energy equal to kT . If an ion significantly influences the water around it to a radius of about 1 nm, this region of its influence would contain about 100 molecules of water.

ever, the dielectric increments of these two amino acids should both be less than that of glycine, yet no anomaly is observed in their dielectric increments compared to the other compounds within the complete series. The behavior of the interaction volumes for the same series of amino acids also shows no evidence of peculiarities that would result from intramolecular ion pairing.³⁹

There is no steric hindrance to the formation of an ion pair between the ammonium ion of the lysine in the peptide *N*^α-acetyl-WLKLL and its carboxy terminus, and such an intermolecular ion pair forms readily when the peptide is dissolved in octanol. When the peptide is dissolved in water, however, no ion pair can be detected.⁴⁰

Although ion pairs do not have net favorable standard free energies of formation in aqueous solution and do not contribute to the stability of a folded polypeptide, the **electrostatic repulsion** of amino acids of like charge can destabilize a particular structure. This distinction is illustrated by the effect of ionic strength on the stability of coiled coils of α helices.⁴¹ A coiled coil of α helices is a stable structure that forms when two α helices coil around each other in a supercoil. This supercoil can stabilize the two α helices sufficiently that they can form in water. Few isolated α helices are stable in aqueous solution, but a coiled coil is one way to circumvent this problem. A series of peptides designed to form coiled coils were synthesized chemically. One of the peptides ($n_{aa} = 30$) had glutamates at the positions flanking its hydrophobic core; the other ($n_{aa} = 30$) had arginines flanking the core. The stability of the heterodimeric coiled coil formed from a positively charged peptide and a negatively charged peptide was not affected by changing the ionic strength, and this result indicated that electrostatic interactions such as ion pairing were not contributing to the standard free energy of formation of that coiled coil. The stabilities of homodimers formed from either two of the positively charged peptides or two of the negatively charged peptides, however, decreased significantly as the ionic strength of the solution was lowered, and this result demonstrates that these complexes were destabilized by charge repulsion. It is this destabilization of the two homodimers, not the formation of ion pairs, that accounts for the fact that the heterodimer forms preferentially.⁴¹

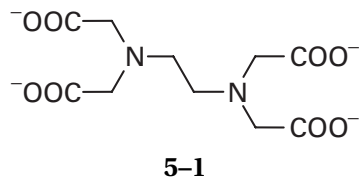
Proteins are generally dissolved in aqueous solutions the **ionic strengths** of which are between 0.1 and 0.3 M, and the effect of such ionic strengths on ionic interactions, although small, should be noted. The activity coefficients of electrolytes decline sharply from a value of 1 at low concentrations to minimum values, depending on the salt, of from 0.05 to 0.8 at concentrations of about 0.3 M.⁴² When activity coefficients of solutes are less than 1, it means that the solute is behaving with a chemical potential less than it would have if it were an ideal solute at the same concentration, and its tendency to leave the solution is less than it should be.

The fact that, at low concentrations, the activity coefficients of ions are near 1 means that as long as they are far enough apart their activities increase in proportion to their concentration as expected for any solute. As their concentrations become high enough that each ion begins to experience the presence of the others, the presence of the others decreases the tendency of that ion to leave the solution. This decreased tendency arises from the departure of all of the ions in the solution from a random distribution in such a way that a region enriched in counterions forms around each individual ion as expected of an ionic double layer (Equation 1-71).⁴² These enriched counterionic layers around each dissolved ion make each of them more stable in the aqueous solution than it would be if it were an ideal solute, and this is what causes the decrease in its activity coefficient.

This effect of ionic strength makes the formation of an ion pair even less likely than it would be in the absence of added salt, because its formation would involve the diminishment of a considerable fraction of the counterionic layer around each separated ion. The presence of these ionic layers also makes it more difficult to remove an ionic functional group from a solution of moderate ionic strength than it would be to remove it from pure water. The activity coefficients for most ionic solutes are between 0.2 and 0.8 at the ionic strengths encountered in biochemical situations, and these values should lead to decreases in the standard free energies of hydration between -4 and -0.4 kJ mol⁻¹, respectively.⁴²

Although ion pairs between simple monovalent cations and anions have positive standard free energies of formation, there are two situations in which ion pairs are favorable. Ion pairs involving **divalent metal ions** often have negative standard free energies of formation. For example, significant concentrations of the ion pairs $\text{Ca}^{2+} \cdot \text{SO}_4^{2-}$ and $\text{Mg}^{2+} \cdot \text{SO}_4^{2-}$ are present in aqueous solutions of the respective salts, and ion pairs between divalent cations such as Ba^{2+} , Ca^{2+} , and Mg^{2+} and hydroxide ion in aqueous solution show appreciable stabilities.²⁹ There are, however, no divalent side chains among the 20 natural amino acids. Phosphorylated amino acids, such as serine phosphate (**2-30**), are divalent at high pH and can readily form ion pairs with divalent cations such as Ca^{2+} .

The other situation in which ion pairs become favorable is encountered when **chelation** can occur. Chelation is the binding of an ion to a molecule, the chelating agent. The chelating agent contains two or more functional groups of opposite charge to the bound ion that can simultaneously associate with it, or it contains two or more dipoles that simultaneously can be favorably directed toward the bound ion, or it contains some combination of such charges and dipoles. The paradigm of chelating agents is *N,N,N',N'*-tetracarboxymethyl-1,2-diaminoethane:



which can wrap its nitrogens and carboxylates around a divalent or trivalent metal ion and form an ion pair of high stability. It has already been mentioned that the binding of monovalent cations and anions by proteins is thought to involve particular binding sites that have advantageous dispositions of functional groups, often with charge opposite to the charge on the bound ion. Chelation, however, assumes a preexisting arrangement of two or more charged groups or dipoles that create a pocket within which an ion can be held, and this arrangement does not exist in an unfolded polypeptide or with isolated anions and cations in solution. Chelation could be important in forming an interface between two already folded polypeptides or binding a charged substrate to an already folded enzyme.

Suggested Reading

Parsegian, A. (1969) Energy of an ion crossing a low dielectric membrane: solutions to four relevant electrostatic problems, *Nature* 221, 844–846.

Problem 5-2: Calculate the standard enthalpy changes for the transfer of a sodium ion ($a_{\text{Na}^+} = 0.097$ nm) and a chloride ion ($a_{\text{Cl}^-} = 0.181$ nm), respectively, from the gas phase ($\epsilon_r = 1$) to water at 25 °C ($\epsilon_{r,\text{H}_2\text{O}} = 78$ at 25 °C). Calculate the standard enthalpy changes for the formation of an ion pair between a sodium ion and a chloride ion in the gas phase and for the transfer of that ion pair from the gas phase to water. By difference, calculate the standard enthalpy of formation of an ion pair between a sodium ion and a chloride ion in water.

Problem 5-3: Consider the series of six compounds $\text{H}_3\text{N}^+(\text{CH}_2)_n\text{COO}^-$, where $n = 1-6$. Within each of these molecules there is a carboxy group and an amino group, which bear opposite charges at pH 7. The elementary negative charge on the carboxy group is located between the two oxygens.

- (A) For each compound, construct with molecular models the conformation of the molecule in which the positively charged nitrogen is positioned as close as possible to one of the negatively charged oxygens.
- (B) In which of the molecules is it possible to form an ion pair that juxtaposes NH_3^+ and O^- ?

The dipole moment (μ) of a particular molecular structure that contains fixed charges is equal to the product of

the magnitude of the charges $z_j e_a$ and the distance, r , that separates them: $\mu = z_j e_a r$. In each structure you have made, $z_j e_a$ is the same but r changes.

- (C) Examine the structures you have drawn and rank them in order of increasing dipole moment. Indicate ranking with the symbols $<$ and $=$.

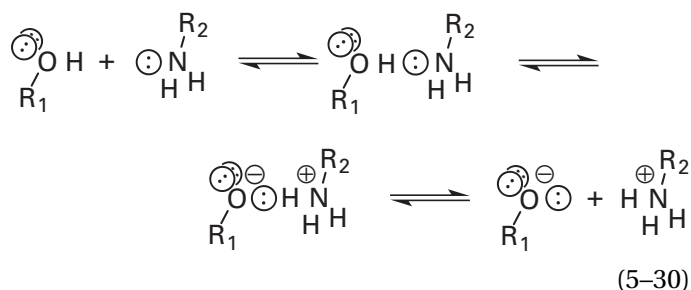
The observed dipole moments for these molecules dissolved in water at pH 7.0 are⁴³

n	1	2	3	4	5	6
μ	12 D	15 D	18 D	20 D	22 D	24 D

- (D) Explain why the actual dipole moments for these molecules fail to agree with the theoretical predictions that you made in part C.

The Hydrogen Bond

A **hydrogen bond** is a noncovalent force that arises between an acid, known as the donor, A-H , and a base, known as the acceptor, :B . The atoms A and B in the case of proteins are the **heteroatoms** oxygen, nitrogen, and sulfur. A hydrogen bond is an intermediate on the trajectory of an acid–base reaction:^{44,45}



The two central complexes in Equation 5-30 are each held together by a hydrogen bond, but the overall reaction is the transfer of a proton between two lone pairs of electrons. An example of this relationship between a hydrogen bond and an acid–base reaction is the self-dissociation of water. In the liquid most of the water molecules participate in hydrogen bonds, and these hydrogen bonds are the intermediate steps in the production of a hydroxide anion and a hydronium cation. The anion and the cation are produced when the proton in a hydrogen bond between two molecules of water moves momentarily from donor to acceptor and the hydrogen bond between the new donor, the hydronium ion, and the new acceptor, the hydroxide ion, dissociates to yield the free species.

Manifestations of hydrogen bonding are the alterations it effects in the physical and chemical properties of liquids, gases, and solids. The nonideal behavior of certain gases can be explained by the existence of hydrogen-bonded oligomers of the molecules composing the

gas. The water dimer (Figure 5–1) is an example of such a situation; its existence lowers the pressure of water vapor. Abnormally positive values for the standard enthalpy of vaporization or abnormally negative values for the standard enthalpy of mixing can often be explained as the result of either the breaking of hydrogen bonds as the molecules depart the liquid or the formation of hydrogen bonds as a donor and acceptor are mixed, respectively. When an acceptor is added to a solution of a donor, the infrared spectrum of the resulting mixture often displays a new absorption band, at a lower frequency than the absorption of the A–H stretching vibration observed with the solution of the donor alone. This new absorption increases in magnitude in proportion to the amount of acceptor added, while the amplitude of the absorption of the unshifted stretching vibration of the A–H bond of the donor decreases in proportion. The new stretching vibration is assigned to that of the A–H covalent bond within a hydrogen bond between the donor and the added acceptor. A similar observation is made in nuclear magnetic resonance spectra of mixtures of donors and acceptors. In this case, two separate absorptions are not observed because the rates at which the hydrogen bonds are interchanging among the molecules in the solution are faster than the time resolution of the method, but the chemical shift of the proton participating in the A–H bond moves down-field until it reaches a maximum value, associated with the chemical shift of the proton within the hydrogen bond.

Taken together, these commonly encountered observations demonstrate three features of a hydrogen bond. First, a hydrogen bond causes two molecules to associate with each other and form a complex that prevents them from changing their relative positions as readily as they would otherwise; in other words, it correlates their movements. Second, there is a release of heat associated with the formation of this complex. Third, the proton in the A–H covalent bond of the donor experiences a change in its environment during the formation of this complex. The results of both infrared and nuclear magnetic resonance spectroscopy are consistent with a lengthening of the covalent bond between A and H concomitant with a movement of the proton away from the electrons of the σ bond.

The arrangement of the atoms in **crystallographic molecular models** of small molecules that display these physical manifestations of hydrogen bonding usually displays a pattern that can be assigned to the hydrogen bond itself. The positions in the unit cell of the atoms of the second and third periods of the periodic table, for example, carbon, nitrogen, oxygen, and sulfur, are determined by X-ray crystallography, and the positions of the protons, often as deuterons, are determined most reliably by neutron diffraction. Whereas a proton has little ability to scatter X-rays, inasmuch as it has no core electrons, a deuteron scatters neutrons as readily as a

carbon or an oxygen;⁴⁶ and deuterons are prominent features in maps of neutron scattering density, as opposed to hydrogens in maps of electron density. Furthermore, a proton has a negative scattering amplitude for neutrons while a deuteron has a positive scattering amplitude. This causes difference maps of neutron scattering density for deuterated against protonated molecules to display sharp maxima where the protons are located in the former.

In crystallographic molecular models of small molecules known to be hydrogen-bonded, the bond is recognized as an enforced orientation of the donor and acceptor (Figure 5–10).⁴⁷ Associated with this orientation are certain bond lengths and bond angles.⁴⁶ It is these bond lengths and bond angles that are the most important property of a hydrogen bond as far as the structures of proteins are concerned. The hydrogen bond provides no net standard free energy to the process of folding a polypeptide, but hydrogen bonds are responsible for aligning atoms and holding them at precise distances and constrained angles to each other in the folded structure. The A–H σ bond of the donor in a hydrogen bond is pointed at the heteroatom B of the acceptor. The distance, d , between A and B is always less than it would be if the proton on the donor atom and the atom acting as the acceptor were simply in van der Waals contact. For example, in a hydrogen bond of the type O–H \cdots N (Equation 5–30), the distance between oxygen and nitrogen is 0.28 ± 0.01 nm,⁴⁶ while the distance between carbon and nitrogen in a van der Waals contact of the type C–H \cdots N would be 0.35 nm. It is this shortened distance between donor and acceptor that reflects the bonding. The **bond lengths**, d , for most types of sterically unconstrained hydrogen bonds (Table 5–1) between neutral donors and neutral acceptors lie between 0.25 and 0.30 nm, but the bond angles are more variable.

In general, the angle α between the axis of the hydrogen bond and one of the σ covalent bonds to the heteroatom of the donor, A (Figure 5–10A), will reflect

Table 5–1: Length of Hydrogen Bonds^a

A–H \cdots B	compounds	average bond length ^b (nm)
OH \cdots O	carboxylic acids	0.26 ± 0.01^c
OH \cdots O	phenols	0.27 ± 0.01
OH \cdots O	alcohols	0.27 ± 0.01
OH \cdots N	all O–H	0.28 ± 0.01
NH \cdots O	ammoniums	0.29 ± 0.01
NH \cdots O	amides	0.29 ± 0.01
NH \cdots O	amines	0.30 ± 0.01
NH \cdots N	all N–H	0.31 ± 0.01

^aThe values in this table are reproduced directly from tables in ref 46. With the exception of the hydrogen bonds involving ammonium cations, these are hydrogen bonds between a neutral donor and a neutral acceptor. ^bThese are the distances between the heteroatoms, nitrogens or oxygens. ^cThese standard deviations may be standard deviations of actual lengths or standard deviations of the measurement or both.

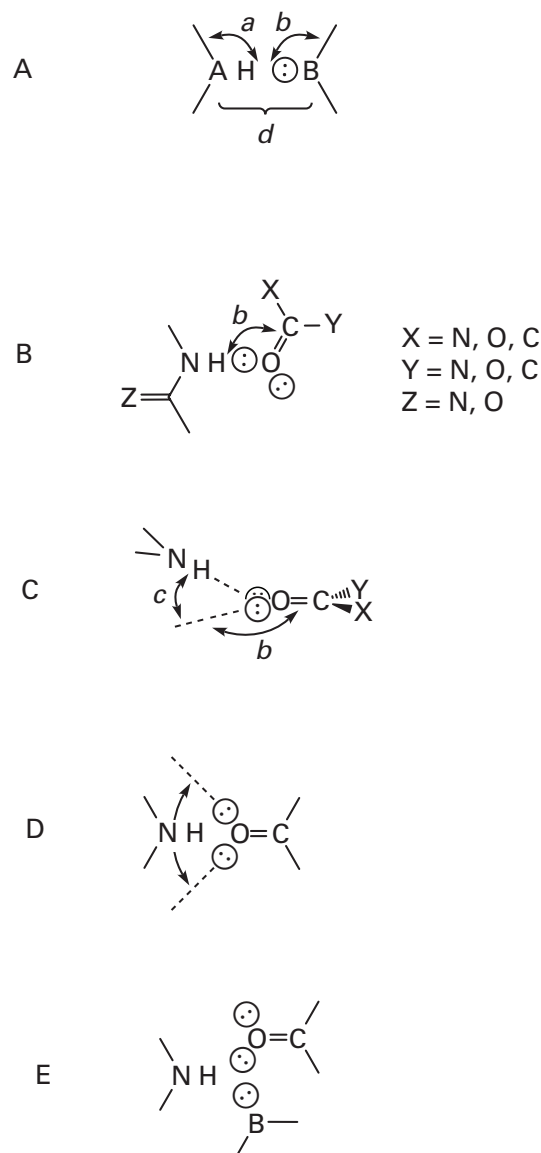


Figure 5-10: Relationships defining bond angles for hydrogen bonds.⁴⁷ (A) In a simple hydrogen bond between a donor AH and an acceptor $\odot B$, the line of center between heteroatoms A and B creates an axis, which is the axis of the hydrogen bond. The angle a is the angle between that axis and a σ covalent bond to the heteroatom A. The angle b is the angle between that axis and a σ covalent bond to the atom B. The distance d is the length of the hydrogen bond. (B, C) For a hydrogen bond between an amide nitrogen and a carbonyl oxygen or an acyl oxygen, the carbonyl or acyl group defines a plane. The bond angle b is the angle in the plane between the carbon–oxygen double bond and the projection of the axis of the hydrogen bond on the plane. In panel C, the lower dotted line shows the projection of the axis of the hydrogen bond on the plane. The bond angle c is the angle between this projection and the axis of the hydrogen bond. For a hydrogen bond between a nitrogen–hydrogen donor and the σ covalent bond of a carbonyl oxygen and acyl oxygen, the bond angle b can vary over a range bounded by the two lone pairs of electrons on the oxygen if the oxygen is otherwise unoccupied. (E) In several instances, the axis of the σ bond between the heteroatom of the donor and the proton lies between two lone pairs of electrons from two different heteroatoms, and one donor interacts with two acceptors in a bifurcated hydrogen bond.

the hybridization of that heteroatom, while the angle b between the axis of the bond and one of the σ covalent bonds to the heteroatom of the acceptor, B, although much more flexible than angle a , will tend to reflect the hybridization of the lone pair of electrons on atom B.

The type of hydrogen bond that accounts for the majority of those in biological macromolecules, both proteins and nucleic acids, is the hydrogen bond between the sp^2 lone pair on an acyl oxygen as an acceptor and the nitrogen–hydrogen bond of an acyl derivative such as an amide or an amidine as a donor (Figure 5-10B). From the crystallographic molecular models of 1500 intermolecular hydrogen bonds between either a carbonyl oxygen or an acyl oxygen and such a nitrogen–hydrogen bond, the bond lengths and **bond angles** were compiled.⁴⁷ For the collection of all such hydrogen bonds examined, the mean nitrogen–oxygen distance, d , was 0.297 nm. If it is assumed that the acyl or carbonyl carbon and its three σ bonds define a plane and that the line of centers between the nitrogen and oxygen of the hydrogen bond defines a line, two angles define the hydrogen bond: angle b , the angle in the plane between the projection upon the plane of the line and the carbon–oxygen double bond (Figure 5-10B); and angle c , the angle that the line of centers between the nitrogen and oxygen makes with the projection of that line of centers on the plane of the acyl group (Figure 5-10C). Angle c determines how far the nitrogen is above or below the plane, and $d \sin c$ is the actual distance the nitrogen is above or below the plane. In the hydrogen bonds examined, the nitrogen atom was usually within 0.1 nm ($\sin c = 0.33$) of the plane defined by the acceptor (Figure 5-11).⁴⁷ If the carbonyl oxygen or acyl oxygen participates as the acceptor in two hydrogen bonds, there is a strong tendency for angle b to be 120° (Figure 5-11B), the angle expected from an sp^2 hybridization of its two lone pairs.

If the carbonyl oxygen or acyl oxygen, however, participates as an acceptor in only one hydrogen bond, angle b will still show a slight preference for 120° but the angle can also assume other values between 120° and 180° with almost equal facility (Figure 5-11A). It is as though the nitrogen–hydrogen bond can **pivot** over the electron cloud formed by both lone pairs when the other one of them is not forming another hydrogen bond (Figure 5-10D), and the location it eventually assumes in the crystal is determined by forces other than the hydrogen bond itself. This apparent ability of a nitrogen donor to associate with two lone pairs on the same atom may be related to its ability to participate in a **bifurcated hydrogen bond**⁴⁸ in which it associates with two lone pairs from separate atoms (Figure 5-10E). The fact that the nitrogen is not confined strictly to the plane of the acyl group (Figure 5-11), although it has a strong preference for that plane, demonstrates that the nitrogen–hydrogen dipole can also pivot up or down out of the plane about a single lone pair (Figure 5-10C) or about two lone pairs (Figure 5-11A).

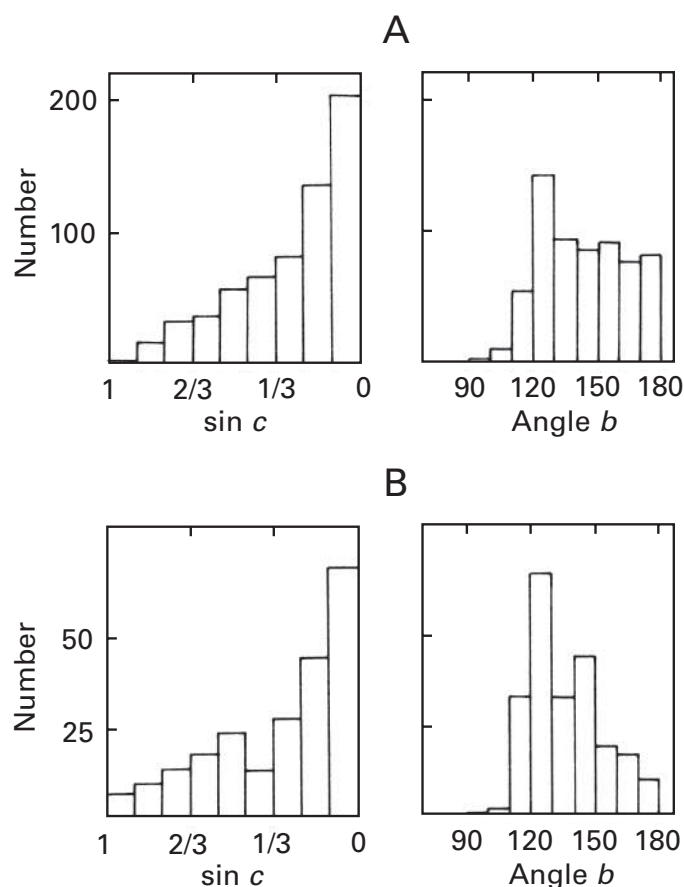


Figure 5-11: Distribution of values for the angle b (Figure 5-10B) and the sine of angle c (Figure 5-10C) over a population of hydrogen bonds between nitrogen-hydrogen donors and carbonyl oxygen acceptors or acyl oxygen acceptors observed in crystallographic molecular models of small molecules.⁴⁷ (A) Hydrogen bonds involving a carbonyl oxygen or acyl oxygen in which the oxygen atom accepts no other hydrogen bonds. (B) Hydrogen bonds involving a carbonyl oxygen or acyl oxygen in which the oxygen atom accepts one other hydrogen bond. In each of the four panels, the number of bonds falling within a range of values is plotted as the value of the ordinate. In the two left panels, the values on the abscissa defining the ranges are values of the sine of angle c ($\sin c$). In the two right panels, the values on the abscissa defining the ranges are values of the angle b in degrees. Adapted with permission from ref 47. Copyright 1983 American Chemical Society.

All of the observations presented in Figure 5-11 are for acyl oxygens that are not in carboxylates. In the case of the oxygens in **carboxylates**, such as those on the side chains of aspartate and glutamate, the tendency for the nitrogen to reside in the plane of the carboxylate is lessened and the tendency for angle b to assume 120° is increased.⁴⁷ Although it has been proposed that a *syn* pair of electrons on a carboxylate (Equation 2-12) should be more basic than an *anti* pair, no preference is shown for one over the other in forming hydrogen bonds in crystallographic molecular models of small molecules⁴⁹ or proteins.⁵⁰

The shift in frequency of the infrared absorption for an oxygen-hydrogen stretching vibration has been used to examine the effect of stereochemistry on the strength

of **intramolecular hydrogen bonds**.⁵¹ A series of pyridines substituted at the *o*-position by $-\text{CH}_2\text{OH}$, $-\text{CH}_2\text{CH}_2\text{OH}$, or $-\text{CH}_2\text{CH}_2\text{CH}_2\text{OH}$ were examined. These functional groups can form hydrogen-bonded rings with the pyridine nitrogen that are four, five, or six atoms in size, respectively, when the proton is not counted. All three compounds display an absorption that could be assigned to a shifted OH stretching vibration. The hydrogen bonds increase in strength, as indicated by the shift in wavelength of the OH stretching frequency ($\Delta\nu = 192$, 203, and 357 cm^{-1} , respectively), as the ring becomes larger. The ring with six atoms, the only one large enough to permit the hydrogen bond to be linear, display the largest frequency shift, an observation consistent with its being the strongest. This result suggests that, in a cyclic hydrogen-bonded structure, the hydrogen bond should be considered as a somewhat longer and more flexible covalent bond between the two heteroatoms, and the proton should not be counted as one of the atoms in the ring.

There has been some disagreement over the ability of **sulfur** to participate as an acceptor in a hydrogen bond because of the poor overlap between its atomic orbitals and those of nitrogen or oxygen. In a survey of crystallographic structures for a number of compounds in which nitrogen donors and sulfur acceptors both appear,⁵² juxtapositions were frequently observed and these were consistent with hydrogen bonds of the type $\text{NH}\cdots\text{S}$. The most telling observation in favor of the existence of such hydrogen bonds was the fact that the nitrogen-sulfur distances ($0.33\text{--}0.35\text{ nm}$) were shorter than the distance expected from purely van der Waals contact.

The pairs of electrons in π bonds in a simple olefin or a phenyl ring are less basic than the σ lone pairs on the acyl oxygen in a secondary amide ($\text{p}K_a = -0.5$)⁵³ or the oxygen of a molecule of water ($\text{p}K_a = -1.7$). The values of $\text{p}K_a$ for the conjugate acids of ethene and propene are -24.3 and -19.3 , respectively.⁵⁴ The values of $\text{p}K_a$ for the conjugate acids in which a carbon in the ring is protonated are -24.3 for benzene, -16.3 for benzofuran, -10 for 3-hydroxy-5-methyltoluene, -7.8 for 3-hydroxyphenol, -5.8 for 3,5-dihydroxytoluene, and -3.1 for 3,5-dihydroxyphenol.^{54,55} From these values, the values of $\text{p}K_a$ for the conjugate acids of a phenylalanyl side chain and a tyrosyl side chain, in which a carbon in the ring is protonated, should be around -20 and -13 . The phenyl ring of a tryptophanyl side chain should have a $\text{p}K_a$ somewhat greater than that of a tyrosyl side chain. The differences between the values of $\text{p}K_a$ for donor and acceptor in the hydrogen bonds between two water molecules or between two molecules of *N*-methylacetamide, however, are already large, around $17\text{--}18$ units, so it would not be surprising if the differences in $\text{p}K_a$ required to use one of these aromatic side chains as an acceptor, although even larger, would still permit the formation of a hydrogen bond.

There are indications that hydrogen bonds can

form between the π clouds of **aromatic rings as acceptors** and biologically relevant donors.⁵⁶ In the complex between water and benzene in the gas phase, the water sits upon the π cloud with the positive end of its dipole oriented towards the ring and its two hydrogens lie 0.1 nm closer to the plane of the ring than van der Waals contact should allow.⁵⁷ All of these features suggest that a hydrogen bond has been formed. There are crystallographic studies of other complexes that also suggest that hydrogen bonds between a hydroxyl group and the π electrons of an aromatic ring do form,^{58,59} and theoretical calculations suggest that a hydrogen bond between an amido nitrogen–hydrogen and a phenyl ring could be as much as half as strong as a normal hydrogen bond between an amido nitrogen–hydrogen and a σ lone pair of electrons.⁶⁰

Associated with any hydrogen bond are two **wells of potential energy** (Figure 5–12), the well of potential energy for a proton within the lone pair of the donor and the well of potential energy for a proton within the lone pair of the acceptor. As there is only one proton between the donor and the acceptor, only one of the two wells is occupied at any given instant. When the proton is located in a particular well, it is participating in a covalent bond with the heteroatom to which the well belongs. When in that covalent bond the proton cannot have an energy less than that of the lowest or first vibrational energy level, and it is usually occupying that level. Because energy is quantized, the energy of the first vibrational level is above the bottom of the well of potential energy. The energy of the first vibrational level is the **zero-point energy** of the bond to which the well applies.

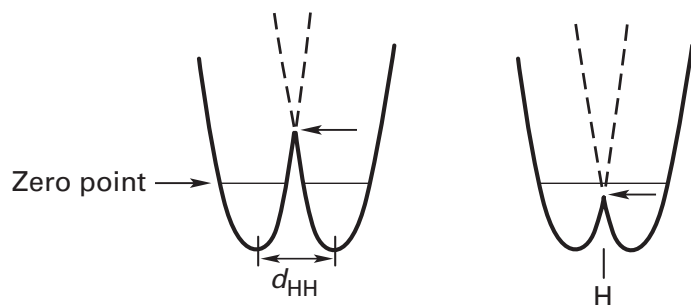


Figure 5–12: Overlap of wells of potential energy for the covalent bonds between the proton and the heteroatom of the donor (left panel) and between the proton and the heteroatom of the acceptor (right panel) in a hydrogen bond. In a case where the values of pK_a for donor and acceptor are matched, the zero-point energies (thin horizontal lines) are the same. (Left panel) If the distance between donor and acceptor is long, the intersection of the two wells of potential energy is above the zero-point energy and there is a barrier to transfer of the proton between the wells (arrow pointing to the left). The proton divides its time between the wells and the two mean positions it assumes are separated by d_{HH} , the distance between the bottoms of the two wells. (Right panel) If the distance between donor and acceptor is short, the intersection between the two wells is below the zero-point energy and the barrier to transfer between the wells (arrow pointing to the left) is no longer effective. The proton (H) is found halfway between donor and acceptor.

In the separated donor and acceptor, these two wells of potential energy are also present and are the wells of potential energy associated with protonating the acceptor or protonating the conjugate base of the donor. As the donor and acceptor approach each other, these wells of potential energy overlap. The point of their intersection (Figure 5–12) is the height of the barrier of potential energy that must be crossed if the proton is to be transferred from donor to acceptor (Equation 5–30). The more closely the heteroatom of the donor and the heteroatom of the acceptor approach each other, the lower will be this barrier.

The difference in the zero-point energies between the well for the donor and the well for an oxygen–hydrogen bond in the hydronium ion is the standard enthalpy change associated with the pK_a of the donor; the difference between the zero-point energies of the well for the conjugate acid of the acceptor and the well for the hydronium ion is the standard enthalpy change associated with the pK_a of the acceptor; and the difference in zero-point energies of the well for the donor and the well for the acceptor is the standard enthalpy change associated with the difference in pK_a (ΔpK_a) between them. If the difference in pK_a between donor and acceptor is small, the two wells of potential energy will have about the same minimum; or better yet, if the acceptor is the conjugate base of the donor, the two wells of potential energy will be mirror images of each other. In such situations, as the heteroatoms are brought closer together, the barrier between them decreases rapidly until it is equal to or less than the zero-point energy (Figure 5–12). When this occurs, the two wells become continuous, as far as the proton is concerned, even though there are still two minima of potential energy. Hydrogen bonds in which the distance between the heteroatoms of donor and acceptor approaches but does not necessarily reach this point at which the barrier vanishes are **low-barrier hydrogen bonds**.

In a hydrogen bond in which the distance between donor and acceptor has become short enough that the barrier has vanished, the two wells have become one and the proton necessarily occupies a position midway between the two heteroatoms.⁶¹ A number of such hydrogen bonds have been observed by neutron diffraction in the crystalline state. When the same hydrogen bond in which the proton is found to be centered in the crystalline state is formed in solution, however, the proton is usually not centered^{62,63} because **solvation** of the bond is more favorable for the situation in which the proton is closer to one of the heteroatoms than to the other.⁶⁴ It is as if solvation has recreated the barrier, probably by biasing the relative energies of the occupied and unoccupied wells at a given instant even if they are identical when unoccupied. When the proton is transferred to the other heteroatom, the change in solvation causes the levels of the wells to switch. Because water strongly solvates dipoles, a barrierless hydrogen bond in

which the proton is centered between the heteroatoms and in which the distinction between donor and acceptor has disappeared rarely if ever exists in water.⁶³

The **length of the covalent bond A–H** between the proton and the heteroatom of the donor is longer when it is in a hydrogen bond than when it is not. In a series of hydrogen bonds of the same type (Figure 5–13),⁶⁵ regardless of whether they are intermolecular or intramolecular examples of the class,⁶⁵ as the distance d_{AB} between the two heteroatoms in a hydrogen bond decreases, the length of the bond between the proton and the heteroatom of the donor increases^{66,67} from its length when it is not hydrogen-bonded (horizontal dashed lines) until the bond becomes so short that the proton sits halfway between donor and acceptor (Figure 5–12). There are several physical measurements that register this increase in the length of the bond between the proton and the heteroatom of the donor as the bond shortens.

The movement of the proton away from the heteroatom of the donor that occurs as the hydrogen bond is formed decreases the electron density of the covalent bond surrounding that proton, and this **deshielding** shifts its peak of absorption downfield in a **nuclear magnetic resonance** spectrum. In a series of hydrogen bonds

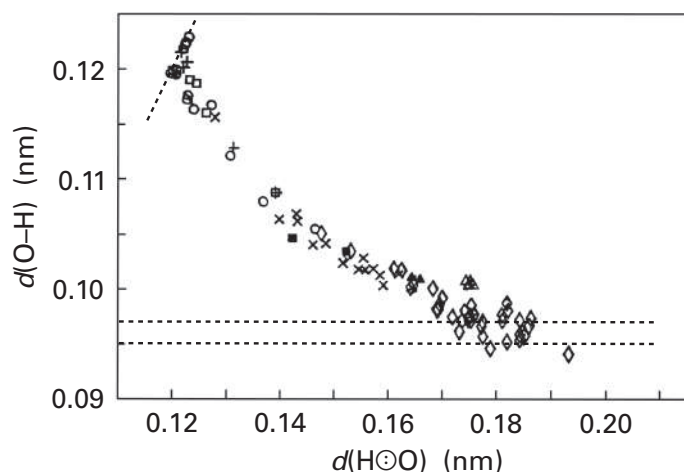
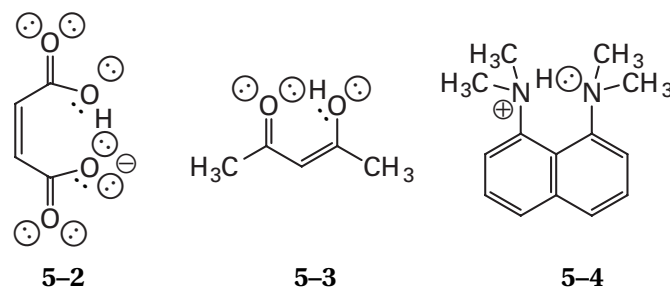


Figure 5–13: Length of the bond between the proton and the oxygen atom of a donor [$d(\text{O}–\text{H})$] in a hydrogen bond between two oxygens as a function of the distance between the oxygen atom of the acceptor and the proton [$d(\text{H}\cdots\text{O})$].⁶⁵ Crystallographic molecular models of complexes containing either intermolecular or intramolecular hydrogen bonds between two oxygens were retrieved from the Cambridge Crystallographic Database. The types of complexes collected were carboxylic acid–carboxylates (○), metal oximes (□), inorganic acid salts (×), hydronium hydroxyls (+), β -diketone enols (■), carboxylic–carboxylics (▲), alcohols (◇), and ice Ih (△). The dashed diagonal line in the upper left-hand corner is drawn for $d(\text{O}–\text{H}) = d(\text{H}\cdots\text{O})$. In the shortest hydrogen bonds, $d(\text{O}–\text{H})$ does equal $d(\text{H}\cdots\text{O})$, the proton sits halfway between donor and acceptor, and donor and acceptor are indistinguishable. As the hydrogen bond increases beyond a length of about 0.24 nm, the proton is closer to the more basic oxygen, and donor and acceptor become distinguishable. The horizontal dashed lines indicate the range of the values for the length of an oxygen–hydrogen bond in an isolated, non-hydrogen-bonded molecule in the gas phase.

of the same type, as the hydrogen bond becomes shorter and the proton moves farther away from the heteroatom of the donor and becomes even more deshielded, its chemical shift becomes even larger.⁶⁸ An absorbance in nuclear magnetic resonance spectroscopy between 16 and 24 ppm for a proton demonstrates that it is in a low-barrier hydrogen bond. For example, chemical shifts of 20.5 ppm for the proton between the two oxygens in hydrogen maleate monoanion (5–2),



of 16.1 ppm for the proton between the two oxygens of the enol of 2,4-dioxopentane (5–3), and 18.5 ppm for the proton between the two nitrogens in hydrogen 1,8-diamino-*N,N,N',N'*-tetramethylnaphthalene monocation (5–4), each measured in organic solvents,⁶⁹ indicate that these are low-barrier hydrogen bonds, as do their lengths (0.241, 0.243–0.251, and 0.258 nm, respectively).^{70–75}

When the donor enters into a hydrogen bond and its A–H bond becomes longer, the force constant for its stretching vibration becomes smaller and the frequency at which it absorbs infrared light becomes lower than when it is not in a hydrogen bond. Consequently, the peak of absorption for the A–H bond of the donor when it is in a hydrogen bond appears in the **infrared spectrum** at a lower frequency than the peak of absorption for the free A–H bond of the donor. The existence of these two distinct peaks of absorption allows the concentrations of bonded and unbonded donor to be quantified (Problem 5–7). In a series of hydrogen bonds of the same type, as the hydrogen bond becomes shorter and the A–H bond becomes longer, the **stretching frequency** of the A–H bond decreases.

The **fractionation factor** ϕ is the equilibrium constant defined by

$$\phi = \frac{[\text{AD}\cdots\text{B}][\text{L}_2\text{O}\cdots\text{HOL}]}{[\text{AH}\cdots\text{B}][\text{L}_2\text{O}\cdots\text{DOL}]} \quad (5-31)$$

where H is protium, D is deuterium, and L is either protium or deuterium. AD \cdots B is the hydrogen bond between the deuterated donor and the acceptor, L₂O \cdots HOL is a hydrogen bond between two molecules of water in which a proton is within the hydrogen bond, AH \cdots B is the hydrogen bond between undeuterated donor and acceptor, and L₂O \cdots DOL is a hydrogen bond between two mol-

ecules of water in which a deuteron is within the hydrogen bond. A fractionation factor of less than 1 indicates that a proton has a greater preference than a deuteron for sitting in the hydrogen bond being examined, relative to the preferences of proton and deuteron for sitting in a hydrogen bond between two molecules of water. The fractionation factor scales the relative preferences of proton and deuteron for any hydrogen bond to their relative preferences for the reference hydrogen bond between two water molecules, much as the acid dissociation constant scales the basicity of any lone pair of electrons to the basicity of a lone pair of electrons on a molecule of water. The fractionation factor is measured by following the concentration of the protonated form of the hydrogen bond of interest as the mole fraction of H₂O is varied in mixtures of H₂O and D₂O.⁷⁶

In a series of hydrogen bonds of the same type, as the distance between the heteroatoms of donor and acceptor decreases, so does the fractionation factor.⁷⁷ This decrease states that as the hydrogen bond becomes shorter, the proton has a greater and greater preference for its occupation relative to that of a deuteron. A value of less than 1 indicates that the hydrogen bond is a short, low-barrier hydrogen bond. The fractionation factors for the hydrogen bonds in hydrogen maleate monoanion (5-2) and hydrogen 1,8-diamino-*N,N,N',N'*-tetramethylnaphthalene monocation (5-4) are 0.84 and 0.90 in water.^{78,79} The fractionation factor for aqueous FHF⁻, which contains one of the shortest hydrogen bonds, is 0.60.⁸⁰ Fractionation factors for hydrogen bonds in organic solvents, however, can be as small as 0.4.⁸¹

The chemical shift, the stretching frequency, and the fractionation factor all monitor the length of the hydrogen bond. They do not, however, provide any indication of its strength.

The strength of a hydrogen bond is expressed in thermodynamic parameters. The **standard enthalpy of formation**, or the heat released when the bond forms, is a measure of the electronic strength of the bond. It is usually the property that is referred to when the **strength of the bond** is discussed indiscriminately. The standard free energy of formation determines the degree to which the hydrogen bond will be favored over the unbonded reactants. Its magnitude is complicated by the fact that it is a function of both the standard enthalpy of formation, the electronic term, and the standard entropy of formation, the quantitative measure of the total change in disorder occurring during the reaction. The standard entropy of formation is usually a negative term because order increases when hydrogen bonds are formed. It is also affected significantly by changes in solvation. In addition, the standard entropy of formation depends on the choice of units for concentration because of the entropy of mixing. The association equilibrium constant, which is usually the quantity that is directly measured, is connected directly to the standard free energy of formation, not to the standard enthalpy of formation.

Ordinarily, the values of these thermodynamic properties are obtained systematically.⁴⁶ A method of measurement, such as infrared spectroscopy, is used to provide values for the molar concentration of free donor, [HA], the molar concentration of free acceptor, [B⊙], and the molar concentration of hydrogen bonds, [B⊙HA], in a solution. The total concentrations of donor and acceptor are systematically varied at a given temperature, and the experimental association equilibrium constants are measured for each set of concentrations:

$$K_{\text{AHB}} = \frac{[\text{B}\odot\text{HA}]}{[\text{B}\odot][\text{HA}]} \quad (5-32)$$

From the **association equilibrium constant** at a particular temperature converted into the proper units (Equation 5-13), the standard free energy of formation of the hydrogen bond can be calculated (Equation 5-14). The variation of the equilibrium constant with temperature is determined experimentally, and from these observations, the standard enthalpy of formation of the hydrogen bond can be calculated:

$$\left(\frac{\partial \ln K}{\partial T^{-1}} \right)_P = - \frac{\Delta H^\circ_{\text{AHB}}}{R} \quad (5-33)$$

Finally, the **standard entropy of formation** is calculated from the experimental results by the relationship

$$\Delta S^\circ = \frac{\Delta H^\circ - \Delta G^\circ}{T} \quad (5-34)$$

The standard enthalpies of formation for hydrogen bonds between uncharged donors and acceptors of biological interest (Table 5-2) lie between -12 and -23 kJ (mol of bond)⁻¹ when the donor and acceptor are dissolved in organic solvents such as CCl₄ or benzene. In spite of these favorable standard enthalpies of formation, the equilibrium constants for the formation of the complexes in a similar set of hydrogen bonds, disregarding those that involve two hydrogen bonds, are quite small when expressed in units of molarity⁻¹ (Table 5-3). When expressed in units of corrected volume fraction (Equations 5-5 and 5-13) to eliminate entropy of mixing, the values are somewhat larger. The small magnitude of these values results from the fact that the negative standard enthalpy of formation is canceled to a considerable degree by a negative standard entropy of formation because even though the correction for volume fraction takes care of the entropy change involved in their finding each other, the two molecules still must reach the proper relative orientations so that the bond can form. Even in the best of circumstances, a hydrogen bond is a weak interaction.

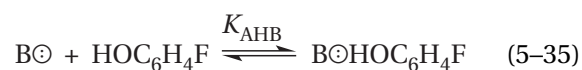
The standard enthalpy of formation of a hydrogen bond is a function of the **difference in pK_a** between the

Table 5-2: Standard Enthalpies of Formation for a Series of Biochemically Important Hydrogen Bonds in Organic Solvents^a

hydrogen bond	solvent	ΔH° (kJ mol ⁻¹)
	CCl ₄	-45 ^b
	CCl ₄	-20
	CCl ₄	-13
	CCl ₄	-18
	CCl ₄	-21
	neat	-15
	benzene	-15
	neat	-14

^aValues are copied directly from tables in ref 46. ^bValue for formation of the two hydrogen bonds of the dimer.

donor and the acceptor.²⁹ For example, a reference donor, *p*-fluorophenol, was chosen, and a series of acceptors were used to form hydrogen bonds with this donor in CCl₄ at 25 °C.^{82,83}



The correlation between standard enthalpy of formation and $\text{p}K_{\text{a}}$ for these hydrogen bonds was⁸³

$$\Delta H^\circ_{\text{AHB}} = (-1.3 \text{ kJ mol}^{-1})\text{p}K_{\text{a}} + b_j \quad (5-36)$$

where b_j (kilojoules mole⁻¹) assumes different values for each category of base examined; for example, b_j has different values for carbonyl oxygens, pyridines, and primary amines. Within a particular category, however, the standard enthalpies of formation are linearly related by Equation 5-36.* This correlation states that as the acceptor becomes a stronger base (as the $\text{p}K_{\text{a}}$ of its conjugate acid becomes larger), the hydrogen bond becomes stronger (the standard enthalpy of formation becomes more negative).

A closely related correlation also exists between the standard free energy of formation and the $\text{p}K_{\text{a}}$ for hydrogen bonds between *p*-fluorophenol and various bases of different $\text{p}K_{\text{a}}$.⁸²

$$\Delta G^\circ_{\text{AHB}} = (-1.3 \text{ kJ mol}^{-1})\text{p}K_{\text{a}} + c_j \quad (5-37)$$

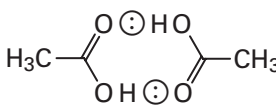
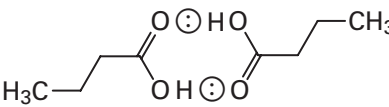
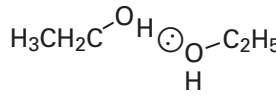
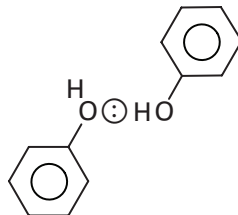
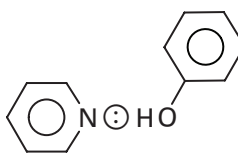
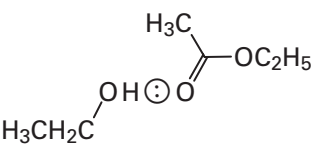
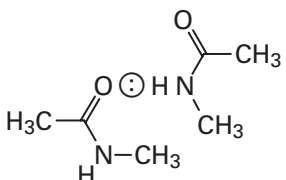
where c_j (kilojoules mole⁻¹) again assumes a different value for each category of bases. The standard enthalpy of formation⁸³⁻⁸⁵ and the standard free energy of formation⁸⁶ of a hydrogen bond are also linearly related to the $\text{p}K_{\text{a}}$ of the donor in similar experiments in which a common acceptor and a systematic series of donors were used (Figure 5-14).⁸⁷

It has been observed that when the infrared stretching frequencies of the A-H bond in a particular donor are examined as a function of the values of $\text{p}K_{\text{a}}$ for the conjugate acids of a set of acceptors, the stretching frequencies of the donor, which monitor the strength of the A-H bond, decrease linearly as the acceptor becomes a stronger base.^{88,89} This observation is consistent with the fact that as an unconstrained, intermolecular hydrogen bond becomes stronger, it also becomes shorter.

If one considers the situation of a hydrogen bond in which the acceptor remains the same and a series of donors of increasing acidity is examined, the standard enthalpy of formation of the hydrogen bond will decrease linearly as the $\text{p}K_{\text{a}}$ of the donor decreases until

* This correlation is an example of the common practice of relating thermodynamic properties to the acid dissociation constant. When the correlation is between the respective values of $\text{p}K_{\text{a}}$ the logarithms of a set of equilibrium and constants (Figure 5-14), the slope is a dimensionless number referred to as the **Brønsted coefficient**. When the correlation is between the values of $\text{p}K_{\text{a}}$ and enthalpies or free energies, the slope of the correlation is the Brønsted coefficient multiplied by $2.303RT$ (kilojoules mole⁻¹). When the correlation is between the values of $\text{p}K_{\text{a}}$ and entropies, the slope of the correlation is the Brønsted coefficient multiplied by $2.303R$ (kilojoules mole⁻¹ kelvin⁻¹). The units on the slope should be consistent with the comparison being made.

Table 5-3: Association Constants of Hydrogen Bonds

hydrogen bond	solvent	temperature (°C)	$K_{\text{AHB}} (\text{M}^{-1})^a$	$K_{\text{AHB}} (\text{cvf})^b$
	benzene	30	130	1680
	benzene	30	430	3420
	CCl ₄	25	0.64	8.1
	CCl ₄	21	2.3	19
	CCl ₄	20	55	480
	CCl ₄	25	1.7	17
	benzene	25	6.2	60

^aValues are copied directly from tables in ref 46. ^bValues for the the association constant given in the dimensionless units of corrected volume fraction (Equation 5-13).

the $\text{p}K_{\text{a}}$ of the donor is equivalent to the $\text{p}K_{\text{a}}$ of the conjugate acid of the acceptor. If the acidity of the donor is increased further, the proton will be transferred between donor and acceptor, the conjugate acid of the former acceptor becomes the new donor, and the conjugate base of the former donor becomes the new acceptor. If the $\text{p}K_{\text{a}}$ of the former donor is decreased below the $\text{p}K_{\text{a}}$ of the conjugate acid of the former acceptor, the standard enthalpy of formation for the hydrogen bond, because donor and acceptor have switched roles, will begin to increase. A corresponding argument could be made for the situation in which the donor remains the same and a series of acceptors, the conjugate acids of which increase in $\text{p}K_{\text{a}}$, is examined. From these considerations and the

fact that $\Delta G_{\text{AHB}}^{\circ}$ tracks $\Delta H_{\text{AHB}}^{\circ}$, it follows that the strength of the hydrogen bond, as measured by its association equilibrium constant K_{AHB} , is determined by the difference in $\text{p}K_{\text{a}}$ between the donor and the conjugate acid of the acceptor (Figure 5-14); the smaller the difference, the stronger the bond. If this is the case, it must follow that the **strongest possible hydrogen bond** in a given series is the one in which the $\text{p}K_{\text{a}}$ of the donor is equal to the $\text{p}K_{\text{a}}$ of the conjugate acid of the acceptor. The most obvious examples of such a hydrogen bond are those between a donor and its conjugate base.

It is important to note, however, that such a **symmetric hydrogen bond**, even one between a donor and its conjugate base, is no stronger than would be pre-

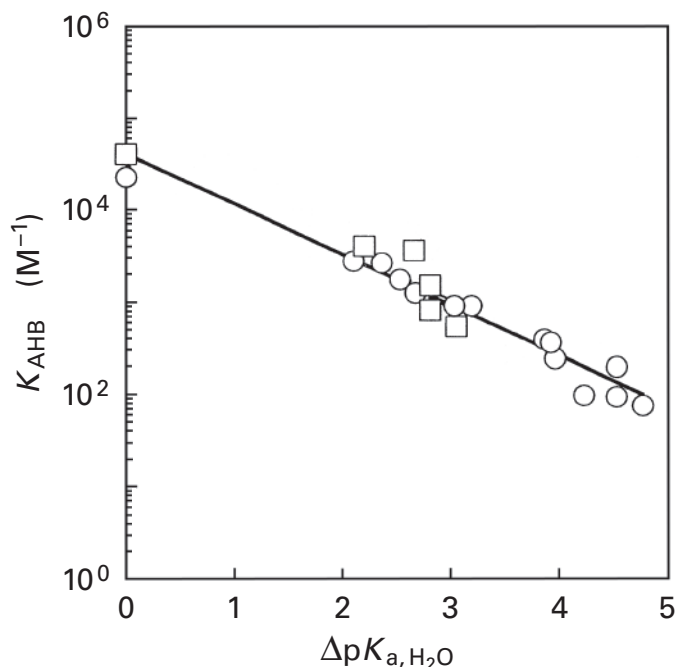


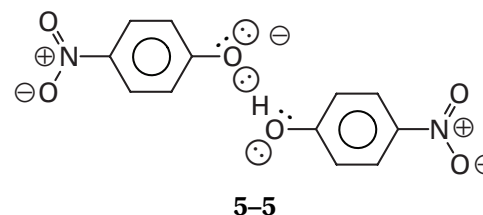
Figure 5-14: Association constants K_{AHB} in tetrahydrofuran for the hydrogen bonds between either 3,4-dinitrophenolate ion (\circ) or 4-nitrophenolate ion (\square) (5-5) and a series of phenols acting as donors as a function of the difference in the values of $\text{p}K_{\text{a}}$ ($\Delta\text{p}K_{\text{a,H}_2\text{O}}$) for the donor and the conjugate acid of the acceptor.⁸⁷ In tetrahydrofuran, the absorptions of both the 3,4-dinitrophenolate ion and the 4-nitrophenolate ion shift from 420 to 388 nm upon formation of a hydrogen bond. As donor is added to the solution, the absorbance at 420 nm decreases and that at 388 nm increases with a clear isosbestic point. The molar concentrations of free donor ($[\text{HA}]$), free acceptor ($[\text{B}\ominus]$), and hydrogen bond ($[\text{B}\ominus\text{HA}]$) were calculated from the ratio of the two absorbances (see Problem 5-7). The logarithms of the association constants (K_{AHB}) in units of molarity^{-1} calculated in this way are plotted as a function of the difference between the respective values of the $\text{p}K_{\text{a}}$ ($\Delta\text{p}K_{\text{a,H}_2\text{O}}$) as measured in water. These values of $\Delta\text{p}K_{\text{a,H}_2\text{O}}$ are expected to be proportional to those in tetrahydrofuran.

dicted by the relationship between the difference in $\text{p}K_{\text{a}}$ and the association constant K_{AHB} (as indicated by the points on the ordinate in Figure 5-14). A similar continuous increase in the strengths of hydrogen bonds has been observed in the gas phase as the difference in the proton affinities of the donors and acceptors decreases again, however, with no evidence for a discontinuity when they are matched.⁸⁰ It has also been observed that the downfield shift of the proton in the hydrogen bonds between 1-methylimidazole and a series of carboxylic acids passes through a maximum with no obvious discontinuity as a function of the $\text{p}K_{\text{a}}$ of the carboxylic acids.⁹⁰

There are at least two ways that the distance between a donor and acceptor in a particular type of hydrogen bond can be shortened. If the hydrogen bond is an intermolecular one, anything that increases its strength will cause it to shorten. If the hydrogen bond is an intramolecular one, the structure of the molecular

scaffold supporting that hydrogen bond can physically compress it. The consequences of these two effects are quite different and should not be confounded.

In a simple, unconstrained, intermolecular hydrogen bond, such as the one between 4-nitrophenolate ion and 4-nitrophenol (5-5 and Figure 5-14)



the distance between donor and acceptor is established by two opposing forces. The short-range, unfavorable potential energies of repulsion between the electrons of the structure and between the three nuclei in the hydrogen bond push donor and acceptor apart. The longer-range, favorable potential energy of attraction between the monopole and dipole and the overlap energies of any covalent bonding pull donor and acceptor together. As with any ionic or covalent bond, it is this tradeoff between longer-range attraction and shorter-range repulsion that defines the **minimum in potential energy** that sets the length of the bond (Figure 5-15).⁹¹ These potential energies are unrelated to the wells of potential energy confining the proton.

Strong intermolecular hydrogen bonds such as the ones in FHF^- ($d_{\text{FF}} = 0.226 \text{ nm}$) and $\text{H}_2\text{OHOH}_2^+$ ($d_{\text{OO}} = 0.236 \text{ nm}$)⁹² are short hydrogen bonds because the monopoles and dipoles are large and the repulsive energies become dominant only at shorter distances. Such

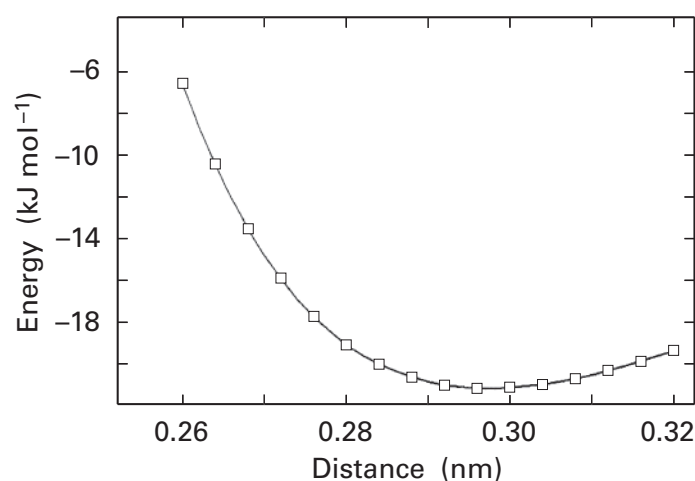


Figure 5-15: An estimate of the energy of formation of a hydrogen bond as a function of the distance between the heteroatom of the donor and the heteroatom of the acceptor.⁹¹ The calculations were for the energy of formation of a hydrogen bond between two molecules of water in the gas phase (Figure 5-1). The variation of that estimate of the energy as a function of the distance between the two oxygens is presented.

hydrogen bonds are special cases that are irrelevant to hydrogen bonds in a molecule of protein at neutral pH. Other hydrogen bonds between acids and their conjugate bases, however, are more relevant to the acids and bases found in a protein (Table 5–4). The hydrogen bond between an acid of a type found in proteins and its conjugate base is about 0.02–0.03 nm shorter than the hydrogen bond between the neutral acid and itself or an equivalent functional group on a different molecule (Table 5–1) because it is a stronger hydrogen bond (Figure 5–14). It seems reasonable that such **strong, short hydrogen bonds**—for example, that between the imidazole of histidine and the imidazolium of another histidine or between the carboxylate of a glutamate and the carboxylic acid of another glutamic acid—should be found in crystallographic molecular models of proteins, but they are rarely observed, probably because the neutralization of the cationic acid or the anionic base required to produce the acceptor or the donor, respectively, requires more free energy at neutral pH than would be gained by forming the stronger hydrogen bond.

It is also possible to shorten a hydrogen bond by physically compressing it within a covalent framework. A number of intramolecular hydrogen bonds are short because of such compression. For example, the hydrogen bond in hydrogen maleate monoanion* (0.241 nm)^{70–72} is shorter than the unconstrained hydrogen bond between two hydrogen fumarate monoanions* (0.247 nm),^{102–105} and that in hydrogen 1,8-diamino-*N,N,N',N'*-tetramethylnaphthalene cation (0.258 nm) is

Table 5–4: Lengths of Hydrogen Bonds between Acids and Their Conjugate Bases^a

acid or conjugate base	bond length (nm)
<i>p</i> -nitrophenol ⁹³	0.246
8-hydroxyquinoline ⁹⁴	0.243
pentachlorophenol ⁹⁵	0.244
	0.248 ^b
1-(<i>p</i> -hydroxyphenyl)thianium ⁹⁶	0.247
cyclohexylamine ⁹⁷	0.280
1,10-diaminodecane ⁹⁸	0.280 ^c
<i>N,N,N</i> -tris(2-aminoethyl)amine ⁹⁹	0.280 ^c
	0.285
<i>N</i> -methylimidazole ¹⁰⁰	0.265
hydrogen succinate ¹⁰¹	0.244 ^d

^aValues presented in this table were gathered during a search of the Cambridge Structural Database by Dr Hens Borkent at the Catholic University of Nijmegen.

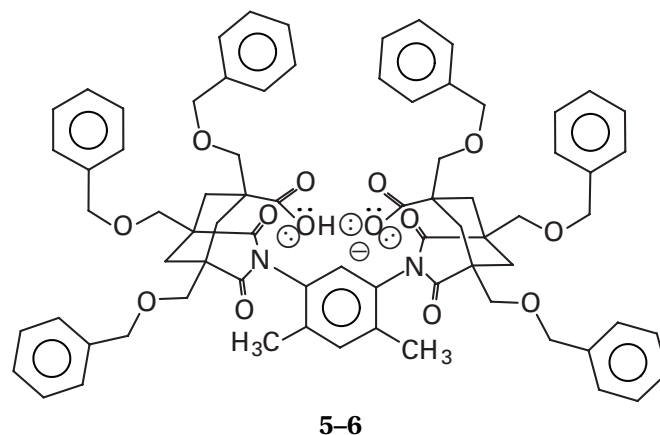
^bTwo different hydrogen bonds in the same unit cell. ^cHydrogen bond between two monocations of the diamine or triamine, respectively. ^dHydrogen bond between two hydrogen succinates.

* The crystallographic molecular models on which these measurements are based were gathered from the Cambridge Structural Database by Dr Hens Borkent at the Catholic University of Nijmegen.

shorter than the unconstrained hydrogen bond between a dimethylalkylamine and its dimethylalkylammonium cation* (0.264 nm).¹⁰⁶ Both of these **short intramolecular hydrogen bonds** display downfield chemical shifts (20.5 and 18.5 ppm) and fractionation factors less than 1 (0.84 and 0.90). In a comparison between intramolecular and intermolecular hydrogen bonds between carboxylate anions and the corresponding carboxylic acid or between enols of β -diketones and the corresponding carbonyl oxygen,⁶⁵ the ranges of lengths of intramolecular hydrogen bonds (0.239–0.242 and 0.243–0.255 nm, respectively) were about 0.006 nm shorter than those of the intermolecular hydrogen bonds (0.244–0.249 and 0.246–0.265 nm, respectively).

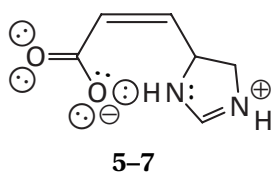
In such intramolecular situations where a hydrogen bond is constrained by the framework of the molecule to be shorter, this **compressed hydrogen bond** must have a less negative standard enthalpy of formation relative to an equivalent, intermolecular, uncompressed one. This conclusion follows from the fact that repulsive potential energy has to be overcome to compress the bond (Figure 5–15). The energy necessary to compress the bond is provided by the covalent framework of the molecule. It follows that an intramolecular hydrogen bond that is shorter than an equivalent intermolecular hydrogen bond must be weaker than that intermolecular hydrogen bond even though it has a lower barrier to proton transfer (Figure 5–12).

The measurements available for the **free energies of formation for such intramolecularly shortened hydrogen bonds** confirm this expectation. The hydrogen bond in hydrogen maleate anion in dimethyl sulfoxide is only -18 kJ mol^{-1} more stable than that in neutral maleic acid,¹⁰⁷ about what one would expect for the difference in standard enthalpy of formation for two hydrogen bonds the acceptors of which differ in $\text{p}K_a$ by 10 units. In water, the difference is only -2 kJ mol^{-1} . The hydrogen bond between the carboxylic acid and the carboxylate anion in **5–6**



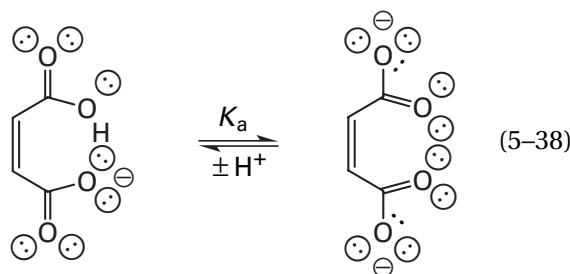
shows a chemical shift (18.0 ppm) in its nuclear magnetic resonance spectrum that is characteristic of a low-barrier hydrogen bond. Yet the standard free energy of its for-

mation differs from the standard free energy of formation for the hydrogen bond in the homologous acid amide, in which an NH_2 replaces the OH , by only -10 kJ mol^{-1} in benzene ($\epsilon_r = 2.3$) and -6 kJ mol^{-1} in dichloromethane ($\epsilon_r = 8.9$).¹⁰⁸ These differences, if anything, are less than expected for the differences in free energies of formation in these two solvents for two hydrogen bonds the donors of which differ by so much in $\text{p}K_a$. The proton in the hydrogen bond in zwitterionic *cis*-urocanic acid



has a downfield shift (18.5 ppm) in nuclear magnetic resonance characteristic of a low-barrier hydrogen bond, yet the standard free energy of formation of this hydrogen bond is only 5 kJ mol^{-1} less than the standard free energy of formation of the hydrogen bond in its conjugate base, *cis*-urocanic acid monoanion,¹⁰⁹ even though the difference in $\text{p}K_a$ between donor and acceptor in the latter hydrogen bond is 13. Consequently, the strength of hydrogen bond 5-7, in which the values of $\text{p}K_a$ for the donor and the conjugate acid of the acceptor are closely matched, cannot be unusually high even though it has the signature of a low-barrier hydrogen bond. The chemical shift and fractionation factor for the proton in the hydrogen bond in the enol of 2,4-dioxopentane (5-3) are those expected of a low-barrier hydrogen bond, but the values of the $\text{p}K_a$ for its donor and the conjugate acid of its acceptor differ so widely ($\Delta\text{p}K_a = 22$) that this cannot be a strong hydrogen bond.

The common practice of equating the unusually high values for $\text{p}K_a$ of the donor in an intramolecular hydrogen bond to its strength is mistaken. The unusually high $\text{p}K_a$ of such an intramolecular hydrogen bond is because of the **repulsion between the lone pairs** of the acceptor and the unprotonated donor that are forced by the framework of the molecule to reside immediately adjacent to each other in the conjugate base



and not because of the strength of the hydrogen bond. The interposition of the proton between the two lone pairs of electrons that are forced into juxtaposition

relieves the repulsion. It comes as no surprise that the activation energy for the exchange of the proton in such a confined location is much higher than that for a proton in an unconstrained, intermolecular hydrogen bond.¹¹⁰

Whether an intermolecular hydrogen bond is shortened by its strength or an intramolecular hydrogen bond is shortened by the compression exerted by the molecular framework to which it is covalently attached, the same increase in the overlap of the wells of potential energy for the proton associated with the heteroatoms of donor and acceptor (Figure 5-12) and the same lowering of the barrier occur. Both strong intermolecular hydrogen bonds and compressed intramolecular bonds can be low-barrier hydrogen bonds because the height of the barrier depends on the length of the bond, not its strength. It follows from these considerations that comparisons of the physical properties of intermolecular hydrogen bonds with those of intramolecular hydrogen bonds are misleading and should be considered guilty until proven innocent. This becomes an even greater offense whenever entropy is involved, as will become more apparent in the next section.

A hydrogen bond is mainly an **electrostatic attraction** between donor and acceptor. The A-H bond of the donor is a dipole, electronegative on the heteroatom, A, and electropositive on the proton. The σ orbital of the acceptor, $\odot\text{B}$, is electronegative on the lone pair and electropositive on the heteroatom, B. These two dipoles attract each other electrostatically when oriented in the same direction. If the donor is positively charged or the acceptor is negatively charged or if both are so charged, the electrostatic attraction is increased.

The hydrogen bond may also have a covalent component and thereby may involve both the overlap of atomic orbitals to form molecular orbitals and the delocalization of valence electrons over the three participating atoms, A, H, and B. Its **covalency** would result from the mixing of the sp^2 or sp^3 orbital on the heteroatom, A, of the donor, the $1s$ orbital of the proton, and the sp^2 or sp^3 orbital of the lone pair of electrons of the acceptor, B, to form a molecular orbital system with three molecular orbitals (Figure 5-16). The covalent component, however, is significantly less important than the electrostatic attraction of the dipoles. It has been proposed that as much as 10% of the hydrogen bonding in ice Ih could be covalent,^{111,112} but this conclusion has been challenged by alternative evaluations of the data suggesting that there is no covalency.¹¹³ The hydrogen bond between two molecules of water, however, is one of the weaker ones, and stronger, shorter hydrogen bonds may have more covalent character.^{65,80}

To the extent that a hydrogen bond is the electrostatic attraction between the dipole of the donor and the dipole of the acceptor, the **relative permittivity of the solvent** should affect its strength just as it affects the strength of an ion pair. The more effectively these dipoles are solvated when they are separated from each other,

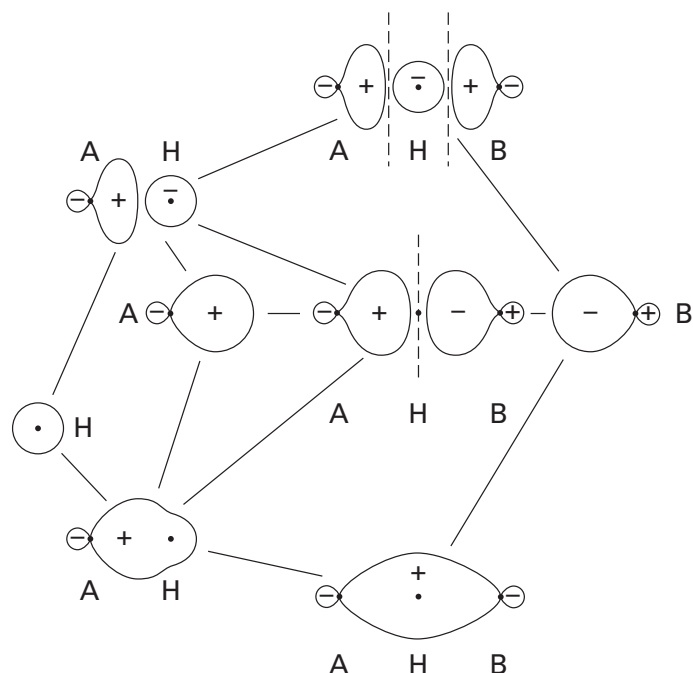


Figure 5-16: Molecular orbitals for a covalent hydrogen bond. The molecular orbitals are for a symmetric hydrogen bond in which the donor and the conjugate acid of the acceptor are equivalent. The covalent molecular orbital system is formed from two sp^2 or sp^3 orbitals, one from atom A and one from atom B, and the s orbital on hydrogen. These three atomic orbitals combine to form the three molecular orbitals—bonding, nonbonding, and antibonding—shown in the middle of the diagram. The final molecular orbital system is constructed formally in steps by first mixing the atomic orbitals of atom A and the hydrogen to form the two molecular orbitals—bonding and antibonding—of the A-H covalent bond and then mixing the A-H molecular orbital system with the atomic orbital on atom B containing the lone pair of electrons.

the less advantage will there be in their combination to form the hydrogen bond. The higher the relative permittivity of the solvent, the weaker will be the bond. One way to quantify this effect⁸⁰ is to compare the slopes of correlations between the free energies of formation of a set of hydrogen bonds and either the values for pK_a of the donor (Figure 5-14) or the values for pK_a of the conjugate base of the acceptor (Equation 5-37). In water ($\epsilon_r = 78$ at 25 °C), the slope for such a correlation for hydrogen bonds between phenolate ion and a set of ammonium ions (Figure 5-17)¹¹⁴ is $0.6 \text{ kJ mol}^{-1} (\text{unit of } pK_a)^{-1}$ and that for a correlation¹¹⁴ of hydrogen bonds between ethylenediammonium dication and a series of phenolate ions is $-0.9 \text{ kJ mol}^{-1} (\text{unit of } pK_a)^{-1}$. The magnitudes of these slopes are less than the $3.1 \text{ kJ mol}^{-1} (\text{unit of } pK_a)^{-1}$ for the correlation of hydrogen bonds between 4-nitrophenolate or 3,4-dinitrophenolate ion (Figure 5-14) and a series of phenols in tetrahydrofuran ($\epsilon_r = 7.5$) or the $-1.3 \text{ kJ mol}^{-1} (\text{unit of } pK_a)^{-1}$ for the correlation (Equation 5-37) of hydrogen bonds between fluorophenol and a diverse set of bases in CCl_4 ($\epsilon_r = 2.2$).

To this point, most of the hydrogen bonds that have been discussed are those formed in aprotic organic sol-

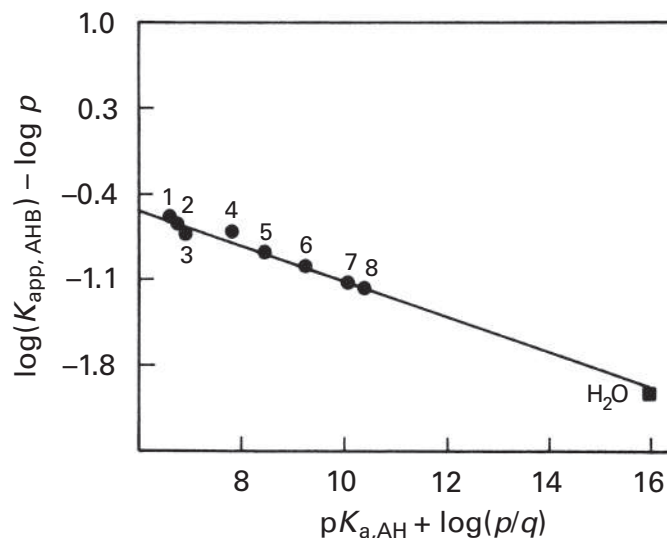
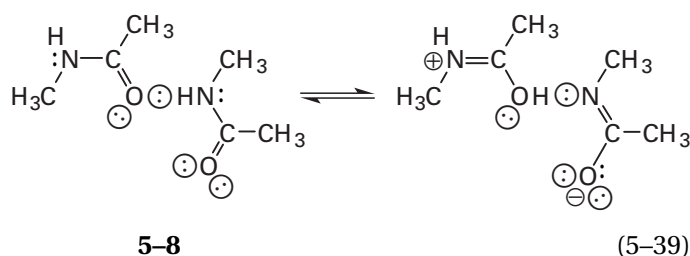


Figure 5-17: The apparent association equilibrium constants ($K_{app,AHB}$) for a series of hydrogen bonds between the phenolate ion and a series of aliphatic ammonium ions in aqueous solution as a function of the acid dissociation constants $K_{a,AH}$ for those ammonium ions.¹¹⁴ The associations between the phenolate ion and the ammonium ions were followed spectrophotometrically by changes in absorbance at 300 nm as ammonium ion was added to an aqueous solution of phenolate ion at 2 M ionic strength and 25 °C. The values of the apparent association equilibrium constants were divided by the number of protons (p) on the respective ammonium ion to convert the molar concentration of the free cation to the molar concentration of donors. The acid dissociation constants were also statistically corrected by multiplying the observed acid dissociation constants by the number of lone pairs on the conjugate base (q) and dividing by the number of protons on the conjugate acid (p), so that the corrected values are for the molar concentration of protons on the respective conjugate acid and the molar concentration of the lone pairs of electrons on the respective conjugate base. The logarithms of the apparent association equilibrium constants (in units of molarity^{-1}) are linearly correlated with the logarithms of the corrected acid dissociation constants (in units of molarity^{-1}) by a line with Brønsted coefficient of 0.15. As the same corrections would be made to both the association constant and the acid dissociation constant in converting to units of corrected volume fraction, the slope of the line would be unaffected. The value of $\log(K_{app,AHB} p^{-1})$ calculated by Equation 5-51 for an acid with a pK_a equal to that of water is indicated by a filled square. The ammonium ions used were (1) hydroxylammonium ion, (2) piperazine dication, (3) *sym*-tetramethylethylenediammonium dication, (4) *N,N,N*-trimethylethylenediammonium dication, (5) ethylenediammonium dication, (6) 2-hydroxy-1,3-diaminopropane dication, (7) 1,3-diaminopropane dication, and (8) (2-hydroxyethyl)ammonium ion. Adapted with permission from ref 114. Copyright 1986 American Chemical Society.

vents such as carbon tetrachloride or benzene. The situation changes dramatically when the donor and acceptor are dissolved in **water** because of the competition of the donors and acceptors of the water molecules themselves. Pauling and Pressman¹¹⁵ noted that the standard free energy of formation of a hydrogen bond in water must be the difference between its own standard free energy of formation and the free energies of formation of the hydrogen bonds of its donor and acceptor with water.

The fact that the concentrations of donors and acceptors in water are both 110 M is a sufficient observation in itself to lead to the conclusion that the hydrogen bond between a solute A-H and a solute \ominus B would be unlikely to form.

The intermolecular hydrogen bond between the nitrogen-hydrogen bond of an **amide** and the lone pair of electrons on the acyl oxygen of another amide can be used again as an example of the majority of the hydrogen bonds in biological macromolecules. When *N*-methylacetamide is dissolved in carbon tetrachloride or dioxane, an absorption appears in the infrared spectra of the two solutions that can be assigned¹¹⁶ to the stretching vibration of the hydrogen-nitrogen bond in a hydrogen bond of the structure



Hydrogen bond 5-8 is the one supposedly holding α helices and β structure together in a molecule of protein and the base pairs together in DNA. From calorimetric measurements, it can be calculated that the standard free energy of formation of this hydrogen bond¹¹⁷ at 25 °C in CCl_4 is -16 kJ mol^{-1} , when an infinitely dilute solution of *N*-methylacetamide is defined as the standard state and the association equilibrium constant is expressed in units of corrected volume fraction (Equation 5-13). When *N*-methylacetamide is dissolved in water, however, the infrared absorption arising from hydrogen bond 5-8 can barely be detected even at a concentration of 12.5 M. From the small absorption that was observed, the standard free energy of formation of the hydrogen bond in aqueous solution was judged¹¹⁶ to be about $+7 \text{ kJ mol}^{-1}$ at 25 °C, again in units of corrected volume fraction.

A more complete picture of the situation is gained by estimating the **standard free energy of transfer** of an amido group from water to CCl_4 at 25 °C.^{118,119} The standard free energy of transfer of *N*-methylacetamide from water to CCl_4 at 25 °C is $+9 \text{ kJ mol}^{-1}$ at infinite dilution in both phases when units are corrected volume fraction (Equation 5-18). If this value is corrected for the hydrophobic effect expressed during the transfer of the three hydrogen-carbon bonds on the methyl group in the acetyl group and the three hydrogen-carbon bonds of the *N*-methyl group from water to CCl_4 ,^{119,120} the standard free energy of transfer for just the unassociated amido group ($-\text{CONH}-$) should be about $+25 \text{ kJ mol}^{-1}$. One amido group contains one donor and an acceptor and by itself represents the two participants in the final hydrogen bond, so its free energy of transfer from water to CCl_4 should be equal to that for the transfer of one

amido nitrogen-hydrogen and one acyl carbon-oxygen (indicated in Figure 5-18 by N-H and $\text{C}=\text{O}\ominus$). The complete standard free energy diagram for the hydrogen bond (Figure 5-18)¹¹⁷⁻¹¹⁹ suggests that the standard free energy of transfer of a hydrogen-bonded amido nitrogen-hydrogen and acyl carbon-oxygen from water to CCl_4 at 25 °C is around $+2 \text{ kJ mol}^{-1}$, a value that registers the polarity of the hydrogen bond. The most significant difference between CCl_4 and H_2O , however, is the high stability of the separated donors and acceptors in the H_2O . The large unfavorable standard free energy of transfer for the amido group from water to CCl_4 reflects the necessity to break hydrogen bonds between it and the water before the transfer can occur.

If this is the case, the **formation of the hydrogen bond in water** must be written, in analogy to Equation 5-1, as

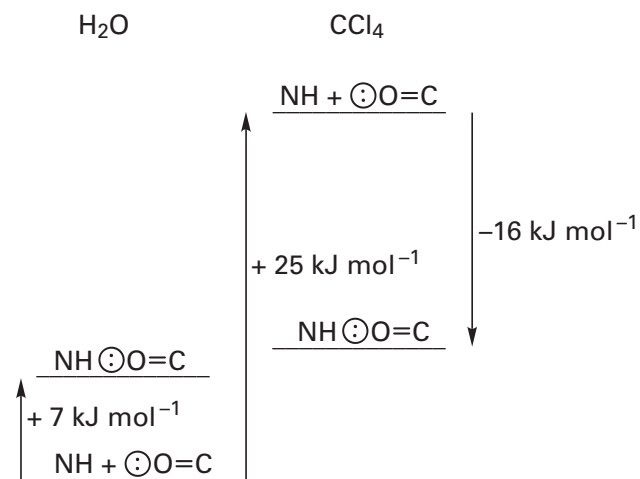
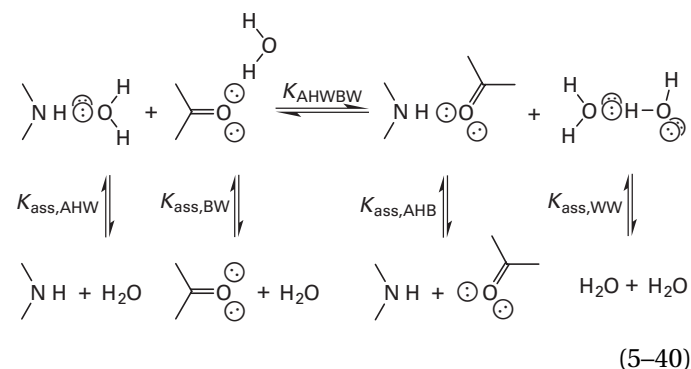


Figure 5-18: Diagram of standard free energy for a hydrogen bond between the amido nitrogen (NH) and acyl oxygen ($\text{C}=\text{O}\ominus$) of *N*-methylacetamide.¹¹⁷⁻¹¹⁹ The standard free energies of association for the hydrogen bond in water ($+7 \text{ kJ mol}^{-1}$) and carbon tetrachloride (-16 kJ mol^{-1}) are positioned in relation to each other on the diagram by an estimate of the standard free energy of transfer ($+25 \text{ kJ mol}^{-1}$) of an unbonded amido group between the two solvents. The standard free energy of transfer for the amido group alone was calculated from the distribution coefficient of *N*-methylacetamide between water and carbon tetrachloride, extrapolated to infinite dilution, with the concentrations of solute in each solvent expressed in units of corrected volume fraction, and an estimate of the standard free energy of transfer of its two methyl groups^{119,120}

where the association constant for any of the complexes, $K_{\text{ass},XY}$, is

$$K_{\text{ass},XY} = \frac{[X\odot Y]}{[X][Y]} \quad (5-41)$$

It is entirely possible⁴⁴ that the concentrations of unbonded donors and unbonded acceptors in aqueous solutions are negligible and that only the upper part of Equation 5-40 is thermodynamically relevant.

The equilibrium constant for the upper part of Equation 5-40, K_{AHWBW} , is defined by

$$K_{\text{AHWBW}} = \frac{[\text{B}\odot\text{HA}][\text{H}_2\text{O}\odot\text{H}_2\text{O}]}{[\text{B}\odot\text{H}_2\text{O}][\text{H}_2\text{O}\odot\text{HA}]} \quad (5-42)$$

where, in the particular case of *N*-methylacetamide, AH is the amido nitrogen–hydrogen bond and $\text{B}\odot$ is a lone pair of electrons on the acyl oxygen. The equilibrium constant actually observed, $K_{\text{app,AHB}}$, is

$$K_{\text{app,AHB}} = \frac{[\text{B}\odot\text{HA}]}{([\text{B}\odot\text{H}_2\text{O}] + [\text{B}\odot])([\text{H}_2\text{O}\odot\text{HA}] + [\text{HA}])} \quad (5-43)$$

from which it follows that

$$K_{\text{app,AHB}} = K_{\text{ass,AHB}} \left[\frac{1}{1 + \frac{K_{\text{ass,AHW}}}{(K_{\text{ass,WW}})^{1/2}} [\text{H}_2\text{O}\odot\text{H}_2\text{O}]^{1/2}} \right] \times \left[\frac{1}{1 + \frac{K_{\text{ass,WB}}}{(K_{\text{ass,WW}})^{1/2}} [\text{H}_2\text{O}\odot\text{H}_2\text{O}]^{1/2}} \right] \quad (5-44)$$

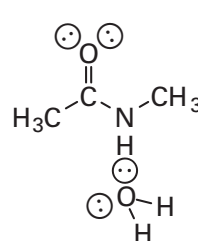
If, as is reasonable, $K_{\text{ass,AHW}} > (K_{\text{ass,WW}})^{1/2}$, $K_{\text{ass,WB}} > (K_{\text{ass,WW}})^{1/2}$, and $[\text{H}_2\text{O}\odot\text{H}_2\text{O}] > 1 \text{ M}$, it follows that

$$K_{\text{app,AHB}} \cong \frac{K_{\text{AHWBW}}}{[\text{H}_2\text{O}\odot\text{H}_2\text{O}]} \quad (5-45)$$

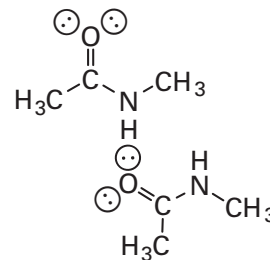
where $[\text{H}_2\text{O}\odot\text{H}_2\text{O}]$ is the molar concentration of hydrogen bonds in the water.

The difference in $\text{p}K_{\text{a}}$ between the nitrogen–hydrogen bond in *N*-methylacetamide as a donor ($\text{p}K_{\text{a}} = 16$) and the oxygen–hydrogen bond in water as a donor ($\text{p}K_{\text{a}} = 15.7$) should be negligible, as should be the differ-

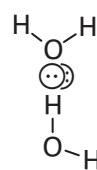
ence between the lone pair of electrons on the acyl oxygen of *N*-methylacetamide ($\text{p}K_{\text{a}} = -0.6$) and the lone pair of electrons on water ($\text{p}K_{\text{a}} = -1.7$) as acceptors. Therefore, the standard enthalpy of formation for the following four hydrogen bonds should be similar



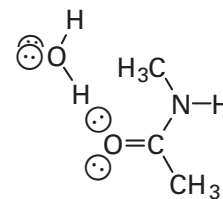
5-9



5-10



5-11



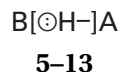
5-12

and the standard enthalpy change for the upper part of Equation 5-40 should be near zero. If the upper part of Equation 5-40 were isoentropic as well as isoenthalpic, so that $K_{\text{AHWBW}} \cong 1$ (Equation 5-42), then $K_{\text{app,AHB}}$ would be equal to $[\text{H}_2\text{O}\odot\text{H}_2\text{O}]^{-1}$. The observed apparent association constant for the formation of hydrogen bond 5-8 in aqueous solution when expressed in units of reciprocal molarity¹¹⁶ is about $(190 \text{ M})^{-1}$, which is in the range expected for the reciprocal of the concentration of hydrogen bonds in pure water, $[\text{H}_2\text{O}\odot\text{H}_2\text{O}] \leq 110 \text{ M}$. The conclusion to be drawn from these considerations is that a hydrogen bond in aqueous solution will always have a **small apparent association constant** and a large apparent standard free energy of formation because the concentration of hydrogen bonds between water molecules in the solution is a hidden and significant term in that apparent association constant (Equation 5-45).

The hydrogen bond represented by that of *N*-methylacetamide accounts for the majority of those found in proteins and nucleic acids, and yet its standard free energy of formation is positive by a considerable degree. From this it follows that each hydrogen bond of this type in a protein or a nucleic acid is an energetic liability rather than an asset. It is possible, however, that some other combination of donor and acceptor might produce a hydrogen bond strong enough to overcome the competition of the water and provide a negative standard free energy of formation. In assessing this possibility, it would be useful to have an equation that could be used to **estimate the apparent**

equilibrium constant, $K_{\text{app,AHB}}$, for the formation of any hydrogen bond in aqueous solution. Such an equation has been derived⁴⁴ and has been demonstrated to be reliable.¹¹⁴

Consider the hydrogen bond



and focus in turn on the portions within the brackets and the portions without the brackets. The portion within the brackets is the same for all hydrogen bonds. The structures without the brackets on either side affect the intrinsic enthalpy of the hydrogen bond by donating or withdrawing electrons from this central structure.¹²¹ The net result of their action is an intrinsic enthalpy, $H_{\text{int,AHB}}$, that is proportional to the product of the two respective σ constants:¹²¹

$$H_{\text{int,AHB}} = v_{\ominus\text{H}} \sigma_{\text{A}} \sigma_{\text{B}} \quad (5-46)$$

where $v_{\ominus\text{H}}$ is a constant of proportionality. These σ constants are the same terms used in physical organic chemistry to provide a quantitative assessment of the ability of any group to withdraw or donate electrons and cause changes in standard enthalpy in any similar situation.

For the upper part of Equation 5-40, when the specific example of *N*-methylacetamide is replaced by the general reaction between A-H and $\ominus\text{B}$, the standard enthalpy change for the reaction should be

$$\Delta H_{\text{AHWBW}}^{\circ} = H_{\text{int,AHB}} + H_{\text{int,WW}} - H_{\text{int,AHW}} - H_{\text{int,BW}} \quad (5-47)$$

or

$$\Delta H_{\text{AHWBW}}^{\circ} = v_{\ominus\text{H}} (\sigma_{\text{A}} - \sigma_{\text{OH}}) (\sigma_{\text{B}} - \sigma_{\text{H}_2\text{O}}) \quad (5-48)$$

where σ_{OH} is the σ constant for OH taking the place of A in 5-13 and $\sigma_{\text{H}_2\text{O}}$ is the σ constant for H_2O taking the place of B in 5-13. As the values of the σ constants are proportional to the values of $\text{p}K_{\text{a}}$ for the appropriate acids

$$\Delta H_{\text{AHWBW}}^{\circ} = 2.303 RT \tau (\text{p}K_{\text{aHA}} - \text{p}K_{\text{aHOH}}) (\text{p}K_{\text{aHB}} - \text{p}K_{\text{aH}_3\text{O}^+}) \quad (5-49)$$

where τ incorporates the constants of proportionality. If it is assumed for the moment that the standard entropy change for the upper part of Equation 5-40 is negligible and that differences in standard enthalpy are the only significant determinants of the relative strengths of the hydrogen bonds being considered, then

$$-\log K_{\text{ass,AHWW}} = \tau (\text{p}K_{\text{aHA}} - \text{p}K_{\text{aHOH}}) (\text{p}K_{\text{aHB}} - \text{p}K_{\text{aH}_3\text{O}^+}) \quad (5-50)$$

and⁴⁴

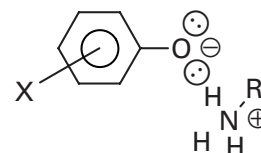
$$\log K_{\text{app,AHB}} = \tau (\text{p}K_{\text{aHA}} - \text{p}K_{\text{aHOH}}) (\text{p}K_{\text{aH}_3\text{O}^+} - \text{p}K_{\text{aHB}}) - \log [\text{H}_2\text{O} \odot \text{H}_2\text{O}] \quad (5-51)$$

where the values of the measured $\text{p}K_{\text{a}}$ have been corrected statistically for the number of protons, p , on the conjugate acid and the number of lone pairs, q , on the conjugate base:

$$K_{\text{a,corr}} = \frac{q}{p} K_{\text{a,meas}} \quad (5-52)$$

As the units of concentration for the acid dissociation constants are in units of molarity, the units for the association constant and the concentration of hydrogen bonds in the water must also be expressed in units of molarity. Equation 5-51 can be used to estimate the association equilibrium constant in units of molarity^{-1} , and hence the standard free energy of formation (Equations 5-13 and 5-14), of any hydrogen bond in aqueous solution.

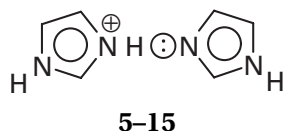
This relationship was validated¹¹⁴ by an examination of the formation of hydrogen bonds between a series of phenolate anions as acceptors and ammonium cations as donors at 25 °C



5-14

where the substituents X and R were various electron-donating and electron-withdrawing groups chosen to vary the values of $\text{p}K_{\text{a}}$ for the donor and acceptor. The logarithms of the association equilibrium constants for the formation of these hydrogen bonds varied with the $\text{p}K_{\text{a}}$ of either the donor (Figure 5-17) or the acceptor as predicted by Equation 5-51. Extrapolating the relationships to either $\text{p}K_{\text{a,HA}} = \text{p}K_{\text{a,HOH}}$ or $\text{p}K_{\text{a,HB}} = \text{p}K_{\text{a,H}_3\text{O}^+}$ gave the same value, 2.0, for $\log [\text{H}_2\text{O} \odot \text{H}_2\text{O}]$. This numerical value is a reasonable estimate for the logarithm of the concentration of hydrogen bonds in liquid water, where $[\text{H}_2\text{O} \odot \text{H}_2\text{O}] \leq 110 \text{ M}$, and it is in reasonable agreement with the results gathered independently with *N*-methylacetamide.

The value of τ (Equations 5-49 and 5-51) at 25 °C and 2 M ionic strength was found to be 0.013, from which it follows that even the strongest possible hydrogen



Because the large **relative permittivity of water weakens** the overwhelming electrostatic component of the hydrogen bond and because the **donors and acceptors of the water molecules themselves compete** for occupation of any other donors and acceptors in solution, any comparisons of hydrogen bonds in organic solvents with hydrogen bonds in water are misleading and should be avoided. It is incumbent on the author to state clearly the solvent in which the property of each hydrogen bond being discussed was measured.

The interaction coefficient τ in Equation 5-49 is a small number because it is the product of the slope of the line relating $\Delta H^\circ_{\text{AHB}}$ to the difference in the values of $\text{p}K_{\text{a}}$ between the donor and water (Figure 5-17) and the slope of the line relating $\Delta H^\circ_{\text{AHB}}$ to the difference in the values of $\text{p}K_{\text{a}}$ between the conjugate acid of the acceptor and conjugate acid of water. Both of these slopes are small because, as a result of its high relative permittivity, water solvates the separated donors and acceptors so strongly. Because both slopes are small, their product is even smaller. Because τ is such a small number (0.013), the standard enthalpy of formation for any hydrogen bond formed in water will be negligible. It follows that a hydrogen bond in aqueous solution cannot be strengthened significantly by any alteration of the acid dissociation constants of donor and acceptor.

Consequently, it is the standard entropy of formation that determines the standard free energy of formation.¹²³ It is the entropic effect of the high molar concentration of the hydrogen bonds between molecules of water ($[\text{H}_2\text{O} \cdots \text{H}_2\text{O}]$ in Equation 5-51) that is usually the overriding contributor to this standard entropy of formation. Equation 5-51 states that the standard free energy of formation of a hydrogen bond in an aqueous solution, when standard state has the units of molarity for concentration, will be increased by about $+11 \text{ kJ mol}^{-1}$ because of the high molar concentration of the hydrogen bonds between molecules of water. There is a way, how-

Suggested Reading

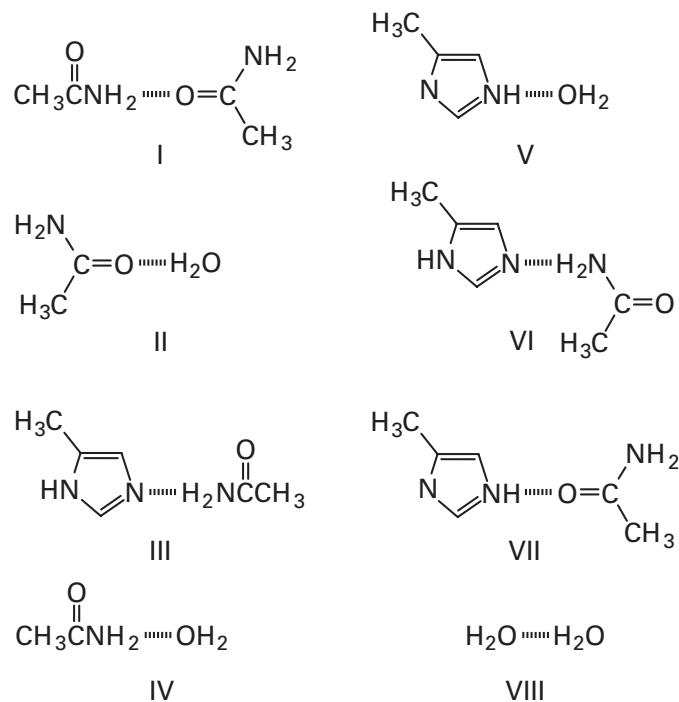
- Jencks, W.P. (1987) Hydrogen Bonds, in *Catalysis in Chemistry and Enzymology*, Chapter 6, pp 323–350, Dover, New York.
- Perrin, C.L., & Nielson, J.B. (1997) “Strong” hydrogen bonds in chemistry and biology, *Annu. Rev. Phys. Chem.* 48, 511–544.

Problem 5–4: Draw structures that represent all of the possible hydrogen bonds that can form between the following pairs of molecules when they are dissolved in a nonpolar solvent. Draw the structures with proper geometry and include all lone pairs of electrons.

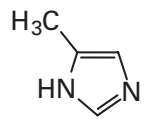
- (A) two molecules of *N*-methylacetamide
(B) *N*-methylacetamide and ethanol
(C) ethanol and water
(D) *N*-methylacetamide and water
(E) urea and *N*-methylacetamide
(F) acetic acid and *N*-methylacetamide
(G) 4-methylimidazole and *N*-methylacetamide
(H) 4-methylimidazole and ethanol
(I) acetic acid and ethanol

Problem 5-5:

- (A) Draw the structure of each of the following hydrogen bonds in the most stable geometry (bond angles and distances). Include all lone pairs of electrons.



- (B) Write the acid dissociations to which the following apparent values of pK_a refer, and correct them for number of protons and number of lone pairs.

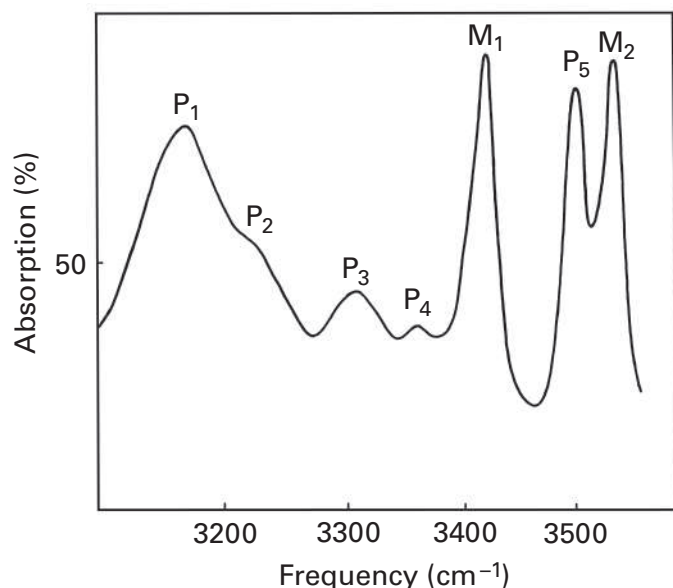
	pK_{a1}	pK_{a2}
$\text{CH}_3\text{C}(=\text{O})\text{NH}_2$	0.6	15.70
H_2O	-1.75	15.75
	-7.51	15.10

- (C) Rank the eight hydrogen bonds in order of strength.

Problem 5-6: Draw the structure of only the most stable hydrogen bond that forms between side chains of the following pairs of amino acids. Include all relevant angles and distances around the various hydrogen bonds.

- (A) glutamic acid and histidine
 (B) serine and tyrosine
 (C) glutamine and histidine

Problem 5-7: The panel below is an infrared spectrum of propionamide in carbon tetrachloride.¹²⁴ The bands marked M_1 and M_2 are the absorptions of the monomeric propionamide, and the bands marked P_1 – P_5 are absorptions of hydrogen-bonded species. As the concentration of propionamide is increased, within each of these two sets [(M_1, M_2) and $(P_1$ – $P_5)$] the amplitudes of the individual absorptions remain in constant ratio to each other and must be different absorptions of the same species.



Below is a table of amplitudes from the infrared spectra for various total concentrations of propionamide at 298 K in CCl_4 .

$[\text{propionamide}]_{\text{TOT}}$ (M)	A_{M_1}	A_{P_1}
1.71×10^{-3}	0.415	0.055
2.15×10^{-3}	0.509	0.083
2.43×10^{-3}	0.566	0.102
4.72×10^{-3}	0.981	0.308
6.90×10^{-3}	1.32	0.558
10.40×10^{-3}	1.79	1.020

Recall that, by Beer's law, $[\text{monomer}] = (\epsilon_{M_1})^{-1}A_{M_1}$ and $[\text{polymeric species}] = (\epsilon_{P_1})^{-1}A_{P_1}$. If the hydrogen bonding is a dimerization, it should be described by the following equation:



$$K_{\text{eq}} = \frac{[\text{propionamide}_2]}{[\text{propionamide}]^2}$$

- (A) Show that the data are consistent with a dimerization.
 (B) Use the data to determine K_{eq} at 298 K in units of corrected volume fraction. (Hint: $[\text{propionamide}]_{\text{TOT}} = [\text{propionamide}] + 2[\text{propionamide}_2]$.)

Values of K_{eq} were determined at a number of different temperatures.

temp (K)	$K_{\text{eq}} (\text{M}^{-1})$
303	35.5
313	24.6
329	13.3

- (C) Convert these three equilibrium constants to units of corrected volume fraction.
 (D) Using these three numbers and your value for 298 K, calculate ΔH° for the dimerization.
 (E) The appearance of the species P_1 – P_5 in the infrared spectrum can be adequately explained only if propionamide has two hydrogen bonds that form a cyclic structure. Draw that structure.
 (F) What is ΔH° for each mole of hydrogen bond?

Problem 5-8: Poly[d(AT)·d(AT)] melts at 67 °C and poly[d(GC)·d(GC)] melts at 102 °C. The usual explanation for this observation is that there is one more hydrogen bond in a G-C pair than in an A-T pair. Consider, however, this explanation in terms of Figure 5-18. Assume

that when the heterocyclic bases in single strands of randomly coiled DNA form a double helix, the interior σ faces of the bases are transferred from water to a completely nonpolar environment (Figure 6–48) and that the proper tautomers for base pairing are present at the pH of the experiment.

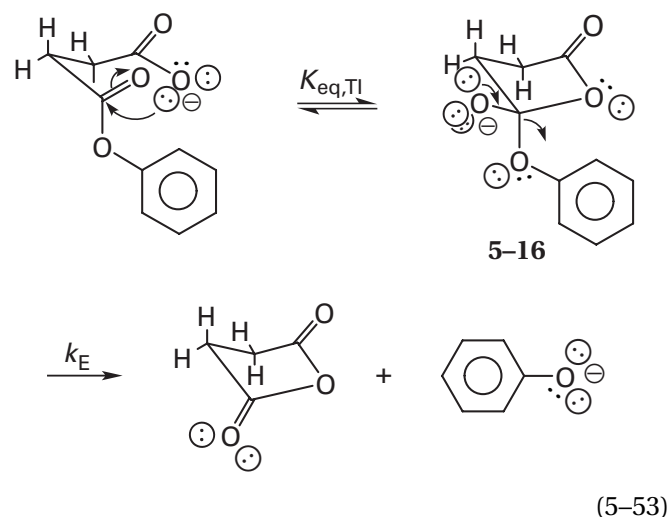
- (A) Write complete equations for the formation of an A·T pair and a G·C pair as the double helix forms, drawing in all hydrogen-bonded waters to all donors and acceptors on each side of the equation and the hydrogen bonds that form between the residual waters. Assume that the donors and acceptors of hydrogen bonds accessible to water in the major and minor grooves retain their hydrogen bonds with H_2O .
- (B) Why is poly[d(GC)·d(GC)] more stable than poly[d(AT)·d(AT)]?

Problem 5–9: There are two possible hydrogen bonds that can form between the neutral form of phenol, a model for tyrosine, and the free base of imidazole, a model for histidine.

- (A) Draw the full structures of both partners in both of the possible hydrogen bonds with proper hybridization on the central atoms and proper bond angles around the hydrogen bond.
- (B) Which of the two possible hydrogen bonds is the more stable? Why?
- (C) Estimate the standard free energy of formation of the more stable hydrogen bond at 25 °C when it is formed in aqueous solution if the statistically corrected values of $\text{p}K_a$ are 9.65 for phenol and 7.35 for the conjugate acid of imidazole.

Intramolecular and Intermolecular Processes: Molecularity and Approximation

Intramolecular chemical reactions often occur at rates much faster than equivalent intermolecular reactions, and intramolecular associations often occur with association equilibrium constants much larger than those of equivalent intermolecular associations. A particularly informative series illustrating such effects can be gathered¹²⁵ from among the reactions involving intramolecular nucleophilic catalysis of the hydrolysis of phenyl esters by the carboxylate anion. The mechanism for this nucleophilic catalysis has been shown to involve the formation of an intermediate anhydride, which in the intramolecular examples such as phenyl succinate would be cyclic:



The tetrahedral intermediate **5–16** that leads to the anhydride has both a phenolate and a carboxylate as potential leaving groups, and because the latter is the better leaving group, the intermediate should decompose to reactant much more frequently than to anhydride. Therefore, the reaction involves a **preequilibrium** between the reactant and the tetrahedral intermediate. Occasionally the phenolate is ejected from the tetrahedral intermediate in a kinetically irreversible step. It is the ejection of the phenolate that is monitored as the reaction progresses. The first-order rate constant, in units of reciprocal seconds, for the appearance of phenolate would be equal to $K_{\text{eq,TI}}k_E$, where $K_{\text{eq,TI}}$ is the equilibrium constant for the formation of the tetrahedral intermediate. If it is assumed that for all compounds in the series the rate constant k_E , which in all cases is for a chemically equivalent first-order reaction, has the same value, then the differences in observed rates result from differences in $K_{\text{eq,TI}}$, an equilibrium constant.

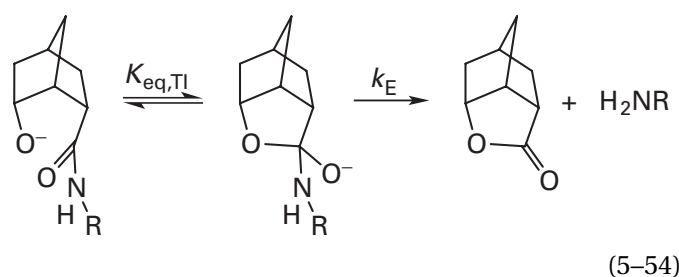
A comparison of the first-order rate constants of phenolate release for a series of intramolecular reactions¹²⁵ to the estimated pseudo-first-order rate constant for the intermolecular reaction between phenyl acetate and excess acetate anion, when the catalysis by acetate anion proceeds through acetic anhydride as an intermediate,^{126,127} indicates how large these **intramolecular increases in an association constant** can be (Table 5–5). The 6×10^5 increase in the association equilibrium constant for the intramolecular formation of succinic anhydride,¹²⁸ when compared to the association equilibrium constant for the intermolecular formation of acetic anhydride expressed in units of corrected volume fraction, is somewhat greater than the increase seen with phenyl succinate (Table 5–5). This fact suggests that the increases listed in Table 5–5 are reasonable.

A situation similar to the intramolecular hydrolysis of phenyl esters is encountered in the alkaline hydrolyses of *endo*-6-hydroxybicyclo[2.2.1]heptane-*endo*-2-carboxamides in which a preequilibrium between reactant and tetrahedral intermediate precedes the expulsion of the amine:¹²⁹

Table 5-5: Relative Preequilibrium Constants^{125,126} for the Formation of a Tetrahedral Intermediate^a

phenyl ester	relative equilibrium constant ^b	$T\Delta S^\circ_{\text{approx}}{}^c$ (kJ mol ⁻¹)
	1.0	
	100	-11
	2×10^4	-24
	1×10^6	-34
	4×10^6	-38

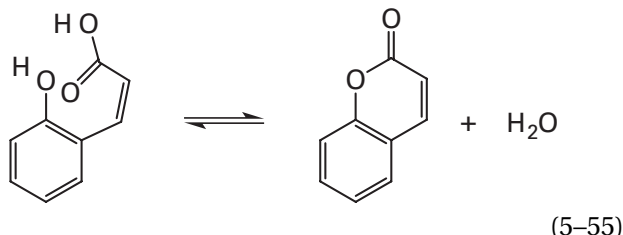
^aThe first-order rate constants for the ejection of phenol from the several phenyl monoesters of dicarboxylic acids (last four entries) were determined¹²⁵ as a function of pH in the range pH 4–8. From the pH-rate behavior of each of these rate constants, the first-order rate constant for intramolecular nucleophilic catalysis of the respective ejection by the appended carboxylate could be calculated. From these values, the first-order rate constants for intramolecular anhydride formation could be calculated. These rate constants were determined for each phenyl monoester at 25, 30, or 35 °C. The values of these rate constants were adjusted to the same temperature and originally presented relative to the first-order rate constant for intramolecular ejection of phenol from phenyl glutarate.¹²⁵ These first-order rate constants were later related to the pseudo-first-order rate constant for the intermolecular formation of acetic anhydride from excess acetate anion and phenyl acetate during the intermolecular nucleophilic catalysis of the hydrolysis of phenyl acetate by acetate anion.¹²⁶ ^bFirst-order rate constants for the formation of anhydride were originally presented relative to the calculated¹²⁶ pseudo-first-order rate constant for the formation of acetic anhydride from phenyl acetate and acetate anion. The latter intermolecular rate constant was in units of molarity⁻¹ second⁻¹. It has been assumed that all of the rate constants for the formation of the anhydrides are directly proportional to the equilibrium constants for the formation of the respective tetrahedral intermediates (Equation 5-53). The resulting units of molarity⁻¹ for the equilibrium constant for the formation of the tetrahedral intermediate from acetate anion and phenyl acetate were converted to units of corrected volume fraction with Equation 5-13. No correction is required for the intramolecular reactions if it is assumed that they involve only negligible changes in molar volume. All equilibrium constants for the formation of the tetrahedral intermediates are presented relative to that for the reaction of acetate anion with phenyl acetate. ^cStandard entropy of approximation, calculated by Equation 5-60 with the intermolecular equilibrium constant in units of corrected volume fraction and with the assumptions that all of the rate constants are directly proportional to the equilibrium constants for the formation of the tetravalent intermediate and that $\Delta\Delta H^\circ \approx 0$. This entropy of approximation was multiplied by 298 K.



In these reactions the intramolecular increase in association constant $K_{\text{eq,TI}}$ is 1×10^6 relative to that for the formation of the equivalent tetrahedral intermediate from hydroxide ion and a bicyclo[2.2.1]heptane-*endo*-2-carboxamide when concentrations are expressed in corrected volume fraction.

Many other examples of intramolecular accelerations of rates or increases in association equilibrium constants have been reported,^{29,126} but the accelerations

presented in Table 5-5 are among the largest that have been noted for each particular size of ring formed during each of these intramolecular reactions. In some of the other instances, electronic effects are difficult to separate from the effects of approximation. For example, in the intramolecular reaction



the two reactants are connected electronically by the π system in addition to being juxtaposed by the σ system.

At one time it was fashionable to refer to these accelerations in rate or increases in association as the result of an increase in the **effective molarity** of one of the reactants brought about by attaching it covalently to the other. As more exaggerated examples of this phenomenon were reported, however, the unreality of discussing concentrations of millions of molar became apparent,¹²⁸ and a more reasonable view of the situation was required.^{128,130}

In all instances in which an intramolecular association, for example, the intramolecular formation of a hydrogen bond in a folding polypeptide, is compared to an equivalent intermolecular association, the difference observed in the two equilibrium constants $K_{eq,intra}$ and $K_{eq,inter}$ is due in large part to an increase in the change in standard entropy caused simply by the fact that a unimolecular reaction is being compared to a bimolecular reaction. This increase in the standard entropy change for the intramolecular association results from the fact that the standard entropy of approximation is missing from the standard free energy change for the intramolecular association. The **standard entropy of approximation**, ΔS°_{approx} , is the change in standard entropy due solely to bringing the separate reactants together into the same molecule or into the same complex, respectively, prior to the beginning of the reaction. It is a negative number because the intrinsic entropy of two separate reactants relative to the unimolecular product of a reaction is greater than the intrinsic entropy of one molecule containing the two reactants or of one complex into which the two reactants have been assembled relative to the product of the intramolecular reaction. For example, the intrinsic entropy of a free acetate anion and a free phenyl acetate is much larger relative to the tetrahedral intermediate formed when they associate than is the intrinsic entropy of a phenylsuccinate anion relative to the cyclic tetrahedral intermediate in Equation 5-53. Because the standard entropy of approximation is missing from the standard entropy change for the intramolecular reaction, owing to the fact that approximation has

already been accomplished synthetically or biologically, the standard entropy change for the intramolecular reaction is more positive than the standard entropy change for the intermolecular reaction.

These relationships can be expressed in equations. These equations are not intended to reflect actual changes in standard entropy, which are often dominated by changes in solvation, but to represent the quantitative consequences of approximation underlying the enhancements in rate or equilibrium constant. The standard entropy change of the intramolecular reaction, ΔS°_{intra} , should be related to the standard entropy change of the intermolecular reaction, ΔS°_{inter} , by

$$\Delta S^\circ_{inter} = \Delta S^\circ_{intra} + \Delta S^\circ_{approx} \quad (5-56)$$

if the same reaction with the same change in standard enthalpy is occurring once the reactants have been approximated. If this were an adequate description of the situation and the same change in standard enthalpy did occur in each reaction, then

$$R \ln \left(\frac{K_{eq,intra}}{K_{eq,inter}} \right) = \Delta S^\circ_{intra} - \Delta S^\circ_{inter} = -\Delta S^\circ_{approx} \quad (5-57)$$

Because $\Delta S^\circ_{approx} < 0$, $K_{eq,intra} > K_{eq,inter}$.

The magnitude of the standard entropy of approximation is determined by the difference between two other standard entropy changes, the standard entropy of molecularity and the standard entropy of rotational restraint.^{128,130} The formation of a unimolecular product during an intermolecular reaction requires that two or more independent molecules become one molecule, and this involves a considerable decrease in standard entropy. The standard entropy change responsible for this decrease, the standard entropy of molecularity, ΔS°_{molec} , has a negative value and is a major, unavoidable, unfavorable term in the change in standard free energy in any intermolecular reaction. In an intramolecular reaction, however, the decrease in standard entropy due to the standard entropy of molecularity does not occur, because reactants are already on only one molecule, and this has the effect of increasing dramatically the change in standard entropy for the intramolecular reaction relative to the intermolecular reaction and hence increasing its yield of product. There is affiliated with an intramolecular reaction, however, a standard entropy of rotational restraint, which, conversely, is irrelevant to an intermolecular reaction. The **standard entropy of rotational restraint** is the increase in standard entropy that results from the fact that the formation of the transition state or product during an intramolecular reaction requires that a portion of the rotational entropy in the molecule be

eliminated because only a fraction of the accessible rotational isomers can participate in the reaction productively. The standard entropy change accompanying this decrease in the number of rotational isomers, the standard entropy of rotational restraint, $\Delta S^\circ_{\text{rot}}$, has a negative value and its inclusion causes the standard entropy change for the reaction to be smaller than it would be if no rotational freedom were lost during the reaction because only a productive rotational isomer was present.

The relationships between the standard entropy of approximation and the standard entropy of molecularity and standard entropy of rotational restraint are

$$\Delta S^\circ_{\text{approx}} = \Delta S^\circ_{\text{molec}} - \Delta S^\circ_{\text{rot}} \quad (5-58)$$

The magnitudes of each of these terms can be discussed in turn.

The **standard entropy of molecularity** is the decrease in standard entropy that should accompany the change of an intermolecular reaction to a rigidly oriented intramolecular reaction.¹²⁸ In the specific case of a bimolecular reaction, the two independent reactants have six translational and six rotational degrees of freedom, but the one molecule, formed by the association of the two others, should have only three translational and three rotational degrees of freedom. The standard entropy change associated with the loss of the three translational and three rotational degrees of freedom during this inescapable association, calculated for the situation in which the two reactants and the transition state or product are dissolved in a solution at 25 °C with a standard state of 1 M in solutes, has been estimated¹²⁸ to be between -190 and $-210 \text{ J K}^{-1} \text{ mol}^{-1}$. This estimate can be compared to the standard entropy change observed for a simple bimolecular reaction, such as the dimerization of cyclopentadiene in the liquid phase, during which the standard entropy change is -130 to $-170 \text{ J K}^{-1} \text{ mol}^{-1}$ with the same choice of standard state. The difference between the calculated standard entropy change and the observed standard entropy change in the particular instance of cyclopentadiene can be completely accounted for by the presence of low-frequency vibrations in the dimer that could not be present in the two monomers because of their smaller size. It has been concluded¹²⁸ that -190 to $-210 \text{ J K}^{-1} \text{ mol}^{-1}$ is an adequate estimate for $\Delta S^\circ_{\text{molec}}$, the maximum decrease in standard entropy change expected from converting a bimolecular reaction into a unimolecular reaction, when molarities are used as units of concentration for standard states. If corrected volume fractions are used as units of concentration for standard states and the partial molar volumes of the solutes are in the range between 40 and 150 mL mol⁻¹, the range for expected standard entropy of approximation for a bimolecular reaction would be -165 to $-195 \text{ J K}^{-1} \text{ mol}^{-1}$ (Equation 5-13).

As the example of the dimerization of cyclopentadi-

ene illustrates, an intramolecular reaction, because it usually involves a larger and more flexible molecule than any of the reactants in an intermolecular reaction, can never realize all of this favorable standard entropy of molecularity. Major factors in decreasing the portion of the standard entropy of molecularity that an intramolecular reaction will enjoy are the internal rotations within the intramolecular reactant itself. These rotations decrease the probability that the necessary juxtaposition of reactants will occur. For example, in the case of phenyl succinate (Equation 5-53) the dihedral angles around three carbon-carbon single bonds must be appropriate if the carboxyl oxygen is to be placed adjacent to the acyl carbon. It has been estimated^{128,130} from the results of thermodynamic and kinetic measurements from a number of intramolecular reactions that the standard entropy of rotational restraint decreases by about $20 \text{ J K}^{-1} \text{ mol}^{-1}$ for every bond that lies between the two atoms participating directly in the reaction and about which free rotation can occur.

When two similar intramolecular associations are compared, for which it is assumed that differences between their standard enthalpies of formation are negligible¹²⁶

$$\Delta \Delta S^\circ = R \ln \left(\frac{K_{\text{eq1}}}{K_{\text{eq2}}} \right) \quad (5-59)$$

where $\Delta \Delta S^\circ$ is the difference between their standard entropy changes and K_{eq1} and K_{eq2} are the respective association equilibrium constants. The change in relative rate of phenoxide release in going from phenyl glutarate to phenyl succinate (Table 5-5) is 230-fold and in going from phenyl succinate to the phenyl ester of the fused ring is also 230-fold. In each comparison, one less carbon-carbon bond around which free rotation is allowed is found in the more constrained member of the pair. The changes in rate, presumably reflecting differences in $K_{\text{eq,TP}}$, are equivalent in each comparison to $45 \text{ J K}^{-1} \text{ mol}^{-1}$. It has been noted, however, that the case of nucleophilic catalysis of the hydrolysis of phenyl esters provides the largest standard entropy of rotational restraint (carbon-carbon bond)⁻¹ yet observed.¹²⁸ It has been proposed that any increase in the apparent entropy of approximation in excess of 20 J K^{-1} that is accomplished by freezing the rotation around a carbon-carbon bond may be due to a **decrease in the strain** encountered by the reaction over and above the decrease in the rotational degrees of freedom.¹³⁰ This decrease in strain would be effected by an unintended improvement in the orientation and alignment of the two reactants in the productive conformation produced by the changes in the structure of the molecule that were required to freeze the rotation.

With small molecules, the effect of approximation on an equilibrium constant is usually significant only when the two central atoms that participate in the asso-

ciation become involved in a **five-membered or six-membered ring** in the product. A four-membered ring is usually too strained, because of the normal bond angles of commonly encountered molecules, to provide any favorable approximation. A seven-membered ring, if there is free rotation about every bond, has too small a value of $\Delta S^\circ_{\text{rot}}$ to exhibit a $\Delta S^\circ_{\text{approx}}$ small enough to overcome the strain of the ring and have a noticeable effect on equilibrium. For example, even in the intramolecular nucleophilic catalysis of phenyl ester hydrolysis, a series of reactions unusually prone to intramolecular catalysis, phenyl adipate would show a rate of phenolate release due to nucleophilic catalysis only 4-fold greater than that for the same reaction of phenyl acetate in 1.0 M sodium acetate. Large, rigid molecules in which the two atoms that must react are more than six atoms apart yet close enough to collide have been synthesized,^{131,132} but proteins and nucleic acids are the ultimate examples.

The magnitude of the actual difference in the change in standard entropy between a given intramolecular reaction and the corresponding intermolecular reaction will be less than the magnitude of the standard entropy of approximation because vibrational degrees of freedom, unavailable to the reactants in the intermolecular reaction, are available to the necessarily larger reactant in the intramolecular reaction and because steric effects that do not apply to the intermolecular reaction are often unavoidable consequences of designing the intramolecular reactant. For example, the intramolecular rates of lactonization for a series of bicyclic γ -hydroxy-carboxylic acids decrease as the strain energies of the rigid five-membered rings of the tetrahedral intermediate (see Equation 5-54) increase,¹³⁰ even though in each case the hydroxy group and the acyl carbon are positioned rigidly in the same orientation and at about the same distance from each other.

If the magnitude of $\Delta\Delta S^\circ$, the actual difference between the standard entropy changes in the reactions, must be less than the magnitude of $\Delta S^\circ_{\text{approx}}$, then

$$T\Delta S^\circ_{\text{approx}} < -RT \ln \left(\frac{K_{\text{eq,intra}}}{K_{\text{eq,inter}}} \right) + \Delta\Delta H^\circ \quad (5-60)$$

where $K_{\text{eq,intra}}$ and $K_{\text{eq,inter}}$ are the intramolecular and intermolecular association equilibrium constants. If the differences in the actual standard enthalpies of formation, $\Delta\Delta H^\circ$, are known, the estimates of the upper limits for $\Delta S^\circ_{\text{approx}}$ can incorporate them. If they are unknown, they can be assumed to be zero, for the sake of argument, but such an assumption can be misleading.

In relating the change observed in an equilibrium constant to the entropy of approximation, the **difference in standard enthalpy change** resulting from the chemical strategy used to accomplish the approximation, $\Delta\Delta H^\circ$, complicates the interpretation (Equation 5-60). In several reactions displaying large increases in the rate of

the reaction or the yield of the product due to covalent approximation of the reactants, the major effect of the approximation is on the standard enthalpy change of the reaction rather than the standard entropy change.²⁹ For example,¹³³ 2,2,3,3-tetramethylsuccinamide at pH 5 displays a rate of aniline release 1200 times greater than that of succinamide itself. This increase in rate, however, which is equivalent to a change in the standard free energy of activation of -18 kJ mol^{-1} , is accompanied by a change in the standard enthalpy of activation, $\Delta\Delta H^\ddagger$, of -25 kJ mol^{-1} . Therefore, in this case the standard entropy of activation actually decreases as the rate of the reaction is enhanced by approximation. It is difficult, however, to interpret such observed changes in the thermodynamic parameters of activation because they are usually dominated by solvent effects that mask the underlying effects of approximation on the rates or equilibrium constants.¹³⁰ In the case of the intramolecular catalysis manifested in the alkaline hydrolyses of *endo*-6-hydroxybicyclo[2.2.1]heptane-*endo*-2-carboxamides (Equation 5-54), it has been concluded that the rate enhancement of 1×10^6 ($T\Delta S^\circ_{\text{approx}} \leq -34 \text{ kJ mol}^{-1}$) "results almost entirely from the entropy effect", probably because there is no strain involved in the formation of the additional five-membered ring of the tetrahedral intermediate.¹²⁹

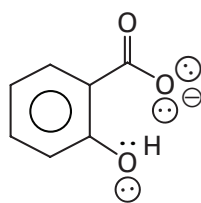
When the upper limits of $T\Delta S^\circ_{\text{approx}}$ are calculated from the relative rates between the intramolecular and bimolecular nucleophilic catalysis of the hydrolysis of phenyl esters, on the assumption that $\Delta\Delta H^\circ$ is equal to 0,¹²⁵ they are all equal to or greater than -38 kJ mol^{-1} (Table 5-5). The upper limit of $T\Delta S^\circ_{\text{approx}}$ calculated from the increase in the equilibrium constant for the formation of the tetrahedral intermediates in the hydrolyses of *endo*-6-hydroxybicyclo[2.2.1]heptane-*endo*-2-carboxamides is -34 kJ mol^{-1} . If it is assumed that the fused rings retain two rotational axes, which is a generous assumption, $\Delta S^\circ_{\text{approx}}$ (Equation 5-58) should be about $-140 \text{ J K}^{-1} \text{ mol}^{-1}$ and $T\Delta S^\circ_{\text{approx}}$ about -40 kJ mol^{-1} when units of corrected volume fraction are used. The upper limit of $T\Delta S^\circ_{\text{approx}}$ calculated from the increase in the equilibrium constant for the formation of tetrahedral intermediates in the hydrolyses of *endo*-6-hydroxybicyclo[2.2.1]heptane-*endo*-2-carboxamides is -34 kJ mol^{-1} .

At least three points are illustrated by this exercise. First, the largest intramolecular increases in rate or degree of association yet measured, with the educational exceptions of the cases involving severe compressive steric effects in the transition states or products, are of a magnitude less than that expected simply for the transformation of an intermolecular reaction into a fully constrained intramolecular reaction. Second, the **maximum decrease in standard free energy** of association to be expected when a bimolecular association such as the formation of a hydrogen bond is turned into an intramolecular association is about -55 kJ mol^{-1} , which would produce an increase in its association equilibrium constant, when the units are corrected volume fraction, of 5×10^9 . Third, the larger the

number of bonds about which rotation can occur between the two atoms—for example the heteroatom of the donor and the acceptor—that must be juxtaposed during a reaction, the smaller will be the decrease in standard free energy to be expected when an intermolecular association becomes an intramolecular association.

With these points in mind, the issue of **intramolecular hydrogen bonds** in aqueous solution can be addressed. The standard enthalpy change for a hydrogen bond forming in water should be quite small, but possibly of a negative value (Equation 5–49). The competition of water molecules for donor and acceptor seems to contribute an entropic effect the magnitude of which is $-38 \text{ J K}^{-1} \text{ mol}^{-1}$, which is $R \ln 100$, when concentrations of donor and acceptor are expressed in units of molarity. In an intramolecular association, the consequent elimination of the standard entropy of approximation should be able to compensate for the entropic deficit caused by the presence of the water.

Evidence for the existence of intramolecular hydrogen bonds within solutes dissolved in aqueous solution has been reported. The extensive lore surrounding involvement of hydrogen bonds in the equilibrium acid–base behavior of the monoanions of **dicarboxylic acids** is by and large equivocal,²⁹ but it has been noted⁴⁵ that decreases in the rates of the reactions of the acidic hydrogens in the monoanions of salicylates (**5–17**) with hydroxide ion



5–17

suggest that they contain intramolecular hydrogen bonds the standard free energies of formation of which are around -15 kJ mol^{-1} . A series of compounds capable of forming intramolecular hydrogen bonds either between the pyrrole nitrogen–hydrogen bond on imidazole as donor ($pK_a = 15$) and a carboxylate as acceptor ($pK_a = 5$) or between the pyridinyl lone pair on imidazole ($pK_a = 7.5$) as acceptor and the nitrogen–hydrogen bond on an ammonium cation as donor ($pK_a = 10$) has been described (Table 5–6). As the standard entropy of approximation was decreased by confining the juxtaposed donor and acceptor more severely, or as the difference in pK_a was decreased, the equilibrium constant for the intramolecular hydrogen bond

$$K_{\text{AHB}} = \frac{[\text{B}\oplus\text{HA}]}{[\text{B}\oplus + \text{HA}]} \quad (5-61)$$

increased in magnitude (Table 5–6).

Table 5–6: Intramolecular Hydrogen Bonds in Water

hydrogen bond	ΔpK_a^a	$K_{\text{intra,AHB}}^b$
	3	2^c
	10	$<1^c$
	10	1.5^c
	13	13^c
	3	100^d

^aDifference in pK_a between intermolecular equivalents of donor and acceptor.

^bValue for $K_{\text{intra,AHB}}$ is for the ratio of the concentration of the hydrogen-bonded species (see column 1) to the concentration of the same tautomer not hydrogen-bonded. The equilibrium constants $K_{\text{intra,AHB}}$ for the formation of the noted intramolecular hydrogen bonds in aqueous solution were estimated from values of the N1–N3 tautomeric equilibrium constants (Equation 2–31) for the respective 4-substituted imidazoles determined by ^{15}N nuclear magnetic resonance.¹³⁴ It was assumed that a difference between the value of the tautomeric equilibrium constant for a 4-substituted imidazole in which a hydrogen bond can form and the value of the tautomeric equilibrium constant for a similar compound in which a hydrogen bond cannot form is due to the formation of the noted hydrogen bond. It was also assumed that the value of the tautomeric equilibrium constant for the form of the hydrogen-bonding species in which the bond is not formed is equal to that of the reference compound and that the observed excess of one of the two tautomers represents entirely the hydrogen-bonded form. ^cEstimated from the difference between the second macroscopic acid dissociation constants of *cis*- and *trans*-urocanic acid and the N1–N3 tautomeric ratios for neutral *cis*-urocanic acid.¹⁰⁹

The question that these results raise is whether or not the hydrogen bond in a conformation such as an α helix, a hairpin of β structure, or a β turn can be made favorable by sufficient standard entropy of approximation. Certain cyclic hexapeptides are rigid enough to enforce the conformation of a β turn in which an intramolecular hydrogen bond is formed between the acyl oxygen of one of the six amino acids and the amido nitrogen–hydrogen of the amino acid three positions to the amino-terminal side of it (Figure 4–16D),¹³⁵ and the conformation of the β turn can be varied by changing the sequence of the cyclic peptide. In fact, the first tight turn

to be observed was in such a cyclic hexapeptide.^{136,137} An α helix or β structure, however, cannot be encompassed so easily.

If it is assumed that either an α helix or a hairpin of β structure has already been initiated, could the standard free energy of formation for the next hydrogen bond (Figure 5-19) during the propagation have a negative value? In Figure 5-19, the next donor and acceptor in each structure are marked with asterisks. Because each of these reactions occurs in aqueous solution, the standard enthalpy of formation for such a hydrogen bond is close to 0 (Equation 5-49). The value for the standard free energy of formation for the hydrogen bond will be determined in part by the difference between the unfavorable competition of the water and the favorable elimination of entropy of approximation that the structures provide. For the α helix, there are two bonds about which rotation

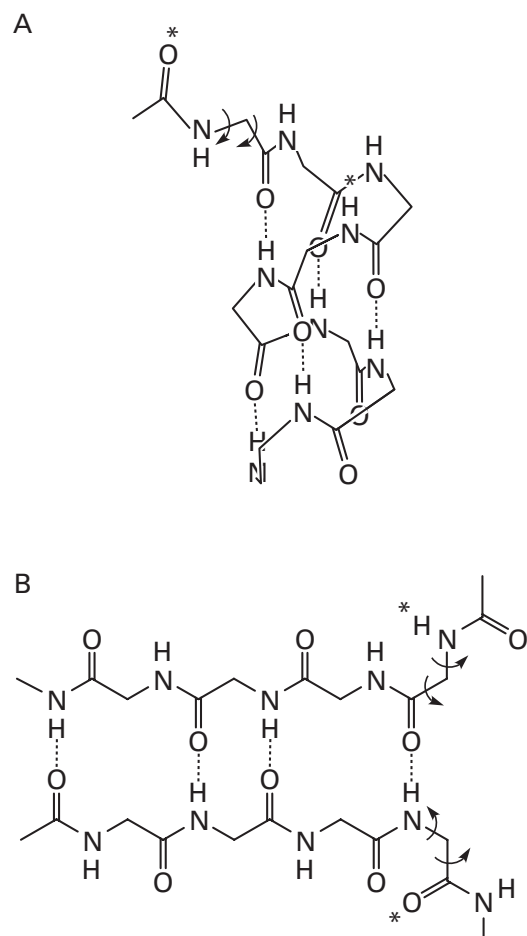


Figure 5-19: Intramolecular formation of a hydrogen bond to elongate an α helix or a β hairpin. (A) To add the next hydrogen bond in an elongating α helix, the acyl oxygen marked with the asterisk must combine with the amide nitrogen-hydrogen marked with the other asterisk. The two bonds about which rotation can occur between the last acyl group fixed in the α helix and the acyl oxygen in question have been highlighted with arrows. (B) To add the next hydrogen bond in an elongating antiparallel β hairpin, the acyl oxygen marked with the asterisk must combine with the amide nitrogen-hydrogen marked with the other asterisk.

can occur between donor and acceptor;¹³⁸ for the hairpin of β structure, there are four. It should, however, be remembered that when the problem is stated in these terms, the difficulties involved in the initiation of either of these structures are ignored, and these can be even more formidable.¹³⁸

Recently, several short peptides (8–16 aa in length) have been shown to form antiparallel β structure in aqueous solution by folding back on themselves to form a hairpin.^{139,140} These hairpins of antiparallel β structure have marginal stability at room temperature and the pairing across the sheet is as yet unpredictable. Short stretches of parallel β structure (encompassing three or four pairs of amino acids) have been observed in aqueous solution in molecules in which two short peptides are coupled at their carboxy-terminal ends to two properly spaced positions in a somewhat rigid molecular template that assists in properly orienting them.¹⁴¹ These results suggest that approximation and the cooperative formation of the hydrogen bonds between the two strands in these parallel and antiparallel β structures can overcome the competition of the water for donors and acceptors, but only barely.

All short, linear peptides examined so far fail to form α helices when they are dissolved at room temperature in water at neutral pH, even if they have the same amino acid sequence as an α helix in a crystallographic molecular model.¹⁴² This is almost certainly due to the difficulty of forming the first few hydrogen bonds required to initiate the α helix rather than the formation of the hydrogen bonds that fall in line after it has been initiated (as in Figure 5-19). When short peptides are attached to rigid structural templates that provide properly oriented acceptors for hydrogen bonds and thereby promote initiation, those peptides display a considerable fraction of α helix even at room temperature.¹⁴³

It was noted that when a cyanogen bromide fragment containing the first 12 amino acids of ribonuclease, KETAAAKFERQHHse (where Hse is homoserine), was dissolved in 33 mM Na_2SO_4 between pH 4 and 5, an equilibrium existed between the structureless form of the peptide and an α -helical form. At 0 °C, 15% of the peptide was in the α -helical form at equilibrium.¹⁴² From this original observation, it was eventually^{144–146} discovered that when the peptide acetyl-AAQAAAAQAAAAQAAY- α -amide is dissolved at 0 °C in 1.0 M NaCl, about 50% of it is α -helical and 50% is structureless at equilibrium.¹⁴⁷ This peptide, however, even with as peculiar a sequence as it has, exhibits significant α -helical content only at low temperature. No other simple peptide examined so far displays a significantly higher amount of α helix at equilibrium.^{144,148,149} Considerably higher α -helical content, however, is displayed even at room temperature by continuous segments of polyalanine (4–19 alanines in length) when they are appropriately isolated from the charged amino acids at the two ends of these hybrid molecules that are required to dissolve the polyalanine in

water.¹⁵⁰ Polyalanine, however, is considerably different from the variable and almost unbiased amino acid sequences of the α helices in proteins. All of the peptides displaying α -helical structure, because of the marginal stability of those α helices and the peculiar sequences required, reemphasize the difficulty of overcoming the competition of the molecules of water for donors and acceptors.

The existence of marginally stable synthetic α helices, albeit at 0 °C, has provided a biochemically relevant framework on which to examine the advantages provided by such a structure and its resulting standard entropy of approximation to the formation of intramolecular hydrogen bonds. Because the side chains protrude from an α helix at intervals of 99°, ^{151–153} the side chain four amino acids to the amino terminus of any position in an α helix lies almost directly (396°) below the side chain at that position (Figure 4–17). A hydrogen bond will form between a donor and an acceptor placed synthetically at the i and $i + 4$ positions of an α -helical peptide.¹⁴⁴ For example, in the peptide acetyl-AAQAAEAQAKAAQAAY- α -amide, a hydrogen bond can form between the glutamate and the lysine when the two are held rigidly above and below each other in the α -helical conformation of the peptide. The free energy of formation of this hydrogen bond can be assessed by measuring the differences between the equilibrium constant at 0 °C for the formation of the α helix of this peptide and those for the α helices of various controls in which the hydrogen bond is unable to form.

A series of such measurements have been made (Table 5–7). None of these free energies of formation is remarkable, again presumably because the standard entropy of approximation in such a situation barely overcomes the competition for donors and acceptors from the water. An important point that should be reiterated is that these small negative free energies of formation do

not state that the hydrogen bond is a net contributor to the stability of the α helix; in fact, each of the α helices that contains hydrogen bonds is less stable than an α helix in which the donor and acceptor are replaced by alanine.¹⁴⁷ Rather, it is the formation of the α helix itself that provides the standard entropy of approximation necessary to make the standard free energy of formation of each of these intramolecular hydrogen bonds less than 0.

The difference in standard free energy of formation between the hydrogen bond of a lysinium cation and a glutamate anion and the hydrogen bond of a lysinium cation and a glutamic acid is -0.3 kJ mol^{-1} ; that between the hydrogen bond of a lysinium cation and an aspartate anion and the hydrogen bond of a lysinium cation and an aspartic acid is -0.2 kJ mol^{-1} ; that between the hydrogen bond of a histidinium cation and a glutamate anion and the hydrogen bond of a histidinium cation and a glutamic acid is -0.5 kJ mol^{-1} ; and that between the hydrogen bond of a histidinium cation and an aspartate anion and the hydrogen bond of a histidinium cation and an aspartic acid is -0.7 kJ mol^{-1} (Table 5–7). Because in each case the carboxylate is more basic than the carboxylic acid, and hence a better acceptor, Equation 5–51 predicts that these differences should be -3.5 , -3.5 , -6.5 and -6.5 kJ mol^{-1} , respectively. The fact that the actual differences are significantly less negative than the expected differences is further evidence for the fact that an **ion pair** is unstable in aqueous solution relative to the separated ions because the conversion of the monocationic hydrogen bond into an ion pair actually destabilizes. This conclusion is reinforced by the fact that the difference in standard free energy of formation between the hydrogen bond of a glutamine and aspartic acid and the hydrogen bond of a glutamine and aspartate anion, neither of which is an ion pair, is -2.3 kJ mol^{-1} , even though glutamine is a much weaker acid than either lysinium cation or histidinium cation (Table 2–2).

One could imagine that the standard **base pairs** between adenine and uracil or between guanine and cytosine might form in water because the formation of the second hydrogen bond or the second and third hydrogen bonds in the respective complexes would be aided by standard entropy of approximation gained by the formation of the first. Such a cooperative enhancement of hydrogen bond strength has been observed in complexes between glutaric acid and a cyclic tetraresorcinol in CHCl_3 . In these complexes the two carboxylic groups of the diacid form hydrogen bonds with the phenolic hydroxyls of resorcinols on opposite sides of the ring.¹⁵⁶ The advantage, however, to the formation of the second hydrogen bond, a cooperative but intermolecular situation, was only -10 kJ mol^{-1} , and the hydrogen-bonded complex could be observed only in aprotic solvents such as CHCl_3 . In a similar fashion, standard and nonstandard base pairs between two nucleic acid bases will form in organic solvents or within micelles, but they do not form

Table 5–7: Standard Free Energies of Formation of Hydrogen Bonds within an α Helix^a

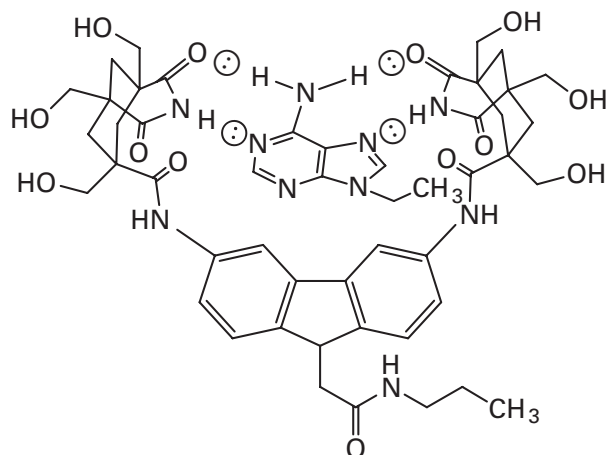
donor/acceptor	standard free energy of formation (kJ mol^{-1})
lysinium cation/glutamic acid ^{147,149}	-1.1
lysinium cation ion/glutamate anion ^{147,149}	-1.4
lysinium cation/aspartic acid ¹⁴⁹	-0.9
lysinium cation/aspartate anion ¹⁴⁹	-1.1
histidinium cation/glutamic acid ¹⁴⁹	-0.6
histidinium cation/glutamate anion ¹⁴⁹	-1.1
histidinium cation/aspartic acid ¹⁵⁴	-2.4
histidinium cation/aspartate anion ¹⁵⁴	-3.1
glutamine/aspartic acid ¹⁵⁵	-1.8
glutamine/aspartate anion ¹⁵⁵	-4.1

^aStandard free energy of formation for a hydrogen bond between the i and $i + 4$ positions of an α -helical peptide.

in water.¹⁵⁷ In fact, the normal bases in DNA itself can be replaced by isosteres that cannot form any hydrogen bonds, and the complementary base pairs form as efficiently from steric complementarity as from base pairing.^{158–161}

The **hydrogen bonds in DNA** contribute to the specificity of the pairing of its bases, but in a negative sense. For example, in order to compensate partially for the strongly unfavorable act of removing the two donors and the one acceptor of guanine and the two acceptors and one donor of cytosine from water, the three hydrogen bonds of the base pair must be formed in a double helix. When only the thymine in the base pair with adenine is replaced by a 2,4-difluoro-5-methylphenyl group so that the two hydrogen bonds cannot form, the base pair is less stable¹⁶² by about 7 kJ mol^{-1} .

There are, however, situations in which hydrogen bonds to nucleic acid bases can be sufficiently assisted by standard entropy of approximation so as to be marginally stable in aqueous solution. The complex between 9-ethyladenine and a specifically designed host



5-18

has an association constant of 28 M^{-1} at 27°C in water at pH 6 and an ionic strength of 0.05 M , but it was estimated that each hydrogen bond in the complex contributed only -0.8 kJ mol^{-1} to its stability even though the standard entropy of approximation must be significant.¹⁶³ It is also possible to form in aqueous solution a hydrogen-bonded dimer of the deoxydinucleotide 5'-phosphodeoxyguanylyl(3'-5')deoxycytidine (pdG-dC). A hydrogen-bonded dimer of this self-complementary dinucleotide forms with an association equilibrium constant of 8 M^{-1} at 2°C and pH 7.5.¹⁶⁴ If it has the common base pairing, six hydrogen bonds hold the two halves together. The rings of the bases are rigid, and once two of the hydrogen bonds in each base pair form, the standard free energy of formation of the third must incorporate the undiluted standard entropy of molecularity. Furthermore, the nucleotide bases on the two monomers are probably already stacked one on top of the other,¹⁶⁵ an association that at least brings all six donors and

acceptors to the same side of the monomer if not in proper alignment. This stacking of the bases results from the hydrophobic effect. It must contribute significantly to this association by roughly aligning the bases before the dimerization occurs. All of these observations demonstrate that the hydrogen bonds between the base pairs of a double-stranded nucleic acid do not contribute significantly to its stability.

Suggested Reading

Page, M.I., & Jencks, W. (1971) Entropic contributions to rate accelerations in enzymic and intramolecular reactions and the chelated effect, *Proc. Natl. Acad. Sci. U.S.A.* 68, 1678–1683.

Problem 5-10:

- Draw the structure of a hydrogen bond between the propionate anion and a proton on N3 of neutral 4-methylimidazole (see Equation 2-31 for numbering). Include all lone pairs of electrons in your drawing.
- Use Equation 5-51 and a value of 0.013 for τ to estimate the apparent equilibrium constant in units of molarity⁻¹ for the formation in water of a hydrogen bond between one of the lone pairs on a propionate anion ($\text{p}K_a = 4.88$) and a proton on N3 of 4-methylimidazole (microscopic $\text{p}K_a = 15.1$).
- The equilibrium constant for the formation of the intramolecular hydrogen bond in the 3-(3*H*-imidazol-4-yl)propionate anion in which carbons 2 and 3 have been locked by a cyclopentene is 1.5 (Table 5-6). Using the estimated value of the apparent equilibrium constant for the intermolecular formation of the same type of hydrogen bond, from section B, calculate an upper limit for the value of $\Delta S^\circ_{\text{approx}}$ for the conversion of the intermolecular reaction into the intramolecular reaction. Convert the apparent equilibrium constant for the intermolecular formation of the hydrogen bond into units of corrected volume fraction (Equation 5-13) before performing this calculation. Assume that $\Delta\Delta H^\circ$ is 0.

Problem 5-11: Why might it be that the peptide with the amino acid sequence SEEEEKKKKKEEEKKKKF displays 35% α helix at pH 8.3 and 4°C ?¹⁶⁶

The Hydrophobic Effect

The hydrophobic effect is exemplified by the fact “that oil and water are hostile”¹⁶⁷ and do not mix. The reason is that water is more stable when the oil is not dissolved in it than when the oil is. This failure of oil and water to mix is only the most extreme manifestation of the tendency of liquid water to expel solutes that are not ions and that

do not have significant numbers of donors and acceptors of hydrogen bonds. An ionic solute is held in water by large, negative standard enthalpies of hydration. Solutes that have donors and acceptors of hydrogen bonds are held in water by the hydrogen bonds they form with it. Solutes that neither are ions nor have donors and acceptors of hydrogen bonds are expelled from liquid water. This expulsion is the **hydrophobic effect**.

The nature of the hydrophobic effect has been succinctly described by G.S. Hartley.¹⁶⁸

The antipathy of the paraffin chain for water is, however, frequently misunderstood. There is no question of actual repulsion between individual water molecules and paraffin chains, nor is there any very strong attraction of paraffin chains for one another. There is, however, a very strong attraction of water molecules for one another in comparison with which the paraffin-paraffin or paraffin-water attractions are very slight.

Aside from the overuse of “very”, it is clear from this description that the term hydrophobic is misleading, if its etymology is examined closely.¹⁶⁹ The oil does not dislike the water. In fact, measurements of interfacial energies suggest that the oil prefers the water to itself.¹⁷⁰ Rather, **water** ejects the oil because water molecules have a greater like for other water molecules.

The hydrophobic effect upon a hydrophobic solute A can be represented formally by the transfer of the solute from water to another solvent *j*.^{171,172}



It has been proposed¹⁶⁹ that the hydrophobic effect can also be represented by the formation of a macroscopic interface between an immiscible phase of hydrocarbon and water. There are, however, significant differences between the physical properties of such a macroscopic interface and those of the microscopic layer of hydration surrounding an isolated molecule of hydrocarbon dissolved in water.¹⁷³ Because, upon the folding of a molecule of protein, individual side chains of the amino acids are transferred from the water to the interior of the folded structure, the transfer of a molecule of solute from water to another phase (Equation 5-62) seems to be the more appropriate process to examine for the present purposes.

In the case of the failure of oil and water to mix, solute A in Equation 5-62 is a molecule of oil and solvent *j* is the liquid oil itself. Because, in general, pure phases of the different solutes A in a comparison may in themselves have unique peculiarities, the transfer of a solute from water to a nonpolar solvent should be studied in a systematic fashion by choosing a common solvent for all of the transfers.^{174,175} The hydrophobic effect can be quantified^{24,176} by measuring a standard free energy for this transfer (Equation 5-18). The **standard free energy of transfer** of solute A between water and solvent *j*, $\Delta G^\circ_{A, \text{H}_2\text{O} \rightarrow j}$, is the change in standard free energy that results only from the

change in solvation of solute A between the other solvent and the water experienced when solute A, dissolved in water at standard state, is transferred from the water into that other solvent at standard state.

As expected from everyday experience, the standard free energies of transfer of hydrophobic solutes between water and an organic solvent such as benzene, carbon tetrachloride, or the liquid solute itself are negative (Table 5-8). It is this negative change in standard free energy that produces the hydrophobic effect. The hydrophobic effect is the only noncovalent force in aqueous solution that proceeds with a net negative change in standard free energy, and it is thought to provide all of the driving force for the folding of polypeptides, the association of ligands with a protein, and the formation of interfaces between subunits in oligomeric proteins.

The explicit reason for the negative standard free energies of transfer at physiological temperatures is that the **standard entropies of transfer** are larger than the standard enthalpies of transfer (Table 5-8). At 25 °C, the standard enthalpy of transfer, which in most cases is positive and thus unfavorable, is overcome by a much larger positive and thus favorable standard entropy of transfer. This peculiarity has led to the maxim that the hydrophobic effect is entropy-driven, but this is a misleading view. Edsall and Scatchard¹⁸⁰ noted that the incremental standard entropies of solution for $-\text{CH}_2-$ groups in water had anomalously large negative values, but they also pointed out that because of the large changes in standard molal heat capacity, these incremental standard entropies of solution would become less and less significant as the temperature was raised (Figure 5-20).¹⁸¹ As the temperature increases, the standard enthalpy of transfer becomes more and more exothermic. At high enough temperatures the standard entropy of transfer passes through zero and becomes endergonic. As a result, at intermediate temperatures the reaction changes from an entropically driven process to an enthalpically driven process, and at high temperatures the standard entropy of transfer is actually unfavorable even though the transfer itself remains favorable because the standard free energy of transfer does not vary significantly with temperature.

This behavior illustrates the fact, first noted by Edsall, that the most characteristic feature of the hydrophobic effect is not its change in standard entropy but its **change in standard heat capacity**.¹⁸² The anomalously large, positive incremental change in standard molal heat capacity of solution for solutes in water remains the most reliable signature of the hydrophobic effect.¹⁸¹

The changes in the standard thermodynamic state functions, such as standard entropy, standard enthalpy, and standard heat capacity, that are associated with the hydrophobic effect (Table 5-8) have been assigned to **changes in the thermodynamic properties of the water** surrounding the solute as it leaves the aqueous phase and changes in the thermodynamic properties of the nonpolar solvent as it enters; in other words, to differences in

Table 5–8: Thermodynamic Properties of Transfer from Water to Another Solvent^a

solute	solvent <i>i</i>	$\Delta G^{\circ}_{\text{H}_2\text{O} \rightarrow i}$ (kJ mol ⁻¹)	$\Delta H^{\circ}_{\text{H}_2\text{O} \rightarrow i}$ (kJ mol ⁻¹)	$T\Delta S^{\circ}_{\text{H}_2\text{O} \rightarrow i}$ (kJ mol ⁻¹)	$\Delta C^{\circ}_p, \text{H}_2\text{O} \rightarrow i$ (J K ⁻¹ mol ⁻¹)
methane ^{7,171}	benzene	-11.0	+11.7	+22.7	
methane ^{7,171}	CCl ₄	-12.1	+10.5	+22.6	
ethane ^{7,171}	benzene	-17.6	+9.2	+26.8	-250
ethane ^{7,171}	CCl ₄	-17.1	+7.1	+24.2	
ethane ¹⁷⁷	ethane	-18.2	+10.4	+28.6	-280
propane ^{7,171,178}	propane	-24.5	+7.5	+32.0	-290
butane ^{7,171,178}	butane	-29.4	+4.2	+33.6	-290
benzene ^{7,177,178}	benzene	-24.2	-2.4	+21.8	-430
toluene ^{7,177,178}	toluene	-29.0	-2.7	+26.3	-450
ethanol ^{7,178,179}	ethanol	-5.3	+10.1	+15.4	-150
1-propanol ^{7,178,179}	1-propanol	-10.3	+10.1	+20.4	-220
2-propanol ^{7,178,179}	2-propanol	-8.7	+13.0	+21.7	-220
1-butanol ^{7,178,179}	1-butanol	-15.3	+9.3	+24.6	-280
1-pentanol ^{177,179}	1-pentanol	-20.7	+7.8	+28.5	-350

^aValues were calculated from the thermodynamic behavior of the partition coefficients (Equation 5–17). Original published values for the partition coefficients were in units of mole fraction at infinite dilution. These units were converted to units of molarity by dividing them by the appropriate molar volumes of the respective solvents. The partition coefficients in units of molarity were then converted to free energies of transfer and entropies of transfer for units of corrected volume fraction (Equation 5–18). All values are for a temperature of 25 °C.

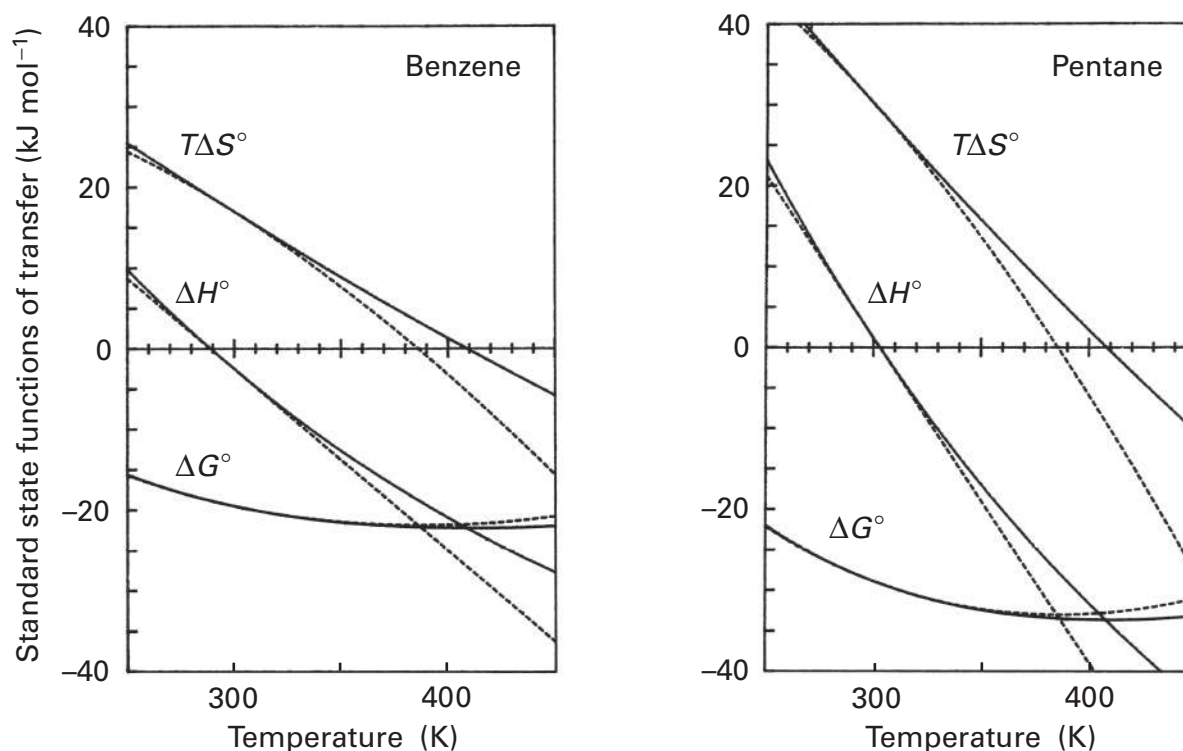


Figure 5–20: Dependence on temperature (Kelvin) of the changes in standard free energy (ΔG°), standard enthalpy (ΔH°), and standard entropy multiplied by temperature ($T\Delta S^{\circ}$) for the transfer of benzene between water and benzene (left panel) and the transfer of pentane between water and pentane (right panel).¹⁸¹ The activities are expressed in units of mole fraction rather than as corrected volume fraction as in Table 5–8; and the three state functions, free energy, enthalpy, and entropy, are expressed in units of kilojoules mole⁻¹. The lines were calculated from the values of the thermodynamic state functions measured in the range from 0 to 40 °C and either the assumption that the observed behavior of ΔC°_p can be fit by an analytic function of temperature and that analytic function can be extrapolated beyond the range of measurement (solid lines) or the assumption that ΔC°_p is independent of temperature and has a value that is the mean of the observed values (dashed lines). The values of ΔC°_p over the range of measurements seem to decrease somewhat with increasing temperature and this deviation is the basis for considering adjusted values of ΔC°_p for changes in temperature. The trends are the same regardless of the assumption. Reprinted with permission from ref 181. Copyright 1988 Academic Press.

the solvation of the solute by the two solvents. The reason that these changes of the thermodynamic state functions are assigned to the respective solvents rather than the solutes is that the solutes in most cases are too small, as in the case of methane, ethane, or propane, or too rigid, as in the case of benzene, to account internally for the significant changes that are observed.

There are several observations which suggest that a more **rigid hydrogen-bonded lattice**, similar to the lattice in ice Ih, surrounds hydrophobic solutes when they are dissolved in water.^{183,184} Macroscopic solids known as clathrates form spontaneously when hydrocarbons are mixed with pure water at proper molar ratios. **Clathrates** are solid, crystalline hydrates that are composed of isolated individual molecules of hydrocarbon encased in rigid hydrogen-bonded networks of water molecules. The thermodynamic parameters associated with these solids are of a magnitude sufficient to lead to the conclusion that similar rigid networks of water molecules should also surround hydrophobic solutes when they are present at dilute concentration.¹⁸⁵ Whether or not clathrates are of relevance to the hydrophobic effect, their very existence has subconsciously influenced our views of the process. The unusually large changes in standard heat capacity (Table 5-8)¹⁸² associated with the hydrophobic effect are believed to result from an increase in the order of the water surrounding the solute.¹⁸⁶ It should be recalled that it is the gradual melting of the hydrogen-bonded lattice that is supposed to be responsible for the anomalously high heat capacity of liquid water itself, and increasing the amount or the degree of structure of this lattice should produce even greater capacity for melting and, hence, greater heat capacities. The **partial molar volume** of a hydrophobic solute in water is usually about 13 cm³ mol⁻¹ less than that of the same solute in other solvents,²³ and this difference has been thought to result from the efficient packing of the solute within a cage of hydrogen-bonded waters that resembles the networks in clathrates. It has already been noted, however, that ice Ih, and presumably liquid water, contain a large amount of vacant space that a solute could occupy (Figure 5-2), and this occupation by itself could explain the smaller values for partial molar volume. The increase in the **dielectric relaxation time**¹⁸⁷ and the anomalously large increase in the **viscosity**²⁹ observed when a hydrophobic solute is added to water and the expansion of the structure of water around hydrophobic solutes detected by **neutron scattering**,¹⁸⁴ however, also indicate that a more rigid and structured shell of water forms around the solute than the water in the bulk solvent.

If it is the case, however, that the water surrounding a hydrophobic solute is more structured and held within a more rigid hydrogen-bonded lattice than water in the bulk phase, there should be **compensatory thermodynamic changes**^{29,186} associated with this increase in structure. Specifically, the standard enthalpy of the aqueous solution of hydrocarbon should be less than the

standard enthalpy of liquid water because the hydrogen bonds in this more structured cage should be stronger. If a noncovalent chemical transformation occurs in any solution, the change in standard entropy observed is usually compensated by a change in standard enthalpy. This observation can be stated mathematically⁷ as

$$\Delta H^\circ(\alpha) = \Delta H' + T_c \Delta S^\circ(\alpha) \quad (5-63)$$

where α refers to any noncovalent process and $\Delta H'$ and T_c are parameters peculiar to that process. Many noncovalent processes occurring in water¹⁸⁸ satisfy this relationship with $T_c = 280 \pm 10$ K.*

In the particular case of the hydrophobic effect, it has been noted²⁹ that the decrease in standard entropy associated with the formation of a rigid cage of hydration as the solute enters water is not accompanied (Table 5-8) by the decrease in standard enthalpy (Equation 5-63) to be expected from such compensation.¹⁷ The argument is that the **missing enthalpy** in this reaction is the enthalpy that was required to crack open the lattice of the liquid water to form a cavity for the hydrophobic solute. Because standard entropy and standard enthalpy should compensate almost completely in the formation of the more rigid shell of hydration and have little effect on the overall reaction because of this cancellation, it is actually this positive enthalpy of opening the lattice to form a cavity or, conversely, the negative enthalpy realized upon collapsing the cavity that produces the hydrophobic effect.

The positive enthalpy required to open the lattice could result from the fact that some of the hydrogen bonds of the fluid lattice within liquid water must be broken irretrievably when a cavity is formed for a hydrophobic solute. The empty donors and acceptors of such **broken hydrogen bonds** can be observed in molecular dynamics simulations of aqueous solutions of hydrophobic solutes.⁵⁶ Water, however, is presumably adept at rearranging around small nonpolar solutes to form cages like those formed in the clathrates and in the process retaining as many hydrogen bonds as there would be in the absence of those solutes, and other simulations indicate that it is only when all of the dimensions of a nonpolar solute are more than twice the diameter of a molecule of water that significant numbers of hydrogen bonds are lost upon the formation of the cavity.¹⁸⁹ Nevertheless, in this view, the magnitude of the expulsion of hydrophobic solutes from aqueous solution should depend on the number of hydrogen bonds that must be broken to form the cavity.

If this picture of the driving force propelling the

* This value for T_c means that values of changes in standard enthalpy or changes in standard entropy for most processes occurring entirely in aqueous solution are monotonously uninformative because they register mainly compensatory changes in the structure of the solvent.

hydrophobic effect were correct, then it would follow that only the **size of the cavity** should determine the magnitude of the hydrophobic effect. It is the case that the larger the solute, the greater the hydrophobic effect (Table 5-8, Figure 5-21),^{23-25,190} but there is a peculiar aspect to this relationship.

The only feature of a molecule that determines the magnitude of the hydrophobic effect exerted upon it is the number of **hydrogen-carbon bonds** that it contains.¹²⁰ This conclusion follows from the following facts.

First, the change in standard heat capacity upon dissolving a molecule in water, which is the fundamental thermodynamic manifestation of the hydrophobic effect,¹⁸² correlates with high precision ($r \geq 0.985$) to the number of hydrogen-carbon bonds that it contains in 15 different sets containing among themselves a total of 120 molecules.¹⁸⁶ Furthermore, the slopes of each of the 15 correlations are all the same [$30 \pm 2 \text{ J K}^{-1} (\text{mol hydrogen-carbon bond})^{-1}$].

Second, it has long been noted^{24,176,191} that the standard free energies of transfer for linear acyclic alkanes (the lines connecting the symbols \circ in Figure 5-21) from water to any solvent (hexadecane in the case of Figure 5-21) correlate with the number of hydrogen-carbon bonds they contain. The high precision of this correlation ($r > 0.9999$ for the data calculated with units of corrected volume fraction, the lower line) is one of the most remarkable facts concerning the hydrophobic effect.¹⁹² This high precision seems to belie explanations of the hydrophobic effect based on the dimensions of the cavity formed by the solute because these dimensions should depend on the particular length and conformational flexibility of the particular alkane.

Third, the standard free energies of transfer from water to hexadecane for branched alkanes, cyclic alkanes, alkenes, alkadienes, cyclic alkenes, cyclic alkadienes, cyclic alkatrienes, arenes, and alkynes¹⁹⁰ all fall close to the line governing the behavior of linear alkanes (lower solid line in Figure 5-21) when they are plotted as a function only of the number of hydrogen-carbon bonds that the molecules contain. The data fall even closer to the line for linear alkanes when molarities are used as units (upper lines in Figure 5-21).^{*} The same

* The scatter in the data for the free energies of transfer plotted in Figure 5-21 when they are calculated from activities with units of corrected volume fraction (lower line) is more pronounced than when they are calculated from activities with units of molarity (upper line). This may be due to the fact that, in the former case, the partial molar volumes of the solutes have a large effect on the final values for free energies of transfer, yet these partial molar volumes are from estimates rather than direct measurements. It does seem, however, that the standard free energies of transfer for most of the other classes deviate systematically from the line correlating the standard free energies of transfer based on corrected volume fraction for linear alkanes. If this is a real effect, not one due only to the fact that the partial molar volumes used are inaccurate, then a phenyl ring is worth about 1.4 hydrogen-carbon bonds during transfer from water to hexadecane.

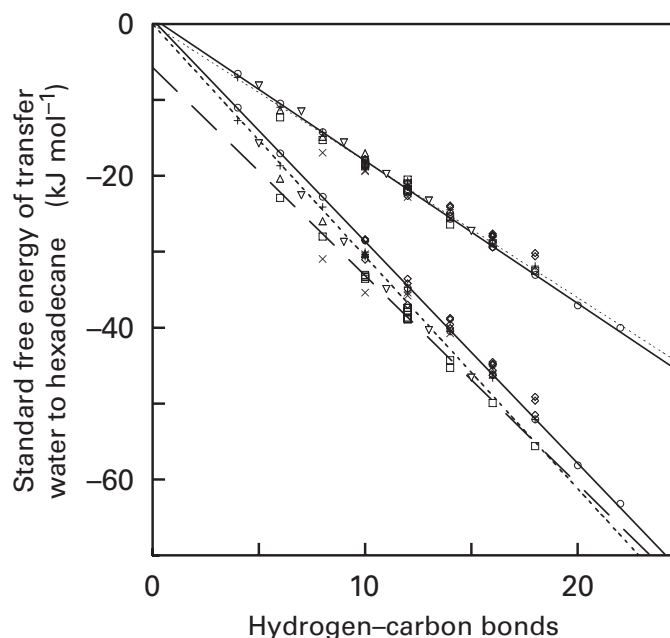


Figure 5-21: Standard free energy of transfer from water to hexadecane as a function of the number of hydrogen-carbon bonds in a hydrocarbon. It was found that when the values for the partition coefficients for transfer from water to hexadecane for the linear alkanes listed by Abraham et al.¹⁹⁰ were treated as the quotients of the molarities of the hydrocarbons in the two phases and inserted into Equation 5-18, standard free energies of transfer resulted that were almost identical to those for the transfer of the same hydrocarbons from water to their own liquid calculated with Equation 5-18 from the solubilities tabulated by McAuliffe.²⁵ Consequently, it was assumed that the units on the dimensionless partition coefficients listed by Abraham et al. were molarity molarity⁻¹. These partition coefficients were either inserted into Equation 5-18 directly to obtain standard free energies of transfer for activities in units of corrected volume fraction (lower set of data) or inserted into Equation 5-18 modified so that it did not include the exponential term within the argument of the natural logarithm to obtain standard free energies of transfer for activities in units of molarity (upper set of data). The partial molar volumes for hydrocarbons in hexadecane were calculated with Equation 5-9, and those in water with the algorithms of Traube.^{23,24} The sets of hydrocarbons included (in descending order on the graph for each number of hydrogen-carbon bonds) branched acyclic alkanes (\diamond ; $n = 18$), linear acyclic alkanes (\circ ; $n = 10$), cyclic alkanes (\times ; $n = 6$), acyclic monoenes ($+$; $n = 13$), acyclic dienes (\triangle ; $n = 3$), primary alkynes (∇ ; $n = 6$), cyclic monoenes (\diamond ; $n = 2$), alkyl arenes containing only one phenyl ring (\square ; $n = 17$), and alkenyl arenes containing only one phenyl ring (\ast ; $n = 2$). Consequently there are 78 independent data points in each set. The acidic hydrogens on the primary alkynes were not counted as hydrogen-carbon bonds. The solid lines drawn are fit to only the respective data for the linear acyclic alkanes (\circ) in each set. The dotted line and the line of short dashes were fit to the respective points for nine representative acyclic alkanes, nine representative acyclic monoenes, the nine acyclic dienes and alkynes, nine representative alkyl arenes, and the two alkenyl arenes in each set of data, the upper set calculated with units of molarity and the lower set calculated with units of corrected volume fraction. Each of these lines was forced to pass through the origin. All of the representatives chosen contained between 4 and 16 carbons, and within each class the representatives chosen spanned the largest possible range of lengths. The line of long dashes was fit to all of the data for the alkyl arenes in the data set calculated with units of corrected volume fraction.

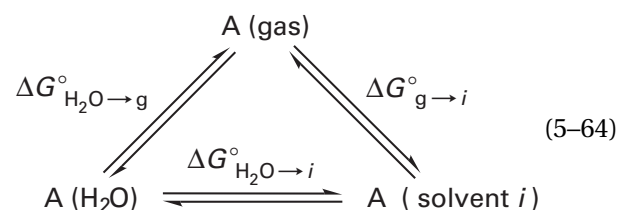
results are observed for the standard free energies of transfer from water to the pure liquid calculated from the solubilities of a similar set of hydrocarbons in water.²⁵

Fourth, both the upper and lower solid lines in Figure 5-21 intersect the ordinate so close to the origin that the hydrophobic effect represented by these standard free energies of transfer is for all practical purposes directly proportional to the number of hydrogen-carbon bonds a molecule contains. The upper dotted line in Figure 5-21 is a line forced to pass through the origin that was fit to a representative set of data for acyclic alkanes, acyclic monoenes, acyclic dienes and alkynes, alkyl arenes, and alkenyl arenes, and this fit is statistically indistinguishable from a line fit to the same representative set but not forced to pass through the origin ($r = 0.992$ and $r = 0.992$, respectively). The line of short dashes in Figure 5-21 is a line forced to pass through the origin that was fit to a representative set of data for acyclic alkanes, acyclic monoenes, acyclic dienes and alkynes, alkyl arenes, and alkenyl arenes, and the fit is almost statistically indistinguishable from a line fit to the same set of representative data but not forced to pass through the origin ($r = 0.973$ and $r = 0.977$, respectively).

The mean of the slopes of the four linear fits of the data spanning the largest range of hydrogen-carbon bonds, those for linear acyclic alkanes (lower solid line in Figure 5-21), branched acyclic alkanes, acyclic alkenes, and acyclic arenes (lower line of long dashes in Figure 5-21), for the standard free energies of transfer from water to hexadecane, based on units of corrected volume fraction, is $-2.80 \pm 0.08 \text{ kJ (mol hydrogen-carbon bond)}^{-1}$, which agrees with the value given by Pace.¹⁹³

The significant advantage of the fact that the magnitude of the hydrophobic effect exerted upon a molecule is determined only by its content of hydrogen-carbon bonds is that the change in standard free energy of transfer [$-2.8 \text{ kJ (mol hydrogen-carbon bond)}^{-1}$] or change in standard heat capacity of transfer [$30 \text{ J K}^{-1} \text{ (mol hydrogen-carbon bond)}^{-1}$] due to the hydrophobic effect for a structural transformation can be estimated simply by counting the number of hydrogen-carbon bonds removed from or inserted into water during that transformation.

The hydrophobic effect can be dissected more finely if the process of transfer itself is dissected further. The most reliable way to measure the transfer of a volatile solute between water and another solvent is to place a vessel of water and a vessel of that other solvent containing the solute into a sealed chamber and allow equilibration to occur. In this way, the solute dissolved in the water and the solute dissolved in the other solvent both come into equilibrium with the vapor of solute that fills the sealed chamber. In practice, this experiment dissects the transfer into the following thermodynamic cycle:^{181,194,195}



Ideally, the two standard free energies of transfer between the gas phase and the two liquid phases have the usual function of measuring the two separate **standard free energies of solvation**.

The partition coefficients for the transfer of a number of alkanes between the gas phase and various solvents have been collected.¹⁹⁵ The standard free energies of transfer calculated from these partition coefficients can be presented graphically by plotting the standard free energy of transfer for a given solute into a given solvent as a function of the standard free energy of transfer for that solute into benzene, which can be used arbitrarily as a reference solvent (Figure 5-22).¹⁹⁵ The solvents hexane, cyclohexane, carbon tetrachloride, toluene, phenyl bromide, and phenyl iodide all yield standard free energies of transfer between those of benzene and decane; the solvents phenyl chloride, *N*-methylpyrrole, 1-octanol, 1-butanol, 1-propanol, and ethanol all give standard free energies of transfer intermediate between those of benzene and methanol; and the solvents acetonitrile, propylene carbonate, and nitromethane all display free energies of transfer between those of methanol and dimethyl sulfoxide.

The solvation of most hydrophobic solutes by most solvents proceeds with a negative change in standard free energy, and all of these solvents show clear decreases in standard free energy of solvation as the size of the solute is increased (Figure 5-22). This is to be expected because the **van der Waals forces** that arise as the solute is surrounded by solvent have negative standard enthalpies of formation and increase in magnitude as the size of the solute increases. Water, however, is the clear exception among the solvents examined because it displays positive standard free energies of solvation for all of the hydrocarbons and these standard free energies of solvation increase as the size of the solute increases. With the exception of ethylene glycol, which is also strongly hydrogen-bonded, the nonaqueous solvent showing the least negative standard free energy of transfer is dimethyl sulfoxide, which has been included in Figure 5-22. Even dimethyl sulfoxide, however, fails to demonstrate the extreme behavior of water. The difference in behavior between water and all of the other solvents is the hydrophobic effect: the exclusion of hydrophobic solutes from aqueous solution.

Water must participate in van der Waals interactions with hydrophobic solutes just as the other solvents do. Nevertheless, the unfavorable hydrophobic effect must be greater in magnitude than these favorable van

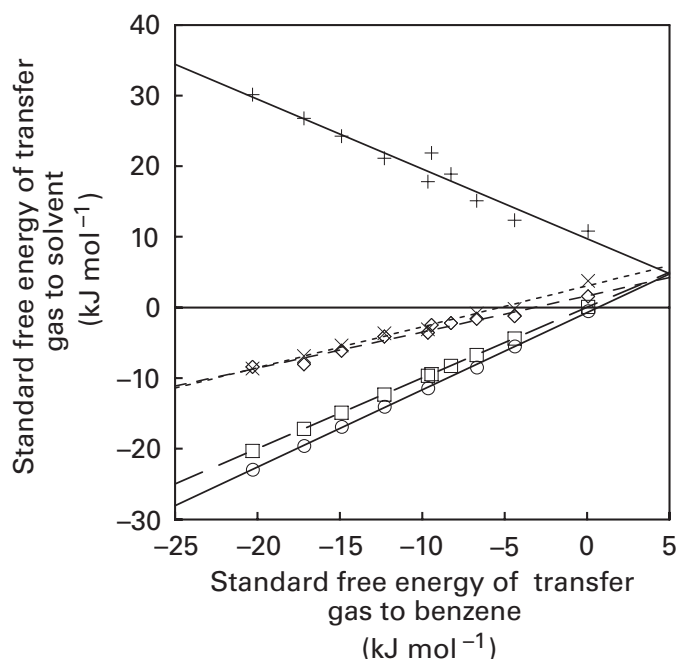


Figure 5-22: Standard free energy of transfer (kilojoules mole⁻¹) for a given alkane from the gas phase into a given solvent, plotted as a function of the standard free energy of transfer for that alkane from the gas phase into benzene.¹⁹⁵ The values tabulated for partition coefficients, originally in units of atmosphere for the vapor and mole fraction for the solute dissolved in a solvent, were converted to moles liter⁻¹ for the vapor and corrected volume fraction for the solute in a solvent (Equation 5-5) by use of estimates of the partial molar volumes for the hydrocarbons in water^{23,24} or Equation 5-9 for the partial molar volumes of the hydrocarbons in the other solvents. The standard free energies of transfer from the gas phase into the various solvents were then calculated (Equation 5-21). The solvents chosen for display were decane (○), benzene (□), methanol (◇), dimethyl sulfoxide (×), and water (+). The alkanes chosen were methane, ethane, propane, *n*-butane, isobutane, *n*-pentane, neopentane, *n*-hexane, *n*-heptane, and *n*-octane. Benzene was chosen as the reference solvent because more values for standard free energy of transfer into benzene were available and benzene has behavior similar to decane.

der Waals forces and more than overcomes them. As the polarities of the solvents increase, the slopes of the lines in Figure 5-22 become less negative. This effect might be explained by noting that as the solvents become more polar, more standard free energy is required to form a cavity within them, and this change is deducted from the favorable standard free energy of interaction between solute and solvent. In this view, water would simply be the extreme example of the difficulty in forming a cavity.

The solutes chosen for the free energies of transfer from the gas phase to the various solvents in Figure 5-22 were all acyclic alkanes. Further insight into the hydrophobic effect is gained when the free energies of transfer from the gas phase into hexadecane and from the gas phase into water are plotted for linear alkanes, branched alkanes, cyclic alkanes, alkenes, alkadienes, cyclic alkenes, alkynes, cyclic alkadienes, saturated arenes, and unsaturated arenes (Figure 5-23)^{175,196} as a

function of the number of their hydrogen-carbon bonds. The difference between each pair of lines, the one for transfer from gas to hexadecane and the one for transfer from gas to water, is the line for the standard free energy of transfer from water to hexadecane for the respective class of compounds (Figure 5-21). As the degree of unsaturation increases, each of these pairs of lines in Figure 5-23 is found at a lower level on the graph [about -2.2 kJ mol⁻¹ (level of unsaturation)⁻¹ for water and -2.4 kJ mol⁻¹ (level of saturation)⁻¹ for hexadecane].* It is the difference between these two values for the incremental free energies of transfer that causes each of the lines for the standard free energies of transfer for the other classes of hydrocarbons presented in Figure 5-21 (lower set of data) to fall progressively below the line for the linear alkanes.

With the notable exception of cyclization, when any two hydrocarbons are compared that have the same number of hydrogen-carbon bonds, the more unsaturated one will be the larger one, and the larger one should display stronger van der Waals interactions with the hexadecane. To the extent that the incremental decreases in standard free energies of solvation by hexadecane for hydrocarbons with the same number of hydrogen-carbon bonds but different levels of unsaturation (offsets of the lower set of lines in Figure 5-23) represent increases in van der Waals interactions, the similar incremental decreases observed for transfer of the same hydrocarbons into water suggest that water also participates in similar van der Waals interactions.

The standard enthalpies of transfer from the gas phase to water for the linear alkanes can be described by the relationship¹⁹⁷

$$\Delta H_{g \rightarrow H_2O}^\circ = -17 \text{ kJ mol}^{-1} - 1.7 \text{ kJ (mol hydrogen-carbon bond)}^{-1} \quad (5-65)$$

This inclusive, exothermic standard enthalpy of transfer must arise from the establishment of van der Waals interactions between water and the alkane during its entry. If so, this also provides evidence that water does participate in van der Waals interactions with hydrocarbon.

In Figure 5-23, the lines for the transfer of the linear alkanes from the gas phase into water and from the gas phase into hexadecane intersect on the ordinate at +4 kJ mol⁻¹, in agreement with the intersection of all of the lines in Figure 5-22 at +5 kJ mol⁻¹. In Figure 5-23, this

* The offsets of the lines for aqueous solutions in Figure 5-23 suggest that it is the only the degree of unsaturation that is of consequence in the solvation experienced by hydrocarbons in water, but the offsets of the lines for solutions in hexadecane suggest that, unlike in water where they are equivalent to only one degree of unsaturation, the solvation of a ring in hexadecane is equivalent to the solvation of two degrees of unsaturation.

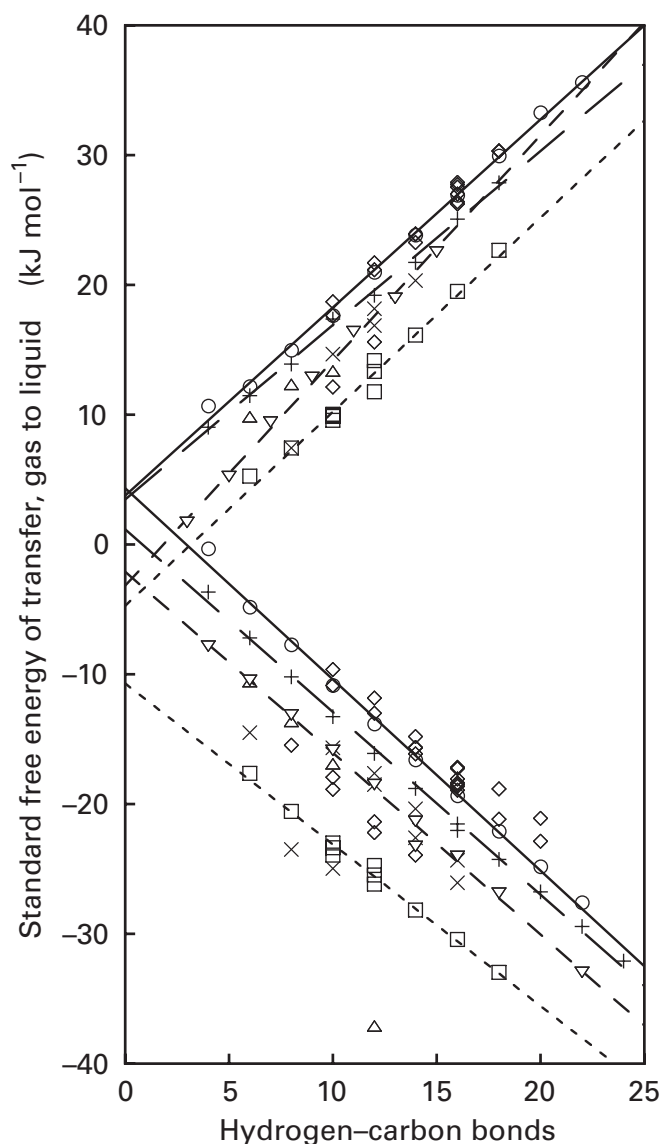


Figure 5-23: Standard free energies of transfer from the gas phase to water (upper set of lines) and from the gas phase to hexadecane (lower set of lines) as a function of the number of hydrogen-carbon bonds in a hydrocarbon. It was found that when the values for the partition coefficients for transfer from hexadecane to the gas phase calculated with Equation 1-10 from the mobility of linear alkanes on gas-liquid chromatography with a stationary phase of hexadecane¹⁷⁵ were treated as the quotients of the molarities of the hydrocarbons in the two phases and inserted into Equation 5-21, standard free energies of transfer resulted that were almost identical to those for the transfer of the same hydrocarbons from the gas phase to their own liquid, calculated with Equation 5-21 from the data of Hine and Mookerjee.¹⁹⁶ Consequently, it was assumed that the units on the dimensionless partition coefficients listed by Abraham were molarity molarity⁻¹, and they were inserted as such into Equation 5-21 directly to obtain standard free energies of transfer from the gas phase to hexadecane for activities in units of corrected volume fraction. The standard free energy of transfer for each hydrocarbon from the gas phase to water was obtained by summing its standard free energy of transfer from the gas phase to hexadecane and the standard free energy of transfer from hexadecane to water (Figure 5-21, lower set of data). The partial molar volumes for hydrocarbons in hexadecane were calculated as in Figure 5-21. The classes of hydrocarbons included, in descending order at each numerical value for hydrogen-carbon bonds, were branched acyclic alkanes (\diamond), linear acyclic alkanes (\circ), acyclic monoenes ($+$), cyclic alkanes (\times), acyclic dienes (\triangle), primary alkynes (∇), cyclic monoenes (\diamond), cyclic dienes (\times), alkyl arenes containing only one phenyl ring (\square), alkenyl arenes containing only one phenyl ring (\times), and diphenylmethane (\triangle). The acidic hydrogens on the primary alkynes were not counted as hydrogen-carbon bonds for transfer to water. The lines drawn are fit to the respective data for the linear acyclic alkanes (solid line), the acyclic monoenes (long dashes), the primary alkynes (intermediate dashes), and the alkyl arenes (short dashes). These four sets contained the largest spreads in the number of hydrogen-carbon bonds.

intersection corresponds to the standard free energy of transfer of a linear alkane with no hydrogen-carbon bonds from the gas phase to either condensed phase, which is the transfer of nothing. It is reassuring that the **transfer of nothing** from water to hexadecane proceeds with no change in standard free energy, but the fact that the standard free energy of transfer of nothing from the gas phase to either of these condensed phases is $+4 \text{ kJ mol}^{-1}$ suggests that, not surprisingly, there remains some difference in standard free energy between the gas phase and either condensed phase unaccounted for by the choices of standard state. If, as has been stated,¹⁹⁷ an additional term equal to RT ($+2.5 \text{ kJ mol}^{-1}$) must be added to all of the standard free energies of transfer, this correction would only increase this difference.

As the amount of unsaturation in the sets of hydrocarbons increases, the standard free energies of transfer at the respective intersections decrease in value. Each of these latter intersections corresponds to the free energy of transfer of the unsaturated hydrocarbon in that set

with no hydrogen-carbon bonds, which would be the particular carbon-carbon double bonds alone. These intersections decrease monotonically in value as the number of carbon-carbon double bonds increases because van der Waals interactions are realized when each of these carbon-carbon double bonds are transferred from the gas phase to either water or hexadecane.

The slope of the line in Figure 5-23 correlating the standard free energies of transfer from the gas phase to water for the linear alkanes is $+1.45 \text{ kJ (mol hydrogen-carbon bond)}^{-1}$, that for branched alkanes is $+1.48 \text{ kJ (mol hydrogen-carbon bond)}^{-1}$, and that for arenes is $+1.50 \text{ kJ (mol hydrogen-carbon bond)}^{-1}$. These values, which are for hydrocarbons related to the side chains of the amino acids, define the magnitude of the **active exclusion of hydrogen-carbon bonds from water**.

It is the separate solvations dissected in Figure 5-23, the one accomplished by hexadecane and the one accomplished by water, that together further illustrate the unique contribution of hydrogen-carbon bonds to

the hydrophobic effect. When any two hydrocarbons are compared that have the same number of hydrogen-carbon bonds, the more unsaturated one will be the one with the larger **surface area**. As the degree of unsaturation and hence the surface area increases at a constant number of hydrogen-carbon bonds, the standard free energy of solvation exerted by the hexadecane becomes more negative. As the degree of unsaturation and hence the surface area increases at a constant number of hydrogen-carbon bonds, the standard free energy of solvation exerted by the water becomes more negative. As the number of hydrogen-carbon bonds and hence the surface area increases at a constant degree of unsaturation, the standard free energy of solvation exerted by the hexadecane becomes more negative. In distinct contrast to these three trends, however, as the number of hydrogen-carbon bonds increases at a constant degree of unsaturation, the standard free energy of solvation exerted by the water becomes more positive. It is only the water that responds to an increase in the surface area of the solute by rejecting it more and more strongly but only when that increase in the surface area is accomplished by adding hydrogen-carbon bonds.

When the standard free energies of transfer from the gas phase to water for an even larger set of organic solutes than those displayed in Figure 5-23 are examined,¹⁹⁸ small differences in the standard free energy of transfer (mole hydrogen-carbon bond)⁻¹ for different types of hydrogen-carbon bond become apparent, and these differences have been quantified by defining a value for the contribution of each type to the overall standard free energy. For pairs of molecules otherwise identical except that one contains $-\text{CH}_2\text{CH}_2-$ and the other $-\text{CH}(\text{CH}_3)-$, the free energies of transfer from gas to water differ by $+0.4 \text{ kJ mol}^{-1}$ (15%), and for pairs of molecules, otherwise identical except that one contains $-\text{CH}_2\text{CH}_2\text{CH}_2-$ and the other contains $-\text{C}(\text{CH}_3)_2-$, the standard free energies of transfer from gas to water differ by $+0.6 \text{ kJ mol}^{-1}$ (15%). These differences might be taken as evidence that branched alkanes are more hydrophobic than unbranched alkanes except for the fact that when standard free energies of transfer from water to hexadecane (Figure 5-21) are examined instead of those from gas to water, branched alkanes (\diamond in Figure 5-21) are less hydrophobic than unbranched. These opposite conclusions bring into focus the unfortunate fact that currently there are two ways of defining the hydrophobic effect: one stressing only solvation by water, and the other, transfer from water to hydrocarbon.

There are **two significant contributions to the hydrophobic effect** (Figure 5-23): the active exclusion of hydrogen-carbon bonds from water and the solvation of those hydrogen-carbon bonds by the new surroundings in which they find themselves. If a hydrogen-carbon bond is transferred from water to the gas phase, the new

surroundings, by definition, do not solvate it; only the former contribution is expressed, and the hydrophobic effect is only the exclusion of the hydrogen-carbon bonds from water. If, however, the hydrogen-carbon bond is transferred to a condensed phase, such as hexadecane, half of the magnitude of the hydrophobic effect is its solvation by the new solvent. The more recent habit of equating the hydrophobic effect only with the transfer from gas to water^{26,198,199} avoids dealing with this half of the hydrophobic effect as it was defined traditionally.^{171,177} Both views, however, the one emphasizing solvation and the other transfer, persist.^{19,198}

It is hard to argue that transfer from water to gas is relevant to biochemical events such as the folding of a protein or the association of a substrate or inhibitor with an enzyme. In such instances the hydrogen-carbon bonds are transferred from water into a condensed phase that bears no resemblance to the gas phase. But the condensed phase in such situations does not resemble hexadecane either; rather, it is the interior of the irregular solid that is the native protein itself. There is no reason to assume that the interior of a molecule of protein behaves as if it were an isotropic solvent.

These considerations bring the argument back to the van der Waals forces between the hydrogen-carbon bond and its new surroundings. Regardless of whether or not water engages in van der Waals interactions with a hydrogen-carbon bond that are of the same magnitude as those of the new surroundings, the results in Figure 5-23 suggest that water *behaves* as though it does not engage in van der Waals interactions with a hydrogen-carbon bond at all. Once the hydrogen-carbon bond has been expelled from water, the standard free energy of its transfer is significantly affected by the standard free energy of the van der Waals forces between it and its new surroundings. As a result, and ironically, it is these van der Waals forces that determine much of the magnitude of the hydrophobic effect in any particular circumstance, and it is the magnitude of the van der Waals force felt by a hydrophobic functional group in the interior of the folded molecule of protein that significantly affects the strength of the hydrophobic effect it is able to exert during the folding of the polypeptide.

When a polypeptide is present in water in its unfolded state, the hydrophobic hydrogen-carbon bonds scattered along its length are unstable relative to any state in which they are in contact with each other and out of contact with water. This is the hydrophobic effect that drives the process of **folding**. Because ions and donors and acceptors of hydrogen bonds are more stable in the hydrated state than in any state in which they are isolated from water, even if they are fully joined in ion pairs and hydrogen bonds, they cannot provide net favorable standard free energy to the process of folding. The hydrophobic effect is the only noncovalent force that provides net favorable standard free energy to drive the folding of a polypeptide.

Suggested Reading

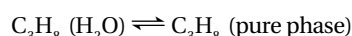
De Young, L. R., & Dill, K.A. (1990) Partitioning of nonpolar solutes into bilayers and amorphous *n*-alkanes, *J. Phys. Chem.* 94, 801–809.

Abraham, M.H. (1979) Free energies of solution of rare gases and alkanes in water and nonaqueous solvents: A quantitative assessment of the hydrophobic effect, *J. Am. Chem. Soc.* 101, 5477–5484.

Problem 5–12: The anomalously high isopiestic heat capacity C_p of liquid water is presumably due to the fact that as the liquid is heated, a certain amount of its hydrogen-bonded structure is lost. This extra heat capacity, beyond that calculated from vibrations and translations of the molecules, is the configurational heat capacity ($C_{p,cf}$). For water, $C_{p,cf} \approx 30 \text{ J K}^{-1} \text{ mol}^{-1}$ ($T = 273\text{--}373 \text{ K}$).

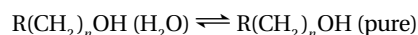
- (A) Calculate ΔS_{cf} and ΔH_{cf} , the configurational standard entropy and standard enthalpy changes, that are associated with heating water from 293 to 313 K.

It is difficult to say with certainty what fraction of the hydrogen bonding is lost over this temperature range, but these numbers give a rough estimate of the ratio of the changes of standard enthalpy and standard entropy ($\Delta H/\Delta S$) to be expected when the structure of water increases or decreases in the liquid. Consider the following phase transfer reaction:



- (B) If all the change in standard entropy (Table 5–8) is due to a decrease in water structure as the more structured regions that surrounded the alkane melt, what standard enthalpy change must accompany this decrease in water structure?
- (C) What is the contribution in percent to the change in standard free energy (ΔG°) in the above reaction that results from the melting of structured water around the alkane?

Problem 5–13: Consider the following transfer reaction:



This chemical equation describes the transfer of an alcohol from water to a pure phase. As such it describes the tendency of the alcohol to remove itself from water, a hydrophobic effect. For any substance A under any circumstances there is associated an intrinsic standard free energy, or chemical potential (μ):

$$\mu_A = \mu_A^\circ + RT \ln \left\{ \phi_{A,j} \exp \left[1 - \left(\bar{V}_{A,j} / \bar{V}_j \right) \right] \right\}$$

where μ_A is the chemical potential of solute A under the experimental circumstances, μ_A° is the chemical potential of solute A at standard state and unit concentration, and $\phi_{A,j}$ and \bar{V} are defined by Equations 5–6 and 5–8. The free energy change for the transfer reaction is

$$\Delta G_{\text{alc,H}_2\text{O} \rightarrow \text{alc}} = \mu_{\text{alc}} - \mu_{\text{alc,H}_2\text{O}} = \mu_{\text{alc}}^\circ - \mu_{\text{alc,H}_2\text{O}}^\circ -$$

$$RT \ln \left\{ \phi_{\text{alc,H}_2\text{O}} \exp \left[1 - \left(\bar{V}_{\text{alc,H}_2\text{O}} / \bar{V}_{\text{H}_2\text{O}} \right) \right] \right\}$$

where μ_{alc} is the chemical potential of pure alcohol and $\mu_{\text{alc,H}_2\text{O}}$ is the chemical potential of the alcohol dissolved in water at a concentration of $\phi_{\text{alc,H}_2\text{O}}$. At equilibrium, when alcohol saturates the water phase, $\Delta G = 0$ and

$$\mu_{\text{alc}}^\circ - \mu_{\text{alc,H}_2\text{O}}^\circ = \Delta G_{\text{alc,H}_2\text{O} \rightarrow \text{alc}}^\circ =$$

$$RT \ln \left\{ \phi_{\text{alc,H}_2\text{O,sat}} \exp \left[1 - \left(\bar{V}_{\text{alc,H}_2\text{O}} / \bar{V}_{\text{H}_2\text{O}} \right) \right] \right\}$$

It is easy to measure the concentration of alcohol in moles liter^{−1} at saturation when H₂O and pure alcohol are shaken in a two-phase system (separatory funnel) and the aqueous phase is clarified and removed. The following results²⁰⁰ were obtained at 25 °C:

alcohol	[alcohol] _{sat} (mol L ^{−1})
<i>n</i> -butanol	0.97
<i>n</i> -pentanol	0.25
<i>n</i> -hexanol	0.059
<i>n</i> -heptanol	0.0146
<i>n</i> -octanol	0.0038

- (A) Change these numbers into

$$\phi_{\text{alc,H}_2\text{O,sat}} \exp \left[1 - \left(\bar{V}_{\text{alc,H}_2\text{O}} / \bar{V}_{\text{H}_2\text{O}} \right) \right]$$

- (B) Calculate $\Delta G_{\text{alc,H}_2\text{O} \rightarrow \text{alc}}^\circ$ for each case, and plot $\Delta G_{\text{alc,H}_2\text{O} \rightarrow \text{alc}}^\circ$ as a function of the hydrogen–carbon bonds (*n*).

- (C) Determine the slope of your plot, $\Delta G_{\text{HC,H}_2\text{O} \rightarrow \text{alc}}^\circ$:

$$\Delta G_{\text{alc,H}_2\text{O} \rightarrow \text{alc}}^\circ = \Delta G_{\text{alc}}^\circ + n \Delta G_{\text{HC,H}_2\text{O} \rightarrow \text{alc}}^\circ$$

- (D) The term $\Delta G_{\text{HC,H}_2\text{O} \rightarrow \text{alc}}^\circ$ is the standard free energy change associated with the transfer of a hydrogen–carbon bond from H₂O to pure aliphatic alcohol. Is this transfer favored or unfavored?

- (E) The term $\Delta G_{\text{alc}}^\circ$ is the standard free energy change due to the fact that the molecule you are using is an alcohol. Is the transfer of the hydroxyl group favored or unfavored?

- (F) Extrapolate to determine $\Delta G^\circ_{\text{alc,H}_2\text{O} \rightarrow \text{alc}}$ for *n*-propanol.
- (G) Determine $\Delta S^\circ_{\text{HC,H}_2\text{O} \rightarrow \text{alc}}$ in the equation $\Delta S^\circ_{\text{alc,H}_2\text{O} \rightarrow \text{alc}} = \Delta S^\circ_{\text{alc}} + n\Delta S^\circ_{\text{HC,H}_2\text{O} \rightarrow \text{alc}}$ and $\Delta H^\circ_{\text{HC,H}_2\text{O} \rightarrow \text{alc}}$ in the equation $\Delta H^\circ_{\text{alc,H}_2\text{O} \rightarrow \text{alc}} = \Delta H^\circ_{\text{alc}} + n\Delta H^\circ_{\text{HC,H}_2\text{O} \rightarrow \text{alc}}$ from the values in Table 5-8.
- (H) Is ΔH° or ΔS° the major contributor to the hydrophobic effect on a hydrogen-carbon bond at 25 °C?

Problem 5-14: Calculate the standard enthalpy of transfer and the standard entropy of transfer of *n*-butane from water to liquid *n*-butane at 50, 70 and 90 °C. Recall that

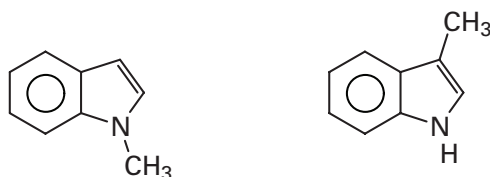
$$d\Delta H = \Delta C_p dT$$

and

$$d\Delta S = \frac{\Delta C_p}{T} dT$$

and assume that all partial volumes and ΔC_p are invariant with temperature over the range 20–100 °C.

Problem 5-15: The partition coefficients of *N*-methylindole and 3-methylindole



between water and cyclohexane were examined²⁰¹ to investigate the effect of the hydrogen-bond donor in tryptophan on its partition between water and the anhydrous interior of a protein. The coefficients for partition from water to cyclohexane in the following table are expressed in units of molarity and have been extrapolated to infinite dilution.

temperature (K)	partition coefficient (M M ⁻¹)	
	3-methylindole	<i>N</i> -methylindole
288	19	300
298	19	290
308	20	270
318	20	260
328	19	230

because

$$\ln \left\{ \frac{\phi_{\text{A,C}_6\text{H}_{12}}}{\phi_{\text{A,H}_2\text{O}}} \exp \left[\left(\frac{\bar{V}_{\text{A,H}_2\text{O}}}{\bar{V}_{\text{H}_2\text{O}}} \right) - \left(\frac{\bar{V}_{\text{A,C}_6\text{H}_{12}}}{\bar{V}_{\text{C}_6\text{H}_{12}}} \right) \right] \right\} =$$

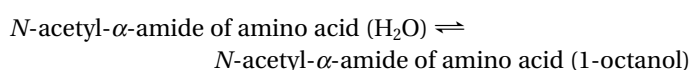
$$\ln \left(\frac{[\text{A}]_{\text{C}_6\text{H}_{12}}}{[\text{A}]_{\text{H}_2\text{O}}} \right) + \ln \left(\frac{\bar{V}_{\text{A,H}_2\text{O}}}{\bar{V}_{\text{A,C}_6\text{H}_{12}}} \right) + \left(\frac{\bar{V}_{\text{A,H}_2\text{O}}}{\bar{V}_{\text{H}_2\text{O}}} \right) - \left(\frac{\bar{V}_{\text{A,C}_6\text{H}_{12}}}{\bar{V}_{\text{C}_6\text{H}_{12}}} \right)$$

if $(\bar{V}_{\text{A,H}_2\text{O}}/\bar{V}_{\text{H}_2\text{O}})$, $(\bar{V}_{\text{A,C}_6\text{H}_{12}}/\bar{V}_{\text{C}_6\text{H}_{12}})$, and $(\bar{V}_{\text{A,H}_2\text{O}}/\bar{V}_{\text{A,C}_6\text{H}_{12}})$ do not change with temperature

$$\left(\frac{\partial \ln K_p}{\partial T^{-1}} \right)_P = \left[\frac{\partial \ln \left(\frac{[\text{A}]_{\text{C}_6\text{H}_{12}}}{[\text{A}]_{\text{H}_2\text{O}}} \right)}{\partial T^{-1}} \right]_P$$

- (A) Calculate the standard enthalpy of transfer for each compound from the behavior of its partition coefficient as a function of the temperature.
- (B) If a hydrogen bond were lost every time a 3-methylindole was transferred from water to cyclohexane and no hydrogen bond were lost every time *N*-methylindole was transferred from water to cyclohexane, which standard enthalpy of transfer should be more negative?
- (C) When a molecule of 3-methylindole leaves water for cyclohexane, what is the net change in hydrogen bonding for the entire system?
- (D) Why isn't the standard enthalpy change for the transfer of 3-methylindole more negative than that for the transfer of *N*-methylindole?

Problem 5-16: The *N*-acetyl- α -amides of a series of amino acids were synthesized, and the partition of each of them between 1-octanol and water was assayed. The measured distribution coefficients for the reaction



in units of (moles of solute) (liter of water)⁻¹ (moles of solute)⁻¹ (liter of 1-octanol) at 20 °C are presented in the following table.²⁰²

<i>N</i> -acetylamide of amino acid	distribution coefficient $\left(\frac{\text{mol of solute (L of 1-octanol)}^{-1}}{\text{mol of solute (L of water)}^{-1}} \right)$
isoleucine	0.93
leucine	0.75
methionine	0.25
valine	0.24
alanine	0.030
threonine	0.027
serine	0.0135
glutamine	0.0089
asparagine	0.0038

- (A) Estimate the partial molar volumes of the *N*-acetylammides in water by using the algorithm of Traube.

Because partial molar volumes in 1-octanol are unavailable, assume that they are equal to those in water + 13 cm³ mol⁻¹. In all cases, the final concentrations of the model compounds in each phase were less than 10 mM.

- (B) Calculate the standard free energy of transfer in kilojoules mole⁻¹ for each of these model compounds from water into 1-octanol, $\Delta G^{\circ}_{A,H_2O \rightarrow \text{octanol}}$.
- (C) Plot $\Delta G^{\circ}_{A,H_2O \rightarrow \text{octanol}}$ against the number of hydrogen-carbon bonds in each *N*-acetylammide.
- (D) Draw a line across your plot with a slope of -2.8 kJ (mol hydrogen-carbon bonds)⁻¹ that passes through the four points for leucine, isoleucine, valine, and alanine on your plot. Why did you do this?
- (E) Why don't the points for glutamine and asparagine lie 52 kJ mol⁻¹ above the line and the points for serine and threonine lie 26 kJ mol⁻¹ above the line?

Hydropathy

As ionic interactions, hydrogen bonds, and the hydrophobic effect are considered in turn, it becomes clear that the functional groups participating in each of these processes—ionic groups, donors and acceptors of hydrogen bonds, and hydrogen-carbon bonds—experience strong favorable and unfavorable interactions with water. Ionic solutes display large negative standard enthalpies of hydration that dominate their behavior in water. When these ions are withdrawn from water, large investments of free energy must be made to strip the shells of hydration from them. Solutes with donors and acceptors form hydrogen bonds with water molecules that have significant negative free energies of formation because of the high molar concentration of the water in

the solution. When donors and acceptors of hydrogen bonds are withdrawn from water, significant investments must be made to counter these free energies. Hydrophobic solutes leave aqueous solution with a preference for almost any other condensed or uncondensed phase because when they are dissolved in water, they cannot form any net favorable interactions with it. When they are withdrawn from water into another phase, significant favorable changes in free energy are realized. Each of these particular outcomes arises from the respective changes in the structure of water that accompany the transfer of the functional groups between water and a nonaqueous phase.

Viewed in this light, few solutes elicit indifferent responses from water. Solutes are either hydrophilic, demonstrating a compatibility with water, or hydrophobic, demonstrating an incompatibility with water. These strong responses together are hydropathy. **Hydropathy** is the continuous spectrum from compatibility to incompatibility with water, at one end of which is hydrophilicity, and at the other, hydrophobicity.

It was suggested by Hine and Mookerjee¹⁹⁶ that the hydropathy of a solute could be judged from its standard free energy of **transfer between water and the gas**. Assembling values from the tables of Hine and Mookerjee¹⁹⁶ and providing several previously unmeasured values, Wolfenden, Andersson, Cullis, and Southgate^{199,203} have tabulated the standard free energies of transfer between water and the gas for model compounds of the side chains of the amino acids in which the α carbon has been replaced by hydrogen (Table 5-9). These values reflect the magnitudes of the standard free energies realized when the various functional groups present in the amino acids are removed from water at pH 7. As previously noted, the hydrocarbons among the side chains are expelled spontaneously from water with standard free energies of transfer between about -5 and -20 kJ mol⁻¹.

The hydroxyl groups on ethanol and methanol increase their respective standard free energies of transfer from water to the gas phase by +27 kJ mol⁻¹ relative to alkanes of the same number of hydrogen-carbon bonds. In part, these unfavorable incremental standard free energies of transfer arise from the fact that a net of one hydrogen bond is lost to the system when a **hydroxyl group** is removed from water into the gas phase. Methanethiol, however, has a standard free energy of transfer only +10 kJ mol⁻¹ greater than an alkane with the same number of hydrogen-carbon bonds, while ethyl methyl sulfide has a standard free energy of transfer +12 kJ mol⁻¹ greater than an alkane with the same number of hydrogen-carbon bonds. A comparison of these two values with those for ethanol and methanol suggests that it is the **sulfur** that is hydrophilic, not the potential hydrogen-bond donor on the methanethiol.

Propionamide and acetamide have standard free energies of transfer +44 kJ mol⁻¹ greater than alkanes of

Table 5-9: Standard Free Energies of Transfer of Model Compounds for the Amino Acids between Water and the Gas Phase at 25 °C and pH 7^a

amino acid	model compound	$\Delta G^\circ_{\text{H}_2\text{O} \rightarrow \text{g}}$ (kJ mol ⁻¹)	amino acid	model compound	$\Delta G^\circ_{\text{H}_2\text{O} \rightarrow \text{g}}$ (kJ mol ⁻¹)	amino acid	model compound	$\Delta G^\circ_{\text{H}_2\text{O} \rightarrow \text{g}}$ (kJ mol ⁻¹)
L	isobutane	-18	C	methanethiol	+1	K	n-butylamine ^b	+28
I	butane	-18	W	3-methylindole	+10	Q	propionamide	+32
V	propane	-15	Y	4-cresol	+14	N	acetamide	+35
A	methane	-10	T	ethanol	+15	E	propionic acid ^b	+35
F	toluene	-8	S	methanol	+18	H	4-methylimidazole ^b	+36
M	ethyl methyl sulfide	-3				D	acetic acid ^b	+40
						R	methylguanidine ^b	+73
			peptide bond	N-methylacetamide	+34 kJ mol ⁻¹			

^aThe values for the standard free energies of transfer from water to the gas phase for the various model compounds were obtained from several tables.^{196,199,203} They were usually presented as the transfer of the compound between the standard state of the real gas at infinitely low partial pressure with concentration expressed in atmospheres and the standard state of the infinitely dilute solution with concentration expressed in molarity. The units of concentration were changed to moles liter⁻¹ for the gas and corrected volume fraction (Equation 5-5) for the solution. The partial molar volumes of the solutes²⁶ were calculated²⁴ for each solute from the formulas developed by Traube.²³ The standard free energies of transfer from water to the gas phase were calculated with Equation 5-21. ^bValues for the pK_a of the various amino acids in a polypeptide (Table 2-2) were used to correct the standard free energies of transfer of the neutral compounds^{199,203} for the standard free energy of neutralization required at pH 7 (Equation 5-66).

the same number of hydrogen-carbon bonds. An argument can be made that these large positive standard free energies of transfer for the **primary amides** arise simply because each of them has two acceptors and two donors so that a net loss to the system of two hydrogen bonds occurs upon their transfer to the gas phase. Consequently, their standard free energies of transfer relative to alkanes of the same number of hydrogen-carbon bonds should be about twice those of ethanol and methanol, which they are. The standard free energies of transfer for acetamide or propionamide are about +12 kJ mol⁻¹ greater than those for neutral acetic acid or neutral propionic acid, respectively. The transfer of the **carboxylic acid** on either of these two acids involves the net loss of only one hydrogen bond to the solution. Although all of these explanations seem reasonable, they leave unexplained the fact that the standard free energy of transfer for *N*-methylacetamide, the model for the **peptide bond**, is actually greater than that for propionamide, which has the same number of hydrogen-carbon bonds, even though a net loss of only one hydrogen bond occurs when the former is removed from water. It was suggested after the fact that in the case of the amides the acceptors may be more important than the donors,²⁰⁴ and subsequently spectroscopic evidence consistent with this suggestion was reported, demonstrating that each of the two acceptors on the acyl oxygen of *N*-methylacetamide interacts with water with about twice the standard enthalpy as that for the interaction of the single donor.^{116,205,206}

The transfers of the **heterocyclic side chains** are dominated by their donors and acceptors of hydrogen bonds. 3-Methylindole and 4-cresol have standard free energies of transfer +18 and +21 kJ mol⁻¹, respectively, greater than an arene with the same number of hydrogen-carbon bonds. These increments arise from the

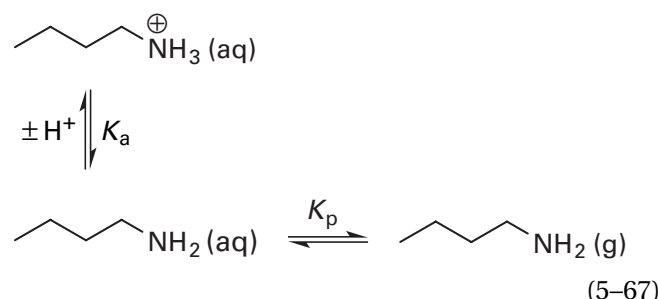
presence of the pyrrole donor and the hydroxyl, respectively, that form hydrogen bonds with the donors and acceptors of water that must be broken during transfer. The standard free energy of transfer for neutral 4-methylimidazole is +39 kJ mol⁻¹ greater than that of an arene with the same number of hydrogen-carbon bonds, in part because its acceptor (pK_a = 7.5) is much stronger than that of either 4-cresol (pK_a = 10.2) or 3-methylindole (pK_a = -2).

In the case of the amino acids that are charged at pH 7, such as glutamic acid, aspartic acid, histidine, lysine, and arginine, the tabulated partition coefficients for transfer of the model compounds between water and the gas are for acidic solutions or basic solutions in which those model compounds are dissolved entirely as the neutral conjugate acids or neutral conjugate bases, respectively. At pH 7 only a fraction of the actual amino acid will be present as the neutral species, and this will decrease the value of the partition coefficient for transfer from water to the gas. This decrease in the partition coefficient can be incorporated into the standard free energy of transfer with the formula²⁰³

$$\Delta G^\circ_{\text{A}_{\text{TOT}}, \text{H}_2\text{O} \rightarrow \text{g}} = \Delta G^\circ_{\text{A}_0, \text{H}_2\text{O} \rightarrow \text{g}} - RT \ln \alpha_{\text{A}} \quad (5-66)$$

where A₀ refers to the un-ionized form of the model compound, A_{TOT} is the sum of the neutral form and the charged form, and α_A is the fraction that is un-ionized at pH 7. The values for α_A were calculated with the values of pK_a for the amino acids in a polypeptide (Table 2-2) rather than with the values of pK_a for the model compounds themselves. The fraction of un-ionized species, α_A, varies from 0.8 for histidine to 10⁻⁶ for arginine. Even the standard free energy necessary to neutralize *N*-methylguanidium and transfer the neutral compound into the gas

(+73 kJ mol⁻¹) is still less than the standard free energy that would be required to transfer it into the gas as a cation (Figure 5-8). Therefore, all of the tabulated values should refer to the most likely reactions, namely, neutralization followed by transfer. For example, the only reasonable reaction for the transfer of *n*-butylammonium ion, dissolved in H₂O at pH 7, to the gas would be



One of the most striking, and perhaps informative, facts about the charged amino acids in a polypeptide is that all of them can be **neutralized** in simple acid-base reactions. This is a common property of almost all organic cations and anions found in biological systems, with the tetraalkylammonium cation among the notable exceptions. Nevertheless, it could be argued that this latter functional group, which does appear in biological settings, has been purposely excluded from among the amino acids by natural selection because it cannot be neutralized. When charged amino acid side chains are transferred to a region with a low relative permittivity, such as the gas phase, they will enter as the uncharged conjugate acids or conjugate bases. The standard free energy required to neutralize the charges of propionic acid, acetic acid, and *n*-butylamine and then remove their donors and acceptors for hydrogen bonds from water is greater by about +45 to +48 kJ mol⁻¹ than the standard free energy of transfer for an alkane with the same number of hydrogen-carbon bonds.

The point emphasized by these experimental results is that water has a high affinity for many of the side chains of the amino acids, and a good deal of standard free energy must be spent whenever they are removed from it. This is the complement of the fact that the only favorable standard free energy gained from removing side chains of the amino acids from an aqueous solution arises from the removal of their hydrogen-carbon bonds from contact with water. At the beginning, every amino acid within a newly synthesized, unfolded polypeptide is completely solvated by water. During the **folding** of the polypeptide, the formation of an interface between subunits, or the binding of a ligand, there is a net transfer of side chains of individual amino acids, segments of polypeptide backbone, and small solutes from water to the interior of a native protein. In analogy with Equation 5-64, the first half of the reaction can be represented by the standard free energies of transfer between water and the gas (Table 5-9). Those

solutes removed from water, however, are transferred into a new environment, the **interior of the protein**. The standard free energies of transfer into this new environment are the second half of the reaction, but the standard free energies of transfer into the interior of a protein from the gas cannot be predicted with any certainty because the interior of a protein is a heterogeneous solid.

Presumably, noncovalent forces with negative standard free energies of formation arise as the amino acids and the polypeptide backbone are packed into the interior. The standard free energy of transfer for the removal of hydrogen-carbon bonds from the aqueous phase is a negative change in standard free energy, but a small one compared to the favorable hydration of the polar side chains (Table 5-9). These hydrogen-carbon bonds, however, are being transferred into a new environment that should not have the aversion displayed by liquid water but should respond with favorable van der Waals forces, perhaps resembling in their magnitude the favorable negative standard free energies of solvation for nonpolar solutes displayed by all solvents other than water (Figure 5-22). The standard free energies of transfer for the removal of the hydrogen-bond donors and acceptors from the aqueous phase, especially those of the polypeptide backbone itself, are large positive standard free energies of transfer (Table 5-9). In the interior of the protein, however, these donors and acceptors participate in hydrogen bonds the standard free energies of formation of which are even more negative than they would be in solution because standard entropies of approximation are often not required for the associations (Figure 5-19). Standard entropies of approximation are at a minimum whenever a set of hydrogen bonds forms cooperatively as in an α helix or β structure.

The interior of a properly folded molecule of protein is tightly packed in a defined conformation. It thus resembles closely the interior of a solid rather than a liquid, and the van der Waals forces arising when amino acids or segments of polypeptide backbone are transferred into it are those that would arise in a solid rather than in a liquid. This is a significant difference because in solids dipoles and polarizable regions remain fixed in their relative orientations rather than being averaged over all orientations as they are in a liquid. Even more troublesome, when the question of reproducing the behavior of this solid is considered, is the fact that the interior of a protein is not a systematic solid, such as a crystalline or microcrystalline mineral or an amorphous glass. Although it is a highly integrated system, shaped by natural selection for the performance of a definite function, little uniformity can be found in the interior of a particular molecule of protein. Because it is the product of evolution by natural selection, the interior has all of the haphazard character of an acre of woodland. It is unlikely that any solvent could reproduce its properties.

Nevertheless, it has been proposed that the standard free energy of transfer of an amino acid, a segment

of polypeptide, or a small solute from water to the interior of a protein should be similar to the standard free energy of transfer for a model compound of that amino acid or segment of polypeptide between water and a solvent the properties of which resemble those of the interior of a protein. **Scales of hydrophathy** for the side chains of the amino acids based on this proposal have been presented. They differ in the personal preferences of their proponents for the type of model compounds and the particular solvent chosen as the basis of the scale.

The first of these was the scale of hydrophobicity proposed by Nozaki and Tanford,²⁰⁷ which, as its name implies, was confined to only one end of the spectrum. It relied on the solubilities of the zwitterionic amino acids in ethanol that had been previously tabulated by Cohn and Edsall.²⁴ By subtracting the standard free energy of transfer for glycine between water and **ethanol** from the standard free energies of transfer for hydrophobic amino acids between water and ethanol, they estimated the standard free energies of transfer for the side chains alone between water and ethanol. The implication in formulating a scale of this type is that the interior of a protein resembled ethanol in its interaction with hydrophobic amino acids, and this may not be far from the truth because all nonaqueous solvents display similar standard free energies of solvation for hydrophobic solutes (Figure 5-22).

Since this first scale was proposed, at least 35 others have appeared,¹⁸⁹ and they have usually been expanded spectra including all 20 of the amino acids, hydrophilic as well as hydrophobic. Most have been based on standard free energies of transfer. The original description of the hydrophobic effect was based on observations of abnormal decreases in the surface tension of water that result when hydrophobic solutes display a preference for the surface of an aqueous solution rather than its interior,^{7,208} and a scale of hydrophathy based on the change in the **surface tension** of an aqueous solution with the change in concentration of the different amino acids has been presented.²⁰⁹ The scale of hydrophobicity based on the solubilities of the amino acids in ethanol has been expanded to include uncharged, hydrophilic amino acids.²¹⁰ The standard free energies of transfer of the model compounds for the amino acids between water and the **gas** (Table 5-9) have also been used to create scales of hydrophathies.^{198,199,203} The standard free energies of transfer of various solutes between water and **1-octanol** have been proposed as the parameters for a general scale for the hydrophobic effect.²¹¹ The standard free energies of transfer of the *N*-acetyl- α -amides of each of the amino acids (Figure 2-1) between water and 1-octanol have been determined, and they have been used to construct a scale of hydrophathies.²⁰² It has been proposed that *N*-cyclohexyl-2-pyrrolidone would be a better solvent to use as reference for standard free energies of transfer into the interior of a protein, and standard free energies of transfer for the amino acids between

water and *N*-cyclohexyl-2-pyrrolidone have been measured and used to construct a scale of hydrophathies.²¹²

In competition with these scales of hydrophathy based on standard free energies of transfer are scales derived from the locations of the various amino acid side chains in crystallographic molecular models of native proteins. The logic in this case is that the purpose of all of these scales is to estimate contributions due to changes in solvation during the folding of a polypeptide and the degree with which particular amino acids are buried in the interior or exposed to the solvent should directly indicate how hydrophobic or hydrophilic, respectively, they are. In these computations, the surface area of each amino acid that is accessible to water^{213,214} in a set of crystallographic molecular models is individually determined. These individual **accessibilities to water** are then grouped by amino acid, and average accessibilities for each amino acid are calculated. The uncertainty in these calculations is in the calculation of these averages, and the three scales of hydrophathy based on the accessible surface area in molecular models of folded polypeptides²¹⁵⁻²¹⁷ are not equivalent, even though they are based on similar raw data.

Finally, there are the scales of hydrophathies for the amino acids that are based on mixtures of the pure scales discussed so far. In one case,²¹⁸ a scale based on the accessible surface area of amino acids in crystallographic models was modified by a theoretical calculation of the standard free energy required to break hydrogen bonds and neutralize charge. In another,²⁶ the standard free energies of transfer between water and the gas and a tabulation of accessible surface areas were combined with personal preference to produce a scale of hydrophathies. In a third case,²¹⁹ a consensus scale of hydrophathies was inferred from two scales based on standard free energy of transfer and three scales based on accessible surface area in folded polypeptides. In a fourth case,²²⁰ a correlation between accessible surface area and the hydrophobic effect, the standard free energy required to neutralize charged amino acids (Equation 5-66), and semi-empirical estimates of the standard free energy for withdrawing each individual hydrogen-bond donor and acceptor from water were combined to obtain a scale of estimated standard free energies of transfer for each of the amino acids, when located in an α helix, from water to a phase of hydrocarbon.

At low resolution all of these scales are similar to each other. The amino acids the side chains of which are alkanes, namely, leucine, isoleucine, and valine, are the most hydrophobic amino acids; the charged amino acids the pK_a of which is farthest from pH 7, arginine, lysine, glutamate, and aspartate, are the most hydrophilic; and neutral but polar amino acids such as serine and threonine reside in the middle; but the details of the ranking and the relative magnitudes of the parameters are dramatically different.¹⁸⁹ At the moment, each of these attempts to estimate the standard free energy of transfer for each of the amino acids between water and the interior of a protein has its particular proponents, some

more forceful than others, and there is no unambiguous way to choose among them or assess whether any of them is realistic or unrealistic.

The usual criterion for the reliability of each scale is to demonstrate either that it correlates with the distribution of amino acids between the surface and the interior of a protein, if it is based on standard free energy of transfer (Figure 5-24),^{199,202,216} or that it correlates with standard free energies of transfer, if it is based on the distribution of amino acids between the surface and the interior.²¹⁶ None of these correlations suggest that any one of the scales is more realistic than any of the others.

Suggested Reading

Kyte, J., & Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol.* 157, 105-132.

Problem 5-17: Consider a saturated solution of the solute A in solvent *j*. In this case, the solution of A is in equilibrium with solid A and

$$\mu_{A,\text{sat},j} = \mu_{A,\text{solid}}$$

$$\mu_{A,j}^{\circ} + RT \ln \left\{ \phi_{A,\text{sat},j} \exp \left[1 - \left(\bar{V}_{A,j} / \bar{V}_j \right) \right] \right\} = \mu_{A,\text{solid}}^{\circ}$$

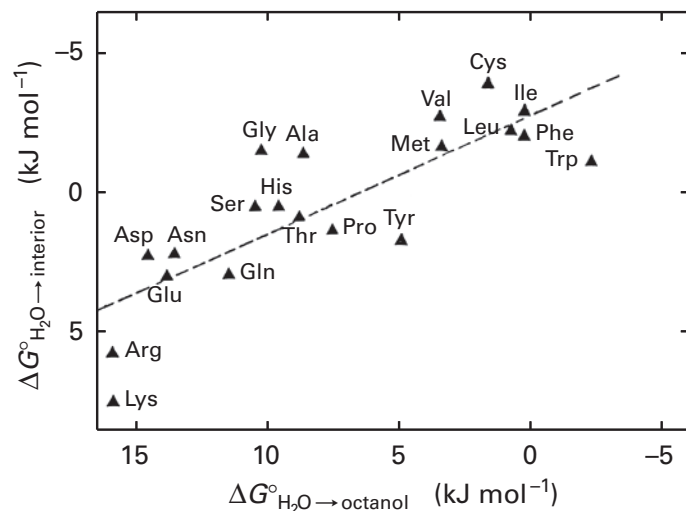


Figure 5-24: Correlation between the standard free energies of transfer of *N*-acetyl- α -amides of the amino acids between octanol and water with the degree to which the amino acids are buried in the interior of a molecule of protein.²⁰² The partition coefficients for the distribution of the *N*-acetyl- α -amides of the 20 amino acids between water at pH 7 and octanol at room temperature were measured (concentrations in molarity) and standard free energies of transfer, $\Delta G^{\circ}_{\text{aa},\text{H}_2\text{O} \rightarrow \text{octanol}}$, were calculated. Each of the 5220 amino acids in the crystallographic molecular models of 22 proteins was identified as either buried (less than 0.2 nm² of accessible surface area) or accessible to water (greater than 0.2 nm² of accessible surface area).²¹⁶ For each type of amino acid a partition ratio [number buried (number accessible)⁻¹] was calculated, and from this partition ratio, a standard free energy of transfer, $\Delta G^{\circ}_{\text{aa},\text{H}_2\text{O} \rightarrow \text{interior}}$ was calculated. Adapted with permission from ref 202. Copyright 1983 Elsevier.

If one wishes to compare two different solvents and their effects on A

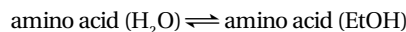
$$\mu_{A,1}^{\circ} + RT \ln \left\{ \phi_{A,\text{sat},1} \exp \left[1 - \left(\bar{V}_{A,1} / \bar{V}_1 \right) \right] \right\} = \mu_{A,\text{solid}}^{\circ}$$

$$\mu_{A,2}^{\circ} + RT \ln \left\{ \phi_{A,\text{sat},2} \exp \left[1 - \left(\bar{V}_{A,2} / \bar{V}_2 \right) \right] \right\} = \mu_{A,\text{solid}}^{\circ}$$

$$\mu_{A,2}^{\circ} - \mu_{A,1}^{\circ} = \Delta G^{\circ}_{A,1 \rightarrow 2}$$

$$\Delta G^{\circ}_{A,1 \rightarrow 2} = RT \ln \left[\frac{[A]_{\text{sat},1}}{[A]_{\text{sat},2}} \exp \left(\frac{\bar{V}_{A,2}}{\bar{V}_2} - \frac{\bar{V}_{A,1}}{\bar{V}_1} \right) \right]$$

where $\mu_{A,1}^{\circ}$ is the chemical potential of A in solvent 1 at a concentration of 1 corrected volume fraction, $\phi_{A,\text{sat},2}$ is the concentration of A in a saturated solution in solvent 2 in units of volume fraction, and $\Delta G^{\circ}_{A,1 \rightarrow 2}$ is the standard free energy change when A is transferred from solvent 1 to solvent 2. The quantity $\Delta G^{\circ}_{A,1 \rightarrow 2}$ is a measure of the change in standard free energy for the following type of reaction:



where ethanol serves as a model for the interior of a protein.

The following data^{24,207} are for 25 °C

amino acid	concn at saturation [g (100 g of solvent) ⁻¹]		partial molar volume (mL mol ⁻¹)
	H ₂ O	EtOH	
glycine	25.16	0.00382	43.5
leucine	2.17	0.0196	108

- Calculate $\phi_{A,\text{sat},j} \exp[(1 - \bar{V}_{A,j}/\bar{V}_j)]$ for these four situations. Assume $\bar{V}_{A,\text{H}_2\text{O}} = \bar{V}_{A,\text{EtOH}}$
- Calculate $\Delta G^{\circ}_{A,\text{H}_2\text{O} \rightarrow \text{EtOH}}$ for glycine and leucine.
- Use the value you have for glycine to subtract away the contribution of $^-\text{OOCCH}_2\text{NH}_3^+$ to the solubility of leucine. The remainder is an estimate of the standard free energy of transfer of the leucine side chain from H₂O to ethanol.
- Draw the structure of the glutamine side chain and divide it into hydrophobic or hydrogen-bonding regions. Label each region on your drawing and indicate all hydrogen-bond donors and acceptors with D or A, respectively.
- Estimate the $\Delta G^{\circ}_{\text{glutamine},\text{H}_2\text{O} \rightarrow \text{ethanol}}$ contributed only by the hydrogen-carbon bonds of the side chain.

- (F) The solubility of glutamine (concentration at saturation) in water is $4.6 \text{ g (100 g of H}_2\text{O)}^{-1}$, the solubility of glutamine in ethanol is $4.59 \times 10^{-4} \text{ g (100 g of ethanol)}^{-1}$, and the partial molar volume of glutamine in water is 96.5 mL mol^{-1} . Calculate $\Delta G^\circ_{\text{transfer, H}_2\text{O} \rightarrow \text{ethanol}}$ of the glutamine side chain.
- (G) Estimate $\Delta G^\circ_{\text{transfer, H}_2\text{O} \rightarrow \text{ethanol}}$ for the $-\text{CONH}_2$ functional group of glutamine.

Problem 5-18: Consider the following table:

A	0.25	G	0.16	P	-0.07
R	-1.76	H	-0.40	S	-0.26
N	-0.64	I	0.73	T	-0.18
D	-0.72	L	0.53	W	0.37
C	0.04	K	-1.10	Y	0.02
E	-0.62	M	0.26	V	0.54
Q	-0.69	F	0.61		

- (A) What are the letters and what is the intention of assigning these numbers to these letters?
- (B) What common property is shared by the letters with the positive numbers?
- (C) What common property is shared by the letters with the negative numbers?
- (D) Divide the letters with the positive numbers into two groups based on differences in chemical properties. Why are the numbers in one of these groups less positive than the numbers in the other?
- (E) On what types of measurements could the numbers assigned to the letters be based?
- (F) Draw the interactions with water that are one of the reasons that R has a value of -1.76 . There are two reasons that R has such a low value: the interactions you have just drawn and another of its properties. What are these two reasons?

References

- Del Bene, J., & Pople, J.A. (1970) *J. Chem. Phys.* 52, 4858-4866.
- Hankins, D., Moskowitz, J.W., & Stillinger, F.H., Jr. (1970) *J. Chem. Phys.* 53, 4544-4554.
- Symons, M.C.R. (1972) *Nature* 239, 257-259.
- Eisenberg, D., & Kauzmann, W. (1969) *The Structure and Properties of Water*, Clarendon Press, Oxford, England.
- Dyke, T.R., & Muentner, J.S. (1974) *J. Chem. Phys.* 60, 2929-2930.
- Dyke, T.R., Mack, K.M., & Muentner, J.S. (1977) *J. Chem. Phys.* 66, 498-510.
- Edsall, J.T., & McKenzie, H.A. (1978) *Adv. Biophys.* 10, 137-207.
- Frank, H.S. (1970) *Science* 169, 635-641.
- Narten, A.H., & Levy, H.A. (1971) *J. Chem. Phys.* 55, 2263-2269.
- Morgan, J., & Warren, B.E. (1938) *J. Chem. Phys.* 6, 666-673.
- Narten, A.H., & Levy, H.A. (1969) *Science* 165, 447-454.
- Narten, A.H., Danford, M.D., & Levy, H.A. (1967) *Discuss. Faraday Soc.* 43, 97-107.
- Lavrov, V.V. (1947) *Zh. Sakh. Promsti.* 17, 1027-1034.
- Poirier, J.P., Sotin, C., & Peyronneau, J. (1981) *Nature* 292, 225-227.
- Wittebort, R.J., Usha, M.G., Ruben, D.J., Wemmer, D.E., & Pines, A. (1988) *J. Am. Chem. Soc.* 110, 5668-5671.
- Walrafen, G.E. (1968) *J. Chem. Phys.* 48, 244-251.
- Silverstein, K.A.T., Haymet, A.D.J., & Dill, K.A. (2000) *J. Am. Chem. Soc.* 122, 8037-8041.
- Sharp, K.A., Nicholls, A., Friedman, R., & Honig, B. (1991) *Biochemistry* 30, 9686-9697.
- Sharp, K.A., Nicholls, A., Fine, R.F., & Honig, B. (1991) *Science* 252, 106-109.
- Deyoung, L.R., & Dill, K.A. (1990) *J. Phys. Chem.* 94, 801-809.
- Huggins, M.L. (1941) *J. Chem. Phys.* 9, 440.
- Flory, P.J. (1941) *J. Chem. Phys.* 9, 660-661.
- Traube, J. (1899) *Samml. Chem. Chem. Tech. Vortr.* 4, 255-332.
- Cohn, E.J., & Edsall, J.T. (1943) *Proteins, Amino Acids and Peptides as Ions and Dipolar Ions*, pp 157-161, Reinhold, New York.
- McAuliffe, C. (1966) *J. Phys. Chem.* 70, 1267-1275.
- Kyte, J., & Doolittle, R.F. (1982) *J. Mol. Biol.* 157, 105-132.
- Moore, W. (1972) *Physical Chemistry*, 4th ed., pp 890-894, Prentice-Hall, Englewood Cliffs, NJ.
- Pauling, L. (1960) *The Nature of the Chemical Bond and the Structure of Molecules and Crystals*, 3rd ed., p 449, Cornell University Press, Ithaca, NY.
- Jencks, W.P. (1969) *Catalysis in Chemistry and Enzymology*, McGraw-Hill, New York.
- Meot-Ner, M. (1984) *J. Am. Chem. Soc.* 106, 1265-1272.
- Benoit, R.L., & Lam, S.Y. (1974) *J. Am. Chem. Soc.* 96, 7385-7390.
- Parsegian, A. (1969) *Nature* 221, 844-846.
- Stokes, R.H. (1964) *J. Am. Chem. Soc.* 86, 979-982.
- Noyes, R.M. (1962) *J. Am. Chem. Soc.* 84, 513-522.
- Cox, B.G., & Parker, A.J. (1973) *J. Am. Chem. Soc.* 95, 6879-6884.
- Leikin, S., Parsegian, V.A., Rau, D.C., & Rand, R.P. (1993) *Annu. Rev. Phys. Chem.* 44, 369-395.
- Rand, R.P., Fuller, N., Parsegian, V.A., & Rau, D.C. (1988) *Biochemistry* 27, 7711-7722.
- Tanford, C. (1954) *J. Am. Chem. Soc.* 76, 945-946.
- Likhodi, O., & Chalikian, T.V. (2000) *J. Am. Chem. Soc.* 122, 7860-7868.
- Wimley, W.C., Gawrisch, K., Creamer, T.P., & White, S.H. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 2985-2990.
- O'Shea, E.K., Lumb, K.J., & Kim, P.S. (1993) *Curr. Biol.* 3, 658-667.
- Moore, W.J. (1972) *Physical Chemistry*, 4th ed., pp 420-476, Prentice-Hall, Englewood Cliffs, NJ.
- Wyman, J., Jr. (1936) *Chem. Rev.* 19, 213-239.
- Hine, J. (1972) *J. Am. Chem. Soc.* 94, 5766-5771.
- Eigen, M. (1964) *Angew. Chem., Int. Ed. Engl.* 3, 1-19.

46. Pimentel, G.C., & McClellan, A.L. (1960) *The Hydrogen Bond*, W.H. Freeman, San Francisco, CA.
47. Taylor, R., Kennard, O., & Versichel, W. (1983) *J. Am. Chem. Soc.* 105, 5761–5766.
48. Taylor, R., Kennard, O., & Versichel, W. (1984) *J. Am. Chem. Soc.* 106, 244–248.
49. Gorbitz, C.H., & Etter, M.C. (1992) *J. Am. Chem. Soc.* 114, 627–631.
50. Thanki, N., Thornton, J.M., & Goodfellow, J.M. (1988) *J. Mol. Biol.* 202, 637–657.
51. Kuhn, L.P., Wires, R.A., Ruoff, W., & Kwart, H. (1969) *J. Am. Chem. Soc.* 91, 4790–4793.
52. Donohue, J. (1969) *J. Mol. Biol.* 45, 231–235.
53. Cox, R.A., Druet, L.M., Klausner, A.E., Modro, T.A., Wan, P., & Yates, K. (1981) *Can. J. Chem.* 59, 1568–1573.
54. McCormack, A.C., McDonnell, C.M., O’Ferrall, R.A.M., O’Donoghue, A.C., & Rao, S.N. (2002) *J. Am. Chem. Soc.* 124, 8575–8583.
55. Kresge, A.J., Chen, H.J., Hakka, L.E., & Kouba, J.E. (1971) *J. Am. Chem. Soc.* 93, 6174–6181.
56. Ravishanker, G., Mehrotra, P.K., Mezei, M., & Beveridge, D.L. (1984) *J. Am. Chem. Soc.* 106, 4102–4108.
57. Suzuki, S., Green, P.G., Bumgarner, R.E., Dasgupta, S., Goddard, W.A., & Blake, G.A. (1992) *Science* 257, 942–944.
58. Atwood, J.L., Hamada, F., Robinson, K.D., Orr, G.W., & Vincent, R.L. (1991) *Nature* 349, 683–684.
59. Allen, F.H., Howard, J.A.K., Hoy, V.J., Desiraju, G.R., Reddy, D.S., & Wilson, C.C. (1996) *J. Am. Chem. Soc.* 118, 4081–4084.
60. Levitt, M., & Perutz, M.F. (1988) *J. Mol. Biol.* 201, 751–754.
61. Tuckerman, M.E., Marx, D., Klein, M.L., & Parrinello, M. (1997) *Science* 275, 817–820.
62. Perrin, C.L., & Thoburn, J.D. (1989) *J. Am. Chem. Soc.* 111, 8010–8012.
63. Perrin, C.L., & Nielson, J.B. (1997) *J. Am. Chem. Soc.* 119, 12734–12741.
64. Perrin, C.L. (1994) *Science* 266, 1665–1668.
65. Gilli, P., Bertolasi, V., Ferretti, V., & Gilli, G. (1994) *J. Am. Chem. Soc.* 116, 909–915.
66. Ichikawa, M. (1981) *Chem. Phys. Lett.* 79, 583–587.
67. Steiner, T., & Saenger, W. (1994) *Acta Crystallogr., Sect. B* 50, 348–357.
68. Berglund, B., & Vaughan, R.W. (1980) *J. Chem. Phys.* 73, 2037–2043.
69. Altman, L.J., Laungani, P., Gunnarsson, G., Wennerstrom, H., & Forsen, S. (1978) *J. Am. Chem. Soc.* 100, 8264–8265.
70. Hsu, B., & Schlemper, E.O. (1980) *Acta Crystallogr., Sect. B* 36, 3017–3023.
71. Madsen, D., Flensburg, C., & Larsen, S. (1998) *J. Phys. Chem. A* 102, 2177–2188.
72. Hussain, M.S., Schlemper, E.O., & Fair, C.K. (1980) *Acta Crystallogr., Sect. B* 36, 1104–1108.
73. Jones, R.D.G., & Power, L.F. (1976) *Acta Crystallogr., Sect. B* 32, 1801–1806.
74. Iijima, K., Ohnogi, A., & Shibata, S. (1987) *J. Mol. Struct.* 156, 111–118.
75. Wozniak, K., He, H.Y., Klinowski, J., Barr, T.L., & Milart, P. (1996) *J. Phys. Chem.* 100, 11420–11426.
76. Markley, J.L., & Westler, W.M. (1996) *Biochemistry* 35, 11092–11097.
77. Edison, A.S., Weinhold, F., & Markley, J.L. (1995) *J. Am. Chem. Soc.* 117, 9619–9624.
78. Kreevoy, M.M., & Liang, T.M. (1980) *J. Am. Chem. Soc.* 102, 3315–3322.
79. Chiang, Y., Kresge, A.J., & More O’Ferrall, R.A. (1980) *J. Chem. Soc., Perkin Trans. 2*, 1832–1839.
80. Perrin, C.L., & Nielson, J.B. (1997) *Annu. Rev. Phys. Chem.* 48, 511–544.
81. Baltzer, L., & Bergman, N.A. (1982) *J. Chem. Soc., Perkin Trans. 2*, 313–319.
82. Taft, R.W., Gurka, D., Joris, L., Schleyer, P., & Rakshys, J.W. (1969) *J. Am. Chem. Soc.* 91, 4801–4808.
83. Arnett, E.M., Mitchell, E.J., & Murty, T.S.S.R. (1974) *J. Am. Chem. Soc.* 96, 3875–3891.
84. Stymne, B., Stymne, H., & Wettermark, G. (1973) *J. Am. Chem. Soc.* 95, 3490–3494.
85. Arnett, E.M. (1963) *Prog. Phys. Org. Chem.* 1, 223–403.
86. Rubin, J., Senkowski, B.Z., & Panson, G.S. (1964) *J. Phys. Chem.* 68, 1601–1602.
87. Shan, S.O., Loh, S., & Herschlag, D. (1996) *Science* 272, 97–101.
88. Arnett, E.M. (1963) *Prog. Phys. Org. Chem.* 1, 223–403.
89. Gordy, W., & Stanford, S.C. (1941) *J. Chem. Phys.* 9, 204–214.
90. Tobin, J.B., Whitt, S.A., Cassidy, C.S., & Frey, P.A. (1995) *Biochemistry* 34, 6919–6924.
91. Singh, U.C., & Kollman, P.A. (1985) *J. Chem. Phys.* 83, 4033–4040.
92. Cotton, F.A., Fair, C.K., Lewis, G.E., Mott, G.N., Ross, F.K., Schultz, A.J., & Williams, J.M. (1984) *J. Am. Chem. Soc.* 106, 5319–5323.
93. Harrowfield, J.M., Sharma, R.P., Skelton, B.W., & White, A.H. (1998) *Aust. J. Chem.* 51, 785–793.
94. Hughes, D.L., & Truter, M.R. (1979) *J. Chem. Soc., Dalton Trans.*, 520–527.
95. Kanters, J.A., Ter Horst, E.H., & Grech, E. (1992) *Acta Crystallogr., Sect. C* 48, 1345–1347.
96. Aleksandrov, G.G., Struchkov, Y.T., Kalinin, A.E., Shcherbakov, A.A., Barykina, L.R., & Karaulova, E.N. (1980) *Kristallografiya* 25, 481–487.
97. Jones, P.G., & Ahrens, B. (1998) *Eur. J. Org. Chem.*, 1687–1688.
98. Hashimoto, M., & Iwamoto, T. (1991) *J. Coord. Chem.* 23, 269–276.
99. Rivas, J.C.M., & Brammer, L. (1998) *Acta Crystallogr., Sect. C* 54, 1799–1802.
100. Therrien, B., & Beauchamp, A.L. (1993) *Acta Crystallogr., Sect. C* 49, 1303–1307.
101. McAdam, A., Currie, M., & Speakman, J.C. (1971) *J. Chem. Soc., A*, 1994–1997.
102. Pei, X.F., Greig, N.H., Flippenanderson, J.L., Bi, S., & Brossi, A. (1994) *Helv. Chim. Acta* 77, 1412–1422.
103. Gupta, M.P., & Ashok, J. (1978) *Cryst. Struct. Commun.* 7, 171–174.
104. Schwartz, A., Madan, P.B., Mohacsi, E., O’Brien, J.P., Todaro, L.J., & Coffen, D.L. (1992) *J. Org. Chem.* 57, 851–856.
105. Amstutz, R., Enz, A., Marzi, M., Boelsterli, J., & Walkinshaw, M. (1990) *Helv. Chim. Acta* 73, 739–753.

106. Philippopoulos, A.I., Bau, R., Poilblanc, R., & Hadjiliadis, N. (1998) *Inorg. Chem.* 37, 4822–4827.
107. Schwartz, B., & Drueckhammer, D.G. (1995) *J. Am. Chem. Soc.* 117, 11902–11905.
108. Kato, Y., Toledo, L.M., & Rebek, J. (1996) *J. Am. Chem. Soc.* 118, 8575–8579.
109. Ash, E.L., Sudmeier, J.L., De Fabo, E.C., & Bachovchin, W.W. (1997) *Science* 278, 1128–1132.
110. Lin, J., & Frey, P.A. (2000) *J. Am. Chem. Soc.* 122, 11258–11259.
111. Isaacs, E.D., Shukla, A., Platzman, P.M., Hamann, D.R., Barbiellini, B., & Tulk, C.A. (1999) *Phys. Rev. Lett.* 82, 600–603.
112. Martin, T.W., & Derewenda, Z.S. (1999) *Nat. Struct. Biol.* 6, 403–406.
113. Ghanty, T.K., Staroverov, V.N., Koren, P.R., & Davidson, E.R. (2000) *J. Am. Chem. Soc.* 122, 1210–1214.
114. Stahl, N., & Jencks, W.P. (1986) *J. Am. Chem. Soc.* 108, 4196–4205.
115. Pauling, L., & Pressman, D. (1945) *J. Am. Chem. Soc.* 67, 1003–1012.
116. Klotz, I.M., & Franzen, J.S. (1962) *J. Am. Chem. Soc.* 84, 3461–3466.
117. Kresheck, G.C., & Klotz, I.M. (1969) *Biochemistry* 8, 8–12.
118. Klotz, I.M., & Farnham, S.B. (1968) *Biochemistry* 7, 3879–3882.
119. Roseman, M.A. (1988) *J. Mol. Biol.* 201, 621–623.
120. Kyte, J. (2003) *Biophys. Chem.* 100, 193–203.
121. Hine, J.S. (1962) *Physical Organic Chemistry*, 2nd ed., pp 81–103, McGraw-Hill, New York.
122. Kresge, A.J., & Chiang, Y. (1973) *J. Phys. Chem.* 77, 822–825.
123. Doig, A.J., & Williams, D.H. (1992) *J. Am. Chem. Soc.* 114, 338–343.
124. Badger, R.M., & Rubalcava, H. (1954) *Proc. Natl. Acad. Sci. U.S.A.* 40, 12–17.
125. Bruice, T.C., & Pandit, U.K. (1960) *J. Am. Chem. Soc.* 82, 5858–5865.
126. Bruice, T.C. (1970) in *The Enzymes: Kinetics and Mechanism*, 3rd ed. (Boyer, P. D., Ed.) Vol. II, pp 217–279, Academic Press, New York.
127. Bruice, T.C., & Turner, A. (1970) *J. Am. Chem. Soc.* 92, 3422–3428.
128. Page, M.I., & Jencks, W.P. (1971) *Proc. Natl. Acad. Sci. U.S.A.* 68, 1678–1683.
129. Morris, J.J., & Page, M.I. (1980) *J. Chem. Soc., Perkin Trans. 2*, 679–684.
130. Page, M.I., & Jencks, W.P. (1987) *Gazz. Chim. Ital.* 117, 455–460.
131. Tadayoni, B.M., Parris, K., & Rebek, J., Jr. (1989) *J. Am. Chem. Soc.* 111, 4503–4505.
132. Tadayoni, B.M., Huff, J., & Rebek, J., Jr. (1991) *J. Am. Chem. Soc.* 113, 2247–2253.
133. Higuchi, T., Ebersson, L., & Herd, A.K. (1966) *J. Am. Chem. Soc.* 88, 3805–3808.
134. Roberts, J.D., Chun, Y., Flanagan, C., & Birdseye, T.R. (1982) *J. Am. Chem. Soc.* 104, 3945–3949.
135. Etzkorn, F.A., Guo, T., Lipton, M.A., Goldberg, S.D., & Bartlett, P.A. (1994) *J. Am. Chem. Soc.* 116, 10412–10425.
136. Karle, I., & Karle, J. (1963) *Acta Crystallogr.* 16, 969–975.
137. Venkatachalam, C.M. (1968) *Biopolymers* 6, 1425–1436.
138. Zimm, B.H., & Bragg, J.K. (1959) *J. Chem. Phys.* 31, 526–535.
139. Gellman, S.H. (1998) *Curr. Opin. Chem. Biol.* 2, 717–725.
140. Searle, M.S., Griffiths-Jones, S.R., & Skinner-Smith, H. (1999) *J. Am. Chem. Soc.* 121, 11615–11620.
141. Fisk, J.D., & Gellman, S.H. (2001) *J. Am. Chem. Soc.* 123, 343–344.
142. Bierzynski, A., Kim, P.S., & Baldwin, R.L. (1982) *Proc. Natl. Acad. Sci. U.S.A.* 79, 2470–2474.
143. Austin, R.E., Maplestone, R.A., Sefler, A.M., Liu, K., Hruzewicz, W.N., Liu, C.W., Cho, H.S., Wemmer, D.E., & Bartlett, P.A. (1997) *J. Am. Chem. Soc.* 119, 6461–6472.
144. Marqusee, S., & Baldwin, R.L. (1987) *Proc. Natl. Acad. Sci. U.S.A.* 84, 8898–8902.
145. Merutka, G., & Stellwagen, E. (1989) *Biochemistry* 28, 352–357.
146. Merutka, G., & Stellwagen, E. (1990) *Biochemistry* 29, 894–898.
147. Scholtz, J.M., Qian, H., Robbins, V.H., & Baldwin, R.L. (1993) *Biochemistry* 32, 9668–9676.
148. Zhou, N.E., Kay, C.M., Sykes, B.D., & Hodges, R.S. (1993) *Biochemistry* 32, 6190–6197.
149. Smith, J.S., & Scholtz, J.M. (1998) *Biochemistry* 37, 33–40.
150. Miller, J.S., Kennedy, R.J., & Kemp, D.S. (2001) *Biochemistry* 40, 305–309.
151. Artymiuk, P.J., & Blake, C.C. (1981) *J. Mol. Biol.* 152, 737–762.
152. Oefner, C., & Suck, D. (1986) *J. Mol. Biol.* 192, 605–632.
153. Pauling, L., Corey, R.B., & Branson, H.R. (1951) *Proc. Natl. Acad. Sci. U.S.A.* 37, 205–211.
154. Huyghues-Despointes, B.M., & Baldwin, R.L. (1997) *Biochemistry* 36, 1965–1970.
155. Huyghues-Despointes, B.M., Klingler, T.M., & Baldwin, R.L. (1995) *Biochemistry* 34, 13267–13271.
156. Tanaka, Y., Kato, Y., & Aoyama, Y. (1990) *J. Am. Chem. Soc.* 112, 2807–2808.
157. Nowick, J.S., Chen, J.S., & Noronha, G. (1993) *J. Am. Chem. Soc.* 115, 7636–7644.
158. Morales, J.C., & Kool, E.T. (1998) *Nat. Struct. Biol.* 5, 950–954.
159. Ogawa, A.K., Wu, Y., McMinin, D.L., Liu, J., Schultz, P.G., & Romesberg, F.E. (2000) *J. Am. Chem. Soc.* 122, 3274–3287.
160. Wu, Y., Ogawa, A.K., Berger, M., McMinin, D.L., Schultz, P.G., & Romesberg, F.E. (2000) *J. Am. Chem. Soc.* 122, 7621–7632.
161. Guckian, K.M., Krugh, T.R., & Kool, E.T. (2000) *J. Am. Chem. Soc.* 122, 6841–6847.
162. Dzantiev, L., Alekseyev, Y.O., Morales, J.C., Kool, E.T., & Romano, L.J. (2001) *Biochemistry* 40, 3215–3221.
163. Kato, Y., Conn, M.M., & Rebek, J., Jr. (1995) *Proc. Natl. Acad. Sci. U.S.A.* 92, 1208–1212.
164. Krugh, T.R., & Young, M.A. (1975) *Biochem. Biophys. Res. Commun.* 62, 1025–1031.
165. Ogasawara, N., & Inoue, Y. (1976) *J. Am. Chem. Soc.* 98, 7054–7060.
166. Lyu, P.C., Marky, L.A., & Kallenbach, N.R. (1989) *J. Am. Chem. Soc.* 111, 2733–2734.

167. Melville, H. (1979) *Moby-Dick or, The Whale*, Chapter 84, University of California Press, Berkeley, CA.
168. Hartley, G.S. (1936) in *Aqueous Solutions of Paraffin-Chain Salts* pp viii, Hermann & Cie., Paris, as quoted in Tanford, C. (1973) *The Hydrophobic Effect: Formation of Micelles and Biological Membranes*, p viii, John Wiley and Sons, New York.
169. Hildebrand, J.H. (1979) *Proc. Natl. Acad. Sci. U.S.A.* 76, 194.
170. Tanford, C. (1979) *Proc. Natl. Acad. Sci. U.S.A.* 76, 4175–4176.
171. Kauzmann, W. (1959) *Adv. Protein Chem.* 14, 1–63.
172. Kauzmann, W. (1954) in *The Mechanism of Enzyme Action* (McElroy, W. D., & Glass, B., Eds.) pp 70–120, Johns Hopkins Press, Baltimore, MD.
173. Scatena, L.F., Brown, M.G., & Richmond, G.L. (2001) *Science* 292, 908–912.
174. Abraham, M.H., Grellier, P.L., & McGill, R.A. (1987) *J. Chem. Soc. Perkin Trans. 2*, 797.
175. Abraham, M.H. (1993) *Chem. Soc. Rev.* 22, 73–83.
176. Cohn, E.J., McMeekin, T.L., Edsall, J.T., & Weare, J.H. (1934) *J. Am. Chem. Soc.* 56, 2270–2282.
177. Tanford, C. (1980) *The Hydrophobic Effect: Formation of Micelles and Biological Membranes*, 2nd ed., Wiley, New York.
178. Franks, F., & Reid, D.S. (1973) in *Water: A Comprehensive Treatise, Water in Crystalline Hydrates: Volume 2, Aqueous Solutions of Simple Non-electrolytes* (Franks, F., Ed.) pp 323–380, Plenum Press, New York.
179. Arnett, E.M., Kover, W.B., & Carter, J.V. (1969) *J. Am. Chem. Soc.* 91, 4028–4034.
180. Edsall, J.T., & Scatchard, G. (1943) in *Proteins, Amino Acids and Peptides as Ions and Dipolar Ions* (Edsall, J. T., & Cohn, E. J., Eds.) pp 177–195, Reinhold, New York.
181. Privalov, P.L., & Gill, S.J. (1988) *Adv. Protein Chem.* 39, 191–234.
182. Edsall, J.T. (1935) *J. Am. Chem. Soc.* 57, 1506–1507.
183. Frank, H.S., & Evans, M.W. (1945) *J. Chem. Phys.* 13, 507–532.
184. Pertsemlidis, A., Saxena, A.M., Soper, A.K., Head-Gordon, T., & Glaeser, R.M. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 10769–10774.
185. Hafemann, D.R., & Miller, S.L. (1969) *J. Phys. Chem.* 73, 1392–1397.
186. Graziano, G., & Barone, G. (1996) *J. Am. Chem. Soc.* 118, 1831–1835.
187. Haggis, G.H., Hasted, J.B., & Buchanan, T.J. (1952) *J. Chem. Phys.* 20, 1452–1465.
188. Lumry, R., & Biltonen, R. (1969) in *The Structure and Stability of Biological Macromolecules* (Timasheff, S.N., & Fasman, G.D., Eds.) pp 65–212, Marcel Dekker, New York.
189. Southall, N.T., Dill, K.A., & Haymet, A.D.J. (2002) *J. Phys. Chem., B* 106, 521–533.
190. Abraham, M.H., Chadha, H.S., Whiting, G.S., & Mitchell, R.C. (1994) *J. Pharm. Sci.* 83, 1085–1100.
191. McMeekin, T.L., Cohn, E.J., & Weare, J.H. (1935) *J. Am. Chem. Soc.* 57, 626–633.
192. Tanford, C. (1973) *The Hydrophobic Effect: Formation of Micelles and Biological Membranes*, Wiley, New York.
193. Pace, C.N. (1992) *J. Mol. Biol.* 226, 29–35.
194. Cramer, R.D. (1977) *J. Am. Chem. Soc.* 99, 5408–5412.
195. Abraham, M.H. (1979) *J. Am. Chem. Soc.* 101, 5477–5484.
196. Hine, J., & Mookerjee, P.K. (1975) *J. Org. Chem.* 40, 292–298.
197. Makhatadze, G.I., & Privalov, P.L. (1993) *J. Mol. Biol.* 232, 639–659.
198. Privalov, P.L., & Makhatadze, G.I. (1993) *J. Mol. Biol.* 232, 660–679.
199. Wolfenden, R., Andersson, L., Cullis, P.M., & Southgate, C.C. (1981) *Biochemistry* 20, 849–855.
200. Kinoshita, K., Ishikawa, H., & Shinoda, K. (1958) *Bull. Chem. Soc. Jpn.* 31, 1081–1082.
201. Wimley, W.C., & White, S.H. (1992) *Biochemistry* 31, 12813–12818.
202. Fauchere, J.L., & Pliska, V. (1983) *Eur. J. Med. Chem., Chim. Ther.* 18, 369–375.
203. Wolfenden, R.V., Cullis, P.M., & Southgate, C.C. (1979) *Science* 206, 575–577.
204. Wolfenden, R. (1978) *Biochemistry* 17, 201–204.
205. Wang, Y., Purrello, R., Georgiou, S., & Spiro, T.G. (1991) *J. Am. Chem. Soc.* 113, 6368–6377.
206. Davies, M., Evans, J.C., & Jones, R.L. (1955) *Trans. Faraday Soc.* 51, 761–774.
207. Nozaki, Y., & Tanford, C. (1971) *J. Biol. Chem.* 246, 2211–2217.
208. Traube, J. (1891) *Justus Liebigs Ann. Chem.* 265, 27.
209. Bull, H.B., & Breese, K. (1974) *Arch. Biochem. Biophys.* 161, 665–670.
210. Segrest, J.P., & Feldmann, R.J. (1974) *J. Mol. Biol.* 87, 853–858.
211. Hansch, C.H., & Leo, A. (1979) *Substituent Constants for Correlation Analysis in Chemistry and Biology*, Wiley, New York.
212. Lawson, E.Q., Sadler, A.J., Harmatz, D., Brandau, D.T., Micanovic, R., MacElroy, R.D., & Middaugh, C.R. (1984) *J. Biol. Chem.* 259, 2910–2912.
213. Hermann, R.B. (1972) *J. Phys. Chem.* 76, 2754–2759.
214. Lee, B., & Richards, F.M. (1971) *J. Mol. Biol.* 55, 379–400.
215. Chothia, C. (1976) *J. Mol. Biol.* 105, 1–14.
216. Janin, J. (1979) *Nature* 277, 491–492.
217. Guy, H.R. (1985) *Biophys. J.* 47, 61–70.
218. von Heijne, G., & Blomberg, C. (1979) *Eur. J. Biochem.* 97, 175–181.
219. Eisenberg, D., Weiss, R.M., Terwilliger, T.C., & Wilcox, W. (1982) *Faraday Symp. Chem. Soc.*, 109–120.
220. Engelman, D.M., Steitz, T.A., & Goldman, A. (1986) *Annu. Rev. Biophys. Biophys. Chem.* 15, 321–353.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Chapter 6

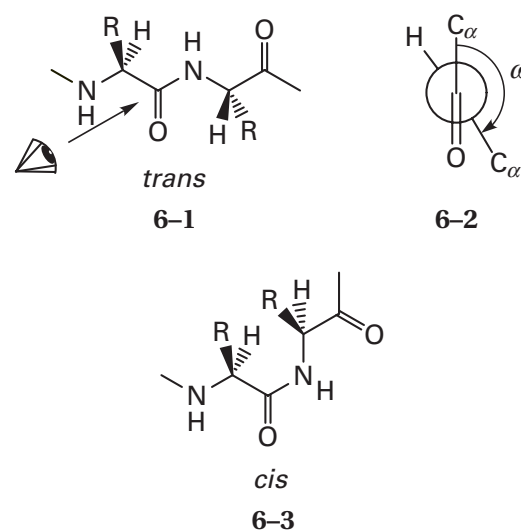
Atomic Details

It is within the crystallographic molecular model of a protein that the consequences of the noncovalent forces just described can be viewed. As the polypeptide folded to form the actual native structure of the protein that is represented by the model, a significant fraction of its backbone had to be withdrawn from water. Secondary structure formed because it offered an efficient way to maintain the total number of hydrogen bonds in the solution. As the polypeptide folded, the donors and acceptors of hydrogen bonds in the side chains of the amino acids, as well as any ionized side chains, remained, by and large, on the surface of the structure to maintain their hydration. The networks of hydrogen-bonded water molecules hydrating the surface of the folded structure are prominent features of the crystallographic molecular model. The core of the model is formed mostly from hydrophobic amino acids, the exclusion of which from the water drove the folding process.

As remarkable as these energetic consequences are in a crystallographic molecular model, it has become clear upon close inspection that they do not, except indirectly, produce the final structure. The most important determinant of the final native structure is the steric effect. In retrospect, this should not be so surprising. Steric effects are always the most overwhelming among the different forces influencing the outcome of a chemical reaction. They are rarely discussed at great length because they are so easy to understand. No two fragments of matter may occupy the same place at the same time. The folding of a polypeptide, however, is a steric nightmare. Not only must all of the functional groups fit together in a confined space without overlapping, but all of the functional groups are connected together by the polypeptide. Although the outcome of any one of these games of packing atoms cannot be predicted, the ultimate solutions to the vast array of steric problems encountered during each game can be appreciated by examining the final native structure. Both the steric effects operating along the polypeptide backbone and those engendered between the side chains as the elements of secondary structure attempt to fit together to produce the final native structure of the protein are represented by the crystallographic molecular model. These steric effects and the noncovalent forces are the players in each of the games.

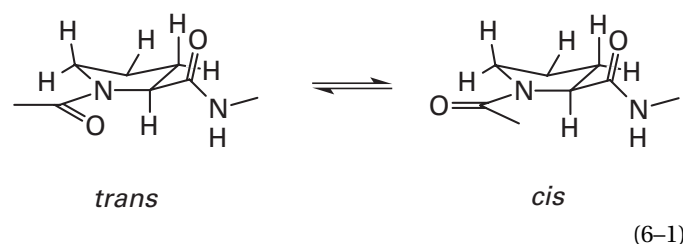
Secondary Structure of the Polypeptide Backbone

Because of its π molecular orbital system (Figure 2-3), the amide of the **peptide bond** is planar. The dihedral angle ω assigned to this amide is defined by looking down the carbon–nitrogen bond from carbon to nitrogen:



The sign of the angle is determined by the right-hand rule. In *trans* peptide bonds, the angle ω is 180° , and the two strands of polypeptide leaving the amide depart in opposite directions from the carbon–nitrogen bond. In *cis* peptide bonds (6-3), the angle ω is 0° , and the two α carbons of the two departing strands are eclipsed. In a protein, at any particular location in the amino acid sequence, the peptide bond is either *trans* in every molecule of that particular protein or *cis* in every molecule of that particular protein. The reason for this is that the stereochemical difference between these two conformers is substantial, and only one will be compatible with the structure at that location in the folded polypeptide.

In the available crystallographic molecular models, only 0.3% of the peptide bonds¹ are ***cis* peptide bonds** (Figure 6-1),² and most of the locations where a *cis* peptide bond is found (87%) have proline as the amino acid on the carboxy-terminal side:¹

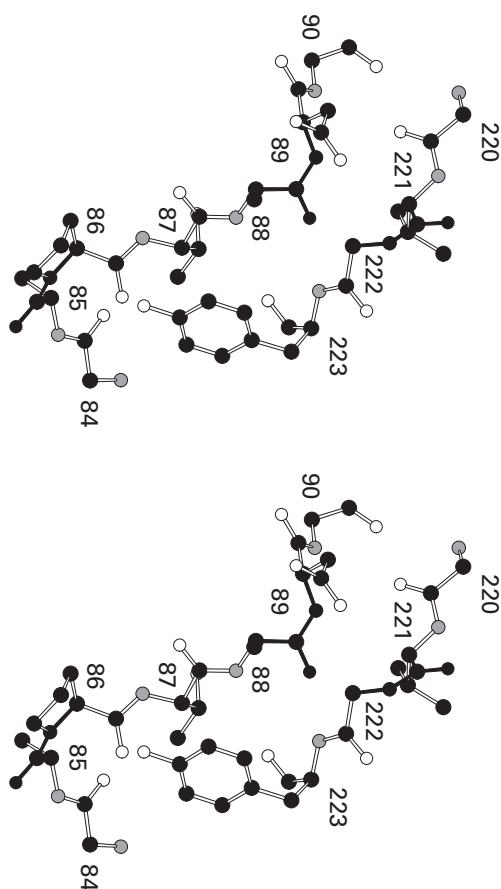


In such peptide bonds, the proline provides the amido nitrogen; consequently, the amide formed is secondary. In this secondary amide, there is little preference for the *trans* stereochemistry because both positions on the amido nitrogen are sterically similar, and the equilibrium constants between the *cis* and *trans* conformations for peptide bonds involving proline are between 0.1 and 1.³ In proteins, about 6% of the peptide bonds in which proline provides the amido nitrogen are *cis* peptide bonds.¹ Unlike proline, the other amino acids form primary amides. Because the hydrogen of a primary amide is much smaller than the alkyl substituent, the equilibrium constant heavily favors the *trans* form. *cis* Peptide bonds in which an amino acid other than proline provides the amido nitrogen are rare (0.04% of all peptide bonds). These are presumably locations where a *cis* peptide bond is unavoidable, but evolution by natural selection has not yet replaced the carboxy-terminal amino acid of that peptide bond with a proline. Other than proline at its carboxy-terminal side, there seems to be little preference for the other amino acids at either location in a *cis* peptide bond.^{1,4} In the rare instances in which there is a cystine between two adjacent cysteines, the peptide bond between them is necessarily *cis*.⁵

Aside from the occasional situations in which *cis* peptide bonds occur, the peptide bonds in crystallographic molecular models are *trans*. In a set of eight crystallographic molecular models,⁶ all built from data sets to Bragg spacings of less than 0.12 nm, the values of the angle ω equal $179^\circ \pm 6^\circ$. Observations from a much larger set of crystallographic molecular models give the same result.⁷ Deviations from 180° as large as 15° have been observed.^{7,8} Were the amide, however, to be distorted too far from planarity by the structure of the protein, it would become susceptible to rapid hydrolysis.⁹

It was noted by Ramachandran, Ramakrishnan, and Sasisekharan,¹⁰ before crystallographic molecular models of atomic resolution became available, that there are severe **steric effects** hindering rotation about the single bonds of a polypeptide. The dihedral angle of the bond between an α carbon and the adjacent amido nitrogen is designated as ϕ ; and that of the bond between an α carbon and the adjacent acyl carbon, as ψ (Figures 6-2 and 6-3).¹⁰ Each amino acid in the protein is assigned a dihedral angle ϕ and a dihedral angle ψ associated with its α carbon. Although it has been noted that it would be more reasonable to assign values of the **dihedral angles ϕ and ψ** to each peptide bond rather than to each amino

Figure 6-1: *cis* Peptide bonds in the crystallographic molecular model (Bragg spacing ≥ 0.20 nm) of lectin IV of *Griffonia simplicifolia*.² By chance, this protein happens to have three *cis* peptide bonds (highlighted with black bonds and atoms) near each other. Only one of these contains a proline (Proline 87). In the other two, the position normally taken by the proline in the *cis* bond is occupied by a glycine (Glycine 222) and an aspartate (Aspartate 89), respectively. In this figure and those that follow, the amino acids are numbered according to their positions in the amino acid sequences in the data base published by the Swiss Institute of Bioinformatics (us.expasy.org). These numbers often differ from those of the preliminary amino acid sequences used by the crystallographers to construct the crystallographic molecular models. This drawing was produced with MolScript.⁵⁷³



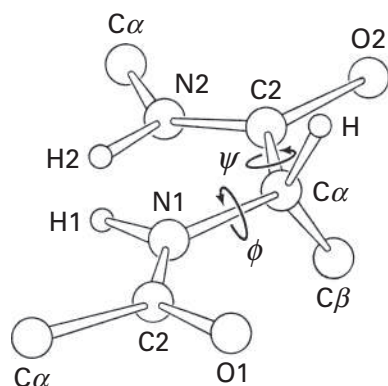
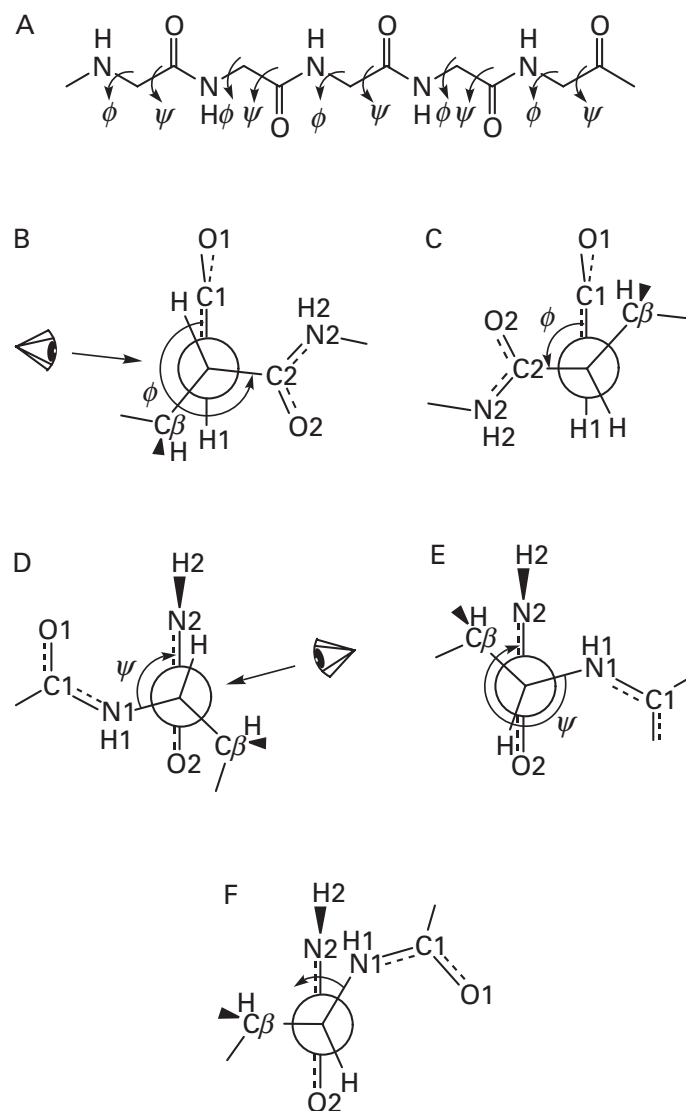


Figure 6-2: Designation of the two dihedral angles ϕ and ψ to the amino-terminal and carboxy-terminal sides, respectively, of the α carbon ($C\alpha$) of an amino acid in a polypeptide.¹⁰ The first carbon of a side chain ($C\beta$) corresponds to that of amino acids in the L-configuration. Adapted with permission from ref 10. Copyright 1963 Academic Press.



acid,¹¹ this suggestion has yet to be adopted. The signs of the dihedral angles ϕ and ψ are determined by the right-hand rule (Figure 6-3). To follow what is about to be described, you should build a model of the structure shown in Figure 6-2.

When the view from an α carbon down the bond to an amido nitrogen is observed, it can be seen that in every *trans* peptide bond, the acyl oxygen O1 of the previous amino acid in the polypeptide leans forward (Figure 6-3B,C). When the dihedral angle ϕ is greater than $+310^\circ$ (-50°), the amido N2-C2-O2 collides with acyl oxygen O1, and when the dihedral angle ϕ is less than $+180^\circ$, the H-C β -C γ of the side chain collides with acyl oxygen O1. Therefore, values of dihedral angle ϕ greater than $+310^\circ$ (-50°) and less than $+180^\circ$, with the exception of a small gap at angle ϕ of $+60^\circ$ (Figure 6-3C), are forbidden.

When the view from an α carbon down the bond to an acyl carbon is observed, it is clear that in all *trans* peptide bonds the hydrogen H2 on the amido nitrogen of the next amino acid leans forward (Figure 6-3D,E). When the dihedral angle ψ is greater than $+200^\circ$ (-160°), the H-C β -C γ of the side chain collides with this hydrogen, and when dihedral angle ψ is between $+300^\circ$ (-60°) and $+30^\circ$, the amido N1-C1-O1 collides with this hydrogen (Figure 6-3D,E). The latter collision is not a serious one so long as the value of dihedral angle ϕ remains around 270° (-90°) or $+90^\circ$ so the amido N1-C1-O1 can squeeze past the hydrogen H2 sideways (Figure 6-3C,F), but values for dihedral angle ϕ of $+90^\circ$ are only permitted for glycine, which lacks a β carbon. When dihedral angle ψ is between $+290^\circ$

Figure 6-3: Definitions of the dihedral angles ϕ and ψ and the steric effects of rotation. (A) Pattern in which the bonds with the dihedral angles ϕ and ψ are distributed along a polypeptide. (B) View down the bond between $C\alpha$ and the amido nitrogen N1 that precedes it along the polypeptide. The dihedral angle ϕ is defined as that between the bond connecting N1 and C1 and the bond connecting $C\alpha$ and C2 (Figure 6-2). Its sign is determined by the right-hand rule. Note that the direction of the arrow is irrelevant to the assignment of the sign of the angle. In the configuration shown, angle ϕ is $+260^\circ$ (-100°). This dihedral angle is in the most sterically free range for angle ϕ (-45° to -180°) because in this range the hydrogen on $C\alpha$ can slip under the acyl oxygen O1. (C) View down the same bond as in panel B but with angle ϕ at $+90^\circ$, produced from the conformation in panel B by rotation about only the bond on the axis of the view. When angle ϕ is $+60^\circ$, the acyl oxygen, O1, sits between the carbon of the next peptide bond at C2 and the first carbon of the side chain, $C\beta$. This would be the value of angle ϕ in a left-handed α helix. (D) The same conformation presented in panel B is viewed along the bond between $C\alpha$ and the acyl carbon C2. The eyes indicate the views interconverting panels B and D. The dihedral angle ψ is defined as that between the bond connecting $C\alpha$ and N1 and the bond connecting C2 and N2 (Figure 6-2). Its sign is determined by the right-hand rule. The configuration shown ($\psi = +105^\circ$) is in the most sterically free range for angle ψ ($+15^\circ$ to $+190^\circ$) because the hydrogen on $C\alpha$ can slip below H2. (E) View down the same bond as in panel D but with angle ψ at $+285^\circ$ (-75°). When angle ψ is $+300^\circ$ (-60°), H2 lies between the first carbon of the side chain, $C\beta$, and the nitrogen of the amino-terminal peptide bond, N1. This is near the value of angle ψ (-39°) in a right-handed α helix. (F) Steric effect between H2 and either N1 or H1 that occurs when angle ψ is near 0° .

(-70°) and $+320^\circ$ (-40°), hydrogen H2 on the amido nitrogen N2 can fit between amido nitrogen N1 and the side chain with little difficulty (Figure 6-3E).

All of these steric effects can be summarized in a **Ramachandran plot** (Figure 6-4A).¹² Using your model, you should verify the noted boundaries on the plot.

Refinements of crystallographic molecular models by use of Equation 4-15 for calculation of the function θ usually do not constrain the values for dihedral angles ϕ and ψ . Even though they are not constrained, however, their values converge upon the **allowed regions** in a Ramachandran plot during the refinement. For example, although many of the values for dihedral angles ϕ and ψ for the various amino acids along the polypeptide in the initial molecular model of deoxyribonuclease I were scattered beyond the allowed regions in a Ramachandran plot before refinement was performed (Figure 6-5A), they clustered within the enclosures after the refinement had been completed (Figure 6-5B).⁸ Because this convergence was not enforced by the choice of $(d_{s,q}^2 - d_{c,q}^2)$ in Equation 4-15, its occurrence can be used as evidence that the refined structure is closer to reality than the unrefined.

When the dihedral angles for the amino acids in eight crystallographic molecular models, all built from data sets to Bragg spacings of less than 0.12 nm, are plotted on a Ramachandran plot (Figure 6-4B),⁶ the points themselves define what should be the **actual regions of lowest energy**. It might have been the case that the three clusters of open squares in Figure 6-4B are more the result of preferences enforced by secondary structures than the steric effects first pointed out by Ramachandran (Figure 6-4A). When, however, dihedral angles are plotted for amino acids not involved in secondary structures, from a much larger collection of crystallographic molecular models (402) but from data sets gathered to minimum Bragg spacings of only 0.2 nm or less, the distribution still shows the same three clusters with the same shapes and extents.^{13,14} Consequently, the extent and magnitude of the actual steric effects in the backbone of a polypeptide are delineated in the distribution of the points in a plot such as that in Figure 6-4B.

With the exception of the region where dihedral angle ϕ is between -70° and -180° and dihedral angle ψ is between 30° and 110° , which should be sterically unhindered anyway, almost all of the amino acids found

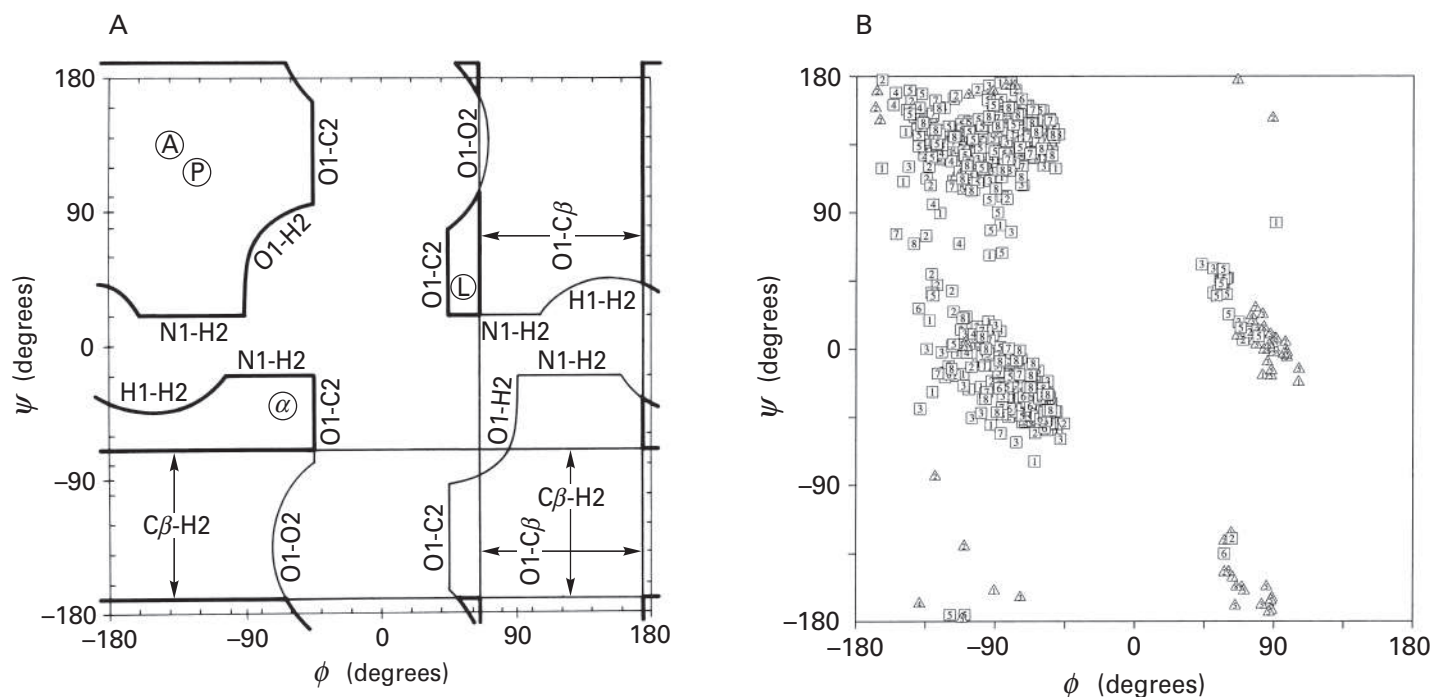


Figure 6-4: Ramachandran plot. (A) Diagram illustrating the steric effects producing the Ramachandran plot.¹² The two dimensions of the plot are the dihedral angles ψ and ϕ (Figure 6-3). Boundaries are drawn between allowed and forbidden regions obtained from a molecular model in which each atom is a hard sphere of the appropriate van der Waals radius. The clashing atoms are identified on the forbidden side of the boundary with the same numbering system as in Figures 6-2 and 6-3. There are only four allowed regions: the large region including the values for parallel (P) and antiparallel (A) β sheet, the region including the values for right-handed α helix (α), the region including the values for left-handed α helix (L), and a small triangle at $\phi = +60^\circ$ and $\psi = +180^\circ$. The clashes can be understood by referring to Figure 6-3. For example, if $\phi = -100^\circ$ and $\psi = +105^\circ$ (Figure 6-3B,D) and angle ϕ is increased to -45° , O1 clashes with C2; if angle ψ is decreased to $+20^\circ$, N1 clashes with H2. If $\phi = -60^\circ$ and $\psi = -60^\circ$ and angle ϕ is decreased to -185° , O1 runs into C β ; if angle ψ is decreased to -70° , H2 runs into C β . Adapted with permission from ref 12. Copyright 1977 *Journal of Biological Chemistry*. (B) Dihedral angles ψ and ϕ from eight crystallographic molecular models of high accuracy.⁶ The crystallographic molecular models and the minimum Bragg spacings of their data sets were cytochrome c_6 (0.12 nm), cutinase (0.10 nm), lysozyme (0.0925 nm), a fragment of protein G (0.11 nm), ribonuclease Sa (0.12 nm), repressor of primer protein (0.11 nm), rubredoxin from *Desulfovibrio vulgaris* (0.092 nm), and rubredoxin from *Clostridium pasteurianum* (0.11 nm). The numbers in the points indicate the respective models. Triangles are glycines, and squares are amino acids other than glycine.

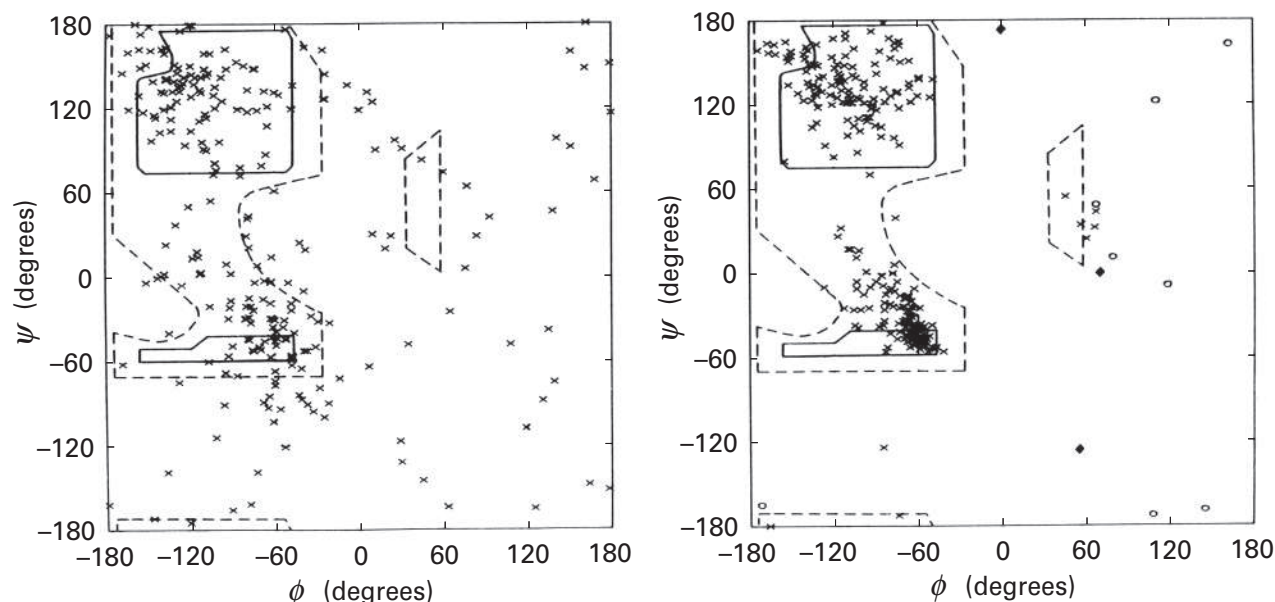


Figure 6-5: Effect of refinement on the values of ϕ and ψ for the amino acids in bovine deoxyribonuclease I.⁸ Each \times in one of the diagrams represents the values for the dihedral angles ϕ and ψ of one of the amino acids in a crystallographic molecular model of the protein. The boundaries in the Ramachandran plot are defined by the steric effects represented diagrammatically in Figure 6-4A. Unbroken lines surround regions of no hindrance; broken lines, regions of little hindrance. (A) Unrefined, initial molecular model. (B) Refined, final molecular model. Glycines are denoted by open circles, cystines by filled squares, and all other amino acids by \times . Reprinted with permission from ref 8. Copyright 1986 Academic Press.

outside of the three clusters of open squares in Figure 6-4B are **glycines**. In addition to being able to reside in cramped locations, glycine lacks a β carbon, and all of the steric collisions involving the β carbon (Figures 6-3 and 6-4A) are irrelevant. Therefore, the dihedral angles around a glycine have a larger compass. In particular, the regions of the Ramachandran plot in which dihedral angle ϕ lies between $+70^\circ$ and $+180^\circ$ or dihedral angle ψ lies between -70° and -180° represent areas where either O1 or H2, respectively, clash with the side chain of any amino acid other than glycine. In Figure 6-4B, the points for glycine (Δ) and the points for all other amino acids (\square) define distinct regions on the plot. In fact, most of the glycines in crystallographic molecular models have dihedral angles ϕ and ψ outside of the traditional enclosures on a Ramachandran plot defined by the dihedral angles ϕ and ψ of the other amino acids.^{15,16} This fact suggests that glycine is selected for situations in which such otherwise unpermitted dihedral angles are unavoidable.

It has been noted that when an amino acid other than glycine has angles ϕ and ψ outside of the enclosures, that amino acid is usually involved intimately in the function of the protein.⁴ For example, Serine 120 is the nucleophile in the active site of cutinase, and Alanine 30 is in the center of the crucial tight turn between the two α helices in the repressor of the primer protein.⁶

The region of the Ramachandran plot bounded by values of dihedral angle ϕ between -140° and -60° and values of dihedral angle ψ between -20° and $+20^\circ$ was originally predicted to be disallowed because when these dihedral angles are within these boundaries, either atom

N1 or atom H1 should be overlapping atom H2 (Figure 6-4A). Nevertheless, the dihedral angles of many amino acids fall within this supposedly disallowed region (Figure 6-4B). One way the overlap between either atom N1 or atom H1 and atom H2 can be prevented is to increase the bond angle $\text{N1-C}\alpha\text{-C2}$ beyond the usual 109.5° of a carbon hybridized sp^3 . In crystallographic molecular models this angle is observed to be wider¹⁷ than expected, with a mean of 112° and deviations up to 120° . In addition, this widening is dependent on the values for the dihedral angles ϕ and ψ . The angle $\text{N1-C}\alpha\text{-C2}$ is equal to 109° for β structure, the dihedral angles ψ and ϕ of which fall in the largest unhindered area of the plot, but is wider by 3° in α helices,¹⁸ the dihedral angles ψ and ϕ of which fall within the lower left cluster in Figure 6-4B immediately adjacent to the supposedly disallowed region. All of these results suggest that the existence of so many amino acids the dihedral angles ϕ and ψ of which fall within a region of the Ramachandran plot originally predicted to be disallowed results from a widening of this bond angle to accommodate the steric effect. The same argument would apply to the glycines the dihedral angles ϕ and ψ of which fall within the boundaries $+60^\circ$ to $+110^\circ$ and -20° to $+20^\circ$, respectively, also previously thought to be disallowed.

In a set of eight crystallographic molecular models, all built from data sets to Bragg spacings of less than 0.12 nm ,⁶ the amino acids within right-handed α helices (Figure 6-6)¹⁹ have dihedral angles of $\phi = -66^\circ \pm 13^\circ$ and $\psi = -39^\circ \pm 10^\circ$ (Figure 6-3E), and these values fall within one of the enclosures in a Ramachandran plot (Figure

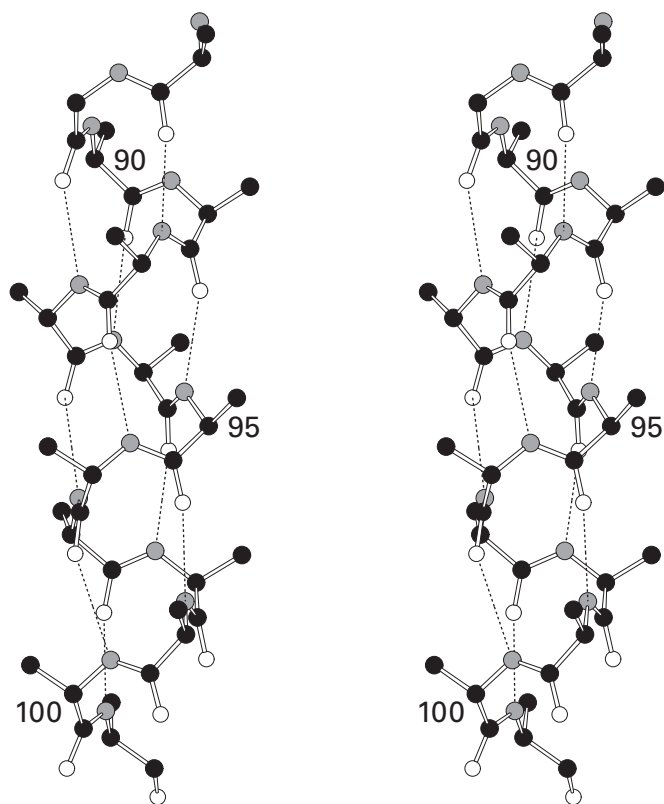


Figure 6-6: α Helix from the crystallographic molecular model (Bragg spacing ≥ 0.15 nm) of cytochrome *c* from *Thunnus alalunga*.¹⁹ Only the peptide backbone and the β carbons are included in the figure. The dihedral angles $\phi = -66^\circ$ (Figure 6-3B turned 34° further) and $\psi = -39^\circ$ (Figure 6-3E turned 36° further) of an α helix are most easily observed at amino acids 99 and 96, respectively. The hydrogen bonds are between the acyl oxygen of amino acid *i* and the amide nitrogen of amino acid *i* + 4. Compare this actual α helix with the one drawn in Figure 4-16A by pushing over the pages in between. This drawing was produced with MolScript.⁵⁷³

6-4A). This location suggests that there are no severe steric problems along the polypeptide in a right-handed α helix. A left-handed α helix of L-amino acids would have dihedral angles of $\phi = +65^\circ$ and $\psi = +40^\circ$, also within one of the enclosures (Figure 6-4A). This latter enclosure, however, is a small one arising from the conformation in which acyl oxygen O1 fits between the side chain and C2 and O2 of the next acyl group (Figure 6-3C). In crystallographic molecular models, about 2% of the amino acids other than glycine usually have dihedral angles ϕ and ψ clustered around this enclosure (Figure 6-4B). This fact demonstrates that these are accessible conformations, yet **left-handed α helices** are almost²⁰ never seen. It may be that an extended sequence of such conformers, which would be required to form a left-handed α helix, in contrast to the few isolated examples that are observed, would be sterically unstable.

The original **right-handed α helix** (Figure 4-16A), built before crystallographic structures were available, was constructed with 3.69 amino acids for every turn, which would have produced a rotational angle for each amino acid of 98° and a rise of 0.147 nm for each amino

acid.²¹ In crystallographic molecular models of proteins,^{8,22} in which these dimensions were not constrained, right-handed α helices display rotational angles for each amino acid of $99^\circ \pm 7^\circ$ and a rise for each amino acid of 0.15 ± 0.02 nm.

The paradigm of an α helix is a linear rod such as the one from cytochrome *c* in Figure 6-6, but only about 15% of those found in crystallographic molecular models of proteins are straight enough to fit the paradigm.²³ A significant proportion (60%) of α helices are **smoothly curved**. The original molecular model of the α helix was built so that each acyl oxygen along the α helix participates as an acceptor to only one hydrogen bond, namely, the one in which the nitrogen-hydrogen bond from the appropriate amide was the donor (Figure 4-16A). The tacit assumption was that the hydrogen bond would be one in which the nitrogen-hydrogen bond would pivot on the lone pairs of the acyl oxygen (Figure 5-10D). The geometric constraints of the α helix itself are such that the carbon-oxygen-nitrogen bond angle²⁴ is about $155^\circ \pm 10^\circ$ (Figure 6-6) rather than 180° . This causes the carbon-oxygen double bond to tilt outward²⁵ from the axis of an α helix (Figure 6-6). This tilt permits the oxygen to form a **second hydrogen bond** with a donor on a side chain of the protein (Figure 6-7A)^{24,26} or a molecule of water (Figure 6-7B).²⁷ This second hydrogen bond can be detected crystallographically (Figure 6-7B) or by a decrease in the frequency of the infrared absorption of the peptide bond.²⁸

The formation of this second hydrogen bond changes the carbon-oxygen-nitrogen bond angle by less than 10° ,^{24,29} and this fact suggests that each acyl oxygen in an α helix presents a second acceptor whether or not it is occupied. The occupation of these second acceptors, however, by donors located on only one side of the α helix, for example, the side facing the solvent (the right side of the α helix in Figure 6-7B), is thought to be a sufficient perturbation to cause the α helix to curve.^{23,29} One type of donor that can occupy these second acceptors on the acyl oxygens in an α helix are the hydroxyl groups on serines and threonines.

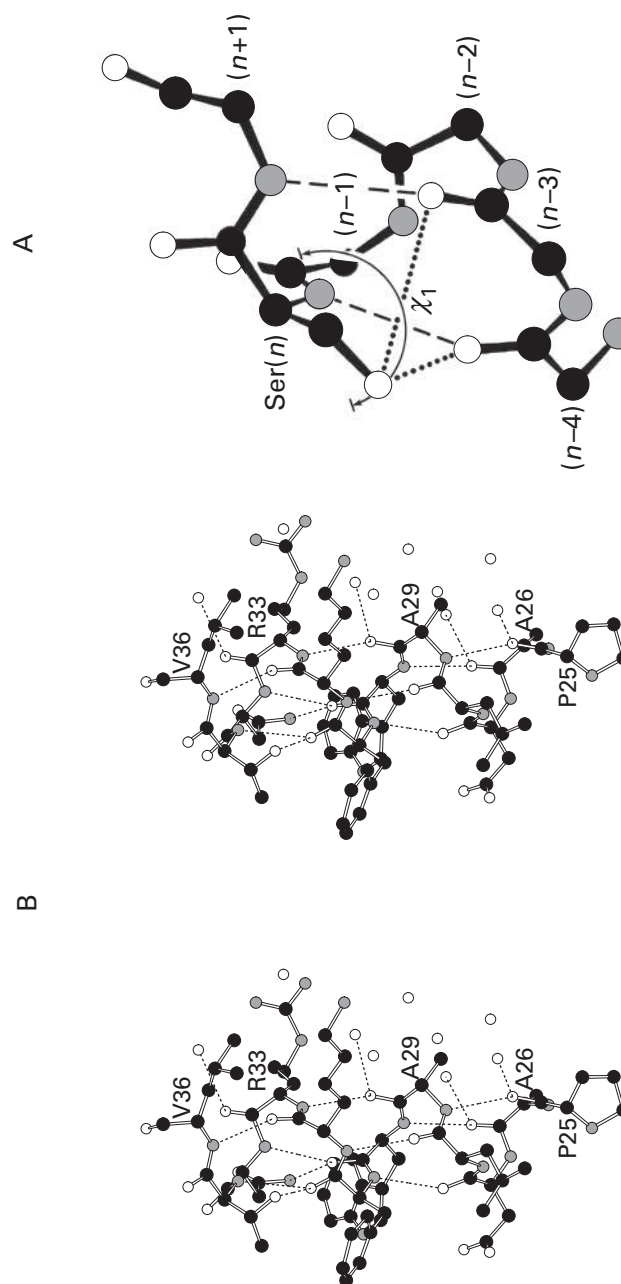
It has been observed that when an α helix contains a **serine or threonine**, the hydroxyl on the side chain has a tendency to be located in a position similar to that occupied by one of the waters in Figure 6-7B with respect to the acyl oxygens on the amino acid three or four positions ahead of that serine in the α helix (Figure 6-7A).²⁶ The proton on the hydroxyl group of the serine acts as a donor in a second hydrogen bond to one of these acyl oxygens just as a molecule of water does in the other situation. Threonine 35 in the α helix in Figure 6-7B participates in such a hydrogen bond with the acyl oxygen of Phenylalanine 31. This hydrogen bond occupies the donor on Threonine 35, which would otherwise be difficult to bury, and consequently turns its side chain from an apathetic one (Table 5-9) into a hydrophobic one, suitable to its surroundings. If Alanine 29 in Figure 6-7B

Figure 6-7: Occupation of the second hydrogen-bond acceptor on the acyl oxygen of a peptide bond in an α helix by the hydroxyl of a serine (A) or water (B). (A) Drawing of a molecular model of an α helix in which the hydroxyl of the side chain of a serine is shown acting as a donor of a hydrogen bond to the acyl oxygen of the amino acid either three or four positions ahead of it in the amino acid sequence.²⁶ The serine hydroxyl group can swing to complete a hydrogen bond to either acyl oxygen. Reprinted with permission from ref 26. Copyright 1984 Academic Press. (B) α Helix between Proline 25 and Valine 36 in the crystallographic molecular model (Bragg spacing ≈ 0.17 nm)²⁷ of dihydrofolate reductase from *E. coli*. The detached circles are positions of water molecules in the vicinity of the α helix. Four of the waters occupy locations consistent with the formation of hydrogen bonds to the adjacent acyl oxygens, those on Proline 25, Alanine 26, Alanine 29, and Arginine 33, in addition to the hydrogen bonds those oxygens accept from the appropriate amido nitrogens. The donor in the side chain of Threonine 35 forms a hydrogen bond with the acyl oxygen of Phenylalanine 31, and one of the donors in the side chain of Asparagine 34 forms a hydrogen-bond with the acyl oxygen of Tryptophan 30. This gradually curved α helix runs along the outer surface of the protein with its left side toward the interior and its right side toward the water. This drawing was produced with MolScript.⁵⁷³

were a serine, its hydroxyl would take the place of one of the waters in the respective hydrogen bonds to acyl oxygens on Proline 25 or Alanine 26. Such intramolecular hydrogen bonds are quite frequently encountered in α helices. In the molecular model of myoglobin, a protein with a large amount of α helix, 6 out of the 11 serines and threonines in the protein form hydrogen bonds with the acyl oxygens on amino acids three or four positions ahead of them in the sequence of the protein.³⁰ Asparagine can also participate in such an intrahelical hydrogen bond (Asparagine 34 in Figure 6-7B).

About 20% of the α helices in crystallographic molecular models are kinked.²³ The most common cause of an abrupt kink in an α helix is a **proline**. For example, Proline 183 in the middle of an α helix 30 amino acids long in citrate (*Si*)-synthase³¹ causes the α helix to bend abruptly by 40° . The mean value for the angles of the abrupt kinks produced in α helices by prolines is 26° .²³ In proteins where such a kink is found naturally, the presumption is that it serves the purpose of fitting the α helix properly into the overall structure. When a proline is inserted into an otherwise straight α helix by site-directed mutation, the α helix, if it tolerates the substitution, displays a kink with a much smaller angle, and the protein becomes significantly less stable.³²

There are also examples of **local distortions** in α helices that seem to be caused by the incompatibility of an undistorted α helix with the surrounding structure of the protein. If the α helix is too short, a gap develops in which molecules of water or donors and acceptors from side chains occupy the acceptors and donors broken by opening the gap;^{33,34} if the α helix is too long, one or more of its amino acids is pushed out of the structure as an aneurysm or loop.^{35,36}



At the **amino-terminal end** of an α helix there are unoccupied nitrogen-hydrogen donors, and at the carboxy-terminal end, unoccupied acyl oxygen acceptors (Figure 6-6). Because each peptide bond has two acceptors for hydrogen bonding but only one donor and because the side chains of the amino acids also have an excess of acceptors over donors, a solution of protein contains more acceptors than donors of hydrogen bonds. Consequently, when a donor remains unoccupied in the native structure, there was a loss of one hydrogen bond from the solution upon folding. Therefore, it comes as no surprise that the donors at the amino-terminal end of an α helix are occupied, or **capped**,^{26,37-40} but the acyl oxygens at the **carboxy-terminal end** are often capped as well.

About half of the time, the side chain of an amino acid such as asparagine, serine, threonine, or aspartate in the position immediately before the beginning of the α helix, the *N*-cap position, provides one or more of the necessary acceptors to occupy the open donors at the amino-terminal end (Figure 6-58). When such an amino acid is replaced by site-directed mutation with one that is of the same size or smaller but that cannot provide an acceptor, the resulting protein is less stable^{41,42} because a hydrogen bond is lost to the solution upon its folding that is not lost when the wild-type protein folds.⁴² Proline often (10%) occurs at the first position in an α helix³⁷ because it does not have a nitrogen-hydrogen donor that requires capping. About a third of the time, the amino acid immediately after the end of an α helix is a glycine, which can readily (Figure 6-4B) adopt the necessary dihedral angles ϕ and ψ ($+70^\circ$ and $+20^\circ$, respectively) that permit the amido nitrogen-hydrogen of the next amino acid to occupy the open acceptor of the first unoccupied acyl oxygen (that on the amino acid 98 in Figure 6-6) at the carboxy-terminal end of the helix.⁴³

The dipole moment of an isolated peptide bond is estimated to be 3.5 D,⁴⁴ and the peptide bonds in an α helix are held with their dipoles almost parallel to the axis (Figure 6-6) so that the positive poles point to the amino-terminal end of the α helix and the negative poles to the carboxy-terminal end. Such an arrangement of dipoles creates an electrostatic field of the respective polarity, the magnitude of which is 1 V at 0.3 nm and 0.5 V at 0.5 nm from each end of the α helix, if the α helix is greater than 10 amino acids long and located in a medium of relative permittivity equal to 2.⁴⁴ These voltages would produce electrostatic potentials equal to about 100 and 50 kJ mol⁻¹, respectively, for a univalent ion. Although these **electrostatic potentials** are less than twice those that would be felt at the same distances from two adjacent, isolated peptide bonds, it is thought that the amplification produced by aligning the peptide bonds in an α helix is significant.

Experimental observations equivocally consistent with this idea have been presented. For example, the upfield shift (+0.4 ppm) in the absorption in a nuclear magnetic resonance spectrum for the proton in a hydrogen bond between an amide and an acyl oxygen in an α helix relative to one in random meander has been attributed to the **α -helical dipole**,⁴⁵ but an unexplained downfield shift of the same magnitude is found in a β sheet. The location of a sulfate ion at the intersection of the amino-terminal ends of three α helices in sulfate binding protein suggests that the positive ends of the dipoles of these α helices stabilize the anion,⁴⁶ but the amino-terminal ends of these helices could simply be providing the properly oriented amido donors that occupy several of the many σ lone pairs of electrons on the sulfate (2-28). The location of Glutamate 35 of lysozyme adjacent to the amino-terminal end of an α helix suggests that this arrangement would stabilize its

anionic conjugate base,¹⁶ but it is usually argued that the acidity of Glutamate 35 must be weakened rather than strengthened so that it will be protonated when substrate binds to the enzyme.

It is unfortunate that the original calculations of the magnitude of the electric field generated by an α helix assumed that it existed in a uniform dielectric with a **relative permittivity** of 2 and no account of the relative permittivity of the medium surrounding it was taken. For example, electrostatic potentials of only 2.5 kJ mol⁻¹ have been observed for univalent elementary charges positioned at the amino-terminal ends of α helices in water ($\epsilon = 78$ at 25 °C),⁴⁷ but even these small potentials disappear when the ionic strength of the solution is increased. Later calculations⁴⁸ have incorporated the contribution of the dielectric surrounding the α helix, and it was found that if the protein was approximated by a solid sphere of relative permittivity 3.5 in a solvent of relative permittivity 80 (water), even when the α helix was completely within the sphere of low relative permittivity, the electric field around the α helix was dramatically less than the electric field in a uniform dielectric with a relative permittivity of 3.5. Furthermore, if the ends of the α helix were at the surface of the sphere, in contact with the solvent, the electric field decreased even further to negligible levels. This effect of the dielectric may explain why the apparent electrostatic potentials exerted on aspartates positioned by site-directed mutation at the amino-terminal ends of the two α helices on the surface of T4 lysozyme were only about -2 kJ mol⁻¹.⁴⁹ If the relative permittivity of the interior of a protein is greater than 3.5, the magnitude of the electric field would decrease accordingly in inverse proportion. Finally, the solution around a molecule of protein always contains electrolytes that would further diminish the electric field.⁴⁷ For all of these reasons, electrostatic free energies of significant magnitude are probably not exerted by an α helix within a protein, although the possibility is often discussed.

When an α helix traverses the surface of a protein as a continuous rod, its face directed toward the protein is hydrophobic and its face directed toward the solution is hydrophilic. The α helix in Figure 6-7B has such an orientation with the surface formed by Leucine 28, Tryptophan 30, Phenylalanine 31, and Threonine 35 facing the protein and the opposite surface facing the solvent, as indicated by the locations for molecules of water. This asymmetry of hydropathy is sometimes reflected in the amino acid sequence of the protein and can be identified by constructing a **helical wheel**.^{50,51} Around a circle, successive amino acids in the sequence are placed at 100° intervals (Figure 6-8). This represents the view down an α helix (Figure 6-46), much as a Newman projection represents the view down a carbon-carbon bond. Any asymmetry in the distribution of hydropathy is easily observed. If a segment of amino acid sequence in a polypeptide, when placed upon a helical wheel, reveals such an asymmetric pattern of

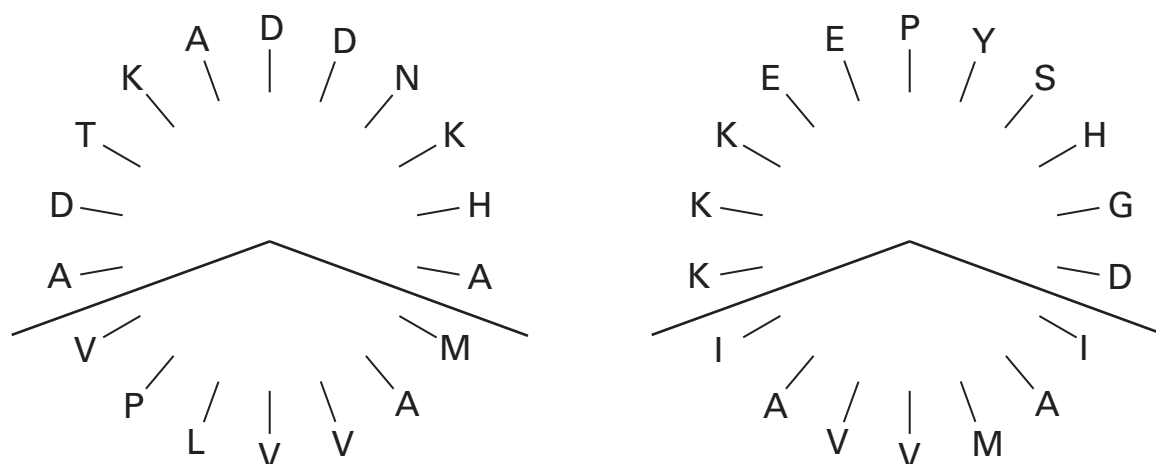


Figure 6-8: Two segments of amino acid sequence displayed on helical wheels. (A) Sequence from Lysine 60 to Proline 77 (KKVADALTNAVAHVDDMP) in the α polypeptide of human hemoglobin. In the crystallographic molecular model of hemoglobin, this sequence is an α helix running across the surface of the protein. In the diagram, the amino terminus is the lysine at the 10:30 position and the sequence is read at 100° intervals. (B) Amphipathic helical sequence from Proline 455 to Lysine 472 (PDVKSAIEGVKYIAEHMK) from the α polypeptide of acetylcholine receptor. The lines in both panels divide the hydrophilic and hydrophobic surfaces of these two amphipathic α helices.

hydropathy, as do those in Figure 6-8, this pattern is evidence that that segment is an α helix in the folded polypeptide. Such α helices are referred to as amphipathic α helices. An **amphipathic α helix** is an α helix that is enriched in hydrophobic side chains on one of its sides and enriched in hydrophilic side chains on the other.

A single-stranded amphipathic α -helical peptide in a mixed solvent of trifluorethanol and water, unattached to a protein, displays an **intrinsic curvature** with the hydrophobic amino acids on the concave face and the hydrophilic on the convex.⁵² In an α helix running across the surface of a protein, the same orientation of curvature is often observed.²³ Whether this curvature is due to the fact that such an α helix is amphipathic,⁵² or to the fact that the acyl oxygens of its peptide bonds on the face exposed to the solvent form hydrogen bonds with water,²³ or to the fact that such curvature simply allows it to adhere more closely to the underlying structure, or to more than one of these reasons is unclear.

It has been proposed that certain amino acids or short sequences of amino acids may impose upon a folding polypeptide biases toward the formation of particular secondary structures at locations where they reside in the native structure of a protein. One hears terms such as “helix-forming” or “helix-breaking” amino acids.⁵³ Originally, these distinctions were based on the observed preferences of homopolymers of the various amino acids to assume α helices or sheets of β structure or to remain structureless at various temperatures, ionic strengths, concentrations of cosolvents, and values of pH.⁵⁴ The propensities of the various amino acids to favor an α -helical conformation have also been examined, either by placing each of them in turn in the center of an α helix in a native protein by site-directed mutation⁵⁵⁻⁵⁷ or by incorporating them in turn into a position in the center

of a peptide that assumes an α -helical conformation in water⁵⁸⁻⁶⁰ and then measuring the changes in stability to that protein or to that unsupported α helix that result.

Although there are significant differences in the various scales that result from these measurements, all agree that, of the 20 amino acids, alanine has the greatest **propensity to stabilize an α helix** and glycine the least and that the differences in free energy of stabilization between these extremes are about 4 kJ mol^{-1} . These preferences presumably explain how antifreeze peptide 3 from the winter flounder, in which 23 of the 37 amino acids are alanines, can naturally assume the conformation of a long, unbroken, unsupported α helix,⁶¹ but in an illustration of the unpredictability of the structure of proteins, all of the alanines in the alanine-rich regions of spider dragline silk are found in β sheets.⁶²

The propensities of the 17 primary amino acids other than alanine, glycine, and proline to stabilize or destabilize an α helix are much less obvious. If the values for their helical propensities in eight different scales⁶³ are averaged, the difference between the mean values for any two of them is rarely as large as the standard deviation of the value for either. Consequently, with the possible exception of methionine and leucine (both at $-2.8 \pm 0.5 \text{ kJ mol}^{-1}$) assigning the rest a value halfway between that of glycine (arbitrarily set at 0 kJ mol^{-1}) and that of alanine (-3.6 kJ mol^{-1}) would be as statistically significant as assigning them each an individual value.

There are several other types of helical structures that occur rarely in crystallographic molecular models of native proteins. The **polyproline helix**, of which there is an example 13 aa long in benzoylformate decarboxylase,⁶⁴ has dihedral angles ϕ and ψ of -75° and 145° , respectively,^{11,65} which places it within the largest allowed region of the Ramachandran plot (Figure 6-4B). It is a much more extended structure than an α helix,

having a rise of 0.31 nm aa^{-1} and 3.0 aa turn^{-1} , and it is a left-handed helix rather than a right-handed one. As their name suggests, polyproline helices in crystallographic molecular models of proteins usually contain a high frequency (25–70%) of proline.⁶⁶ Even though they contain no internal hydrogen bonds and usually occur in situations where they are supported by surrounding structures, there are sequences of amino acids in naturally occurring proteins that form unsupported polyproline helices.⁶⁷ The π helix, of which there is an example 13 aa long in arachidonate 15-lipoxygenase,⁶⁸ is a wider, squatter version of the α helix in which the hydrogen bonds are between the acyl oxygen of amino acid i and the nitrogen–hydrogen bond of amino acid $i + 5$ rather than amino acid $i + 4$.

The values for the dihedral angles ϕ and ψ found in ideal β structure ($\phi = -130^\circ$, $\psi = +120^\circ$) lie within one of the two largest allowed regions of the Ramachandran plot (A and P in Figure 6-4A). These dihedral angles place the hydrogen on the α carbon under the preceding acyl oxygen, O1 (Figure 6-3B), and under the next amide hydrogen, H2 (Figure 6-3D), respectively. This is the least hindered of all the conformations, and β structure experiences no serious steric problems around its α carbons. Because β structure is usually found in the most deeply buried regions of a protein, its polypeptide backbone usually displays the least thermal motion⁶⁹ even though its dihedral angles ϕ and ψ are the least sterically constrained. Nevertheless, it is obvious from an examination of the polypeptide backbones of the proteins presented in Chapter 4 that, because of the size of this region, β structure is far more pliant and unpredictable than an α helix, and efforts to define regular patterns have been less informative than time spent looking at different crystallographic molecular models. The original β -pleated sheets (Figure 4-16B,C) have turned out to be highly idealized. There are, however, several notable structural features of β structure.

When a number of β strands do form a sheet, the sheet usually has a negative, **left-handed twist** to its surface (Figure 6-9).⁷⁰ This is supposed to arise from the fact that the enclosure on the Ramachandran plot in which the dihedral angles ϕ and ψ for parallel and antiparallel β sheets reside has more open area for smaller values of dihedral angle ϕ and larger values of dihedral angle ψ beyond the values of these two dihedral angles that would give a flat sheet. Deviations tend to be biased toward these smaller values of dihedral angle ϕ and larger values of dihedral angle ψ , and this bias creates the twist in the sheet.⁷¹ It may simply be the case, however, that twisted β sheets have surfaces against which other segments of secondary structure, such as α helices, can be more efficiently packed and that packing efficiency dictates the hand and magnitude of the twist because β sheets almost as flat and regular as the idealized version (Figure 4-16B,C) have been observed.⁷²

Another feature of β structure is the **β bulge**.⁷³ In

this arrangement one of the amino acids is skipped in the regular pattern of hydrogen bonding between two antiparallel strands. The hydrogen bond that would have incorporated the nitrogen–hydrogen bond of the skipped amide incorporates the nitrogen–hydrogen bond of the next amide instead. This causes the β structure to bulge at the location of the skipped amino acid (Figure 6-10),⁷³ and the bulge is located where the strands change direction. This change in direction can take two forms. If the β structure remains as a sheet in roughly the same plane, the β bulge puts a bend in the structure. A β bulge, however, also can occur at a location where a large sheet of β structure folds over upon itself to form a sandwich of two opposed β sheets.

As with α helices that contain gaps where a turn is pulled apart, a β sheet can contain a **gap** between two strands. In such a gap, the donors and acceptors that have been pulled apart from each other are occupied by acceptors and donors on the side chains of their amino acids or by ordered molecules of water filling the gap.⁷⁴

Most β structure is **buried** in the middle of a protein, but even in a small protein such as fatty-acid-binding protein from *Escherichia coli*,⁷⁵ that is only a sandwich of two β sheets, there is only a very weak amphipathic pattern of alternating hydrophobic and hydrophilic amino acids along the β strands. Consequently, β structure cannot be identified in an amino acid sequence.

There are three **cylindrical arrays** formed from β structure: a **β barrel** (Figure 6-11)⁷⁶ of 4–12 strands,⁷⁷ a **β helix** (Figure 6-12),^{78,79} and a **β propeller** (Figure 6-13)⁸⁰ with 6–8 blades.^{81,82} In a β barrel, the hydrogen bonds between the β strands are perpendicular to the axis of the cylinder and perpendicular to its radius; in a β helix, they are parallel to the axis of the cylinder and perpendicular to its radius; and in a β propeller, they are parallel to the radius of the cylinder and perpendicular to its axis. Therefore, each of the three orthogonal axes of a cylinder is represented.

The most common type of β barrel has eight β strands (Figure 6-11). Usually β barrels are of eight strands or fewer so that the core can be tightly packed with side chains, but there is a β barrel of 11 strands through the core of which runs an α helix.^{83,84} The β strands in a β barrel reside in a surface that can be approximated quite closely by a twisted hyperboloid.⁸⁵ A hyperboloid is an ellipsoidal cylinder that is narrowest at its center and gradually and continuously widens away from its center in both directions (notice the flare to the hyperboloid in Figure 6-11). In a β barrel of eight strands, the strands are tilted^{77,86} with respect to the axis of the hyperboloid by a mean angle of -34° to -47° (the mean angle of tilt in Figure 6-11 is -34°), but in β barrels of less than eight strands, the angle of tilt gradually increases to -43° to -59° when there are only five.⁷⁷ As in a normal β sheet (Figure 6-9), the sheet that forms the hyperboloid has a negative twist and the mean angles of twist between adjacent strands are between -21° and -30° .

Figure 6-9: Parallel β -pleated sheet within the crystallographic molecular model (PDB filename 20HX; Bragg spacing ≥ 0.18 nm) of alcohol dehydrogenase.⁷⁰ The 12-stranded β sheet is composed of six parallel strands from each of the subunits of the dimer joined in an antiparallel orientation. The two identical series of numbers are those for the respective amino acid sequences of the two identical polypeptides comprising the protein. This drawing was produced with MolScript.⁵⁷³

in barrels of eight strands (the mean angle of twist in Figure 6-11 is -25°), but this angle increases to between -28° and -44° in β barrels of five strands. β Barrels can be constructed from parallel β strands of polypeptide of identical sequence,^{87,88} from an antiparallel β sheet wrapped into a cylinder,^{83,84,89} or from two identical sheets of parallel β strands arranged antiparallel to each other,⁹⁰ but the most common arrangement is parallel β strands of non-identical sequence. In such parallel β barrels the strands are often distributed around the barrel in the order in which they occur in the sequence of the polypeptide, and the carboxy-terminal end of one strand is connected to the amino-terminal end of the next by an α helix. Such β barrels are designated $(\alpha\beta)_n$, where n is the number of β strands.

The β helix displayed in Figure 6-12 is one in which there are three β sheets running up the tube at roughly 60° angles to each other. This configuration seems to be the most common type, but there are β helices in which only two β sheets run up the tube on opposite sides and the two sheets are flattened against each other.^{91,92} Extrusions of random meander (amino acids 167–175 in Figure 6-12) are common features of β helices. There is also an example of a hybrid structure in which each of the β strands in one of the three β sheets in a β helix is replaced by an α helix.⁹³

A third regular structure, in addition to α helices and β structure, universally encountered in the crystallographic molecular models of proteins is the β turn. A **β turn** is any structure that has a hydrogen bond between the acyl oxygen of the first amino acid in the turn and the amido nitrogen–hydrogen of the fourth amino acid in the turn (Figure 6-14).⁹⁴ Usually such a hydrogen bond

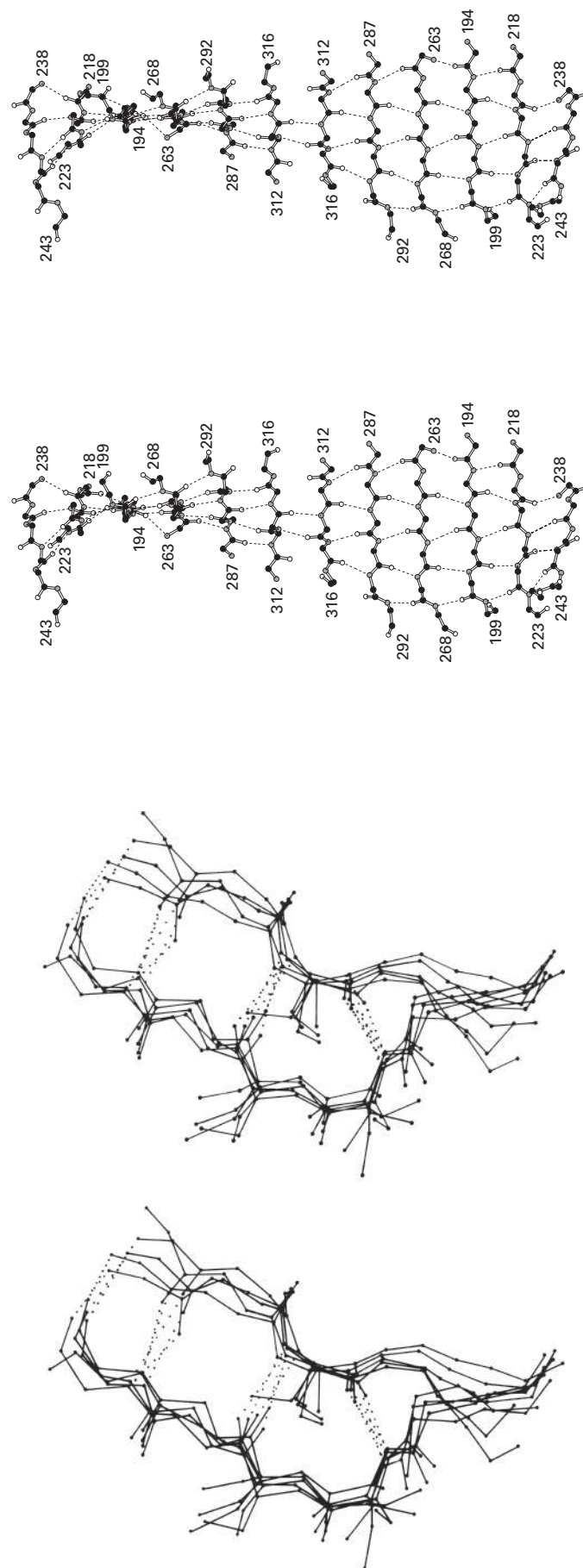
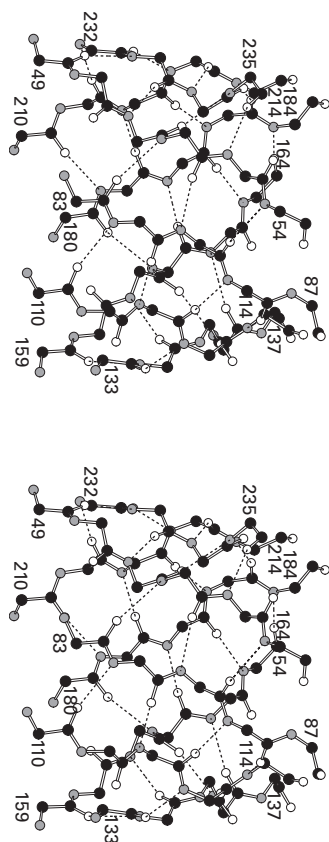


Figure 6-10: Five examples of β bulges from the crystallographic molecular models of various proteins, superposed upon themselves to indicate their uniformity.⁷³ The skipped amido nitrogen–hydrogen in each case is in the center left of the structure. Hydrogen bonds are indicated by dotted lines. The five β bulges are formed by Phenylalanine 41, Cysteine 42, and Leucine 33 of bovine chymotrypsin; Alanine 86, Lysine 87, and Lysine 107 of bovine chymotrypsin; Leucine 107, Serine 108, and Alanine 196 of con-canavalin A from *Canavalia ensiformis*; Isoleucine 90, Glutamine 91, and Valine 120 of human carbonate dehydratase II; and Isoleucine 15, Lysine 16, and Lysine 24 of micrococcal nuclease from *Staphylococcus aureus*, where the first two amino acids listed flank the vacant amido nitrogen–hydrogen and the third provides the amido nitrogen–hydrogen and acyl oxygen from the other strand. Only the side chains of the three central amino acids are included in the figure. Reprinted with permission from ref 73. Copyright 1978 National Academy of Sciences.

Figure 6-11: β Barrel within the crystallographic molecular model (Bragg spacing ≥ 0.20 nm) of indole-3-glycerol-phosphate synthase from *E. coli*.⁷⁶ The β barrel contains eight parallel β strands. The numbers are those for the amino acid sequence of the protein, and the positions of the β strands around the β barrel are in the same order in which they are found in the sequence of the protein. In the crystallographic molecular model between the end of each β strand and the beginning of the next is an α helix that lies across the outer surface of the β barrel running antiparallel to the two β strands that it connects. This drawing was produced with MolScript.⁵⁷³



causes the polypeptide to reverse its direction (Figure 4-15D). The original description of β turns by Venkatachalam⁹⁵ was based on structures built with molecular models. He proposed that there would be six types of β turns. A fundamental distinction is that between type I and type II (Table 6-1). A β turn of type I (Figure 6-14) is represented by the β turn in Figure 4-15D with its second acyl oxygen into the page, and a β turn of type II is represented by the β turn in Figure 4-15D with its second acyl oxygen out of the page. Each of these two fundamental types could also be built with molecular models in such a way that each of their four dihedral angles, ϕ_2 , ψ_2 , ϕ_3 , and ψ_3 , had the opposite sign, respectively (Table 6-1). These alternative conformations with opposite signs to their dihedral angles are called type IA and type IIA, respectively. In each of these two latter types, the polypeptide backbone is the mirror image of the polypeptide backbone in the corresponding β turns of type I and type II in Figure 4-15D, but the amino acids must remain L-amino acids.

Two other types of β turns, of historical significance, were originally defined by Venkatachalam⁹⁵ as type III and type IIIA. The prototype for β turns of type III, however, is the **3_{10} helix**¹⁰¹ originally proposed by Taylor¹⁰² and included by Bragg, Kendrew, and Perutz¹⁰¹ in their catalogue of all helices that would have rotational angles for each amino acid that were integral quotients of 360° . This is a helix that has hydrogen bonds between the acyl oxygen of amino acid i and the nitrogen-hydrogen bond of amino acid $(i+3)$, the pattern that is the primary definition of the β turn. The smaller repeat produced by this shorter connection makes a 3_{10} helix narrower than an α helix. Short segments of 3_{10} helix are occasionally seen in crystallographic molecular models,²³ but they are never more than five or six amino acids in length. Short segments of 3_{10} helix five or six amino acids in length have distorted bond angles along the polypeptide, and this observation suggests that the strain in such a tight helix is considerable.¹⁰³ This steric effect would explain their rarity. Segments of synthetic poly(2-amino-2-methylpropionic acid), however, crys-

Table 6-1: Frequency and Dihedral Angles of the Most Common Types of β Turns

type	frequency ^a (%)	dihedral angles ^b (deg)			
		ϕ_2	ψ_2	ϕ_3	ψ_3
I	41 ^c	-64 ± 8	-19 ± 8	-90 ± 9	-2 ± 11
IA	6	52 ± 3	41 ± 4	87 ± 6	-11 ± 14
II	26	-61 ± 6	132 ± 1	82 ± 12	3 ± 15
IIA	6	63 ± 5	-126 ± 5	-80 ± 9	-11 ± 10

^aThese are the frequencies in which these types of β turn occur in 59 crystallographic molecular models built from data sets gathered to Bragg spacing of ≤ 0.2 nm.⁹⁶ ^bMean and standard deviations of the dihedral angles for amino acids $i+1$ and $i+2$ in the β turns from the crystallographic molecular models of lysozyme,²² α -lytic protease,⁹⁷ deoxyribonuclease I,⁸ and penicillopepsin.⁹⁸ Values from crystallographic molecular models built from data sets gathered to even narrower Bragg spacing^{99,100} fall within these ranges. ^cDoes not include segments judged to be 3_{10} helix. If these had been included, the frequency of β turns of type I would rise to 50%.

tallize as 3_{10} helices.¹⁰⁴ The most frequent location for a short segment of 3_{10} helix is at the end of an α helix. A turn of 3_{10} helix in the middle of an α helix can put an elbow into it. For example, a turn of 3_{10} helix at Serine 143 and Leucine 144 in the center of an α helix in deoxyribonuclease I⁸ causes an abrupt bend of 22° .

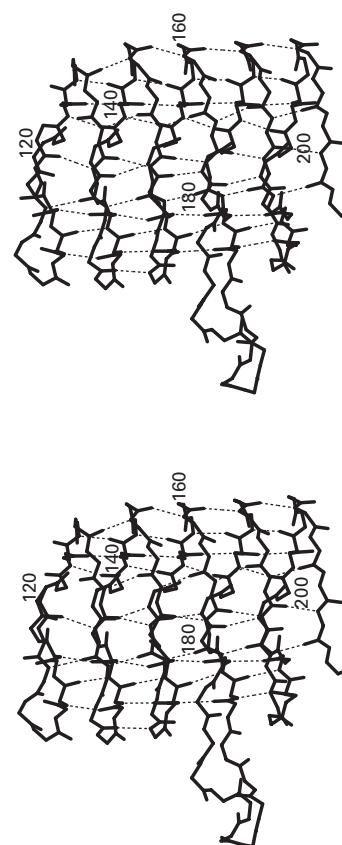
A 3_{10} helix of four amino acids with one hydrogen bond is similar to a β turn of type I because it has mean values for its dihedral angles ϕ and ψ of -71° and -18° , respectively, but with wide ranges²³ that include those for β turns of type I, and it performs the same role as a β turn of type I. Almost all (96%) of the stretches of amino acids in crystallographic molecular models of proteins assigned as 3_{10} helix²³ are four or less amino acids in length, so most instances of 3_{10} helix could as easily be assigned as β turns of type I. Usually, however, they are not classified as such, and if not they are assigned as either 3_{10} helix or β turns of type III, depending on the preferences of the crystallographer. For every segment of amino acids assigned as 3_{10} helix or β turn of type III instead of β turn of type I, there are about 4.5 β turns of type I^{23,96} so the confusion is not a major one.

In crystallographic molecular models, β turns are designated both by the existence of a hydrogen bond between the acyl oxygen on the first amino acid and the amido nitrogen-hydrogen on the fourth amino acid and by the proximity of the α carbons of the first and fourth amino acids. In general, these two α carbons are 0.5–0.6 nm apart.¹⁰⁵ Those configurations designated by these rules as β turns can be grouped into the categories proposed by Venkatachalam (Table 6-1) as well as several other minor categories.⁹⁶ It was only after refined crystallographic molecular models became available that the clear tendency of these structures to fall into specific categories became apparent, because in unrefined structures the orientation of the polypeptide backbone could not be defined with sufficient accuracy.

β Turns of type I are the most common (Table 6-1). The dihedral angles at both of the α carbons in β turns of type I fall in the enclosure on the Ramachandran plot between dihedral angles of $\phi = -50^\circ$ and -130° and $\psi = 20^\circ$ and -30° (Figure 6-4B). This is the region in which the two successive amides are squeezed against each other (Figure 6-3F). Presumably the return on the investment of energy necessary to squeeze them against each other and widen the tetrahedral bond at the α carbon is the efficient reversal of the direction of the polypeptide. It is probably the case that β turns of type I and segments of 3_{10} helix²³ account for most of the amino acids that fall in this well-populated region of the Ramachandran plot.

The values for the dihedral angles ϕ_2 and ψ_2 for **β turns of type II** fall in the largest enclosure on the Ramachandran plot, but those for the dihedral angles ϕ_3 and ψ_3 fall in a region that can be occupied only by an amino acid without a β carbon (Table 6-1, Figure 6-4A), so only glycine should occupy the third position in a β turn of type II. Although this is usually the case (74%),⁹⁶

Figure 6-12: β Helix within the crystallographic molecular model (Bragg spacing ≈ 0.22 nm) of 2,3,4,5-tetrahydroxy-2,6-dicarboxylate *N*-succinyltransferase from *Mycobacterium bovis*.⁷⁹ There are five complete turns of the left-handed β helix, which is an equilateral prism, the three faces of which are parallel pleated β sheets. Protruding from one of the edges of the prism is a loop of random meander. This drawing was produced with MolScript.⁵⁷³

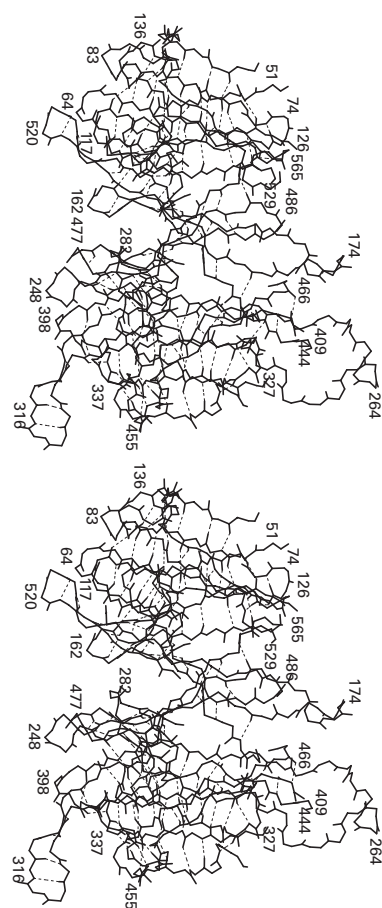


there are exceptions, about half of which are asparagines such as Asparagine 69 in α -lytic endopeptidase.⁹⁷

The mirror image conformations, in which the polypeptide backbone mirrors the respective basic β turn but the amino acids remain, of necessity, L-amino acids, are rare. The third amino acid in a **β turn of type IA** and the second amino acid in a **β turn of type IIA** should be a glycine, but again a few exceptions have been observed, such as Cysteine 170 in deoxyribonuclease I.⁸ It has been noted^{106,107} that when an antiparallel β hairpin reverses itself in the tightest possible β turn, where the hydrogen bond of the β turn is also the last hydrogen bond between the tines of the hairpin, the β turn is usually type IA or type IIA.

Several **minor classes** of β turn have been defined. β Turns of types VIA and VIB with dihedral angles ϕ and ψ of $(-60^\circ \pm 30^\circ, 120^\circ \pm 30^\circ, -90^\circ \pm 30^\circ, 0^\circ \pm 30^\circ)$ and

Figure 6-13: β Propeller within the crystallographic molecular model (Bragg spacing ≥ 0.19 nm) of methanol dehydrogenase from *Methylophilus methylotrophus* W3A1.^{80,569} This β propeller is composed of eight blades, each of which is an antiparallel β sheet. The left-handed twists of the β sheets forming each blade have been incorporated into the propeller. Each blade is composed of four antiparallel β strands that occur consecutively in the sequence; in each, the amino-terminal strand is toward the center and the carboxy-terminal strand is toward the outer edge. The order in which the blades occur around the propeller is also the same order in which they occur in the amino acid sequence with the exception of the outermost β strand of the last blade, which is the amino-terminal β strand of the polypeptide. With the exception of the sixth and eighth blades, each blade is presented in its entirety, its polypeptide unbroken. The missing segments of polypeptide connecting the blades and forming connections within blades six and eight are mostly long stretches of random meander.⁵⁷³ This drawing was produced with MolScript.⁵⁷³



$(-120^\circ \pm 30^\circ, 120^\circ \pm 30^\circ, -90^\circ \pm 30^\circ, 0^\circ \pm 30^\circ)$, respectively, together account for less than 3% of all β turns. β Turns of type VIII, however, with dihedral angles ϕ and ψ of $(-60^\circ \pm 30^\circ, -30^\circ \pm 30^\circ, -120^\circ \pm 30^\circ, 120^\circ \pm 30^\circ)$ ⁹⁶ are somewhat more common (12%), but the dihedral angles ϕ and ψ required to encompass this class are

much less tightly clustered than those for types I and II.¹⁰⁸

Aside from the requirement that glycine occupy certain positions of a β turn for steric reasons, there are some clear **preferences**^{96,103,109} for other amino acids. Because β turns are almost always at the surface of a protein, they contain hydrophilic amino acids more frequently than hydrophobic amino acids. About 25% of all β turns have proline at their second position (Figure 6-55). About 30% of all β turns of type I have either aspartate, asparagine, or cysteine at their third position. Each of these three amino acids has a hydrogen-bond acceptor that is properly situated to accept a hydrogen bond from the amido nitrogen-hydrogen of the amino acid in the next position just beyond the β turn.¹¹⁰

A **γ turn** is another type of turn that occurs rarely in crystallographic molecular models of proteins.^{111,112} A γ turn has a hydrogen bond between the nitrogen-hydrogen of the amide of the first amino acid in the turn and the acyl oxygen of the third amino acid in the turn, causing the dihedral angles ϕ and ψ of the central amino acid of the three to be around 80° and -60° , respectively, which is presumably why such structures are so rare (Figure 6-4B).

In every protein there are also segments of polypeptide that do not assume the configuration of an α helix, β structure, or β turn. These segments of **random meander** pass about the protein as would an α helix or a strand of β structure. They are usually found on the surface of the molecule, and occasionally one of them will loop out a significant distance from the core of the structure. Although there is no regular pattern to this configuration, each of the amino acids in a segment of random meander usually assumes a fixed position in the crystallographic molecular model and has specific values for its dihedral angles ϕ and ψ . These values, however, are still confined to the minima in the Ramachandran plot because these minima are defined by inescapable local steric effects. This places the angles for random meander within the same regions defined by the clusters in Figure 6-4B. The distinction between α helix, β structure, and random meander cannot be made by comparing single values of the dihedral angles ϕ and ψ but only by identifying repeating patterns that extend over several amino acids or several strands of polypeptide. In random meander, no such pattern is evident.

Almost all of the regular structures in which the polypeptide participates are given their regularity by **hydrogen bonds** between the amido nitrogen-hydrogens and acyl oxygens of the backbone. These hydrogen bonds can be readily identified in crystallographic molecular models, and it can be safely assumed that they exist. The **bond length** for such unambiguous hydrogen bonds, expressed as the distance between nitrogen and oxygen, is 0.29 ± 0.015 nm.^{20,97,98,113}

The **angular dependence** of these hydrogen bonds can be expressed either in reference to the nitrogen-hydrogen bond of the one amide (Figure 6-15A,B)^{22,97} or

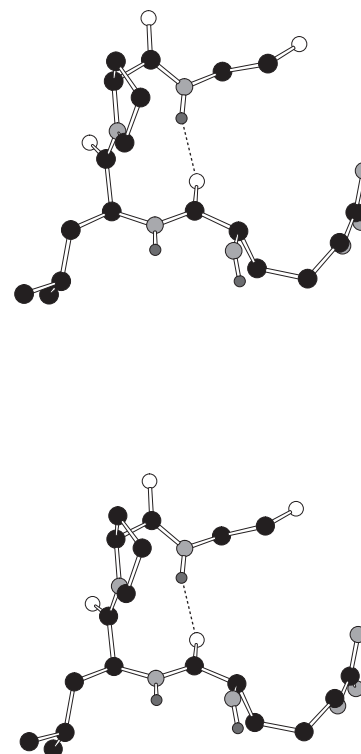
the carbon–oxygen bond of the other amide (Figure 6–15C–G).¹¹³ As expected, the angles around the amido nitrogen–hydrogen of the donor are much more confined than those around the lone pairs on the acyl oxygen of the acceptor. A deviation from 0° of either of the angles around the nitrogen–hydrogen bond (Figure 6–15A) places the hydrogen off the line of centers between the nitrogen and the oxygen and bends the bond.

The angles of the hydrogen bond relative to the carbon–oxygen bond of the acceptor (Figure 6–15C–G) vary over a greater latitude. In keeping with the rigidity of α helices and flexibility of β structure, the angles around the acyl oxygens in β structure are much more variable (Figure 6–15F,G) than those around the acyl oxygens in α helices (Figure 6–15D). In none of these regular structures, however, is there any tendency for the values of the angles around the acyl oxygen to cluster at $\beta = 0^\circ$ and $\gamma = \pm 60^\circ$, the positions at which the nitrogen–hydrogen bond would point directly at one or the other of the lone pairs on the acyl oxygen (Figure 5–10). It has, however, already been noted that, even in crystallographic molecular models of small, unconstrained molecules, the preference for these angles is not remarkable (Figure 5–11), and there seems to be little energetic cost in pivoting the donor over the surface of the acyl oxygen distal to the acyl carbon (Figure 5–10D). Therefore, in regular structures such as an α helix or β structure, it is the steric requirements of these structures themselves that easily take precedence.

Even in refined molecular models from data sets of narrow Bragg spacing, the identification of a hydrogen bond is subjective. It is often based on the fact that two heteroatoms are simply within a certain distance of each other. The dimensions of unquestionable hydrogen bonds in regular structures, however, suggest a more objective definition of a hydrogen bond.²² It has been proposed that a hydrogen bond is an arrangement in which the heteroatoms of the donor and the acceptor are less than 0.34 nm from each other; the angle A–H–B, where A is the donor and B the acceptor, is between 150° and 180° (Figure 6–15B); and the distance between the theoretical position of the hydrogen and the heteroatom of the acceptor is less than 0.24 nm. When these definitions are applied to the rather featureless distribution of distances between nitrogens and oxygens in a crystallographic molecular model, that distribution can be divided into hydrogen bonds and non-hydrogen bonds (Figure 6–16).²²

The regular structures assumed by the polypeptide serve the purpose of maintaining the total concentration of hydrogen bonds in the solution. It seems highly unlikely that a protein could fold without withdrawing considerable numbers of its peptide bonds from contact with water. Were the donors and acceptors on these peptide bonds withdrawn from the solvent without subsequently participating in hydrogen bonds in the inte-

Figure 6–14: β Turn of type I from the crystallographic molecular model (Bragg spacing ≥ 0.089 nm) of crambin from *Crambe abyssinica*.⁹⁴ The β turn contains a hydrogen bond between the acyl oxygen of Arginine 17 and the amido nitrogen–hydrogen of Glycine 20 in the amino acid sequence of the protein. Leucine 18 and Proline 19 occupy the two central positions. The molecular model was from a data set gathered to such narrow Bragg spacing and at such a low temperature (130 K) that the electron density for hydrogen atoms could be readily discerned, and those on the amides are included in the drawing. The C γ -endo pucker of Proline 19 was also clearly defined. This drawing was produced with MolScript.⁵⁷³



rior of the protein, there would be an unavoidable loss of standard enthalpy (Figure 5–18). This loss is cancelled when they find new partners (Table 5–2). α Helices and β structure are simply efficient mechanisms for accomplishing this energetic imperative.

Suggested Reading

- James, M.N.G., & Sielecki, A.R. (1983) Structure and refinement of penicillopepsin at 1.8-Å resolution, *J. Mol. Biol.* 163, 299–361.
- Wilmot, C.M., & Thornton, J.M. (1988) Analysis and prediction of the different types of β turn in proteins, *J. Mol. Biol.* 203, 221–232.

Problem 6–1: Build a space-filling model of the structure displayed in Figure 6–2 with a methyl group at C β .

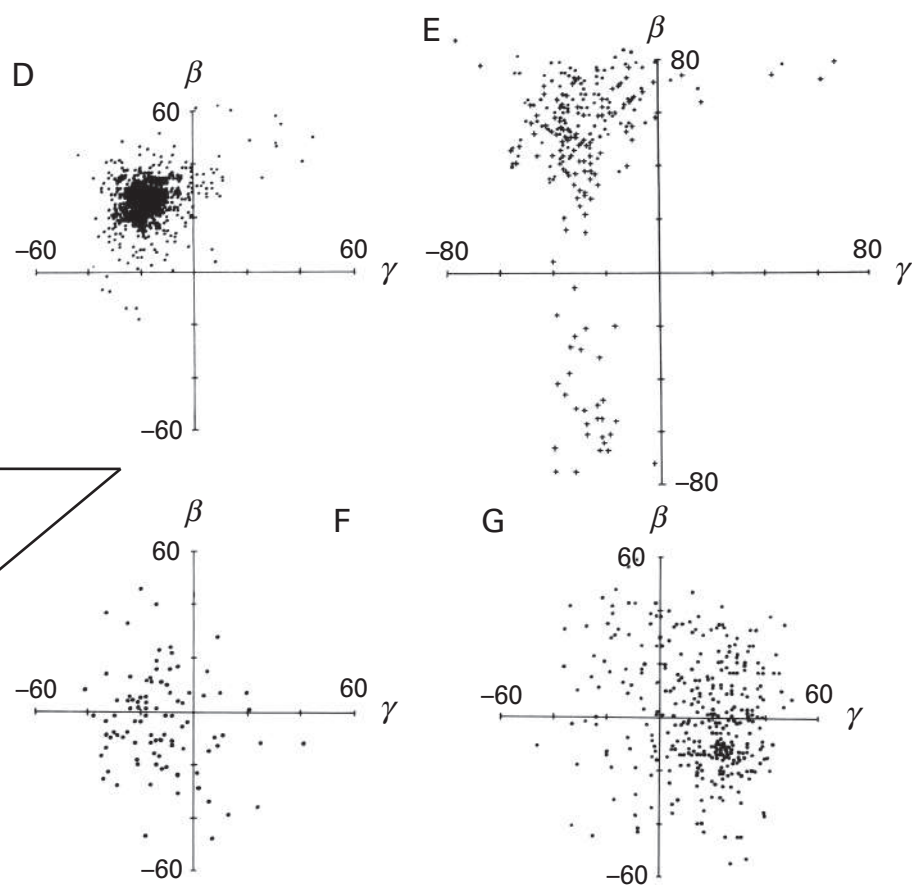
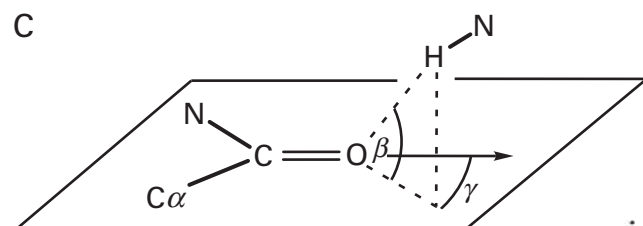
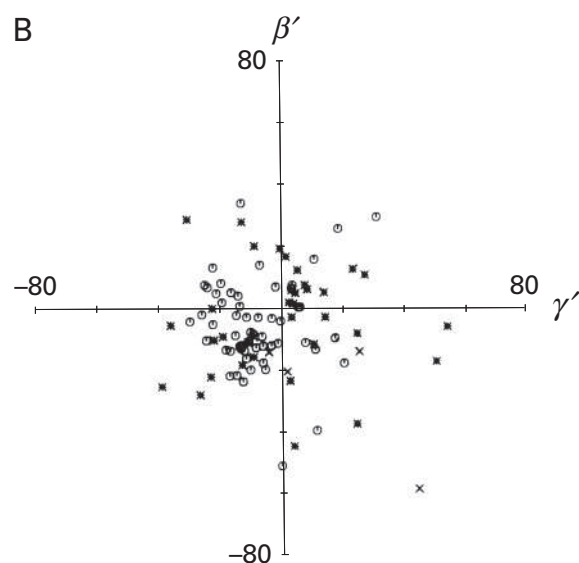
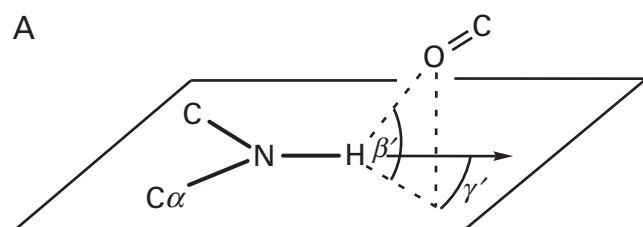


Figure 6-15: Bond angles for the hydrogen bonds in regular structures formed by a folded polypeptide. (A) Bond angles at the amido nitrogen-hydrogen. Two angles are defined, the angle γ' within the plane of the amide and the angle β' out of the plane of the amide.²² When γ' is 0° , the acyl oxygen is in the plane that is normal to the plane of the amide and that contains the nitrogen-hydrogen bond. When β' is 0° , the acyl oxygen is in the plane of the amide. (B) Distribution of these angles. The plot is for all of these bond angles for the hydrogen bonds between the peptide bonds in the crystallographic molecular model of α -lytic endopeptidase.⁹⁷ Symbols are (○) β structure, (×) α helix, and (*) random meander. Reprinted with permission from ref 97. Copyright 1985 Academic Press. (C) Bond angles at the carbon-oxygen bond of the amide. Two angles are defined, the angle γ within the plane of the amide and the angle β out of the plane of the amide. Angles are defined relative to the axis of the carbon-oxygen bond and the plane of the amide. (D-G) Distribution of these angles. These angles at each hydrogen bond involving the peptide bonds in the crystallographic molecular models of 15 proteins¹¹³ are plotted for hydrogen bonds in α helices (D), β turns (E), parallel β structure (F), and antiparallel β structure (G). Each mark is for the angles β and γ of one of the hydrogen bonds included in the set. Reprinted with permission from ref 113. Copyright 1984 Pergamon Press.

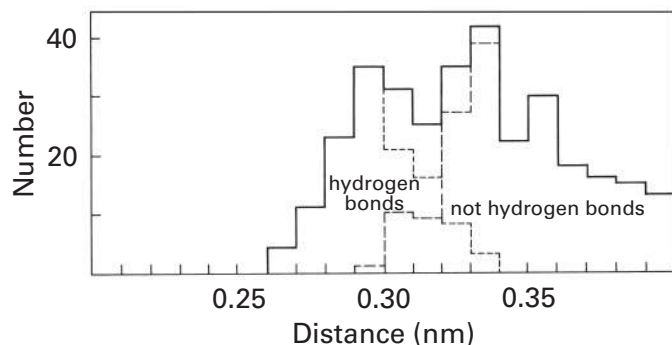


Figure 6-16: Histogram of all of the distances (nanometers) between the centers of nitrogen atoms and oxygen atoms in direct contact with each other in the crystallographic molecular model of human lysozyme.²² The number of pairs of nitrogens and oxygens that are a given distance apart in the molecular model is plotted as a function of those distances. If a hydrogen bond is defined as a nitrogen and an oxygen less than 0.34 nm apart, hydrogen and oxygen less than 0.24 nm apart, and the angle N–H–O between 150° and 180°, the histogram can be divided into nitrogen–oxygen contacts that are hydrogen bonds and contacts that are not hydrogen bonds. Reprinted with permission from ref 22. Copyright 1981 Academic Press.

- Take this molecular model of the polypeptide backbone, and adjust it so that $\phi = 60^\circ$ and $\psi = 120^\circ$. What atoms are colliding?
- Adjust it to $\phi = 60^\circ$ and $\psi = -60^\circ$. What atoms are colliding?
- Adjust it to $\phi = 180^\circ$ and $\psi = 60^\circ$. What atoms are colliding?
- Adjust it to $\phi = -60^\circ$ and $\psi = 60^\circ$. What atoms are colliding?

Use the numbering system of Figure 6-2 for your answers.

Stereochemistry of the Side Chains

It was pointed out in a discussion of the crystallographic molecular model of chymotrypsin, which was one of the first refined crystallographic molecular models,¹⁷ that certain **rotational conformations** of the side chains of the amino acids seemed to be preferred. As more highly refined crystallographic molecular models of proteins built from data sets of narrower Bragg spacing have become available, it has become clear that most of the side chains of the amino acids in the interior of these models assume only one of the several rotational conformations available to them,^{114,115} a remarkable fact that illustrates the confinement exercised by the efficient packing in the interior of a protein.

In maps of electron density calculated from data sets to narrow Bragg spacing, side chains displaying **alternative conformations** (Figure 4-25) are infrequent (12 of 478 side chains in human glutathione reductase;¹¹⁶

35 of 584 side chains in human β_4 hemoglobin^{117,118}) and most of those are side chains found at or near the surface of the molecule of protein, such as lysines, serines, threonines, glutamates, aspartates, and asparagines.^{74,116,118} The valine of Figure 4-25 in the interior of ribonuclease T₁ is an example of an exception to the preference of buried side chains for only one conformation.¹¹⁵ When side chains do assume two alternative conformations, each of them is usually at one of the normal minima of rotational energy,¹¹⁹ as are the unique, fixed conformations of most of the side chains.

Consequently, most of the observed rotational conformations are **staggered** rather than eclipsed, a fact that is reflected in the strong tendency (Figure 6-17)⁹⁸ for dihedral angles along carbon–carbon bonds between saturated carbons and other atoms that are hybridized sp^3 to assume values near 60°, 180°, and 300° (–60°). In extensive tabulations^{120–122} of the dihedral angles for all of the side chains in sets of refined crystallographic molecular models from data sets of narrow Bragg spacing, most of the values for the dihedral angles of carbon–carbon bonds connecting atoms that are hybridized sp^3 are clustered within 10° of one of these three values.

It has been pointed out, however, that a significant fraction of the amino acid side chains have at least one dihedral angle that falls more than 20° (5–30% depending on the side chain)¹²¹ or even more than 30° (1–19% depending on the side chain)¹²² away from the mean. If these are real **deviations**, conformations of side chains exist in which substituents are partially or fully eclipsed,¹²² situations in which considerable steric strain (15–40 kJ mol^{–1}) must be accommodated. Most of these unexpected and sterically strained dihedral angles in these crystallographic molecular models, however, are probably artifactual,¹²³ arising from unresolved alternative conformations,¹²⁴ the inaccuracy of the crystallographic molecular model, incorrect insertion into the map of electron density, and errors accumulated during refinement. Again it must be remembered that a crystallographic molecular model is not the actual structure of the molecule of protein. Nevertheless, some of these deviations may reflect the actual adjustments of some of the side chains to the impossibility of packing as complicated a molecule as a polypeptide into a compact globular structure.

The dihedral angle along the bond between C α and C β in an amino acid is designated χ_1 . It is the dihedral angle between the bond to the amido nitrogen, the most massive of the three atoms around C α , and the bond to the atom attached to C β that has the highest priority in the Cahn–Ingold–Prelog system (Figure 6-18). The sign of **dihedral angle** χ_1 is determined by the right-hand rule. The stereochemistry about this bond between C α and C β is dominated by the polypeptide rather than the rest of the side chain.

Valine is the logical place to begin the discussion of this stereochemistry because its two methyl groups can

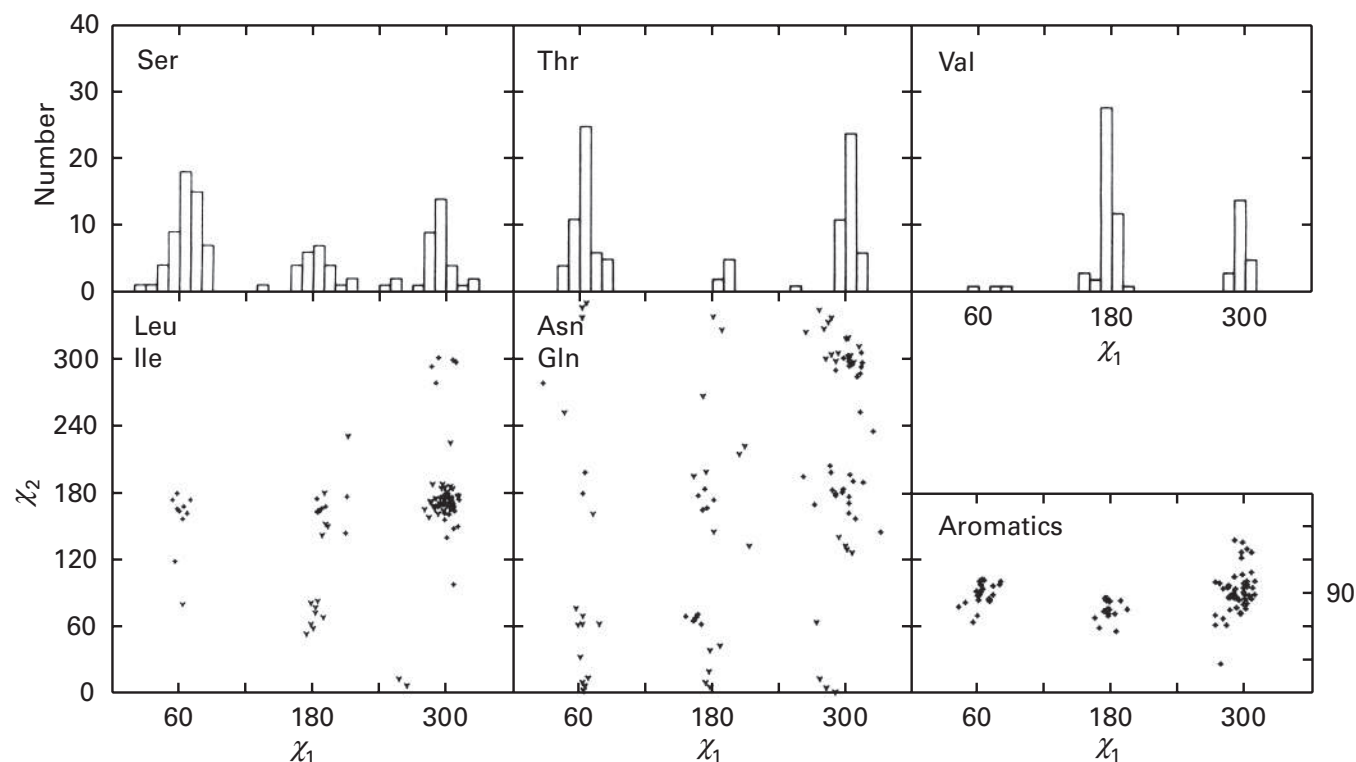


Figure 6-17: Histograms and scatter plots of the distributions of the values for the dihedral angles χ_1 and χ_2 for the first two carbon-carbon bonds of the side chains of amino acids in the crystallographic molecular models of five proteins: penicillopepsin, streptogrisin A from *Streptomyces griseus*, streptogrisin B from *S. griseus*, the third domain of the ovomucoid inhibitor, and α -lytic endopeptidase from *Lysobacter enzymogenes*.⁹⁸ The abbreviation of each side chain appears in the upper left-hand corner of the panel. Serine, threonine, and valine had no observable dihedral angles χ_2 , so in these instances frequency is plotted as a function of the only value of the dihedral angle χ_1 . Leucine, isoleucine, asparagine, and glutamine had observable dihedral angles χ_1 and χ_2 , and in these cases each mark (γ for leucine, $+$ for isoleucine, γ for asparagine, and $+$ for glutamine) represents the value of these two angles for one of these side chains in these molecular models. Because of symmetry, the values for χ_2 for the aromatic amino acids tyrosine, phenylalanine, and tryptophan (listed together as aromatics) fall only between 0° and 180° . Reprinted with permission from ref 98. Copyright 1983 Academic Press.

assess the steric bulk of the three substituents on the α carbon because in each of the three staggered conformations (two of which are displayed in Figure 6-18), one of these three substituents must reside between the two methyl groups, a most hindered location. Because the smallest functional group should occupy this position most frequently, the distribution of the dihedral angles χ_1 of the valines in molecular models (Figure 6-17) states that the hydrogen on the α carbon is smaller (73% of χ_1 are within 30° of 175°)* than the nitrogen of the preceding amide (20% are within 30° of -64°), which is smaller

than the acyl carbon of the following amide (6% are within 30° of 63°). This behavior is completely consistent with the assessment of steric bulk based on preferences of various substituents on cyclohexane for equatorial over axial locations. The increase in free energy¹²⁵ for placing an acetoxy group in an axial location rather than an equatorial location is 2.9 kJ mol^{-1} , but the increase in free energy for placing a methoxycarbonyl group in an axial location rather than an equatorial location is 5.4 kJ mol^{-1} .

Isoleucine (Figure 6-18) reinforces these preferences by showing a similar distribution¹²⁶ of analogous stereochemical conformations (76% within 30° of -64° , 14% within 30° of 61° , and 10% within 30° of -173° , respectively). It has been suggested that isoleucine ($+$ in Figure 6-17) has different preferences from leucine (γ in Figure 6-17) for the dihedral angles χ_1 and χ_2 because these two geometric isomers should be able to satisfy in turn different steric requirements.¹²⁷

Threonine is isosteric with valine, but the designation of the dihedral angle χ_1 of threonine is 240° out of phase with that of the dihedral angle χ_1 of valine because of the precedence of the (S)-oxygen over the (R)-methyl group (Figure 6-18). The conformation of threonine (χ_1 within 30° of 59° , 49% of all threonines) in which the two substituents on $C\beta$ surround the nitrogen of the preceding amide (Figure 6-18) is about 4 times more frequent than the analogous conformation of valine (χ_1 within 30° of -64° , 20% of all valines) relative to the respective conformations (43% and 73%) in which hydrogen is surrounded. The most likely explanation for this difference is the fact that a hydroxyl group is significantly smaller than a methyl group. Another possibility, however, is that

* Values of dihedral angles χ_1 designated as within 30° of the angle at the maximum of the distribution are from the tabulation derived from an analysis of 240 crystallographic molecular models built from data sets all to Bragg spacing less than or equal to 0.17 nm .¹²²

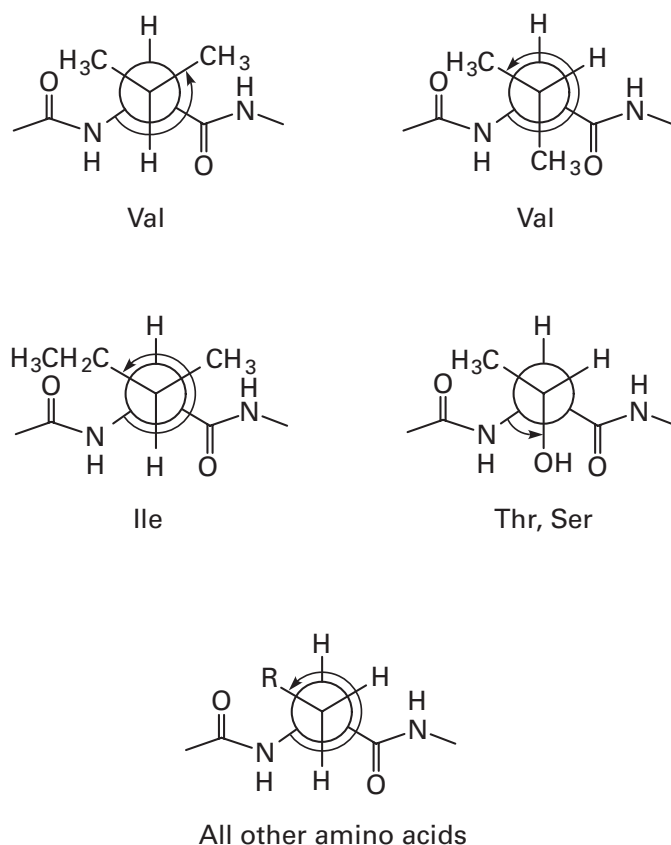


Figure 6-18: Definition of the dihedral angle χ_1 for the carbon-carbon bond between the α carbon and the β carbon of an amino acid in a polypeptide. All dihedral angles follow the right-hand rule. For valine, angle χ_1 is the dihedral angle between the carbon-nitrogen bond and the bond to the *pro-R* methyl group. For isoleucine and threonine, angle χ_1 is the dihedral angle between the carbon-nitrogen bond and the bond to the substituent of higher priority, namely, the ethyl group or the hydroxyl group, respectively. For all other relevant amino acids, the dihedral angle χ_1 is the angle between the carbon-nitrogen bond and the bond to the remainder of the side chain.

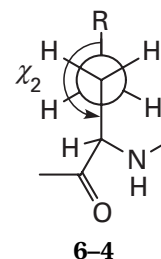
a dihedral angle χ_1 within 30° of 59° places the hydroxyl of threonine in the proper position to form either a hydrogen-bonded ring with its own acyl oxygen or a hydrogen-bonded ring with the acyl oxygen of the amino acid that precedes it in the sequence.¹²⁸ Such hydrogen-bonded rings would explain why **serines** show such a high percentage of dihedral angles χ_1 within 30° of 64° (48%), in contrast to all other amino acids with only one substituent on the β carbon, which have a low percentage of dihedral angles χ_1 near 60° (11% of all side chains with one and only one substituent on $C\beta$).^{120,121,126}

The amino acids, other than serine, with one and only one substituent on $C\beta$ have a preference (55%)^{120,126} for dihedral angles χ_1 of $-65^\circ \pm 10^\circ$.¹²⁶ This preference⁶ is understandable because an angle χ_1 of -65° places the single substituent on the β carbon between the two least bulky substituents around the α carbon (Figure 6-18). The other two maxima occur at $-177^\circ \pm 10^\circ$ and $66^\circ \pm 8^\circ$. The bias in the direction of the α hydrogen is reflected in

the values of -65° and -177° . Of this group of side chains, those with a methylene at the γ position have dihedral angles χ_1 within 20° of 66° ,¹²¹ which would put the methylene between the two bulkiest substituents (Figure 6-18), only 7% of the time, but the side chains with an aromatic ring at the γ position, which is less bulky than a methylene, have dihedral angles χ_1 within 20° of 66° 16% of the time.¹²¹

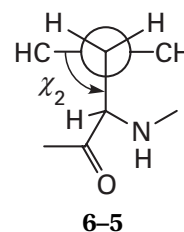
Because the preference for a certain value for the dihedral angle χ_1 usually depends on a choice among three staggered conformers (Figure 6-18), the energies of which differ by only a few kilojoules mole⁻¹, it comes as no surprise that this choice depends on the structure of the immediate surroundings. In particular, the dihedral angles ψ and ϕ of the amino acid can affect the choice of its dihedral angle χ_1 . As might be expected, valine displays the most dramatic effects of the backbone on the dihedral angle χ_1 . For valines within either α helices or β sheets, values of dihedral angle ψ more positive than the respective ideal values cause χ_1 to switch from a value of exclusively $175^\circ \pm 8^\circ$ to a value of exclusively $-64^\circ \pm 7^\circ$.¹²³

The **dihedral angle χ_2** is assigned to the bond between $C\beta$ and $C\gamma$ in an amino acid. In **linear amino acids** such as glutamine, glutamic acid, lysine, arginine, and methionine, the majority (70%) of the dihedral angles χ_2



are clustered⁶ at $176^\circ \pm 11^\circ$, which is the angle expected for a *trans* conformation at this carbon-carbon bond.^{120,126} The remainder of the dihedral angles χ_2 for these amino acids are split equally between the two *gauche* conformations at dihedral angles χ_2 near 60° and -60° .

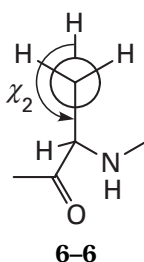
For the **aromatic amino acids** phenylalanine, tyrosine, histidine, and tryptophan (Figure 6-17), the conformations with dihedral angles χ_2 within 30° of 90° or -89° , where the plane of the ring is approximately perpendicular to the bond between the α carbon and the β carbon



are preferred (71% of these side chains).¹²¹ This orientation (Figure 6-21) places the polypeptide most distant from the two *ortho* substituents of the rings and also avoids eclipse. A significant fraction (29%) of these side chains, however, have values for χ_2 outside of these ranges. The aromatic rings are large, bulky substituents and each of these outliers is pushed out of the ideal range by unavoidable steric clashes with atoms from the backbone or other side chains.¹²¹ In spite of their bulk, however, the symmetric rings of tyrosine and phenylalanine have been observed to flip over slowly and continuously.¹²⁹

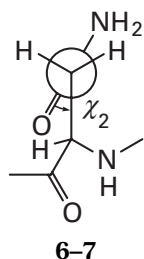
It has also been observed in refined maps of neutron scattering density from crystallography by neutron diffraction that the hydrogens on all methyl groups are staggered.¹³⁰ Although this seems to be the expected result because **methyl groups** in proteins should be free to rotate and assume freely a staggered conformation, there are indications that packing in the interior of a protein is so tight that even methyl groups are confined¹³¹

The dihedral angles χ_2 for the **hydroxyl groups** of serines and threonines



although they define the position only of a hydrogen, can also be observed by neutron diffraction.¹³² There is a strong tendency for the hydroxyl to be staggered (χ_2 near 60° , 180° , and -60°) with the *trans* conformation (χ_2 near 180°) slightly preferred over either *gauche* conformation. The location of an acceptor forming a hydrogen bond with the proton often seems to dictate the dihedral angle assumed by the hydroxyl. Because of conjugation, the oxygen-hydrogen bond of the hydroxyl of tyrosine is within the plane of the ring (2-23).

The distribution of the values of the dihedral angle χ_2 for **asparagine**

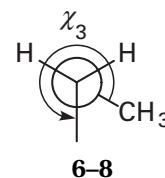


has two maxima at -21° and $+32^\circ$ (82% within 30° of these two maxima) that define two respective classes,^{122,133} the membership of which is determined by

the value of the dihedral angle χ_1 for that asparagine, the type of secondary structure to which it belongs, and the type of hydrogen bond formed between it and the backbone. Asparagine 34 in Figure 6-7B serves as an example of one of these choices. **Aspartate** shows the same preference for values of the dihedral angle χ_2 . The majority (82%) of the dihedral angles χ_2 for aspartates are within 60° of 0° .¹²²

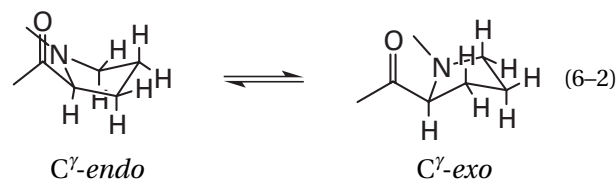
Only glutamine, glutamate, methionine, lysine, and arginine have carbon-carbon bonds with **dihedral angles χ_3** . Both **glutamine** and **glutamate** show the same preferences for dihedral angles χ_3 near 0° that are shown by asparagine and aspartate for the analogous dihedral angle χ_2 .¹²² The *trans* conformation with dihedral angle χ_3 within 30° of 180° is the preferred (66%) conformation for lysine, as expected.¹²²

Methionines are usually buried and confined to one or two overall conformations in crystallographic molecular models, so the dihedral angles χ_1 ($C\alpha-C\beta$), χ_2 ($C\beta-C\gamma$), and χ_3 ($C\gamma-S$) are usually fixed. The normal preferences for dihedral angles χ_1 (59% within 30° of -67°) and χ_2 (55% within 30° of 178°)¹²² are observed, but the value for angle χ_3



has a significantly higher frequency for the two *gauche* conformers (39% within 30° of -72° and 32% within 30° of 75°).^{122,131} It has been pointed out that because the two carbon-sulfur bonds (0.18 nm) are longer than two carbon-carbon bonds (0.15 nm), the steric clashes within methionine in such *gauche* conformations should be less severe and the dihedral angles χ_3 should be less confined.¹³⁴ It seems that this unexpected preference for the *gauche* conformations arises from the fact that when methionine assumes this conformation, it is more compact.

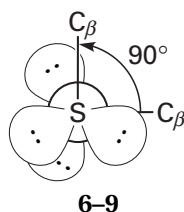
Because the amido nitrogen is planar, it occupies a position in the puckered cyclopentyl ring of a **proline** (Equation 6-1) at which eclipse would occur if it were occupied by a methylene. As a result, only the C^γ -*exo* and C^γ -*endo* conformations of proline



can be significantly populated,¹³⁵ but it is difficult to distinguish crystallographically between these two conformations.

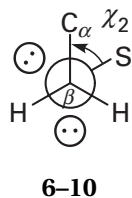
mations even with a data set gathered to narrow Bragg spacing (< 0.17 nm). Nevertheless, decisions can often be made either directly (Figure 6–14) or indirectly¹³¹ as to the proper conformation, and sometimes a map of electron density shows that both conformations are present in equilibrium with each other.¹³⁶

Cystine is an amino acid under peculiar steric constraints (Figure 6–19).^{137,138} The distribution of the two dihedral angles χ_1 of cystine^{26,121} shows the same order and frequencies of preferences (56% within 20° of -65° , 24% within 20° of -175° , and 12% within 20° of 64°)¹²¹ as those of any other amino acid with only one uncomplicated substituent on the β carbon. The disulfide itself, because it is a dithioperoxide, is electronically required to have a dihedral angle χ_3 along the sulfur–sulfur bond similar to the dihedral angle of hydrogen peroxide, which is 94° or -94° . If the dihedral angle χ_3 in a cystine were exactly 90° or -90° , the four lone pairs, two on each sulfur, would be as far from being parallel to each other as is possible, and this orientation would be the most stable electronically.¹³⁹



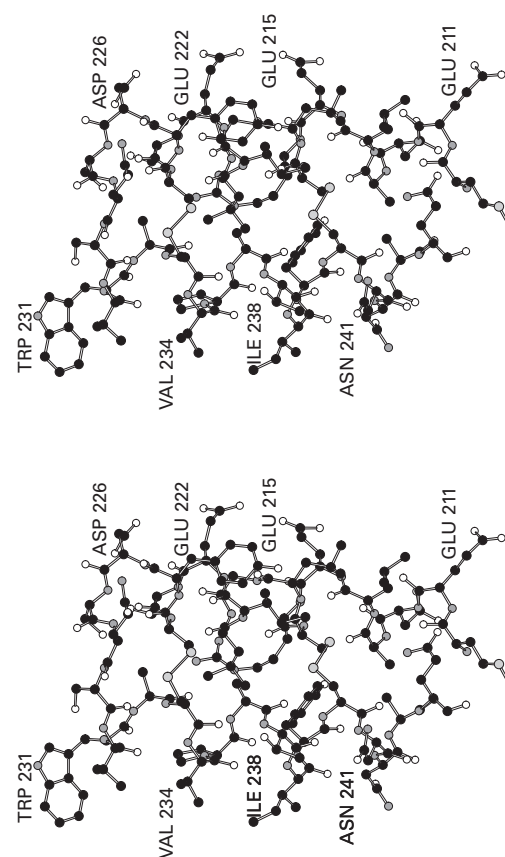
The angles observed for the dihedral angles of cystines in crystallographic molecular models are $97^\circ \pm 15^\circ$ and $-86^\circ \pm 11^\circ$ (indistinguishable from $+90^\circ$ and -90°), with little preference for the positive over the negative.¹⁴⁰ There are instances in which these two discrete, alternative conformations are both populated significantly by the same cystine (Figure 6–40B).^{141,142} The dihedral angle χ_3 of 97° or -86° in a cystine is peculiar enough to attract attention (Figure 6–19).

The bonds between the β carbons and the sulfurs in a cystine might be expected to have a preference for dihedral angles χ_2



equal to 180° as with most other amino acids, but in crystallographic molecular models, values around broad maxima of 60° and -60° (Figure 6–19) are heavily preferred.^{103,143} Each of the two bonds between the β carbons and the sulfurs should be fairly long and not severely confined by the adjacent sulfur or the lone pairs on the immediate sulfur itself, and they are probably the

Figure 6–19: A pair of cystines, Cystine 224/233 and Cystine 217/240, in the amino acid sequence of the crystallographic molecular model (Bragg spacing ≥ 0.26 nm) of carboxypeptidase C from *Saccharomyces cerevisiae*.^{137,138} The two cystines cross-link two α helices in an antiparallel coiled coil. The drawing is aligned to permit the dihedral angles χ_2 of Cystines 217 and 224 to be viewed effectively. The dihedral angles χ_3 for these cystines are both close to $+90^\circ$.



most compliant of the bonds in the cystine. Because a cystine connects two strands of polypeptide engaged in many other interactions, it probably sustains considerable torque. Presumably it is the dihedral angles χ_2 that must accommodate this torque. Because of the electronic requirements, the dihedral angle of the disulfide itself must always be near 97° or -86° , and it is only the two dihedral angles χ_2 that can adjust to allow the entire cystine to fit the distance required to be spanned between the two otherwise fixed α carbons of a cystine in a protein. For example, in Figure 6–19, if the two α helices connected by the two cystines had to move farther apart from each other or closer together, their relative movement could be accommodated by increasing

or decreasing, respectively, the dihedral angles χ_2 of Cysteines 217 and 224 or Cysteines 233 and 240 or all of them together.

If both of the dihedral angles χ_2 were 180° , the angle preferred by other amino acids with a single substituent on the β carbon, then the two α carbons in a cysteine would be about 0.9 nm apart, a rather distant connection. Because the distances between the two α carbons of cysteines in native proteins fall between 0.45 and 0.7 nm,¹⁰³ the dihedral angles χ_2 assume many other values, and rarely 180° .

So far, the dihedral angles χ have been considered independently. Usually, however, the individual rotamers of a side chain are tabulated.^{121,122} A **rotamer** is a rotational conformation of a side chain in which each carbon-carbon bond assumes a dihedral angle χ within a particular range. The range is within a certain number of degrees, for example, within 20° ¹²¹ or within 30° ,¹²² of one of the maxima for the distribution, which are usually close to the staggered dihedral angles of 60° (gauche⁺), 180° (trans), or -60° (gauche⁻). For example there are nine rotamers of isoleucine, which are, in order of their frequency, g^-t , g^-g^- , g^+t , tt , tg^+ , g^-g^+ , g^+g^+ , g^+g^- , and tg^- . Such tabulations of rotamers emphasize the dependence of one dihedral angle on the adjacent dihedral angles. There is, however, no agreement as to the statistical method that should be used to determine rotamers, their variances, or their distributions.

All of the stereochemical observations discussed so far are either consistent with the behavior of small molecules or otherwise make sense. Some of this agreement is probably illusory. During refinement, **constraints on dihedral angles** are imposed either advertently or inadvertently, and the fact that they are near ideal values in the final crystallographic molecular model may not reflect reality. Careful corrections using properly calculated **omit maps** should eliminate this bias, and crystallographic molecular models derived from data sets gathered to narrow Bragg spacing, for which few constraints need to be imposed during refinement, can avoid this problem entirely.

If there are only a few conformations that are preferred for each side chain, then there is far less flexibility involved in the folding of a protein than there seems to be at first glance. Conformations of the folded protein that demand dihedral angles to assume values other than those of lowest energy, which are normally the conformations most heavily populated in the unfolded polypeptide, require that extra energy be spent to occupy those conformations. It turns out that there is not much extra energy to go around.

Suggested Reading

Schrauber, H., Eisenhaber, F., & Argos, P. (1993) Rotamers: To be or not to be? An analysis of side-chain conformations in globular proteins. *J. Mol. Biol.* 230, 592–612.

Lovell, S.C., Word, J.M., Richardson, J.S., & Richardson, D.C. (2000) The penultimate rotamer library, *Proteins: Struct., Funct., Genet.* 40, 389–408.

Problem 6–2: Turn the alanine into a valine in your space-filling molecular model from Problem 6–1 by replacing two of the hydrogens on the β carbon with methyl groups. Rotate around the appropriate bonds until the dihedral angles ϕ and ψ have the mean values for an amino acid in parallel β structure.

- What atoms run into the two methyl groups on the side chain of the valine as rotation occurs around the bond between the α and β carbons? Use the numbering system of Figure 6–2.
- What is the value for the dihedral angle χ_1 that has the most sterically favorable disposition of the side chain in β structure?

Rotate around the appropriate bonds until the dihedral angles ϕ and ψ in your model have the mean values for a right-handed α helix.

- What atoms run into the two methyl groups on the side chain of the valine as you rotate around the bond between the α and β carbons? Again, use the numbering system of Figure 6–2.
- What is the value for the dihedral angle χ_1 that has the most sterically favorable disposition of the side chain in a right-handed α helix?

The theoretical values for the dihedral angles ϕ and ψ for a left-handed α helix should be 65° and 40° , respectively. Rotate around the appropriate bonds until the dihedral angles ϕ and ψ in your model have these values.

- What atoms run into the two methyl groups on the side chain of the valine as you rotate around the bond between the α and β carbons?
- What is the value for the dihedral angle χ_1 that has the most sterically favorable disposition of the valine side chain in a left-handed α helix?
- On the basis of these observations, why is a left-handed α helix unstable relative to a right-handed α helix?

Problem 6–3: Draw Newman projections to explain why the dihedral angles χ_2 for leucine and isoleucine display such a strong preference for 180° (Figure 6–17).

Hydropathy of the Side Chains

There is an obvious bias in the distribution of its constituent amino acids between the **surface** of a molecule of protein, which remains in contact with the water, and

its **interior**, which is more or less withdrawn from the water. This bias reflects their hydrophathy.

The **accessible surface area** of a molecule of protein can be estimated by asking a digital computer to perform a calculation equivalent to rolling a sphere of a particular size, the probe, over the surface of a space-filling crystallographic molecular model (Figure 4-17E) of that protein.¹⁴⁴ The center of the spherical probe will trace a surface, and the area of that surface, removed from that of the surface of the protein by a distance equal to the radius of the sphere, is defined to be the surface area of the protein accessible to the probe. Each portion of the irregular surface defined by the center of the probe can be assigned to a particular atom in the crystallographic molecular model by noting with which atom the probe was in contact when that portion was being created.

The surface of a molecule of protein is not smooth, but highly **irregular**, covered with cracks, crevasses, cavities, and ridges (Figure 4-17E).^{144,145} One way to demonstrate this fact is to vary the radius of the probe (Figure 6-20). When the probe is large (≥ 1.5 nm in the example chosen) the crystallographic molecular model is indistinguishable from a hard sphere (a sphere of radius 4.55 nm in the example chosen), but as the radius of the probe is decreased, much more surface area becomes accessible (the difference between the points and the curve in Figure 6-20) as the probe becomes small enough to enter the irregularities of the surface. The radius usually chosen for the probe,¹⁴⁴ in an attempt to mimic a molecule of water, is 0.15 nm (the arrow in Figure 6-20). The

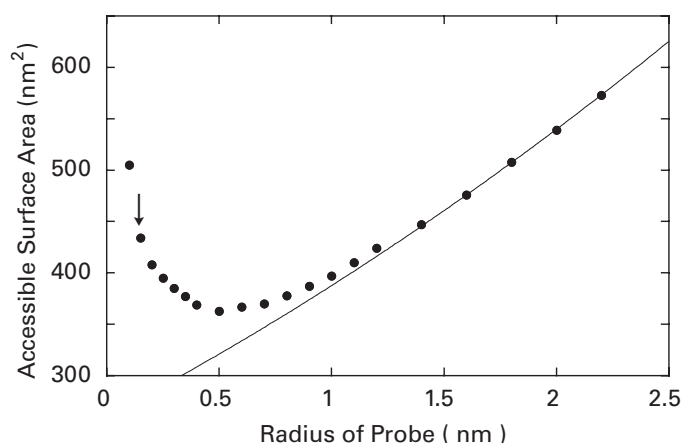


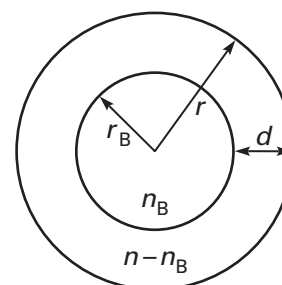
Figure 6-20: Irregularity of the surface of a protein. The radius of the probe used to determine accessible surface area¹⁴⁴ was assigned a particular length (nanometers) and the accessible surface area (nanometers²) of the crystallographic molecular model (Bragg spacing ≥ 0.18 nm) of glyceraldehyde-3-phosphate dehydrogenase from *Bacillus stearothermophilus*⁵⁷⁰ was computed (points on graph). The smooth curve on the graph is the accessible surface area that a sphere of radius 4.55 nm would give as a function of the radius of the probe rolled over its surface. The arrow is at a radius for the probe of 0.15 nm. The program used in the calculations was adapted from that of Lee and Richards¹⁴⁴ by Dr Ilya Shindyalov of the Protein Data Bank.

choice is arbitrary, especially because locations to which only a single molecule of water can gain access experience diminished effects of the solvation arising from the bulk properties of water.

Usually a particular amino acid in a crystallographic molecular model is designated as **buried** if less than a certain amount of its surface area is accessible. It has been shown¹⁴⁵ that when the radius of the probe is set at 0.15 nm

$$n^{1/3} - n_B^{1/3} = \kappa \quad (6-3)$$

where n is the total number of amino acids in a protein, n_B is the number designated as buried by a particular rule, for example, every amino acid with an accessible surface area less than 0.2 nm^2 , and κ is a constant. This equation states that, in a globular protein, the amino acids defined by a rule as buried are found within a roughly spherical solid of radius r_B that is smaller than the roughly spherical solid of radius r containing all of the amino acids.¹⁴⁵



6-11

The spherical shell of width d between these two roughly spherical solids contains all of the amino acids ($n - n_B$) that are accessible by the **rule** that has been chosen. This shell can be considered to be the depth of penetration of the probe, which represents water, into the interior of the protein owing to its irregular surface. If buried amino acids are defined as only those completely inaccessible to a probe of radius 0.15 nm, d is fairly large (1.0 nm) and the buried amino acids are deep in the interior. This rule would require that in a small protein (100 aa) almost no amino acid would be completely inaccessible. If the rule, however, is that any amino acids having accessible surface area of less than 0.2 nm^2 are defined as buried, then d is only 0.5 nm and far more of them are considered to be buried.

Accessible surface areas of the amino acids in crystallographic molecular models of proteins have been calculated by use of a radius for the probe of 0.15 nm, and for each type of amino acid, the **fraction scored as buried** by at least three different rules has been separately tabulated (Table 6-2). When the most stringent rule is used, namely, that a buried amino acid in the molecular model of the protein must have no accessible surface area, the frequencies with which most of the

Table 6-2: Removal of Amino Acids from Water in Molecular Models of Proteins

amino acid	accessible surface area ^a (nm ²)	fraction buried ^b			mean of fraction surface buried ^c
		buried 100% ^c	buried 95% ^c	less than 0.2 nm ² accessible ^d	
hydrophobic					
Ile	1.80	0.18	0.60	0.76	0.90
Val	1.60	0.18	0.54	0.74	0.88
Phe	2.20	0.14	0.50	0.69	0.89
Leu	1.80	0.16	0.45	0.71	0.87
Met	2.05	0.11	0.40	0.66	0.83
Ala	1.15	0.20	0.38	0.63	0.78
Trp	2.60	0.04	0.27	0.62	0.88
apathetic					
Ser	1.20	0.08	0.22	0.46	0.62
Thr	1.45	0.08	0.23	0.41	0.60
His	1.95	0.02	0.17	0.44	0.78
Tyr	2.30	0.03	0.15	0.34	0.74
hydrophilic					
Glu	1.85	0.03	0.18	0.24	0.74
Asp	1.50	0.04	0.15	0.27	0.67
Asn	1.60	0.03	0.12	0.30	0.61
Gln	1.90	0.01	0.07	0.23	0.61
Lys	2.10	0	0.03	0.05	0.60
Arg	2.40	0	0.01	0.10	0.51

^aFor entire amino acid, both its side chain and its contribution to the backbone, in the tripeptide Gly-X-Gly.^{146,147} ^bFraction of the total number of that amino acid in a series of crystallographic molecular models that are buried by the noted criterion. ^cReference 148. ^dReference 145.

amino acids are buried is very low and the statistics become unreliable. When the rule is relaxed, more amino acids are scored as buried, and discriminations become more dependable.

Half of the amino acids, when they are in an unfolded polypeptide and freely accessible to a probe of 0.15 nm, have total accessible surface areas between 1.5 and 2.0 nm² and therefore are of similar **size** (Figure 4-14). Small amino acids such as alanine are probably buried more often simply because they are easier to surround, and large amino acids such as tryptophan are harder to surround completely and bury, especially in the smaller proteins. These stereochemical problems must contribute to the observed distributions.

Nevertheless, it has already been noted that the frequencies with which the various amino acids are buried are correlated¹⁴⁹ with the **free energies of transfer** for their model compounds from water to the gas phase (Table 5-9) and also with many of the other scales of hydrophathy (Figure 5-24). This correlation is actually established by the three main groups of side chains (Table 6-2): those that are hydrophobic, those that are apathetic, and those that are hydrophilic (note the three clusters in Figure 5-24). Within each of these groups, however, there is no significant correlation between extent of burial and any **scale of hydrophathy** derived from free energies of transfer. Presumably, the reason for the lack of correlation within the main groups is that

stereochemically and energetically protein folding is not a transfer between solvents. With this in mind, it still can be stated that if an amino acid is hydrophobic, it is more likely to be buried, and if it is hydrophilic, it is more likely to remain in contact with the water in the folded polypeptide.

The results of the hydrophobic effect are most readily appreciated by examining the internal **core of a crystallographic molecular model** (Figure 6-21).¹⁵⁰ This region is enriched in definitively hydrophobic amino acids such as leucine, isoleucine, valine, and phenylalanine. An even more dramatic example of a hydrophobic core is the center of a β helix, which is completely walled off from the water by the backbone of the helix and which is composed exclusively of aliphatic side chains.¹⁵¹

Each of the **hydrogen-carbon bonds** on the side chains that are removed from water provided favorable **hydrophobic free energy** to drive the folding of the polypeptide. This fact can be verified by performing site-directed mutation. So that no adverse steric effects are encountered, either an isoleucine found in the core of a crystallographic molecular model of the protein is shortened by converting it to a valine or an alanine, a leucine or a valine in the core is shortened by converting it to an alanine,¹⁵²⁻¹⁵⁶ or a position in the core next to a cavity is chosen and the mutants designed so that they expand into the cavity.¹⁵⁷ The change in the standard free energy of folding produced by the various mutations is then

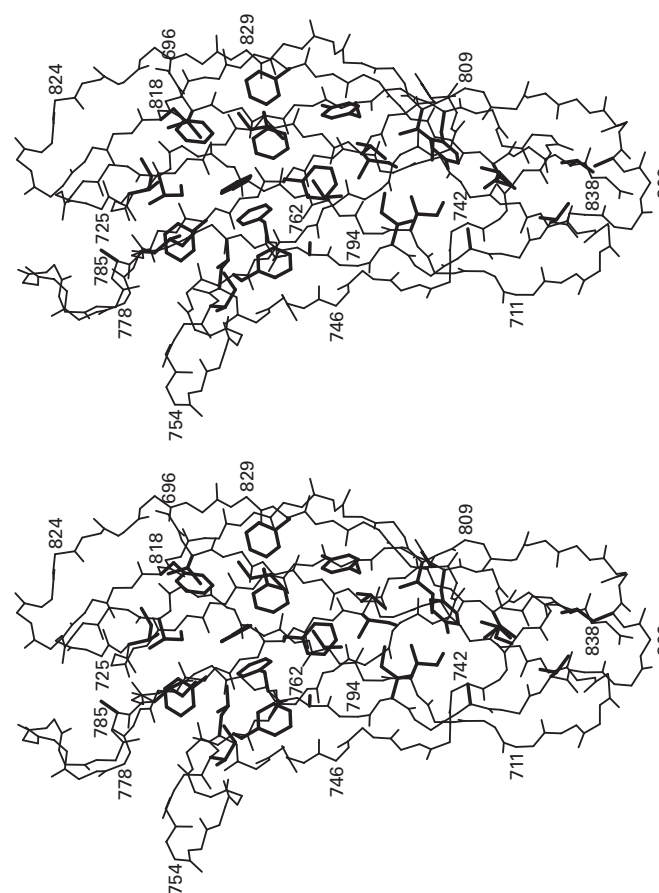
measured. In such studies, it has been found that whether hydrogen-carbon bonds are added to or removed from the core relative to the number present in the core of the native protein, each one contributes between -1 and -5 kJ (mol of hydrogen-carbon bond) $^{-1}$ to the standard free energy of folding, with most of the values clustered^{155,156,158-161} between -2.5 and -3.5 kJ (mol of hydrogen-carbon bond) $^{-1}$. These values encompass* and are indistinguishable from the value of -2.8 kJ (mol of hydrogen-carbon bond) $^{-1}$ for the transfer of hydrocarbon from water to liquid hydrocarbon (Figure 5-21).¹⁵⁴ The interior of a protein is enriched in hydrophobic amino acids because this is the only way to obtain the free energy necessary to drive its folding.

When a **hydrophilic amino acid** such as lysine or glutamate is introduced into the interior of a molecule of protein by site-directed mutation, the protein becomes significantly less stable. For example, when Methionine 102 and Leucine 133 in lysozyme from bacteriophage T4, which are both buried in its interior, were replaced in turn with a lysine and an aspartate, respectively, the protein became less stable by 29 and 24 kJ mol $^{-1}$. The region surrounding the new lysine at position 102 became much more mobile to permit limited access of the lysine to the solvent, and its pK_a was found to be 6.5, a shift indicating that the neutral conjugate base had become more stable than it would have been if it were fully exposed to the water.¹⁶³ When Valine 16 of ribonuclease T₁ from *Aspergillus oryzae* is replaced with its isostere threonine, the stability of the protein decreases by 15 kJ mol $^{-1}$, and the side chain of the threonine has rotated 120° relative to that of the valine around the C α -C β bond to direct the hydroxyl group toward the solvent rather than toward the interior as the -CH₃ group of the valine was directed.¹⁶⁴ Often, however, when valine is replaced by threonine or alanine is replaced by serine in the interior of a protein, the hydroxyl group finds a nearby acceptor for its hydrogen bond, and the change in stability is small (<4 kJ mol $^{-1}$).¹⁶⁵

When a hydrophobic amino acid ends up fully exposed on the surface, this exposure is neither energetically unfavorable nor favorable because it was fully exposed before the polypeptide folded. When a phenylalanine was placed in turn by site-directed mutation at a number of fully accessible sites on the surface of micrococcal nuclease, the stability of the protein decreased on

* The broad range of these values¹⁵⁶ may be due to the fact that each of these hydrophobic chains that has been mutated in the respective experiments is located in its own particular surroundings within the native structure, and each of those surroundings produces its own particular steric effects and standard free energy of solvation.¹⁶² It has been shown that the magnitude of the change in standard free energy of folding produced by removing hydrogen-carbon bonds by mutation is directly proportional to the number of other hydrogen-carbon bonds surrounding the point of the mutation in the crystallographic molecular model of the protein.¹⁵³

Figure 6-21: Hydrophobic core in the crystallographic molecular model (Bragg spacing ≥ 0.17 nm) of a type I cohesin domain from the cellulosomal scaffolding protein A of *Clostridium thermocellum*.¹⁵⁰ The backbone (thinner lines) of the entire domain, except for the first five amino acids that were missing from the map of electron density, is presented. Only the side chains (thicker lines) of the amino acids in the core of the molecule are drawn. Note the different rotamers of isoleucine and the fact that dihedral angles χ_2 for the phenylalanines and tyrosines are almost all near 90°. The numbers are those for the amino acid sequence of the intact cellulosomal scaffolding protein. This drawing was produced with MolScript.⁵⁷³



average for each of these single substitutions by only 2 kJ mol $^{-1}$, and at least half of that decrease in stability resulted from steric effects.¹⁶⁶ No consistent changes in stability were observed when several different hydrophobic amino acids were substituted by site-directed mutation for amino acids on the surface of lysozyme from bacteriophage T4.¹⁶⁷

Clusters of **aromatic amino acids**, such as the one in the type I cohesin domain (Figure 6-21), are often found in the hydrophobic core of a crystallographic molecular model. The tendency for aromatic amino acids, cystines, and arginines to cluster has been explained in terms of favorable overlaps of their π mole-

cular orbitals,¹⁶⁸ but just as often aromatic and aliphatic amino acids intermingle. One interesting aspect of the cluster of aromatic amino acids in type I cohesin domain (Figure 6–21) is that it illustrates the tendency of two phenyl rings, in isolation from water, to form a complex in which the planes of the two rings are at around 90° to each other.¹⁶⁹ This orientation is commonly observed in the crystallographic molecular models of proteins, and theoretical calculations for benzene dimers in the gas phase suggest that it is the energetically favored arrangement.¹⁷⁰

One of the most notable features of the accessibilities of the amino acids in the native structure of a protein is that, in the folded polypeptide, the accessible surface areas of all types are less than they were in the unfolded polypeptide (Table 6–2). The accessible surface areas tabulated are for the entire amino acid in a polypeptide, both side-chain and backbone segments. Usually, the backbone is buried before the side chain, so a significant portion of the mean fraction of surface buried for each type of amino acid is accounted for by this fragment common to all of the amino acids. This backbone portion, however, cannot account for more than 0.6–0.7 nm² of buried surface area because the accessible surface area of glycine in a polypeptide¹⁴⁸ is only 0.75 nm². Therefore even the most accessible side chains, arginine, lysine, and glutamine (with a mean buried surface of 1.2 nm²), have more than 0.5, 0.55, and 0.45 nm² of the surface area of their side chains buried, respectively. The regions of each of these hydrophilic side chains that are buried are usually their hydrogen–carbon bonds. For example, in the crystallographic molecular model of the complex between the *Ha-ras* oncogene product p21 and its substrate, the amino group of Lysine 117 is engaged in several hydrogen bonds, but its butyl group is fully buried just as would be the side chain of a leucine.¹⁷¹

There should be a normal hydrophobic effect associated with the removal of the butyl group of lysine, the propyl group of arginine, and the ethyl groups of glutamine and glutamate from exposure to water even though these side chains in their entirety are among the most hydrophilic on all of the scales of hydrophathy. To assess the hydrophobic effect that was contributed by burying these portions of each of these amino acids, as well as all of the others, the **contribution of each atom** in an amino acid to its overall hydrophathy should be extracted. From these atomic parameters, the free energy of transfer for only those portions of each amino acid that are actually buried could be calculated. It was noted¹⁷² that free energies of transfer for individual solutes between water and the gas could be dissected into the individual contributions of each covalent bond that they contained. A similar dissection has since been performed upon the set of free energies of transfer for the *N*-acetyl- α -amides of the amino acids between water and 1-octanol.¹⁷³ In this latter dissection, a series of parameters based on individ-

ual atoms, rather than covalent bonds, has been presented that can reproduce the original scale of hydrophathy with acceptable precision. Presumably every scale of hydrophathy presently in use can be so dissected.

Free energies of transfer for model compounds of **tryptophan** or **tyrosine** between water and a solvent such as 1-octanol (Figure 5–24)¹⁷⁴ or ethanol,¹⁷⁵ as opposed to free energies of transfer between water and the gas phase (Table 5–9), have always suggested that tryptophan and tyrosine should be more hydrophobic than they seem to be when they are found in a protein (Table 6–2). The explanation for this is probably that solvents such as ethanol and 1-octanol are able to form hydrogen bonds with the one donor on the indole and the donor and acceptor on the phenol, making them more soluble in these solvents than they would be in a hydrocarbon and making them seem more hydrophobic than they are. This would be consistent with the observation that it is the hydroxyl on tyrosine that usually remains in contact with the water in crystallographic molecular models.¹⁷ Because of the requirement that the nitrogen–hydrogen bond in the indole of tryptophan also be hydrogen-bonded, the portion of the side chain containing this bond is also usually in contact with the water. But indole is large and the rest of it is usually buried. As a result, it is only in the last column of Table 6–2 that the hydrophobicity of tryptophan is manifest.

The neutral, but hydrophilic, amino acids **glutamine** and **asparagine** are straightforward examples of the effect of the hydrogen bonds formed with water in the unfolded polypeptide on the location of that amino acid in the folded polypeptide. Complete withdrawal of the two hydrogen-bond donors on glutamine or asparagine from water during folding would result in a net disappearance of two hydrogen bonds from the solution. The difficulty of simultaneously regaining both of these lost hydrogen bonds in the interior of the protein seems to be great enough that the primary amides in the side chains of most of the glutamines and asparagines in a protein end up in the folded polypeptide fully exposed to the aqueous phase.¹²⁶

It might be supposed that **buried hydrogen bonds** between side chains on different segments of secondary structure would be important factors because these would be capable of organizing significant regions of the protein.¹⁴⁸ Of the rarely buried hydrogen bonds between side chains,⁴⁰ however, only about 20% are the type that connect different segments of secondary structure; the other 80% connect donors and acceptors within the same α helix, β sheet, or β turn.¹⁴⁸ The steric requirements of packing the secondary structures efficiently and avoiding empty space are far more important than hydrogen bonds in positioning the segments of secondary structure and organizing the overall structure of the protein, and the few buried hydrogen bonds that do occur between segments of secondary structure are probably adventitious. It is the interdigitation of the side

chains protruding from β sheets and α helices that orient these secondary structures.

Suggested Reading

Chothia, C. (1976) The nature of the accessible and buried surfaces in proteins, *J. Mol. Biol.* 105, 1–14.

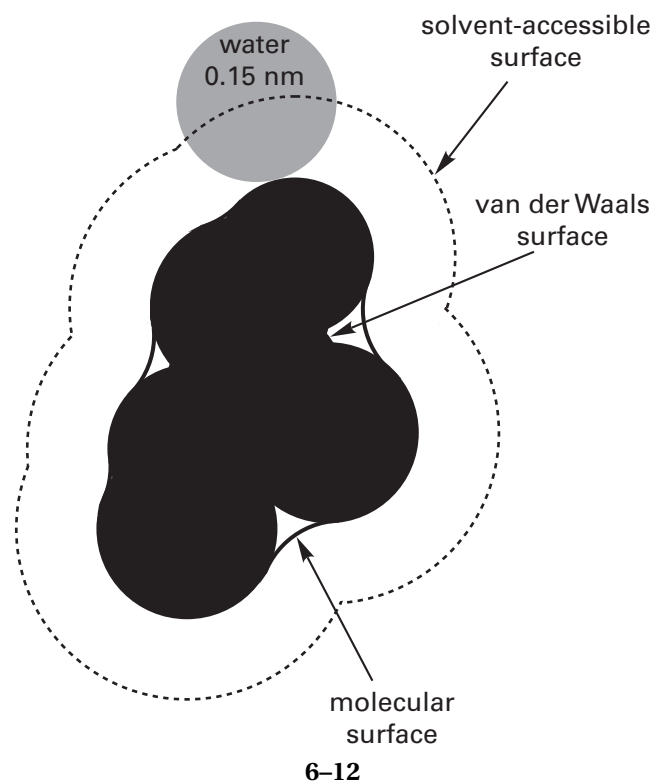
Packing of the Side Chains

Although there often are flexible loops, sometimes quite a long one,¹⁷⁶ bulging out from a globular protein and occasionally there is a tunnel passing through its interior that is required for its function,^{177,178} most of its mass is formed by compactly layering secondary structures one against the other. In the crystallographic molecular model of a protein, the space between these layered α helices, β sheets, and random meander is filled completely by the side chains of the amino acids that protrude from the polypeptide at each α carbon. The arrangement of these side chains is organized in such a way that the amino acids in the interior of the protein are **packed with admirable efficiency**. This tight packing is illustrated by the fact that space-filling crystallographic molecular models of proteins rarely contain empty spaces large enough to accommodate even a molecule of water.

A **space-filling crystallographic molecular model** is constructed by placing spheres of the appropriate van der Waals radii (Table 6–3) at the positions of the centers of atoms in the model. The **van der Waals radius** of an atom is the radius that produces a sphere the surface of which is coincident with the boundary of that atom to penetration by the boundary to penetration of another atom.¹⁷⁹ These van der Waals radii are significantly greater than the covalent radii of the various atoms.

It is difficult to obtain reliable values for van der Waals radii, so those tabulated are averages of values estimated in several different ways. Because the van der Waals radii define the **boundaries of impenetrability**, when matter is packed together in a condensed phase, the centers of no two atoms can approach closer than the

sum of their van der Waals radii, and examining the details of the packing of side chains in the interior of a crystallographic molecular model of a protein is one way to estimate values for van der Waals radii.¹⁸³ The spheres defined by the van der Waals radii define the **van der Waals surface** of a molecule.*



which is distinct from either the **molecular surface** or the **accessible surface**.

When atoms in the interiors of a set of crystallographic molecular models, from data sets each of which is to Bragg spacings of 0.17 nm or less, are assigned their respective van der Waals radii and hydrogens added at their van der Waals radii, one discovers these atoms almost always approach each other as closely as possible and do so in such a way that there is almost **no empty space** among them.¹³¹ The packing is often so tight that methyl groups are confined in equilaterally triangular spaces clamping their hydrogens. Consequently, it is not surprising that the alkyl groups in the center of a molecule of protein are confined to specific conformations rather than being free to rotate as they would be in a liquid. One result of this tight packing is that at some locations there is not enough space for a side chain. Such locations are occupied by glycines. Even in these instances, the fit is often so tight that it is steric interactions with the hydrogens of the glycine that position the α carbon and determine the values for dihedral angles ψ

Table 6–3: van der Waals Radii

atom	van der Waals ^a radius (nm)
H _{aliphatic}	0.115
H _{aromatic}	0.100
C	0.175
N	0.160
O	0.150
S	0.180
P	0.180

^aValues are averaged from various tabulations^{179–183} and expressed to nearest 5 pm, which may overstate their accuracy.¹⁷⁹

* Reprinted with permission from ref 184. Copyright 1996 Academic Press.

and ϕ .¹³¹ Even though it might seem to be the case from an examination of Figure 6-4B, not all of the glycines in a protein, however, are handling unavoidable steric problems because many can be replaced with larger amino acids by site-directed mutation without affecting the function of the protein.¹⁸⁵

Another consequence of the economy in arranging the side chains of the amino acids is that the **volume of a molecule** of protein is quite small relative to its molecular mass. The volume occupied by the atoms in a molecule of protein can be calculated by summing all of its individual atomic volumes defined by the van der Waals surface, and the actual volume of the molecule can be calculated from its partial specific volume and its molecular weight. From this calculation it is learned that 75% of the volume of a molecule of protein is occupied by atoms.^{180,184,186} By comparison, in most organic liquids only about 45% of the volume is occupied by atoms and in water only 36%, but in a solid of hexagonally packed spheres 75% of the volume would be filled by atoms.¹⁸⁶ In anhydrous crystals of small organic molecules, 70–80% of the volume is filled by atoms.¹⁸⁰

The high density of the packing in the interior of a molecule of protein is also reflected in its **compressibility**. The compressibility of the interior of a molecule of protein has been estimated from two different interpretations of the available experimental measurements^{184,187} to be about 20 Gbar⁻¹. This value is intermediate between those for liquids (CCl₄, 100 Gbar⁻¹; C₁₀H₂₂, 105 Gbar⁻¹; H₂O 46 Gbar⁻¹) and solids (ice Ih, 12 Gbar⁻¹; quartz, 2.7 Gbar⁻¹; NaCl, 4 Gbar⁻¹).

At first glance, all of these results seem incompatible with the observation that the partial specific volume of a protein (usually 0.72–0.75 mL g⁻¹) can be calculated* quite accurately from the sum of the molar volumes of its constituents^{188,189} because each protein has a unique structure. The accuracy of this calculation suggests that each structure, although it is unique, incorporates the requirement that its volume be as small as possible. The **minimization of molecular volume** is an important noncovalent force in the folding of a molecule of protein, and it dictates many of the features of the structure. This

* This calculation does not treat the constituents as independent solutes in free solution. In fact, if each side chain were an independent solute, each of their partial molar volumes would include a covolume,¹⁹⁰ which is a volume that arises simply because a particular constellation of atoms is an independent molecule dissolved in a given solvent. These covolumes are substantial. For water¹⁸⁹ the covolume of a solute is 14 cm³ mol⁻¹, and for organic solvents¹⁹⁰ it is 25 cm³ mol⁻¹. Therefore, the sum of the partial molar volumes of the components of a protein, were they each separate molecules in solution, would be significantly greater than its actual partial molar volume. To the extent that its covolume arises from the fact that a solute is in free solution in a given solvent, the fact that the partial molar volume of a protein is the sum of the atomic volumes of its substituents with no added covolume states that those substituents are not in free solution. This of course is true; they are economically packed into a solid.

noncovalent force minimizing the empty space within a molecule of protein can be considered to be a consequence of the hydrophobic effect, if the hydrophobic effect is defined as the tendency of water to minimize the volume of the cavity occupied by any solute.

Although there are a few proteins in which the folded polypeptide forms a knot,^{191,192} the packing of every other protein appears to result from the **consecutive layering** of one element of secondary structure upon another, much as one would fold a cloth or a hinged rod. α Helices pack upon α helices, β sheets pack upon β sheets, and α helices pack upon β sheets. In all of these situations the secondary structures take up orientations with respect to each other that permit the side chains that protrude from each of them to interdigitate (Figure 6-22).^{193,194} This **interdigitation** is the reason that there is very little vacant space in the interior of a molecule of a protein. If it can be assumed that the configuration of minimum volume is the preferred configuration in the condensed phase, then these interdigitations promote the achievement of such a minimum volume. In order to form as many interdigitations as possible, the individual segments of secondary structure are required to assume preferred orientations with respect to each other. Viewed in this perspective, packing is a structural force just as the formation of hydrogen bonds between buried donors and acceptors on the side chains of the amino acids would be a structural force, but packing is more important.

The **orientation** between two α helices, two sheets of β structure, or an α helix and a sheet of β structure can be assigned an **angle Ω** .^{195,196} The angle Ω between two α helices is the angle between their two axes (Figure 6-23).¹⁹⁷ The sign on Ω is given by the right-hand rule. Consequently, the angle Ω in Figure 6-23 has a negative sign. Because the pattern in which the amino acids protrude from an α helix has a 2-fold rotational axis of pseudosymmetry at each position in the α helix (focus on position i at the right of Figure 6-23), the axis of the α helix has no direction associated with it and a value of $\Omega = -50^\circ$ is equivalent to a value of $\Omega = +130^\circ$. The angle Ω between two sheets of β structure is the angle between the direction of the parallel or antiparallel strands in one sheet and the direction of the strands in the other sheet (Figure 6-24).¹⁹⁵ The right-hand rule determines the sign of angle Ω , and the angle Ω in Figure 6-24B is, therefore, negative. No distinction is made between parallel and antiparallel relationships of the strands or the amino- and carboxy-terminal ends of a given strand of β structure because all combinations of these distinct stereochemistries produce almost the same pattern in which the side chains are distributed across the face of the sheet (Figure 4-16B,C). The angle Ω between an α helix and a sheet of β structure is the angle between the axis of the α helix and the direction of the β strands (Figure 6-25D).¹⁹⁵

The most frequently observed angle Ω between two

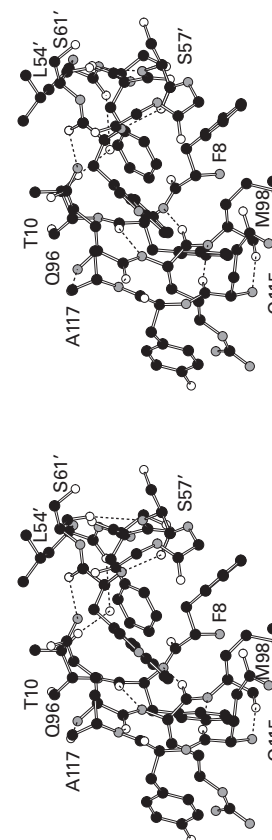
α helices¹⁹⁷ is around -50° , the angle used to construct Figure 6-23. One-third of all adjacent α helices in molecular models of globular proteins are inclined with respect to each other by an angle Ω between -60° and -40° . There is a steric explanation for this preference. When an α helix is observed from one side (Figure 6-26),¹⁹⁷ it can be seen that the side chains are arranged in sets of parallel ridges and grooves. For example, side chains 8 and 4; 19, 15, and 11; and 22 and 18 in Figure 6-26 form a set of ridges and grooves, but so do side chains 4 and 1, 11 and 8, 18 and 15, and 22 and 19. If one α helix is opposed to another, the first set of these ridges will fit into the second set of these grooves, and conversely when the angle between the two helices is -50° (Figure 6-23). An example of the interdigitation that occurs in such situations is found between two adjacent α helices in the molecular model of bovine carboxypeptidase A (Figure 6-27).¹⁹⁷

There are a number of other values for the angle Ω between two α helices that promote less favorable interdigitations of the side chains, and examples of all of them have been observed.¹⁹⁷ Because several possibilities exist and because α helices can tighten or loosen to **accommodate different angles** close to the ideal values, the distribution of angle Ω between -90° and $+90^\circ$ is fairly uniform¹⁹⁷ with the exception of the striking and sharp peak at -50° . For example, in the crystallographic molecular model of cytidine deaminase, two α helices in the interior cross at an angle Ω of 90° , but the side chains protruding from them at the interface do not pack together well.¹⁹⁸ In the distribution of angles Ω , however, there is another preferred angle represented by a broad maximum in the distribution at $+20^\circ$.¹⁹⁹ This angle Ω defines the orientation of two α helices in a coiled coil.

Suppose that the amino acids emerged from a right-handed α helix at successive angles of precisely $102\frac{1}{2}^\circ$ instead of about 99° . Every seven amino acids ($7 \times 102\frac{1}{2} = 720$) in such an α helix, the angular dispositions of the side chains would repeat precisely. Two such tightened α helices could be placed next to each other, with their axes parallel, in such a way that their side chains would interdigitate regularly along the interface (Figure 6-28).^{50,200} Every seventh side chain in one helix would sit to one side of every seventh side chain of the other, and every side chain four amino acids to the carboxy-terminal side of every seventh side chain in one polypeptide would sit to the other side of every side chain four positions to the carboxy-terminal side of every seventh side chain in the other. As a result of these interdigitations, the two α helices could comfortably fit together side by side for an indefinite length because the topography of the interface would repeat precisely every seven amino acids.

α Helices, however, do not have angles between successive side chains of $102\frac{1}{2}^\circ$ but less than that. Crick²⁰¹ pointed out that such an interface, permitting the advantageous interdigitation and repeat every seven

Figure 6-22: Interdigitation of the side chains in the core of the crystallographic molecular model (Bragg spacing ≥ 0.14 nm) of human HLA class I histocompatibility antigen A-2.^{193,194} The three segments of polypeptide to the left of the figure are from three strands of an antiparallel β sheet and together form one of its pleats. The segment to the right is a β turn of type I the side chains of which interdigitate with the side chains on the two edges of the pleat. The donors for hydrogen bonds on the side chain of Glutamine 96 find acceptors on the acyl oxygen of Tryptophan 60' and the aromatic π system of Phenylalanine 56', and the donor on the side chain of Threonine 10 finds an acceptor on a buried molecule of water. Although not shown, the donor on the side chain of Tryptophan 60' finds an acceptor on the side chain of Aspartate 122. The pleated β sheet is from the α subunit of the protein, numbered (unprimed) according to its amino acid sequence, and the β turn is from the β subunit of the protein, numbered (primed) according to its sequence. This drawing was produced with MolScript.⁵⁷³



side chains, nevertheless could be retained if the two α helices, instead of being parallel to each other, twisted around each other in a left-handed coiled coil such that the twist of the coiled coil exactly compensated for the difference between the actual angle between successive amino acids in the α helix and $102\frac{1}{2}^\circ$. If the actual angle between successive amino acids in a right-handed α helix is 99° ,^{8,22} the two α helices would have to twist around each other in a left-handed sense at $-3\frac{1}{2}^\circ$ for

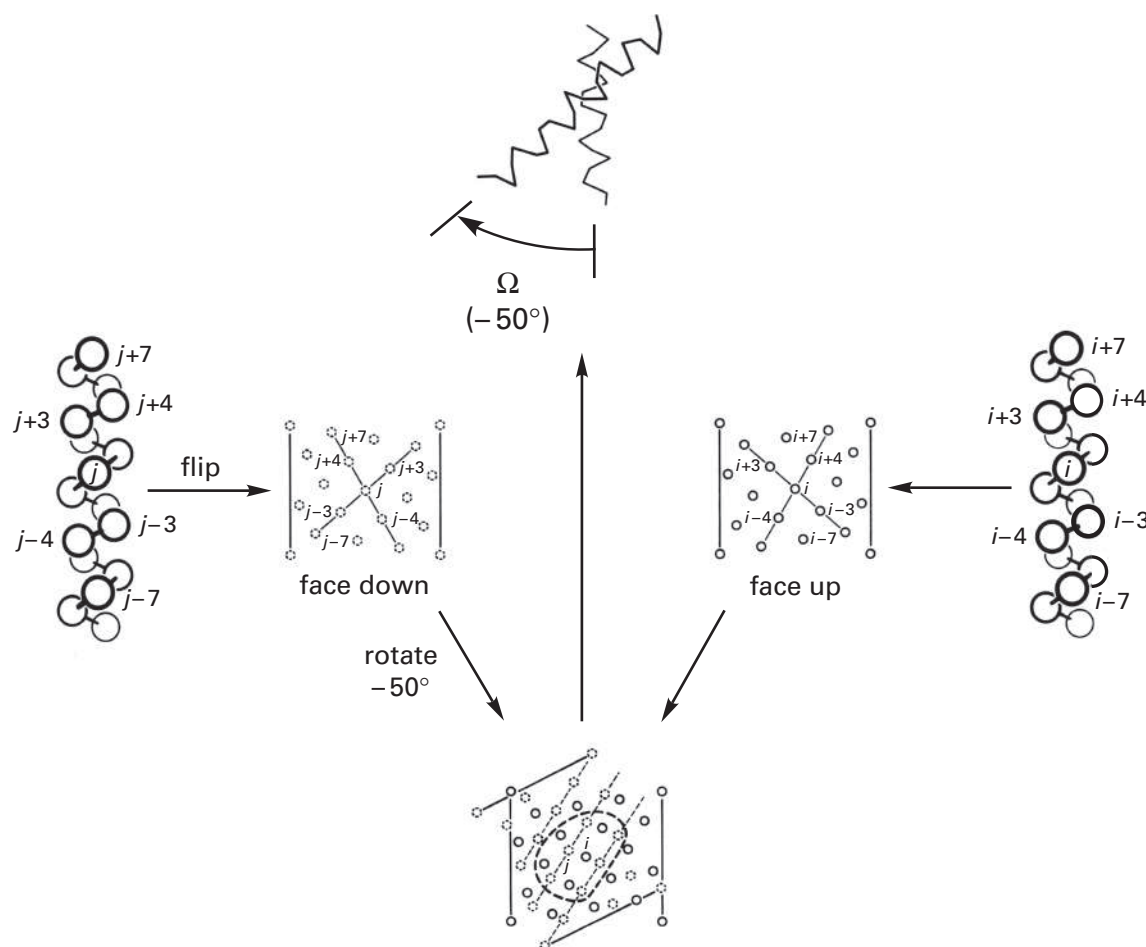


Figure 6-23: Use of superimposed helical nets¹⁹⁶ to describe the contacts at an interface between two α helices.¹⁹⁷ (Top) The angle between two adjacent α helices, i and j , is defined as the angle Ω between the two axes; its sign is determined by the right-hand rule. (Right) α Helix i in a vertical orientation is numbered out from its center. The central amino acid is given the designation i ; those below are designated by negative integers, and those above, by positive integers. The relative orientations in which amino acids $i-7$, $i-4$, $i-3$, i , $i+3$, $i+4$, and $i+7$ are distributed can be projected onto a plane tangential to the position of amino acid i . These seven projected points define a unique lattice, or helical net. For α helix i the lattice is face-up, in the orientation of the original α helix. (Left) α Helix j also defines the same lattice as that defined by α helix i , but, because α helix j is to be opposed to α helix i , face to face, the helical net for α helix j is flipped over, face-down. (Bottom) The two helical nets, the one for α helix i face-up and the one for α helix j face-down, are then opposed and rotated with respect to each other until maximum interdigitation of the lattice points is achieved. The angles at which maximum interdigitation occurs in the helical nets will be the angles at which maximum interdigitation occurs between the amino acids of the two α helices. Adapted with permission from ref 197. Copyright 1981 Academic Press.

every position in one of the sequences. Although the actual twist of a coiled coil should reflect a tradeoff between energy required to tighten the α helix and energy required to bend the α helix into the supercoil, in the coiled coil of tropomyosin the twist is -3.4° to -3.9° for each position in the sequence,^{50,202,203} and in the one from general control protein GCN4 (Figure 6-29)²⁰⁴⁻²⁰⁶ it is -3.6° to -3.9° , values that seem almost too close to the expected one.

The original **coiled coil of α helices** predicted from these geometric arguments contained two parallel α helices. The coiled coils formed by tropomyosin and general control protein GCN4 are coiled coils of two parallel α helices of identical sequence. There are, however, examples of coiled coils of two parallel α helices of nonidentical sequence,²⁰⁷ two antiparallel α helices of nonidentical

sequence,²⁰⁸ three parallel α helices of identical sequence,²⁰⁹ three parallel α helices of nonidentical sequence,^{200,210} three antiparallel α helices of identical sequence,²¹¹ three antiparallel α helices of nonidentical sequence,²¹² four parallel α helices of identical sequence,^{205,213,214} four antiparallel α helices of nonidentical sequence,^{215,216} five parallel α helices of identical sequence,²¹⁷ five antiparallel α helices of nonidentical sequence,²¹⁸ and 12 antiparallel α helices of nonidentical sequence producing a cylinder with a hollow center.²¹⁹ There is even an example of a coiled coil of four antiparallel α helices that coils around another copy of itself to form a coiled coil of coiled coils (Figure 6-30).²²⁰

That both **parallel and antiparallel** arrangements are observed must follow from the two facts that an α helix has a **pseudo-2-fold axis of symmetry** with

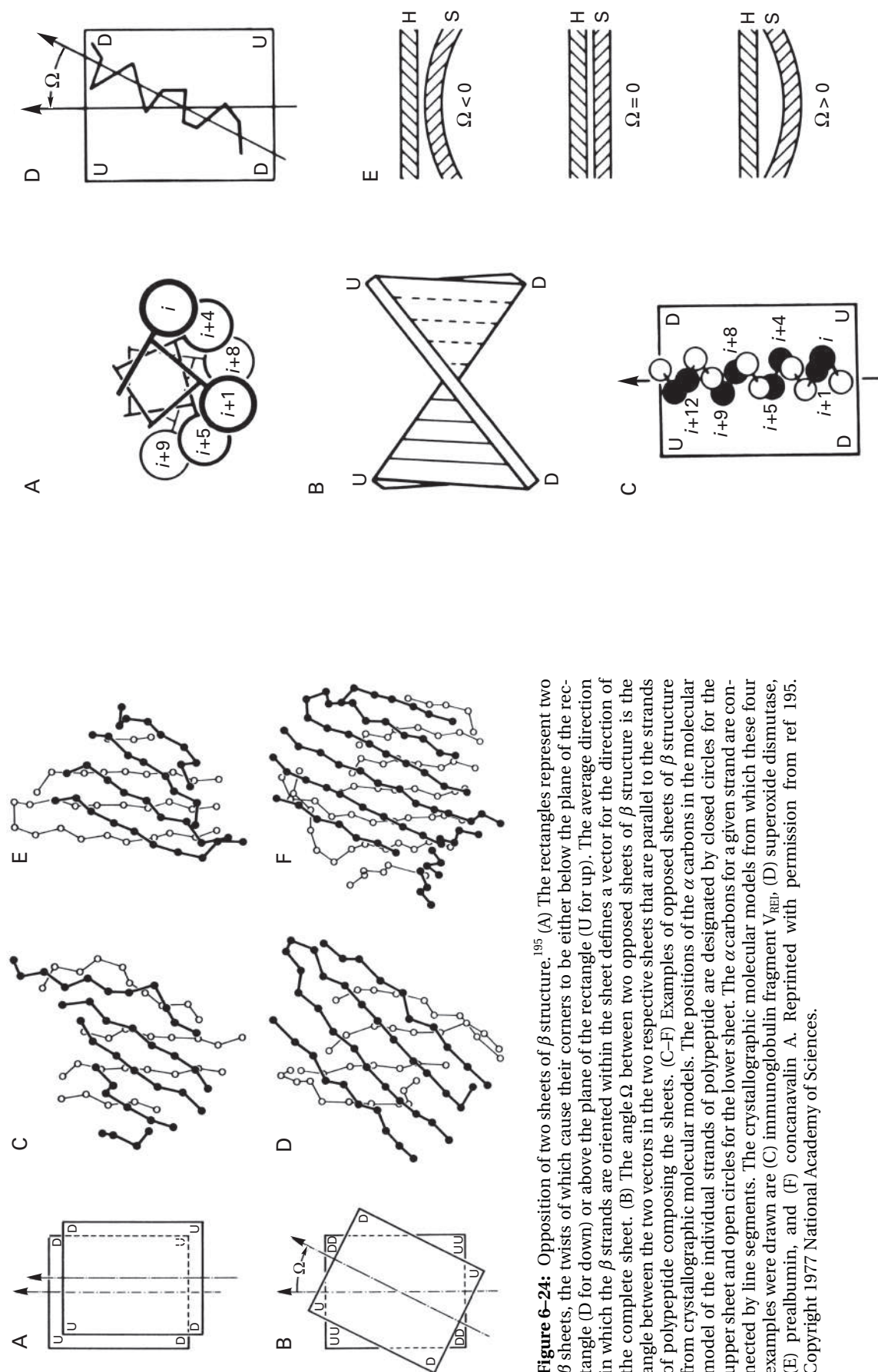


Figure 6-24: Opposition of two sheets of β structure.¹⁹⁵ (A) The rectangles represent two β sheets, the twists of which cause their corners to be either below the plane of the rectangle (D for down) or above the plane of the rectangle (U for up). The average direction in which the β strands are oriented within the sheet defines a vector for the direction of the complete sheet. (B) The angle Ω between two opposed sheets of β structure is the angle between the two vectors in the two respective sheets that are parallel to the strands of polypeptide composing the sheets. (C–F) Examples of opposed sheets of β structure from crystallographic molecular models. The positions of the α carbons in the molecular model of the individual strands of polypeptide are designated by closed circles for the upper sheet and open circles for the lower sheet. The α carbons for a given strand are connected by line segments. The crystallographic molecular models from which these four examples were drawn are (C) immunoglobulin fragment V_{HET} , (D) superoxide dismutase, (E) prealbumin, and (F) concanavalin A. Reprinted with permission from ref 195. Copyright 1977 National Academy of Sciences.

Figure 6-25: Packing of an α helix on a β sheet.¹⁹⁵ (A) View down an α helix of the orientations of the amino acids on one face. (B) The usually observed twist of a β sheet (Figure 6-9) with the corners designated as up or down as in Figure 6-24. (C) An α helix sitting on the surface of a twisted β sheet. The amino acids on the lower face of the α helix are in black and designated by number. The displacements of the four corners of the twisted β sheet are designated by letters. (D) The angle Ω between the α helix and the β sheet is the angle between the axis of the α helix and the vector parallel to the strands of polypeptide in the β sheet. (E) The three relationships that are possible between the straight axis of the α helix (H) and the various curvatures in a twisted β sheet (S). The curvature encountered by the α helix is determined by the value of angle Ω . Reprinted with permission from ref 195. Copyright 1977 National Academy of Sciences.

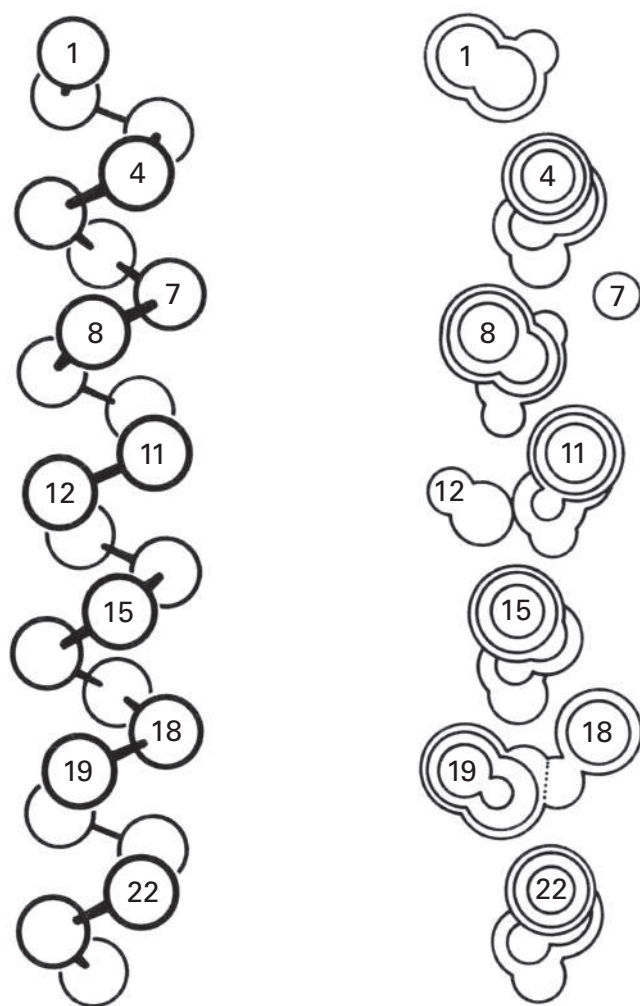


Figure 6-26: Orientation of the side chains in an α helix.¹⁹⁷ (A) The amino acids in an α helix are numbered consecutively from top to bottom with the front face accentuated. (B) A topographic map of the front face of an α helix of polyalanine. The contours are at intervals of 0.05 nm and the topmost contour, that surrounding the numbers 4, 15, and 22, is at 0.475 nm from the axis of the α helix. The numbers are on the methyl carbons of the appropriate alanine. Reprinted with permission from ref 197. Copyright 1981 Academic Press.

respect to the emergence of its side chains and that the packing of these side chains governs the existence of a coiled coil. The twist in a coiled coil of three α helices is -3.0° to -4.0° for each position in the sequence of one if its α helices; that in one of four α helices, -1.9° to -3.0° ; and that in one of five α helices, -2.6° ,^{209,210,213,214,217} but the situation seems to be much less constrained for those with four and five α helices, for which there are examples of coiled coils with right-handed supercoiling.²²¹

Crick calculated the **diffraction pattern** of X-radiation expected from a macroscopic fiber constructed of aligned coiled coils of α helices and was able to explain why the meridional reflection in the pattern that would normally arise from the pitch of 0.54 nm for an untwisted α helix should shorten to a pitch of 0.51 nm when the α helix becomes twisted into a coiled coil. A prominent

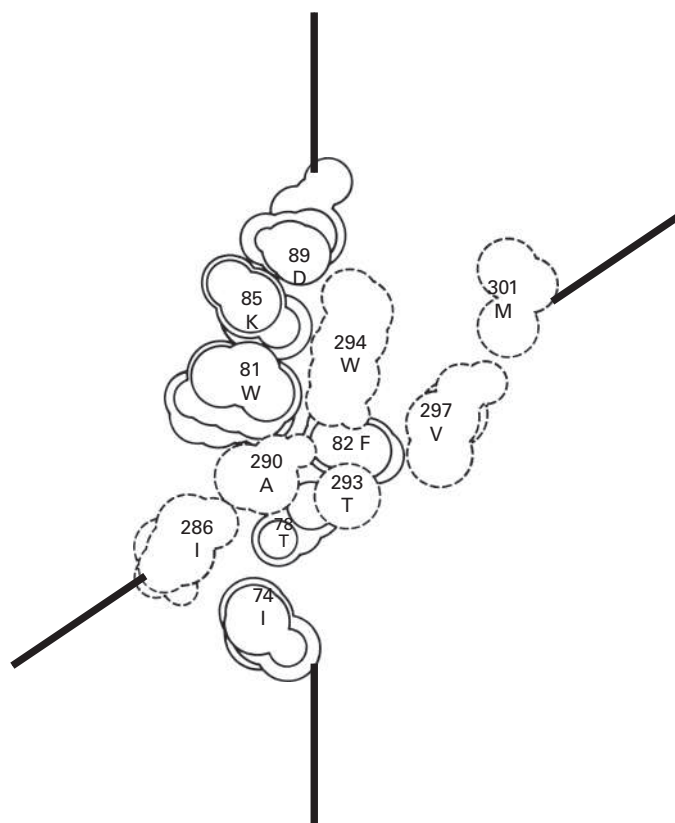


Figure 6-27: Packing of two opposed α helices in the crystallographic molecular model of bovine carboxypeptidase A.¹⁹⁷ An α helix in the molecular model comprising amino acids 72–90 is tightly opposed to an α helix comprising amino acids 285–307. The molecular model was converted into a space-filling representation and only the amino acid side chains along these two α helices were displayed. A set of three parallel planes was cut through the interface between the two helices and superposed to create a topographic map through the interface. The planes were approximately parallel to the two helical axes and at intervals of 0.1 nm. The contours of the side chains in the lower α helix are solid; those in the upper are broken. Each amino acid in the map is designated by its number in the amino acid sequence, and lines designating the axes of the two α helices are included. The angle Ω between these two helices¹⁹⁷ is -56° . Reprinted with permission from ref 197. Copyright 1981 Academic Press.

meridional reflection, representing a repeat of 0.51 nm, had been observed previously²²² in the diffraction patterns of fibers of keratin, myosin, and fibrinogen, and it is now known that such a reflection is indicative of the coiled coils of α helices in these proteins. The infrared spectra of coiled coils of α helices are also characteristic.²²³

The sequence of the polypeptides in any coiled coil of α helices can be divided into successive units, or **heptads**, each seven amino acids in length. The first and fourth amino acid in each heptad (positions *a* and *d* in Figure 6-28) are the most deeply buried amino acids in the interface between the two or more α helices in the coiled coil (Figure 6-29). These most deeply buried locations are isolated from the water surrounding the coiled coil, and the side chains sequestered there are usually

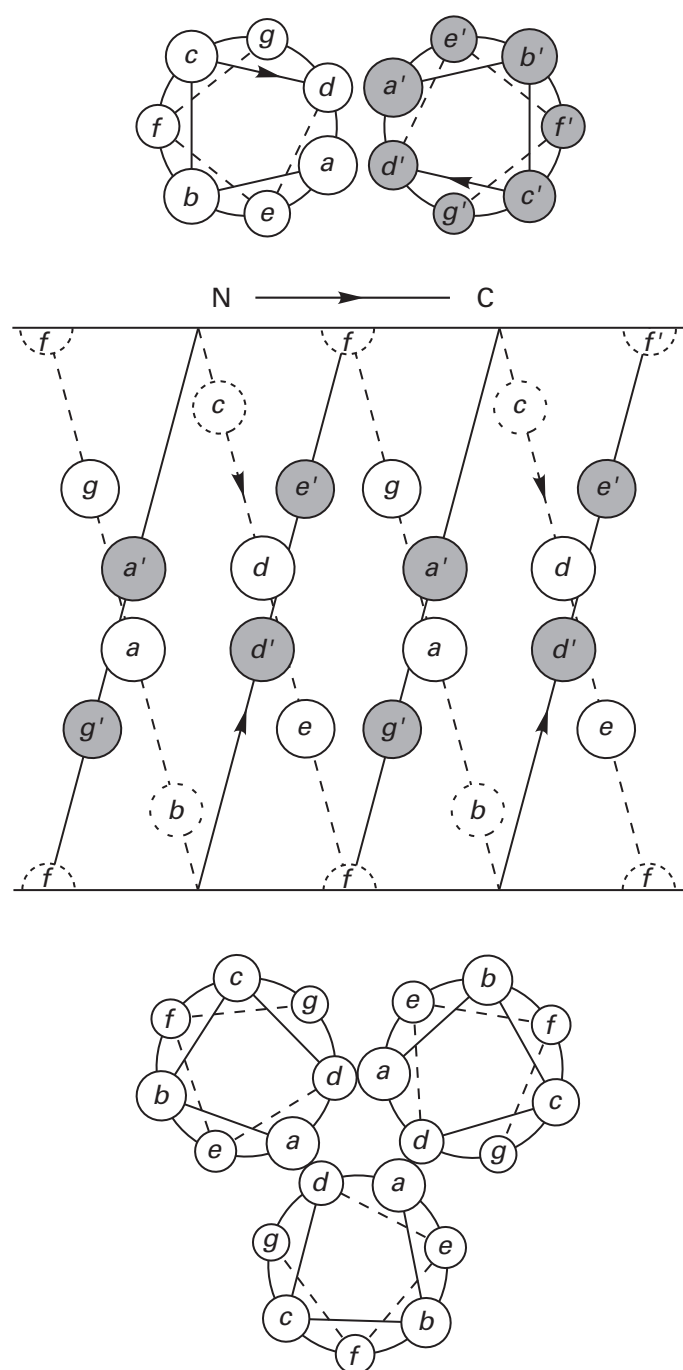
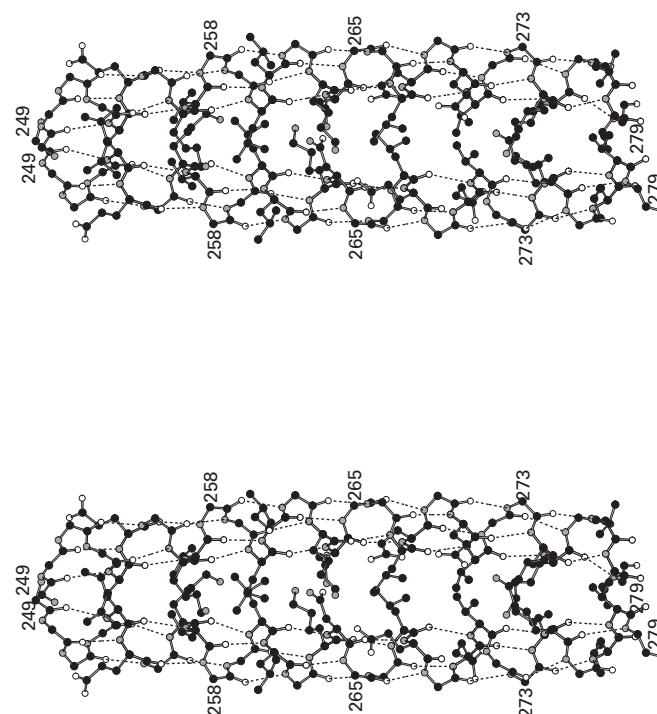


Figure 6-28: Interaction between α helices in a coiled coil.^{50,200}

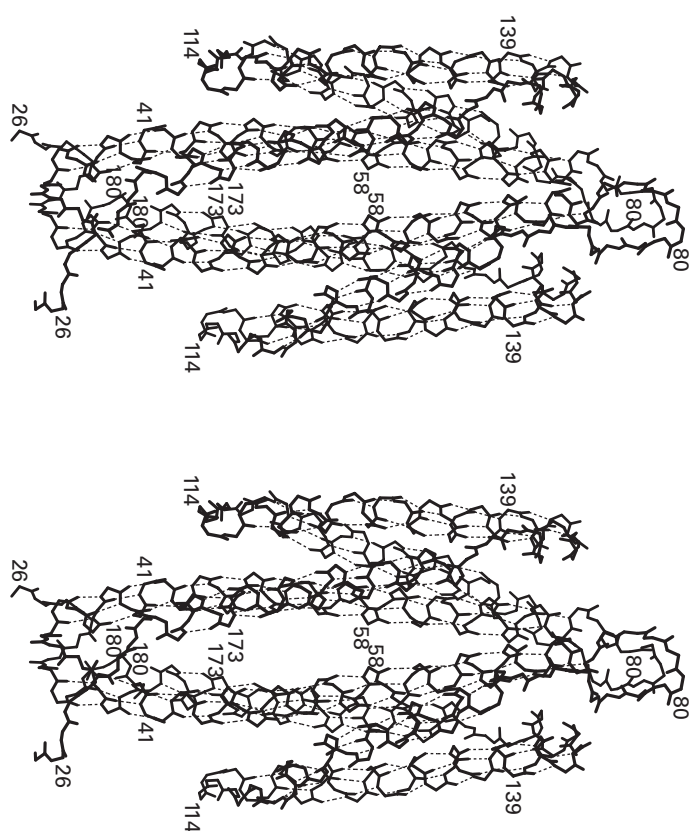
(Top) Alignment of two tightened α helices with 3.5 amino acids for each turn rather than 3.6. Amino acids from amino-terminal to carboxy-terminal are designated as *a, b, c, d, e, f*, and *g*; the view is end-on looking from amino-terminal to carboxy-terminal amino acid. Every seven amino acids the orientations would repeat, and this would place the amino acid after amino acid *g* precisely below amino acid *a* and so forth. (Middle) The two α helices in the top panel are cut along two respective lines normal to the plane of the page and passing through amino acids *f* and *f'* and then flattened, one against the other. The two resulting planes are then turned together -90° about a vertical axis so that the gray positions end up above the white. This view illustrates the interdigitations of amino acids *a* and *d*. (Bottom) Three tightened α helices running parallel to each other. In this arrangement also, amino acids *a* and *d* can interdigitate.

Figure 6-29: Drawing of the crystallographic molecular model (Bragg spacing ≥ 0.18 nm) of a coiled coil of α helices.²⁰⁴ A peptide 31 amino acids long with the amino acid sequence from Arginine 249 to Glycine 279 of general control protein GCN4 from *S. cerevisiae* was synthesized. In solution two copies of this peptide spontaneously form a coiled coil of α helices, which is the native structure of this portion of the full-length protein. The coiled coil was crystallized, and a crystallographic molecular model was built from a map of electron density calculated from a data set gathered from these crystals. The numbering in the figure is that of the amino acid sequence of the full-length transcription factor, only a segment of which is represented by the synthetic peptides. Only the side chains of the amino acids in the *a, d, e, g, a', d', e', g'* positions of the coiled coil (Figure 6-28) are included in the figure to emphasize the interface between the two α helices.⁵⁷³ This drawing was produced with MolScript.



hydrophobic amino acids such as leucine, valine, isoleucine, alanine, phenylalanine, tyrosine, and methionine.⁵⁰ The hydrophobic amino acid can also be a cystine as in the antiparallel coiled coil of α helices in carboxypeptidase C from *Saccharomyces cerevisiae* (Figure 6-19).¹³⁸ An α helix that has a heptad repeat of hydrophobic amino acids is an amphipathic α helix (Figure 6-8). There are a few interesting exceptions to the rule that coiled coils are formed from amphipathic α helices, such as the chloride ion chelated by five symmetrically arrayed glutamines in the center of the coiled coil of five parallel α helices in extracellular matrix protein COMP.²¹⁷

Figure 6-30: Coiled coil of two parallel coiled coils of four antiparallel α helices comprising the complete crystallographic molecular model (Bragg spacing ≥ 0.20 nm) of the ligand binding portion of the methyl-accepting chemotaxis protein II from *Salmonella typhimurium*.²²⁰ The complete protein is built from two copies of a folded polypeptide 553 amino acids long. The portion of the protein containing the two identical segments of polypeptide between Glycine 26 and Arginine 188 was expressed as a fragment, purified, and crystallized. The drawing is of the crystallographic molecular model of this fragment of the entire protein. The numbering is for that of the complete protein. Each of the two polypeptides forms a structure built around a coiled coil of four antiparallel α helices. These two coiled coils are in turn coiled around each other in parallel.



Each of the amino acids in the **core of the coiled coil** (amino acids *a* and *d* in Figure 6-28) on one α helix is sandwiched between its partner on the opposite coiled coil (*a'* and *d'*) and the amino acid to the amino-terminal (*g*) or carboxy-terminal (*e*) side, respectively, of its partner (Figure 6-28, middle panel). For example, Leucine 253 in one of the α helices in the coiled coil of general control protein GCN4 (Figure 6-29)²⁰⁴ is sandwiched between Leucine 253 and Glutamate 254 on the other; Valine 257, between Lysine 256 and Valine 257 on the other; Leucine 260, between Leucine 260 and Leucine 261 on the other; Leucine 267, between Leucine 267 and Glutamate 268 on the other; and so forth. The side chains of the amino acids on the flanking positions (*g* and *e*, respectively) are often ones that are hydrophobic distally

and hydrophilic peripherally, such as glutamate, lysine, arginine, and glutamine,²⁰⁸ that can provide hydrogen-carbon bonds to cover the hydrophobic side chains in the core before they enter the water fully.

In the crystallographic molecular model of the coiled coil from transcription factor GNC4, the hydrophobic amino acids of the heptad repeats interdigitate along the interface between the two α helices to form the hydrophobic core of the structure and to produce the supercoil. They and those that flank them pack so closely together and so efficiently that there is almost no vacant space in the core of the structure. The two identical α helices are parallel to each other and packed in precise register so that each central hydrophobic side chain packs against its twin from the other α helix. The asparagines at position 16 pack side by side in the core of the coiled coil, and because the hydrogen bond between them lies upon the axis of symmetry of the coiled coil, the two possible hydrogen bonds between the respective amido protons and acyl oxygens are present in the map of electron density as alternative conformations.

Because a coiled coil has such a regular structure, it is possible to design **synthetic peptides** that assume a coiled coil of α helices spontaneously^{211,224} by incorporating heptad repeats into their sequences. Almost any amphipathic α helix has the potential to form a coiled coil. It is not possible, however, to predict what type of coiled coil they will form²¹¹ because this decision seems to be made on the basis of the details of the packing in the hydrophobic core between the two α helices. For example, the packing in the interior of a coiled coil of three α helices is so tight that the inability of the three equivalent leucines to adopt the proper angles χ_1 in a parallel arrangement causes the three copies of one of these synthetic peptides to form a coiled coil of antiparallel α helices instead.²¹¹ Likewise, when mutations were made to hydrophobic amino acids in the heptad repeat of the portion of general control protein GCN4 that normally forms a symmetric coiled coil of two parallel α helices (Figure 6-29), the mutant peptides unexpectedly formed coiled coils of three and four parallel α helices.²⁰⁵ Interactions among the amino acids facing the water also influence the type of coiled coil formed.²²⁵ All of these results suggest that the different types of coiled coils have similar structural requirements.

Although there are exceptions in which a coiled coil has a right-handed twist,^{221,226} the left-handed twist of most coiled coils causes the angle Ω between any two of the α helices to be $+18^\circ$ to $+24^\circ$ (Figures 6-29 and 6-30).^{209,210,217} Such structures contribute to the broad maximum at 20° in the distribution of angle Ω between two α helices. As the number of α helices bundled together becomes larger and as the constraints of the overall tertiary structure of the protein are exerted, the regular packing of the coiled coil breaks down. Nevertheless, in a bundle of α helices stacked next to each other at angles Ω around $+20^\circ$, such as the one in

protein R2 from ribonucleoside-diphosphate reductase of *E. coli*²²⁷ and the one in δ -endotoxin CryIIIA(a) from *Bacillus thuringiensis* subspecies *tenebrionis*,²²⁸ there are hints of coiled coils.

Although there are examples of β sheets fully exposed on both faces to solvent,^{229–232} almost all β sheets are found packed against other β sheets or sandwiched between layers of α helices, usually in the most buried regions of a molecule of protein. When β sheets pack against each other, there are only two orientations observed with significant frequency. In a sheet of β structure, the side chains of the amino acids along a given strand alternate in protruding above the sheet and below the sheet (Figure 4–16B,C). On a given side of the sheet, the protrusions form an approximately square array aligned with the β strands (Figure 6–31).²³³ There are two ways for two square arrays to interdigitate if they are flat, one with angle $\Omega = 0^\circ$ and the other with angle $\Omega = 90^\circ$.

There are examples of two β sheets packing against each other with an angle Ω of 0° , each strand of the one sheet running almost exactly between two of the strands of the other,²³⁴ as the closed fingers of one hand fit into the grooves of the closed fingers of another, but because β sheets twist (Figure 6–9) and because the array is not exactly square (Figure 6–31), sheets of β structure usually pack at angles Ω of $-30^\circ \pm 15^\circ$ (Figure 6–24).^{195,233,235} Were two twisted β sheets to be **packed in parallel** to each other with angle Ω equal to 0° , their two (indicated by U and D in Figure 6–24A,B) twists would match. The twist, however, causes the side chains protruding from the bottom of the top sheet to lean one way and those protruding from the top of the bottom sheet to lean the other,^{233,235} and when the side chains interdigitate the sheets cannot be parallel to each other but are forced to assume an angle Ω around -30° . Even at an angle Ω of -30° , however, the twists of the two β sheets fit together effectively.⁸¹ The associations between the two parallel β sheets in prealbumin illustrate the stereochemistry at such an interface between two twisted sheets (Figures 6–24E and 6–32).²³⁵ The preference for an angle Ω of -30° , however, is not a strong one because local disruptions in the β strands can cause the angle to shift to positive values.²³⁶ Even β sheets as narrow as two β strands in an antiparallel β hairpin will stack against each other with the strands parallel.²³⁷

An angle Ω equal to 90° , the other value predicted for the stacking of two square arrays, is also a preferred orientation observed for two stacked β sheets.²³⁸ When two sheets of β structure are aligned at 90° to each other, the twists instead of fitting together oppose one another, and one pair of diagonal corners is closer together than the other pair of diagonal corners (Figure 6–33).²³⁸ It is usually at the close corners that the polypeptide connects one sheet to the other.^{238,239} The interdigitations of the amino acids in the three **layered orthogonal β sheets** of penicillopepsin illustrates the packing observed between orthogonal β sheets (Figure 6–34).²³⁸

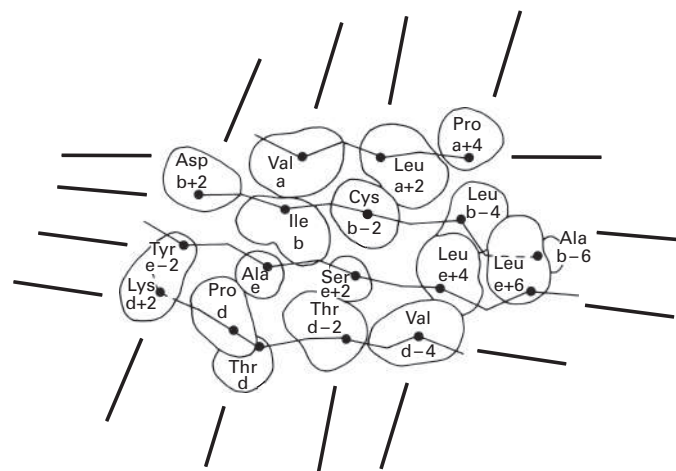


Figure 6–31: Distribution of side chains on one side of a β sheet.²³³ The representation is that of a four-stranded antiparallel β sheet in the crystallographic molecular model of the light constant domain from γ globulin (λ) New. The four strands of polypeptide of the β sheet in the crystallographic molecular model were placed in the plane of the page, and the α carbons were connected with line segments. The α carbon of each amino acid the side chain of which protruded below the plane of the page was marked with a dark dot. The side chains protruding below the page from these positions were displayed in space-filling format, and their projections on the plane of the page were traced. Each projection is labeled with the name of the particular amino acid. The amino acids in one column, valine, isoleucine, alanine, and proline and threonine, were assigned letters determined by the order in which their strands occur in the amino acid sequence. Proline d and Threonine d, which are adjacent in the amino acid sequence, were treated as the same amino acid to retain the pattern. The other amino acids were numbered in the direction in which the particular strand ran across the page. The view is from above and of the amino acids hanging down from the sheet. Lines indicate a lattice on which the α carbons of these amino acids lie. Reprinted with permission from ref 233. Copyright 1981 Academic Press.

Regardless of whether the angle Ω between β sheets is -30° or 90° , the side chains between the two sheets are **well buried**, are usually hydrophobic (Figures 6–32 and 6–34), and fit together tightly with a minimum of empty space. An interesting exception to these rules is found in a family of small proteins that use the interior of an orthogonal sandwich of two β sheets to bind a fatty acid. In such proteins, space is made within the interface into which fits the hydrophobically compatible fatty acid.²⁴⁰

The packing in the center of a **β barrel** (Figure 6–11) can be considered to be a special case of the packing of β sheets. The alternate amino acids along each β strand that enter the core of the β barrel occur in layers roughly perpendicular to the axis of the hyperboloid.⁷⁷ In a β barrel of eight strands, the four members of each of these layers come from alternate strands (Isoleucine 51, Leucine 112, Leucine 161, and Isoleucine 212; Serine 85, Leucine 135, Glycine 182, and Leucine 234; and Glutamate 53, Lysine 114, Glutamate 163, and Glutamate 214 in Figure 6–11). Within each layer of this cylindrical cake, the four side chains pack against each other and each successive layer packs against the layer below it (the

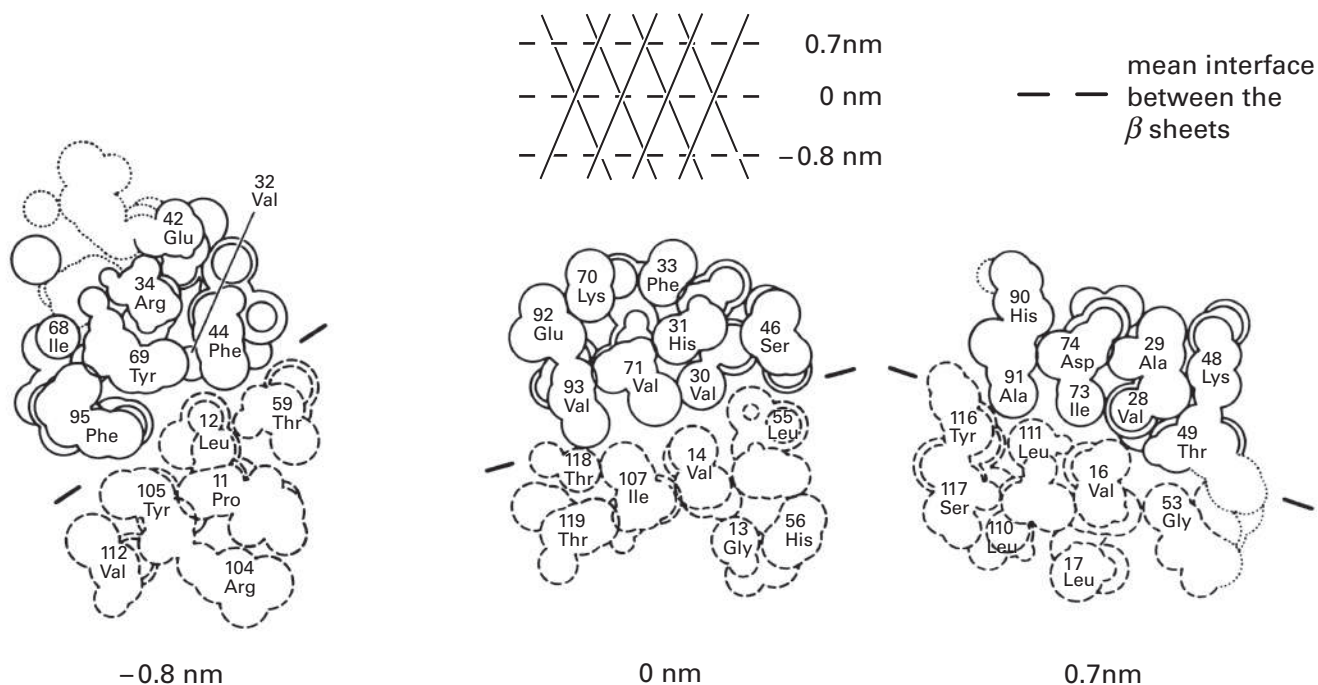


Figure 6–32: Packing of the amino acids at the interface between two opposed β sheets in the crystallographic molecular model of prealbumin.²³⁵ A tracing of the α carbons of the strands in this structure is presented in Figure 6–24E. This is represented diagrammatically in the upper inset, where the locations of the horizontal sections through the structure are designated by their positions in nanometers relative to the central section. The sections are normal to the two β sheets, and the strands run approximately normal to the planes of the sections. The amino acids in the upper sheet (in which the four strands run parallel, antiparallel, parallel, antiparallel) are enclosed in solid lines; those in the lower sheet (in which the four strands run parallel, antiparallel, antiparallel, parallel) are enclosed in broken lines. The straight lines indicate the orientation of the interface, which twists in a right-handed sense as the sections proceed through the structure. Reprinted with permission from ref 235. Copyright 1981 National Academy of Sciences.

layer of 85, 135, 182, and 234 is sandwiched between that of 51, 112, 161, and 212 and that of 53, 114, 163, and 214 in Figure 6–11). In the top layer of the β barrel in Figure 6–11, the hydrophobic portions of the side chains pack against the layer below, and the four hydrophilic atoms, the nitrogen and the three oxygens, are pointed upward out of the end of the barrel.

In β barrels of five or six strands, the radii of the hyperboloids are small enough that the cylinder can

remain circular while the side chains tightly fill the central cavity, but in β barrels of eight strands, the hyperboloid is usually **flattened** into an ellipse to pack the side chains in each layer as tightly together as possible.⁷⁷ In ribonucleoside-diphosphate reductase from *E. coli*, a β sheet of five parallel strands antiparallel to a second β sheet of five parallel strands together form a β barrel of ten strands, but it is flattened so that the two respective sheets are opposed to each other across the minor axis of the ellipse.⁹⁰ This structure is drifting in the direction of a sandwich of two β sheets at an angle Ω equal to 90° . Again, an interesting set of exceptions is that of β barrels in which a cavity is

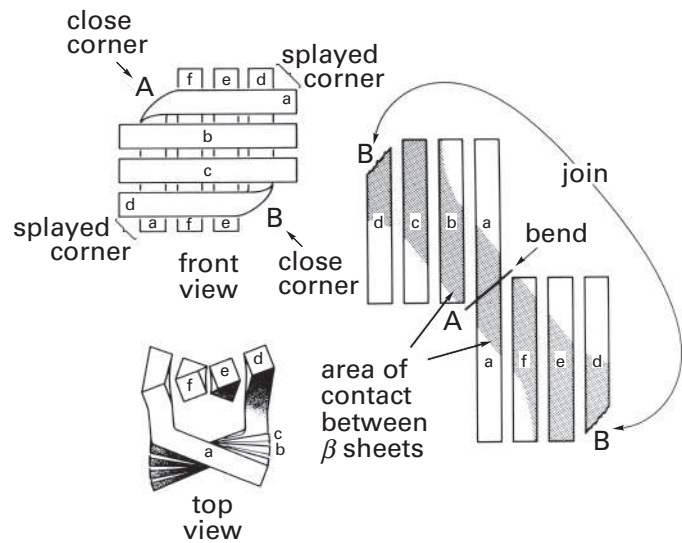


Figure 6-33: Abstract representation of the orientation of two orthogonal β sheets packed against each other that incorporates the left-handed twist usually associated with such structures.²³⁸ The front view illustrates the orthogonal disposition of the strands of the two respective sheets. The top view illustrates the twist, the fact that the strands can join at the two corners (A and B) that are brought together by the separate twists, and the fact that two of the corners are splayed by the twist. The view of the opened structure demonstrates how the stack can be produced by folding over two coplanar parallel sheets of β structure to produce an arrangement joined at the two corners by two continuous strands. A typical example of such a close corner occurs in murine adipocyte lipid-binding protein.²³⁹ The bottom two sheets of β structure in the three-layered sandwich of penicillopepsin (Figure 6-34) also have this arrangement. They are connected at two diagonal corners and splayed at the other two. Reprinted with permission from ref 238. Copyright 1982 American Chemical Society.

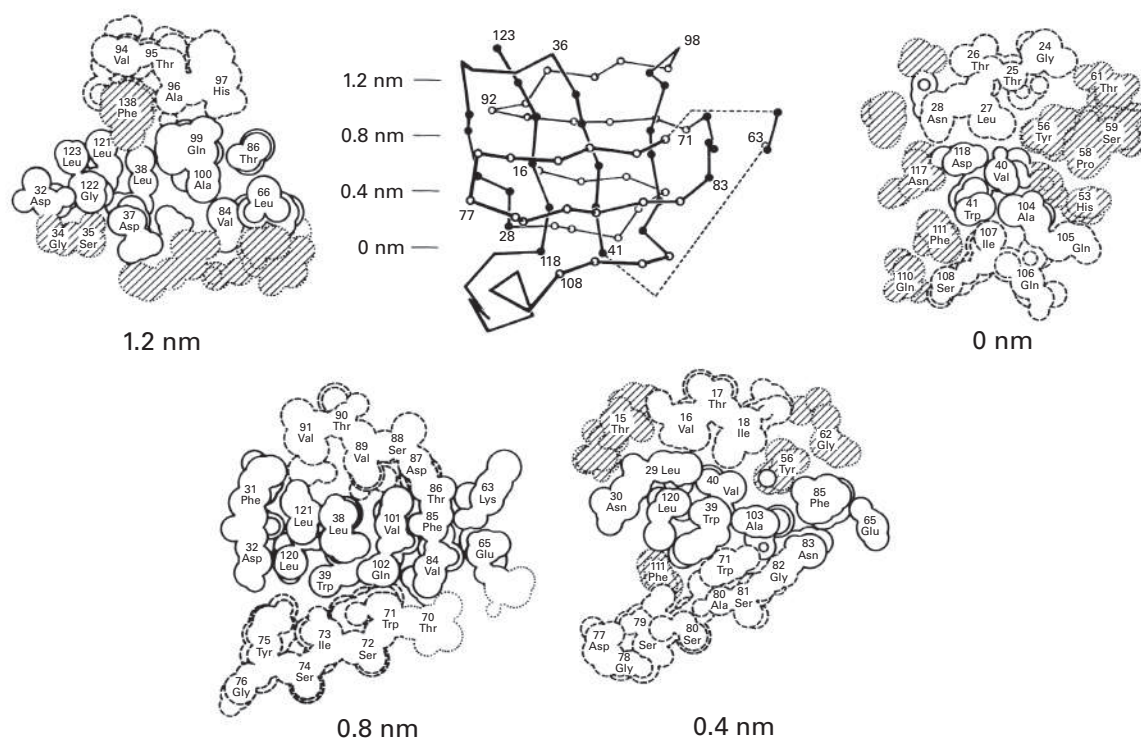


Figure 6-34: Packing of amino acids at the interface between three alternately orthogonal sheets of β structure in the crystallographic molecular model of penicillopepsin.²³⁸ In the center of the figure, the α carbon atoms of the molecular model between amino acids 16 and 123 are connected by line segments to provide a tracing of the polypeptide. The three-layered sandwich is viewed from above and consists of a three-stranded sheet of β structure (parallel, antiparallel, parallel) on top of a four-stranded sheet of β structure (parallel, parallel, antiparallel, antiparallel) on top of a second four-stranded sheet of β structure (parallel, antiparallel, parallel, antiparallel). The indicated sections 0.4 nm apart were cut through the three-layered sandwich in a space-filling representation of the molecular model. The planes of the sections were horizontal and normal to the page as indicated. The packing between the sheets can be viewed in respective cross sections arrayed in counterclockwise order from top to bottom. Amino acids are numbered, and all amino acids in a given pleated sheet are in either solid outline or broken outline. Amino acids in hatched outline are not in the sheets of β structure. In the cross sections, the strands of the sheets at the top and the bottom run parallel to the page while the sheet in the middle runs perpendicular to the page. Reprinted with permission from ref 238. Copyright 1982 American Chemical Society.

required for the function of the protein, such as the cavity in the middle of the β barrel of retinol-binding protein in which the retinol is bound.²⁴¹ In this β barrel, the ligand provides enough extra hydrophobic mass that the barrel can remain circular even though it has eight strands. The β barrel of red fluorescent protein from *Discosoma* is circular even though it is composed of 11 strands because there is an α helix running through its center.²⁴²

An α helix lying upon a sheet of β structure usually has its axis almost parallel to the strands of the sheet because the α helix is straight, the sheet is twisted, and a straight rod can contact a twisted surface only when it is either parallel or perpendicular to the axis of the twist (Figure 6-25).¹⁹⁵ The angle Ω observed²⁴³ between α helices and adjacent sheets of β structure is usually around 0° , and almost all values fall between -20° and $+10^\circ$. The exceptions are usually instances in which the angle Ω is close to 90° .^{244,245} In one of these instances, a long β sheet of four strands wraps around an α helix²⁴⁵ as one's four fingers would wrap around a cylindrical rod 3–4 cm in diameter. This grip is yet another example of the elasticity of β structure.

The interface (Figure 6-35)²⁴³ between three of the α helices and one of the β sheets in lactate dehydrogenase illustrates the fit between an α helix and a twisted β sheet in a parallel orientation. Note that the side chains from the α helices lie upon the gaps between the side chains in the sheet of β structure. Because a sheet of β structure twists appropriately, the α helices lying across its surface parallel to its strands are aligned next to each other with angles Ω of about -40° between adjacent pairs even when they cleave tightly to the surface of the sheet.²⁴³ This value for angle Ω is sufficiently close to the -50° that produces the most frequently observed type of interdigitation between two α helices. Therefore, both the interfaces between the α helices and the sheet of β structure and the interfaces among the α helices themselves can exist simultaneously in almost optimum orientations. It is also possible, however, that the twist of such a sheet of β structure arises from the requirement that the α helices upon it be positioned at the proper angles to maximize the interdigitation of their amino acids. For all of these reasons, a **twisted β sheet sandwiched between two layers of**

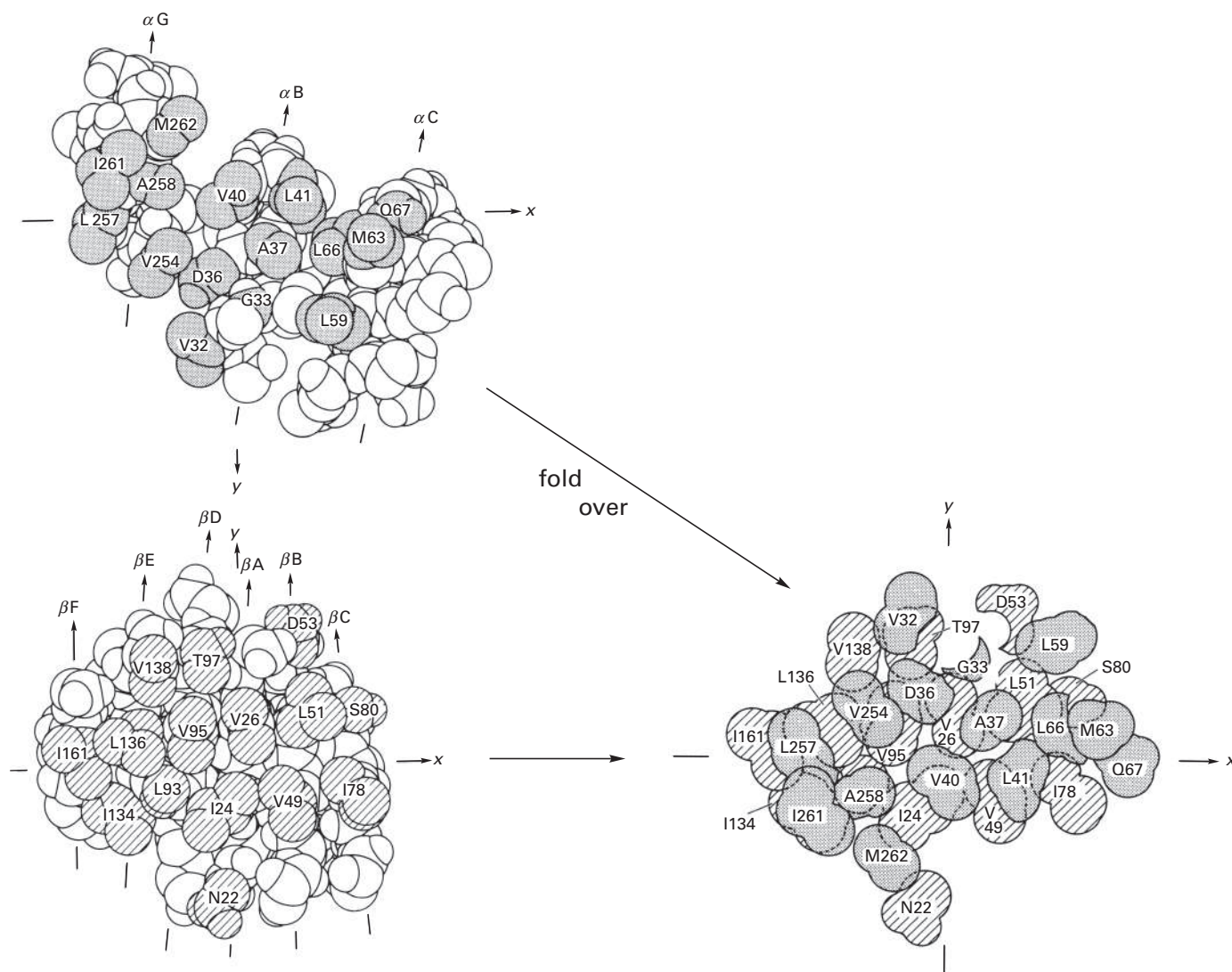


Figure 6-35: Schematic formation of the interface between three α helices and a sheet of β structure found in the crystallographic molecular model of L-lactate dehydrogenase.²⁴³ The sheet of β structure is presented in the bottom left of the figure with its strands running vertically in the y direction and the axis of the right-handed twist parallel to the x -axis. Side chains of the amino acids on the upper face of the sheet are identified and numbered by the amino acid sequence of the protein. The three α helices that will form the interface with the sheet of β structure are presented in the upper left of the figure with the face that will participate in the final interface directed upward. The axes of the three helices (αG , αB , and αC) are almost parallel to the vertical y -axis. Side chains that will participate in the interface are identified and numbered. When the three α helices are rotated 180° around the x -axis and placed upon the sheet of β structure as they are in the molecular model, the interface is produced by the interdigitation of the highlighted amino acids from the sheet and the highlighted amino acids from three α helices, respectively. It is these interdigitations that position the three α helices upon the sheet of β structure. Adapted with permission from ref 243. Copyright 1980 Academic Press.

parallel α helices is one of the most common tertiary structures.

This arrangement is also the one assumed by the **coating on a β barrel**. The most common type of β barrel is one in which all of the strands run consecutively and in parallel (Figure 6-11), and usually within the polypeptide connecting the amino-terminal end of one strand to the carboxy-terminal end of the next (for example, the connection between Cysteine 54 and Alanine 83 in Figure 6-11) there will be an α helix running along the outer surface of the β barrel parallel to the β strands of the β barrel. Consequently, the strands occur consecutively around the barrel, and the underlying repeating pattern

is β strand, α helix, β strand, α helix and so forth with extraneous segments of other secondary structure thrown in at random. Even in the hybrid β barrel in ribonucleoside-diphosphate reductase, the five parallel β strands in each of the two antiparallel β sheets forming the barrel occur consecutively and are each connected to the next by a segment of polypeptide containing an α helix running across the outside of the barrel.²⁴⁶ There is a peculiar variant of this α -helically wrapped β barrel in which each of the strands of β structure in the central barrel is replaced by an α helix to form an α -helically wrapped α -helical barrel.²⁴⁷

It is useful to imagine the interior of a molecule of

protein created by all of these arrangements as a three-dimensional **jigsaw puzzle** because this is an image that emphasizes the interdigitations among the side chains of the amino acids driving the various orientations of the secondary structures. The pieces of this puzzle, however, are neither inelastic nor invariant,²⁴⁸ and there can be flaws in its mosaic.

The **elasticity of the packing** of the amino acids in the interior of a protein is most readily demonstrated by performing site-directed mutation. When Alanine 129 in lysozyme from bacteriophage T4 is replaced with leucine, the stability of the protein decreases²⁴⁹ by 6 kJ mol⁻¹, but its structure is affected only in the vicinity of the mutation. There it expands, most notably at Leucine 121, in response to the increase in the size of the side chain at position 129 (Figure 6-36).²⁵⁰ When Valine 30, located between two β sheets in the core of human transthyretin is replaced with methionine, the β sheets move apart by 0.1 nm to accommodate the consequent steric effect.²⁵¹ The usual response to mutations such as these that increase the volume of matter in the hydrophobic core of a protein is that the structure expands in response to the local increases in volume and the stability of the protein decreases,²⁵² occasionally catastrophically.²⁵³ The strain of the increase in size can also be accommodated by a conformational change of an adjacent side chain to a significantly different rotamer.²⁴⁹

There is **never only one invariant arrangement** of side chains that can solve the problem of filling the space between segments of secondary structure in the hydrophobic core of a protein. For example, in the plastocyanin from *Enteromorpha prolifera* the surface of one β sheet uses Isoleucine 19, Isoleucine 96, and Valine 82 to conform to the surface of a neighboring β sheet that is formed from Valine 3, Phenylalanine 29, and Isoleucine 39, while the plastocyanin from *Populus nigra* uses Phenylalanine 19, Valine 96, and Phenylalanine 82 to conform to the same surface composed of Valine 3, Phenylalanine 29, and Isoleucine 39.²⁵⁴ Leucines 84, 91, 99, 118, 121, and 133 and Phenylalanine 153 of lysozyme from bacteriophage T4 can all be replaced simultaneously with methionines, and although the mutant is 21 kJ mol⁻¹ less stable than the wild type, it has the same overall structure with the exception that the hydrophobic cluster formed by these seven side chains of almost equal volume to the cluster that it replaced is now a different puzzle, just as compact as the first.²⁴⁸

Many proteins in their native states have **cavities** of various sizes* in their hydrophobic cores that are flaws in

the mosaic of the puzzle,²⁵⁷ some large enough to bind random hydrophobic ligands.²⁵⁸ When larger amino acids are replaced by smaller ones through a site-directed mutation, an unnatural cavity is formed, and the contraction of the structure surrounding the artificially created cavity^{256,259-261} again illustrates the elasticity of the puzzle. Although this contraction is usually incomplete, leaving a definite cavity where one was not present before, when Isoleucine 29 in lysozyme from bacterio-

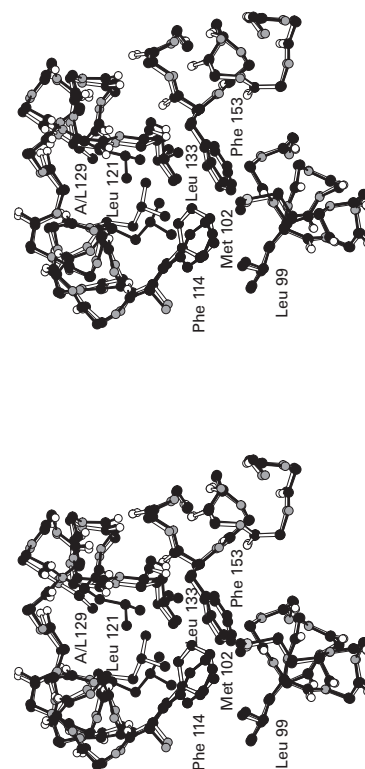


Figure 6-36: Elastic expansion within the hydrophobic core of the crystallographic molecular models (Bragg spacing ≥ 0.19 nm) of lysozyme from bacteriophage T4.^{250,571} Alanine 29 in the wild-type protein was replaced with a leucine by site-directed mutation. When the crystallographic molecular model of the portion of the mutant protein surrounding the mutation (black bonds) is superposed on the crystallographic molecular model of the same portion of the wild-type protein (white bonds), the details of the expansion of the molecule in the vicinity of the mutation are readily observed. The greatest displacement is that of Leucine 121, but Leucine 133 also moves outward as well as the entire α helix containing Leucine 129 and Leucine 133. Phenylalanine 114 is forced to assume a less advantageous rotamer (see 6-5 and Figure 6-21). This drawing was produced with MolScript.⁵⁷³

* The calculation of the volume and extent of cavities inside crystallographic molecular models is not straightforward.^{255,256} The volume accessible to a spherical probe is usually calculated, and the radius chosen for the probe has a dramatic effect on what is regarded as a cavity, what is regarded as its volume, and what is regarded as its dimensions because the smaller the probe, the more easily it slips between the atoms.

phage T4 is replaced with alanine, the structure surrounding the site of the mutation collapses to such a degree that no discernible cavity remains.²⁵⁶ When an unnatural cavity is formed by site-directed mutation, the stability of the protein usually decreases.^{252,259} It has been proposed that this instability produced upon the mutation of a larger amino acid to a smaller suggests that naturally occurring cavities in proteins destabilize their structure,¹²⁷ but it is not possible to extrapolate from the effects resulting from artificial changes performed by site-directed mutation to the effects of changes produced by natural selection.

It has been argued that because the ability of a particular polypeptide to form a particular tertiary structure is not drastically affected by extensive replacement of amino acids in its core by site-directed mutation^{248,262} and because the volume in the interior of a molecule of protein can be filled with a number of different arrangements of the normally available hydrophobic side chains,²⁶³ the packing of the amino acids cannot dictate the tertiary structure that results when the polypeptide folds. Such arguments, however, ignore the fact that it is the **overall pattern in which the side chains emerge** from the secondary structures, not the identity of those side chains, that dictates the values of the angles Ω and hence the tertiary structure. The fact that the details of the packing beyond these dictations display such tolerance is not remarkable because it has long been known that evolution by natural selection frequently performs similar replacements. From a consideration of the logic of the interdigitations that are observed in naturally occurring proteins, it can be concluded that it is such interactions among the side chains that produce the relative orientations assumed by the secondary structures and that these orientations are crucial to creating the tertiary structure of a protein.

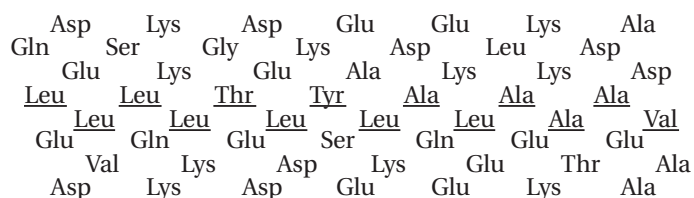
Suggested Reading

- Chothia, C., Levitt, M., & Richardson, D. (1977) Structure of proteins: packing of α -helices and pleated sheets, *Proc. Natl. Acad. Sci. U.S.A.* 74, 4130–4134.
- Word, J.M., Lovell, S.C., LaBean, T.H., Taylor, H.C., Zalis, M.E., Presley, B.K., Richardson, J.S., & Richardson, D.C. (1999) Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms, *J. Mol. Biol.* 285, 1711–1733.
- Baldwin, E., Xu, J., Hajiseyedi, O., Baase, W.A., & Matthews, B.W. (1996) Thermodynamic and structural compensation in “size-switch” core repacking variants of bacteriophage T4 lysozyme, *J. Mol. Biol.* 259, 542–559.

Problem 6–4: The following is a segment of amino acid sequence from the coiled coil region of human epidermal keratin.²⁶⁴

T A A E N E F V T L K K D V D A A Y M N K
V E L Q A K A D T L T D E I N F L R A L Y
D A E L S Q M Q T

- (A) Write out this sequence in the same format⁵⁰ as the following diagram of a portion of the amino acid sequence from the coiled coil of α -tropomyosin:



In your diagram place the appropriate amino acids from the sequence of human epidermal keratin along the center line as was done in the diagram of the sequence of α -tropomyosin.

- (B) What is the role of the amino acids placed along the center line?
- (C) Circle the two amino acids in your diagram that do not seem to fit this role.
- (D) How may they be excused?

Problem 6–5: In the coiled coil of α helices shown in Figure 6–29, why do Lysine 263 and Glutamate 268 and Lysine 275 and Glutamate 270 form hydrogen bonds?

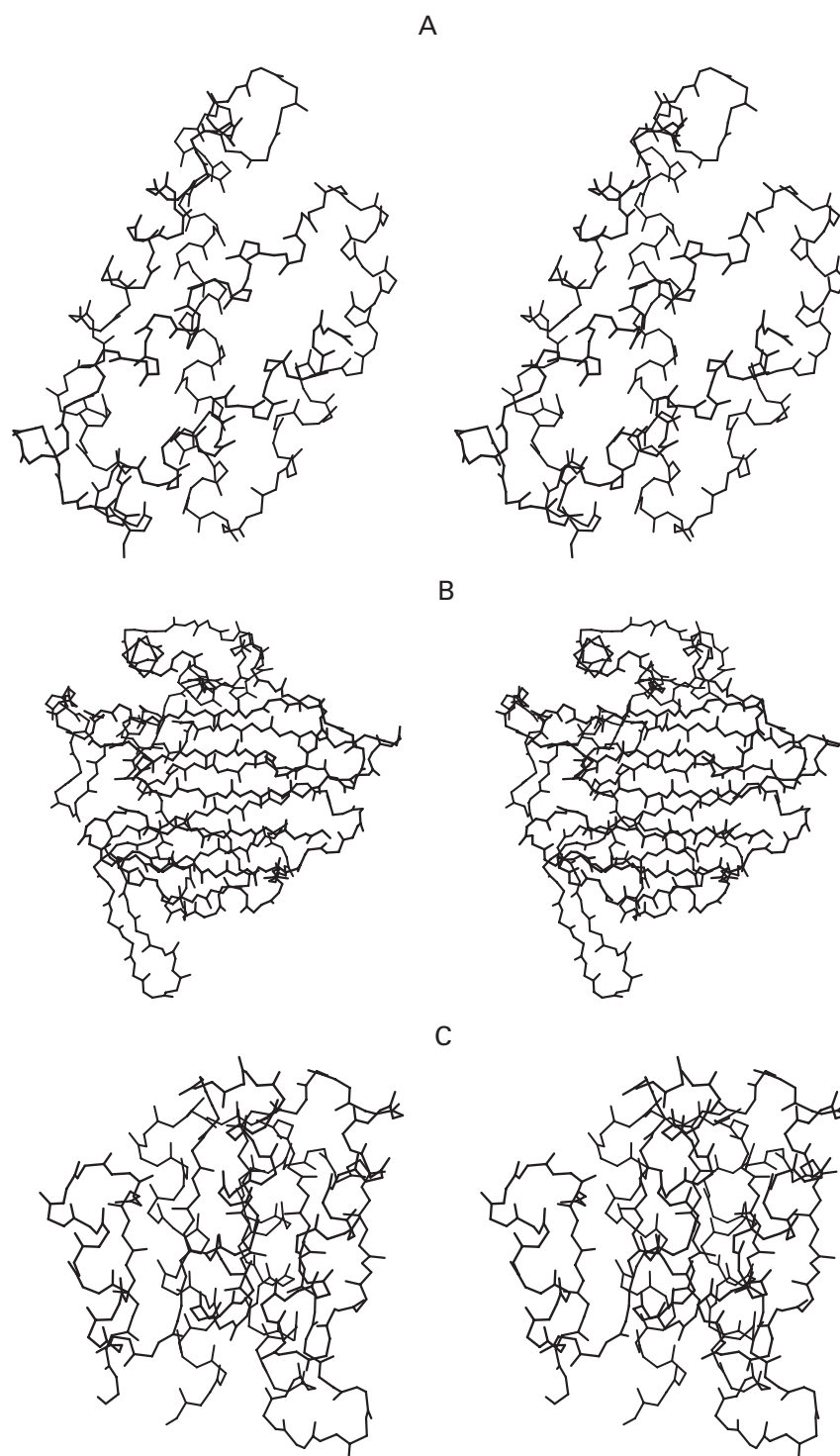
Problem 6–6: The drawings on the next page of three crystallographic molecular models^{265–267} illustrate aspects of the packing between segments of secondary structure. These drawings were produced with MolScript.⁵⁷³ Discuss each molecular model separately and describe the points illustrated by each in turn.

Water

About 40–70% of the volume of a crystal of protein is occupied by water.²⁶⁸ It fills the **large vacant spaces** among the folded polypeptides. The majority of the molecules of water in a crystal of protein are **liquid and disordered** over the time required to collect a data set. Regardless of the degree of refinement or the minimum Bragg spacing, the regions containing this disordered water remain featureless and have a mean electron density similar to that of liquid water.²⁶⁹ These regions of the unit cell are treated as solids of uniform electron density that have the shape of the disordered regions peculiar to the particular crystal. These irregular solids can be used for refining phases by solvent flattening and are always incorporated as such into the molecular model of the unit cell from the first cycle of the refinement because when these water-filled lacunas are added explicitly to the molecular model, the *R*-factor decreases significantly.²⁶⁹

During a refinement, maps of difference electron density are frequently calculated. In addition to the large

Problem 6-6



spaces filled with disordered water, small **discrete peaks of positive electron density** become regularly recurring features of these maps. Because no reasonable rearrangement of the atoms of the molecular model of the protein is able to erase these features and because they are unaccompanied by adjacent peaks of negative electron density indicative of a misalignment of the molecular model, these peaks are assumed to represent either individual molecules of water or individual molecules of solutes from the solution in which the crystals

were formed, which is usually a concentrated solution of ammonium sulfate, poly(ethylene glycol), or a smaller glycol. Sulfate is an ion with a large number of electrons, and any peaks of electron density representing sulfate can usually be recognized with little difficulty.³⁰ Molecules of glycols or other polyols are also easily recognized. The ammonium cation, however, is indistinguishable in its electron density from a molecule of water, but proteins at neutral pH rarely bind many cations so it is usually assumed that the smaller isolated

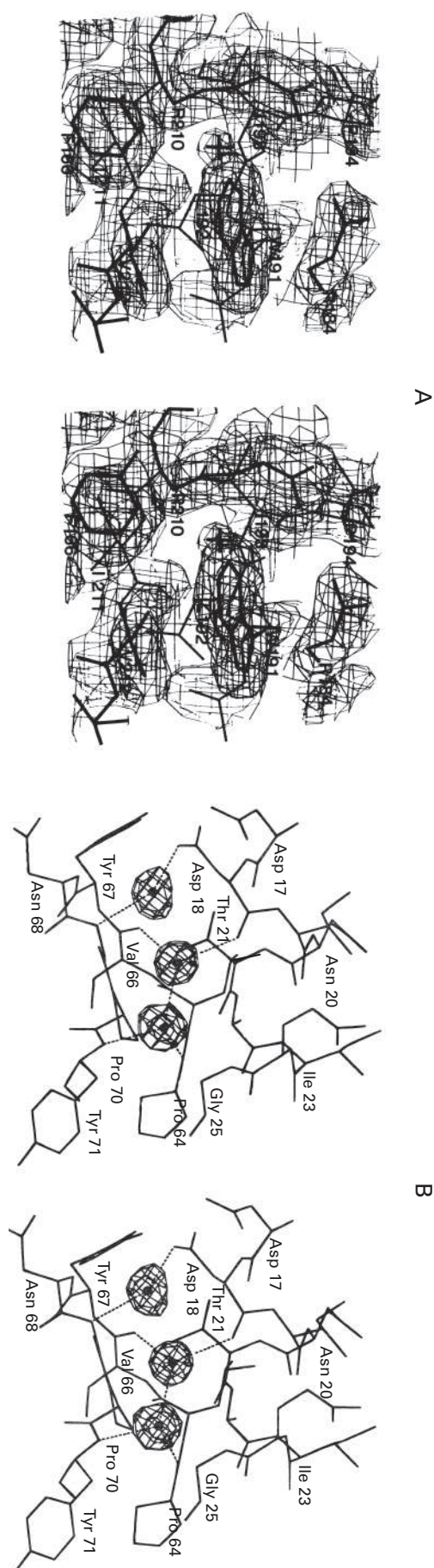


Figure 6-37: Molecules of water buried in the interior of crystallographic molecular models of proteins. (A) A single peak of positive electron density (designated by the cross) assigned to the location for a molecule of water, unconnected to any other peak of positive density, within the interior of the crystallographic molecular model (Bragg spacing ≥ 0.20 nm) of deoxyribonuclease I.⁸ The map of difference electron density is for $(2F_o - F_c)$ with α_c , so the peak is prominent. The molecular skeleton is the refined molecular model itself. The molecule of water is hydrogen-bonded to the amido nitrogen-hydrogens of Isoleucine 193 and Isoleucine 211 and the acyl oxygens of the same two amino acids. Reprinted with permission from ref 8. Copyright 1986 Academic Press. (B) Three peaks of positive electron density buried deeply within the crystallographic molecular model (Bragg spacing ≥ 0.18 nm) of ribonuclease U₂ from *Usilago sphagnetorum*.²⁷⁰ The peaks were assigned as locations for molecules of water at an early cycle in the refinement, but the electron density presented is an omit map. The three molecules of water were omitted from the final molecular model, which after the omission was submitted to five additional cycles of refinement. The map of difference electron density is for $F_o - F_c$ with F_c and α_c calculated from the model produced by these five cycles. These three locations for molecules of water are completely cut off from the bulk water by the protein and participate in hydrogen bonds only with the backbone and side chains of the surrounding amino acids.

peaks of electron density are locations at which molecules of water are situated.

A **location for a molecule of water** in the crystallographic molecular model of a protein is a positive peak of electron density (Figure 6-37)^{8,270} that persists in its location in maps of difference electron density as the refinement progresses and that has a magnitude 0.2–1.0 times the magnitude expected for a stationary molecule of water. The reason for the generous range is that the magnitude of the peak of electron density arising from a more or less fixed molecule of water decays to less than 20% that of a stationary molecule of water when its vibrational amplitude is only 0.06 nm.²⁷¹ For example, the magnitudes of the peaks of electron density in Figure 6-37B vary significantly even though these locations are heavily chelated. The sometimes vague peaks of electron density present in a difference map at the moment when the decision is made by the crystallographer that they represent molecules of water should be distinguished from the more solid peaks of electron density that appear at the same positions in the map of electron density calculated with F_o and α_c after the contributions of these molecules of water have been included in α_c . These more solid peaks are only there because waters have been added to the model, not because the peaks have become more well defined. There is an additional drawback of this strategy for identifying fixed locations occupied by molecules of water. Such a peak of positive electron density sometimes is the result of the alternative conformation of a side chain, and what was assumed to be water at low resolution turns out to be protein atoms from minor conformations at higher resolution.

A portion of the peaks of electron density assigned

as representing molecules of water in a map are observed in the same locations in maps of electron density for the same protein in different crystals or for the same protein from a different species, and such locations are considered to be **conserved**.²⁷² It is assumed that a conserved location is occupied consistently, in particular, when the protein is in solution rather than in the crystal. Molecules of water in the map that are not conserved are assumed to be peculiar to that protein in that crystal, and such locations may or may not be occupied to a significant extent when the protein is free in solution.

There are different **degrees of conservation**. Peaks representing molecules of water may be found at the same locations in two different molecules of the protein in the same unit cell. For example, 25 positions in thioredoxin from *E. coli*,²⁷³ 46 positions in cytochrome *b*₅₆₂ from *E. coli*,⁹⁹ and 26 positions in β -lactamase from *E. coli*²⁷⁴ are occupied by molecules of water in both of the crystallographic molecular models of the respective protein in the same unit cell. Peaks representing molecules of water may be found at the same locations in the same molecule of protein in different crystals. For example, 30 positions for molecules of water in ribonuclease T₁ are conserved in four different crystals of the protein.²⁷⁵ They may be found at the same locations in crystallographic molecular models of the same protein from different species. For example, a string of five molecules of water is found at the same locations in the interiors of crystallographic molecular models of both cytochrome *f* from *Phormidium laminosum* and cytochrome *f* from *Brassica rapa*.²⁷⁶ They may even be found at the same locations in different but related proteins. For example, two positions for molecules of water are found at the same locations in crystallographic molecular models of ferredoxin-NADP⁺ reductase, phthalate-dioxygenase reductase and a fragment of nitrate reductase (NADH).²⁷⁷

The molecules of ordered water included in the final refined molecular model surround the molecule of protein, fill deep clefts in its surface, and are found in its interior (Figure 6–38).⁹⁸ They represent locations in the actual molecule of protein in the crystal that are consistently occupied by a molecule of water. Each **location is fixed and static** because it is observed in the map of electron density, but the actual molecules of water at those locations on the surface of the molecule of a protein change sites and exchange with molecules of water in the disordered regions of the unit cell as rapidly as **molecules of water change positions** in the hydrogen-bonded lattice of liquid water itself. A lower limit of 1 ns⁻¹ for the rate constant for the exchange of molecules of water at such locations on the surface has been established experimentally,²⁷⁸ and the rate constant for their rotation has been observed²⁷⁹ to be about 50 ns⁻¹. The latter rate constant is indistinguishable from that for a molecule of water in the bulk phase. It is the locations for the molecules of water that remain fixed relative to the molecule

of protein, not the molecules of water themselves. There are also locations for molecules of water on flexible portions of the protein that change their conformations so widely that the locations for those molecules of water cannot appear in the map of electron density.

One of the more unexpected observations has been the discovery of molecules of **water buried in the interior** of proteins with no direct contact with the solvent. These occur as single forlorn molecules (Figures 6–37A and 6–39),^{8,280} or as small clusters of two or more mole-

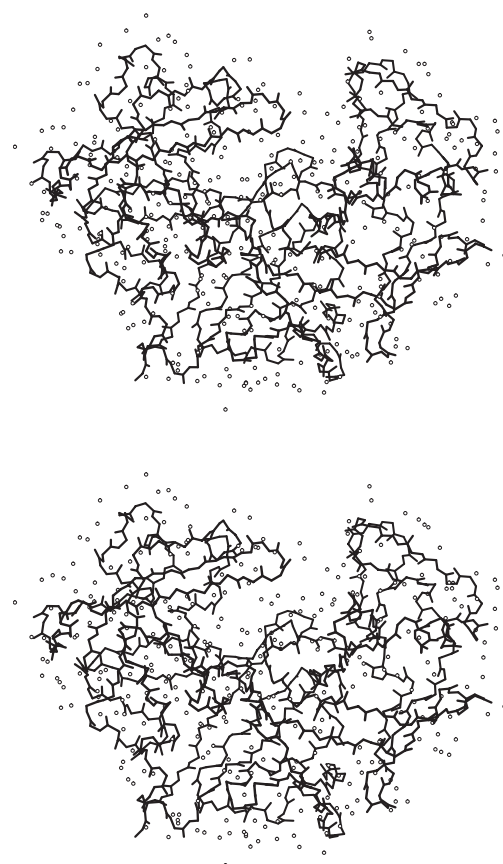
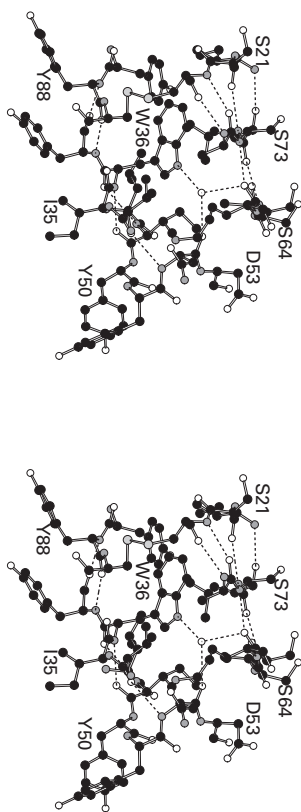


Figure 6–38: Locations for molecules of water in the crystallographic molecular model (Bragg spacing ≥ 0.18 nm) of penicillopepsin.⁹⁸ At various cycles of the refinement of this crystallographic molecular model, it was decided that certain members of the array of as yet unassigned peaks of positive density in the map of difference electron density were locations for molecules of water, and a molecule of water was placed at each of these positions in the model. The 319 unique locations for molecules of water so positioned are designated in the figure with open circles (oxygen atoms) relative to the polypeptide backbone without the side chains. The drawing of the crystallographic molecular model is presented in the same orientation as that in Figure 4–17. This drawing was produced with MolScript.⁵⁷³

Figure 6-39: Structures surrounding Tryptophan 36 in the crystallographic molecular model (Bragg spacing ≥ 0.16 nm) of human Bence-Jones protein Rhe.²⁸⁰ This protein is a compact globular structure formed from a folded polypeptide 114 amino acids in length. Tryptophan 36 lies in the center of the molecule, well sandwiched between two antiparallel β -pleated sheets (above and below Tryptophan 36 in the figure). Segments from Isoleucine 20 to Cysteine 22, Valine 34 to Tyrosine 37, Leucine 48 to Aspartate 53, Serine 64 to Lysine 67, Alanine 72 to Leucine 74, and Tyrosine 87 to Cysteine 89 are included in the figure. Together their side chains and backbone completely surround Tryptophan 36 with mostly hydrogen-carbon bonds, making this an example of a well-buried side chain. The cysteine is also hydrophobic. A position for a molecule of water, represented by an unattached oxygen, is buried with the tryptophan and forms a hydrogen bond to the nitrogen-hydrogen of the indole. Its two donors in turn form hydrogen bonds with the acyl oxygens of the backbone at positions 52 and 62. This drawing was produced with MolScript.⁵⁷³



cules (Figure 6-37B)²⁷⁰ surrounded by donors and acceptors of hydrogen bonds from the protein itself. For example, in the crystallographic molecular model of equine hepatic alcohol dehydrogenase ($n_{aa} = 374$), 12 molecules of water making no contact with the solvent have been located in the interior of the protein,²⁸¹ three as a triplet, two as a doublet, and seven as singlets; in the model of α -lytic endopeptidase ($n_{aa} = 198$), nine molecules of water making no contact with the solvent have been located, three as a triplet, four as two doublets, and two

as singlets;⁹⁷ in the model of fatty-acid-binding protein from *Manduca sexta* ($n_{aa} = 131$), nine molecules of water making no contact with solvent have been located, four as a quartet, two as a doublet, and three as singlets;²⁸² and a string of five molecules of water making no contact with solvent have been located in the model of cytochrome *f* from *Phormidium laminosum*.²⁷⁶ A pair of water molecules is located in the center of the β barrel in the middle of the crystallographic molecular model of chitinase B from *Serratia marcescens*, deeply buried in the core of the protein.²⁸³

Even such locations occupied by molecules of water buried within the structure of the protein exchange rapidly with water in the bulk phase, presumably as the result of breathing movements in the folded polypeptide. For example, a molecule of water at one of the buried locations in bovine pancreatic trypsin inhibitor exchanges with water in the bulk phase at a rate constant²⁸⁴ of 6 ms^{-1} .

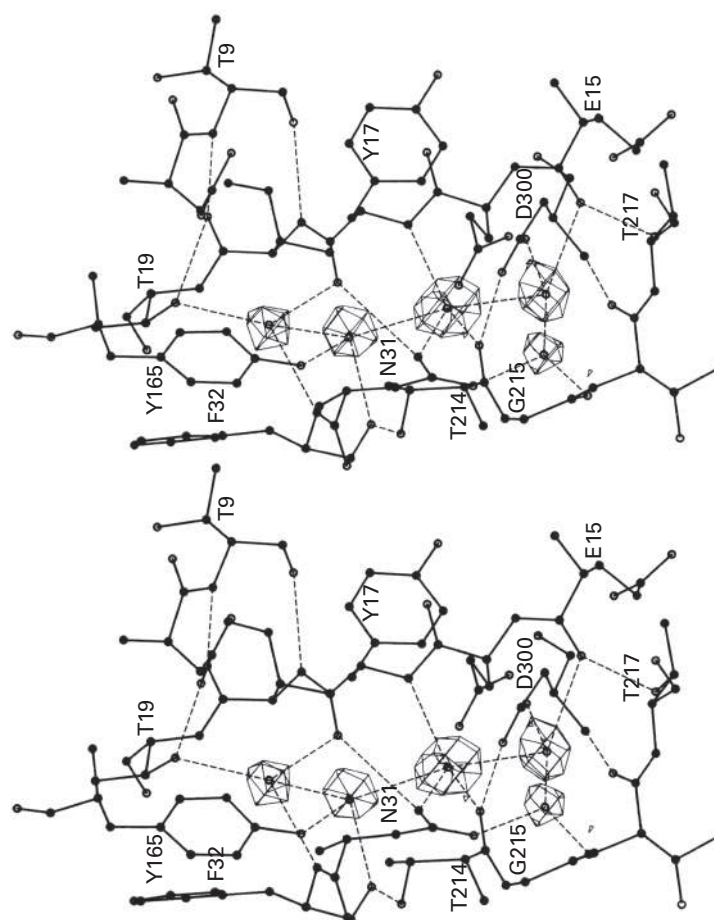
Molecules of water are also buried in **cracks and fissures** connected to the solvent. For example, in penicillopepsin a finger of five water molecules extends from the surface into the interior (Figure 6-40A).⁹⁸ In the crystallographic molecular model of unoccupied chloramphenicol *O*-acetyltransferase, there is a channel 2.5 nm in length filled with a continuous chain of water molecules extending from one side of the molecule of protein to the other.²⁸⁵ A more common situation, however, is for a fairly broad fissure to be filled with an extended cluster of water (Figure 6-40B).^{286,287}

Some proteins have **sizeable cavities** in their interiors. Although there is electron density in the cavity in the crystallographic molecular model of human interleukin 1 β , it is featureless.²⁸⁸ Nuclear Overhauser effects in nuclear magnetic resonance spectra on the absorptions of protons in the side chains lining this cavity demonstrate that it contains molecules of water.²⁸⁹ Because the electron density is featureless, the two or three molecules of water in the cavity must be mobile, which would make sense because it is lined with the side chains of valines, leucines, and phenylalanines that provide no donors or acceptors to pin the molecules of water. The 22 molecules of water in the large internal cavity of unoccupied rat fatty-acid-binding protein, however, are all represented by peaks of electron density in the map and together form a large hydrogen-bonded cluster pinned by the donors and acceptors of tyrosines and glutamates lining the cavity.⁷⁴

Often, a particular role can be assigned to a molecule of water observed in a map of electron density. For example, in the crystallographic molecular model of the triacylglycerol lipase from *Geotrichum candidum*, 17 of the molecules of water at fixed positions donate hydrogen bonds to the vacant acyl oxygens at the carboxy-terminal ends of α helices (Figure 6-6), and 14 of the molecules of water at fixed positions accept hydrogen bonds from the vacant nitrogen-hydrogens at the amino-terminal ends of α helices.²⁹⁰ In the crystallo-

Figure 6-40: Molecules of water penetrating the interior of the molecule of protein but remaining in direct connection with the bulk water. (A) A cluster of four peaks of positive electron density (the four arranged vertically one above the other) assigned as the locations for four molecules of water in the interior of the crystallographic molecular model (Bragg spacing ≥ 0.18 nm) of penicillopepsin.⁹⁶ These four locations are well-defined and the molecules of water occupying them are held in an extensive array of hydrogen bonds with donors and acceptors in the interior of the model. The molecule of water at the bottom of the stack, however, forms one hydrogen bond with a molecule of water at a location to the left that is poorly occupied and in full contact with the bulk water. As such, this structure is a finger of water passing into the interior. This drawing was produced with MolScript.⁵⁷³ (B) Locations for molecules of water in a fissure in the surface of the crystallographic molecular model (Bragg spacing ≥ 0.13 nm) of toxin II from *Androctonus australis*.^{286,287} The fissure is formed from three segments of the polypeptide in which each of the amino acids is numbered by its position in the sequence of the protein. The fissure is in contact with the bulk water (situated to the right) over all of its length, and many of the hydrogen bonds are between molecules of water, but the entire cluster is held in place by hydrogen bonds to donors and acceptors on the protein that surrounds it. Cystine 12/63 assumes two alternative conformations in which the disulfide is the right-handed and left-handed rotamer, respectively. This drawing was produced with MolScript.⁵⁷³

A

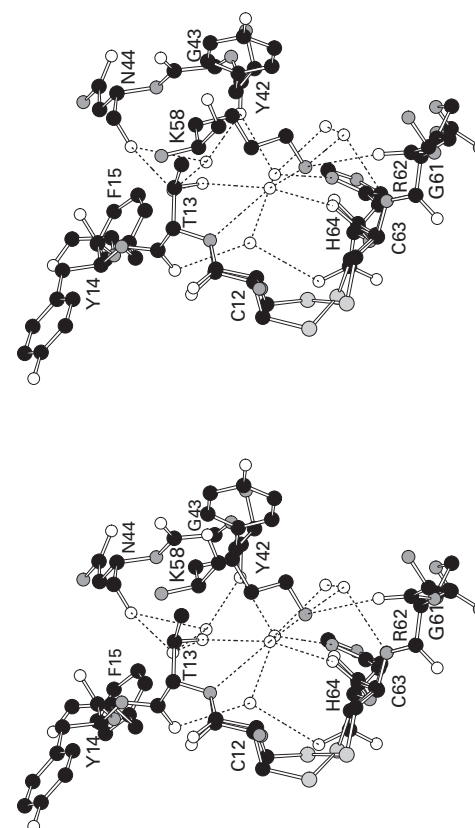


graphic molecular model of cholesterol oxidase from *Brevibacterium sterolicum*, a number of water molecules are involved in linking one segment of β structure to another.¹⁵ Usually, however, the molecules of water at fixed positions are chelated at random by the donors and acceptors of the protein.

The ordered water found covering the open surface of a crystallographic molecular model (Figure 6-38) is held in its discrete locations by **hydrogen bonds** to donors and acceptors on the surface. This pinning is accomplished by particular asparagines, glutamines, lysines, aspartates, arginines, and glutamates distributed over the surface of the individual molecules of protein. These donors and acceptors hold the molecules of water in extended, fixed hydrogen-bonded networks.²⁹¹ It is the donors and acceptors on the protein that anchor these networks in the unit cell. At their peripheries the networks contain molecules of water attached only to other molecules of water, and at the far edges the networks fade into the disorder of the bulk solvent in the crystal. The number of water molecules included as individuals in the refined molecular model is probably a function more of the minimum Bragg spacing of the data set, the peculiarities of the refinement, and the subjective decisions of the crystallographer than of any discrete distinction between ordered and disordered water in the actual crystal, if any such distinction can ever be made.

Networks resembling clathrates are rarely^{292,293} found around hydrophobic functional groups on the surface of crystallographic molecular models,²⁷ and those networks that are found usually do not seem to be the

B



result of the fact that they surround a hydrophobic side chain.²⁹¹ If such networks around hydrophobic amino acids are present in the crystal but are not pinned to the same locations in each unit cell or if they are rearranging continuously while the data set is being gathered, they would not be seen in the crystallographic molecular model. If the charged functional groups on the surface of a protein are surrounded by spheres or semispheres of hydration, as the paradigm associated with the hydration of spherical ions suggests (Figure 5–9), the molecules of water in these shells of hydration are not pinned, because no indication of their presence is seen in the maps of difference electron density. Charged functional groups on the side chains of the amino acids, however, often have one or two molecules of water forming hydrogen bonds to their donors and acceptors. All of these points reemphasize the fact that only specific locations occupied by molecules of water for long periods of time appear as distinct features in maps of electron density.

Of the water molecules that are attached directly to the molecule of protein in a refined crystallographic molecular model, about two-thirds make only one hydrogen bond to the protein^{269,294} and one-third make two or more hydrogen bonds. The **mean number of hydrogen bonds** between molecules of water in this first layer of hydration and the protein is 1.7.⁹⁸ The average distance between one of these waters and a donor nitrogen is 0.29 ± 0.02 nm and between one of these waters and an acceptor oxygen is 0.29 ± 0.02 nm.^{98,269} Of these water molecules directly bound to protein, 42% act as donors to acyl oxygens of the polypeptide backbone, 16% act as acceptors from nitrogen–hydrogen bonds of the polypeptide backbone, and 42% are in hydrogen bonds to donors and acceptors on the side chains.^{8,98,269} In the crystallographic molecular model of lysozyme,²⁶⁹ of the waters bound to functional groups of side chains on the surface of the protein, 24% were donors to carboxylates, 13% were donors to primary amides, 13% were acceptors for primary amides, 14% were acceptors for primary alkyl ammoniums, 14% were acceptors for guanidiniums, 14% were hydrogen-bonded to alkyl hydroxyls, and 6% were hydrogen-bonded to phenolic hydroxyls.

The **disordered side chains** in a crystallographic molecular model are usually at its surface in the most accessible locations. Because they are at the surface, they are usually hydrophilic and are probably even more hydrated than the ordered side chains. Any molecules of water bound to these disordered side chains are never seen in the crystallographic molecular model. Therefore, the mean number of waters bound to each side chain is probably an underestimate of the actual values. Nevertheless, of the side chains to which bound water can be assigned in lysozyme,²⁶⁹ aspartic acids have a mean of 2.0 waters; lysines, a mean of 1.8 waters; asparagines, a mean of 1.6 waters; glutamic acids, a mean of 1.5 waters; threonines, a mean of 1.2 waters; tyrosines, a mean of 1.0 water; serines, a mean of 0.7

water; and glutamines, a mean of 0.7 water bound to their side chains. More than 50% of the arginines are disordered, but those that can be observed have 1.5 waters bound.

When a set of 16 crystallographic molecular models from data sets gathered to Bragg spacings of 0.17 nm or less were examined,²⁹⁵ side chains that had two or three heteroatoms that can participate in hydrogen bonds (aspartate, arginine, glutamate, histidine, and glutamine) frequently (>70%)²⁹⁵ had one or more waters bound to them, while those with only one (tyrosine, tryptophan, threonine, and lysine) less frequently had water bound to them (60–70%). Asparagine (61%) and serine (51%) fall out of these ranges, probably because they are often hydrogen-bonded to the backbone.

It is the molecules of water hydrogen-bonded to the donors and acceptors of these side chains that produce the sharp maximum at around 0.29 nm in the solvent distribution function around a molecule of protein.²⁹⁶ The hydrophobic side chains (methionine, alanine, phenylalanine, isoleucine, leucine, and valine) much less frequently (10–30%) have fixed locations for molecules of water adjacent to them, but these side chains are usually buried in the interior of the protein. Those hydrophobic groups that do have fixed locations for molecules of water adjacent to them are surrounded by a layer of water in which the centers of the oxygen atoms are about 0.4 nm from the centers of the carbon atoms of the side chains,²⁹⁶ as expected from their van der Waals radii (Table 6–3).

There are a number of **physical measurements** which register the fact that each molecule of protein in solution has water bound to it in an irregular network creating a **shell of hydration**. The molecules of water in this shell of hydration differ from the molecules of water in the bulk of the solution away from the molecule of protein in several of their physical properties. For example, neutron scattering has revealed that the layer of hydration immediately adjacent to the surface of a molecule of protein has a density about 10% greater than that of the bulk water.²⁹⁷ From a dissection of the compressibility of this layer of hydration, it has been concluded that it contains extensive hydrogen-bonded networks¹⁸⁴ similar to those observed in crystallographic molecular models.^{286,291} While it is unable to distinguish the water in this layer from water in bulk solution because its rate of relaxation is too fast, nuclear magnetic resonance is able to detect the buried waters in a molecule of protein because their rates of relaxation are so much slower.²⁹⁸

Each of the molecules of water surrounding a molecule of protein at a given instant is in a different situation (Figure 6–38), and the relationship between each one of these molecules of water and the molecule of protein depends upon the respective situation. It is the contribution of each one of these molecules of water to the statistical behavior that produces the value of the physical property measuring the **hydration** of the protein, $\delta_{\text{H}_2\text{O}}$

[grams of H₂O (gram of protein)⁻¹]. Each contribution will be a unique function of the situation of the respective molecule of water, and the physical measurement will be only an average over all of these contributions. Mathematically, this heterogeneity in the situations of the waters participating in the shell of hydration can be expressed as a **weighted mean**:²⁹⁹

$$\delta_{\text{H}_2\text{O}} = \frac{M_{\text{H}_2\text{O}}}{M_p} \sum_{i=1}^n w_i \quad (6-4)$$

where $M_{\text{H}_2\text{O}}$ is the molar mass of water (18.0 g mol⁻¹), M_p is the molar mass of the protein, and the sum is over a set of statistical weights w_i .

There are two ways to think of the meaning of the **statistical weights** w_i . It can be assumed that there are n sites for the binding of water molecules around the protein, the positions of which move through the solution in lock step with the protein. The statistical weight w_i for a given site is then the occupancy of that site, which is the fraction of the time that the site is occupied by a molecule of water.²⁹⁹ It is also possible to consider all n molecules of water in the vicinity of the protein at a given instant. The statistical weight w_i in this case expresses the degree of influence the molecule of protein has over the behavior of water molecule i . When $w_i = 1$, the location occupied by water molecule i is fixed, as if covalently, to the molecule of protein. When $w_i = 0$, the water molecule i is uninfluenced in its behavior by the presence of the molecule of protein.

Under no circumstances should the layer of hydration surrounding a molecule of protein be pictured as a uniform layer clearly distinguished from the water in the bulk of the solution by some discontinuous boundary. Rather, the layer of hydration gradually fades from fixed defined locations for molecules of water adjacent to the surface of the molecule of protein to molecules of water distant from the surface that are only marginally affected by its presence.

It is almost always the case that physical measurements of hydration yield a simple number, $\delta_{\text{H}_2\text{O}}$, the **grams of water bound for every gram of protein** (Table 6-4). It is not surprising that this number varies with the method used to obtain it, as more or less of the molecules of water surrounding the protein differ more or less from the water in the bulk solvent in the particular behavior measured by the particular procedure.

The **self-diffusion of water** decreases when protein is added to the solution,³¹⁰ and this decrease can be explained if it is assumed that the water of hydration, being less mobile than the water in the bulk phase, does not participate significantly in self-diffusion. With this assumption, the amount of water bound to the protein can be calculated.

The relative permittivity of a solution of protein decreases discontinuously when the frequency of the

alternating electric field used to measure that relative permittivity becomes greater than the ability of the molecules of protein to reorient in response to its alterations,³¹¹ and another discontinuous decrease is observed when the frequency becomes greater than the ability of the water in the bulk solvent to reorient. Between these two extremes, there is a third **dielectric relaxation** that is assigned³⁰¹ to the waters of hydration bound to the protein. These waters have dielectric relaxations 10–100-fold slower than the waters in the bulk solution. From the spectrum of these dielectric relaxations, the concentration of these relatively immobilized molecules of water and hence the amount of water bound to the protein can be calculated. These molecules of water, however, are not fixed to the protein, or they would be required to rotate with it, and their dielectric relaxation would be indistinguishable from that of the protein itself.

Solid **powders of dry protein** always have water incorporated in them and the amount of this water of hydration can be chemically determined. A more systematic approach is to equilibrate the dry powder, either as a precipitate, as a microcrystalline solid, or as visible crystals, with air of a certain relative humidity. It has been proposed that air at 90% relative humidity is the appropriate choice.³⁰³ Below this value the powders tend to become glasses,³⁰³ and above this value they become hygroscopic. The amount of water bound by a solid powder of a given protein at 90% relative humidity can be taken as its hydration.

If it is assumed that the water hydrating a protein is entirely unable to dissolve salting-out solutes that are otherwise freely soluble in water,³⁰⁴ the negative of the **preferential solvation** of a particular protein in a particular solution (Equation 1-57) can be multiplied by the molarity of the water in that solution and the molar mass of water to obtain a value for the grams of H₂O (gram of protein)⁻¹ in the layer of hydration. To perform measurements of preferential solvation, a solution of the protein is usually brought into equilibrium with a solution containing only water and the salting-out solute, for example, glucose,³⁰⁴ lactose,³⁰⁴ or sucrose³⁰⁵ (Table 6-4). It is also possible to equilibrate crystals of a protein with solutions of salting-out solutes and from the dependence of the density of the crystal on the density of the solution to determine the amount of water in the crystal that excludes the solute.^{299,302,306}

When a solution of protein is frozen, the water of hydration freezes below the freezing point of the water in the bulk solution. Not until the temperature is lowered to below 180 K does it all become frozen.³¹² For example, at -3 °C, 0.51 g of water (g of protein)⁻¹; at -5 °C, 0.46 g of water (g of protein)⁻¹; and at -7 °C, 0.41 g of water (g of protein)⁻¹ remained unfrozen in a solution of ovalbumin.³¹³ **Unfrozen water** is more mobile than frozen water, and the two can be distinguished by nuclear magnetic resonance.^{307,312} The amount of unfrozen water in a

Table 6–4: Hydration of Proteins^a

protein	self-diffusion ³⁰⁰ of H ₂ O ¹⁸	dielectric relaxation ³⁰¹	solid at RH = 90% ^{302,303}	excluded volume		NMR frozen solution ³⁰⁷	frictional coefficient		scattering of X-radiation at small angles ³⁰⁹
				sugar ^{299,304,305}	(NH ₄) ₂ SO ₄ ^{299,302,306}		diffusion ³⁰⁸	viscosity ³⁰⁸	
ribonuclease			0.35	0.18 0.46					0.27
lysozyme			0.25	0.45			<0.9		0.36
myoglobin		0.25	0.42		0.46		<0.4	<0.6	
chymotrypsinogen			0.29	0.31 0.50	0.26	0.37	<0.5		0.18
γ chymotrypsin					0.18				
α -lactalbumin									0.36
β -lactoglobulin			0.32 0.34	0.56 0.23	0.29		<0.7	<0.6	0.24
ovalbumin	0.18		0.29			0.31	<0.5		
hemoglobin			0.37 0.30		0.31 0.27 0.13	0.45	<0.4	<0.7	
pepsin									0.24
serum albumin			0.32	0.31	0.50 0.43 0.27 0.08	0.43	<1.1	<0.8	0.15

^aAll units are grams of water (gram of protein)⁻¹.

frozen solution of protein at 238 K (-35°C) has been designated as water of hydration.

Upper limits on the amount of water that migrates with a molecule of protein through the solution can be calculated from the **frictional coefficient**.³⁰⁸ The radius of a hard sphere the same volume as a molecule of protein can be calculated from its molar mass and partial specific volume, and the radius of the sphere that would have the same frictional coefficient as the molecule of protein can be calculated with Equation 1-66. The latter sphere is always larger than the former. If it is assumed that the entire difference in volume is water forced to move with the molecule of protein, an upper limit to the amount of bound water can be calculated (Table 6-4). It is an upper limit because molecules of protein are not spheres and a particle with the same volume as a given sphere but a different shape will always have a larger frictional coefficient than that sphere. How much of the difference between the two radii is due to hydration and how much to differences in shape has never been ascertained unambiguously for any protein. The numbers tabulated are not intended to be estimates of hydration, but upper limits of the hydration.

It is also possible to estimate the hydration of a protein from the **scattering at small angles** of X-radiation from a solution of that protein as a function of the angle of that scattered radiation.³⁰⁹

There are several remarkable features of this tabulation (Table 6-4). The values for bound water are all similar, and each technique produces values that, although they do not agree, are in the same range (**0.2–0.4 g g⁻¹**), which is about 2 mol of water (mol of amino acid)⁻¹. There seems to be no significant difference in the amount of bound water for every gram of protein over a 5-fold range in size of the proteins, between ribonuclease ($n_{\text{aa}} = 124$) and serum albumin ($n_{\text{aa}} = 581$). For a small protein such as lysozyme ($n_{\text{aa}} = 129$) or dihydrofolate reductase ($n_{\text{aa}} = 162$), these results indicate that there should be 200–300 molecules of bound water for every molecule of protein. In a crystal of lysozyme, 140 molecules of water had locations that were sufficiently distinct to be incorporated into the refined molecular model.²⁶⁹ In a crystal of dihydrofolate reductase, 264 molecules of ordered water had sufficiently distinct locations to be incorporated in the refined molecular model.²⁷ Whether these ordered molecules of water bear any relation to the bound water detected by the physical measurements is uncertain.

There is a highly significant correlation between the **accessible surface area** of a crystallographic molecular model of a protein and the total number of amino acids it contains, regardless of whether it is a monomer or an oligomer.³¹⁴ As a result of this correlation, the mean accessible surface area for each amino acid falls gradually and monotonically from 0.53 nm^2 (amino acid)⁻¹ when the protein contains 100 amino acids to 0.30 nm^2 (amino acid)⁻¹ when the protein contains 2000 amino

acids. From the definition of accessible surface area (6-12) it follows that a molecule of water, held by hydrogen bonds at 0.28 nm from its nearest neighbors, can cover about 0.07 nm^2 of accessible surface area if waters are assumed to pack in hexagonal array or 0.09 nm^2 if they are in a tetrahedral lattice (Figure 5-2). This means that there are about 7 waters (amino acid)⁻¹ immediately adjacent to the surface of a protein containing 100 amino acids and 4 waters (amino acid)⁻¹ immediately adjacent to the surface of a protein containing 2000 amino acids. These limits would be equivalent to 1.2 and 0.7 g of water (g of protein)⁻¹, respectively. For the proteins gathered in Table 6-4, which all contain less than 600 amino acids, the span would be 1.2–0.9 g of water (g of protein)⁻¹. Therefore, the water of hydration determined by physical measurements is considerably less than the amount of water required to cover the surface of a molecule of protein with a continuous rigidly fixed layer.

Part of the reason for this discrepancy may be that, as with the networks of water covering the surface of a molecule of protein in a crystallographic molecular model, the layer of hydration is patchy and discontinuous^{286,291} but the **heterogeneity** that must exist among the waters of hydration is probably most of the reason. Some waters at the surface are held tightly ($w_i \cong 1.0$), but most are only loosely influenced by the protein ($w_i < 1$) and contribute only partially to the weighted mean (Equation 6-4). Therefore it is not surprising that the weighted mean is less than the limit calculated by simply counting every immediately adjacent molecule of water and presuming it to be fixed to the protein. The range over which the amount of immediately adjacent molecules of water ($0.9\text{--}1.2 \text{ g g}^{-1}$) varies among the proteins of the size of those contributing to Table 6-4 is narrow, and this fact explains why all of the proteins seem to have about the same degree of hydration, within the variation of the measurements.

The molar concentration, and hence the thermodynamic activity, of the water in the bulk phase of a solution of protein can be changed without changing its concentration in the layer of hydration by adding a salting-out solute such as sucrose, triethylene glycol, dioxane, stachyose, or poly(ethylene glycol) that is excluded from the layer of hydration.^{315–317} Because the water in the layer of hydration is in rapid equilibrium with the water in the bulk phase, changing the activity of the water in the bulk phase changes its activity in the layer of hydration, and this change affects any chemical reaction in which the amount of hydration of the protein changes. As one might expect, the binding of substrates to an enzyme,^{316,318} the binding of a protein to DNA,³¹⁵ a large conformational change of a protein,^{317,318} or the binding of one protein to another causes significant changes in hydration. From the magnitude of the effect of changing the concentration of water on the dissociation constant for these reactions, the number of molecules of water removed from or added to the layer of hydration during

the reaction can be estimated. These range from 9 molecules of water for binding of a substrate to a hydrated active site³¹⁶ to 60 molecules of water for a significant conformational change.^{317,318} In the latter transformation, a portion of the water detected as leaving the shell of hydration is thought to be molecules beyond the first layer.

Suggested Reading

Blake, C.C.F., Pulford, W.C.A., & Artymiuk, P.J. (1983) X-ray studies of water in crystals of lysozyme, *J. Mol. Biol.* 167, 693–723.

Problem 6–7: Assign the hydrogens to donors and acceptors in Figures 6–37B and 6–40A,B.

Ionic Interactions

Almost all of the charged amino acids—glutamate, aspartate, histidinium, lysinium, and arginine—are found on the surface of the crystallographic molecular model of a protein, so that they retain their hydration. Aside from the few that have roles as acids and bases in the function of the protein, the reason that these charged amino acids are present on the surface of a protein is to permit it to dissolve in water at high concentrations. For example, the concentration of hemoglobin in an erythrocyte is 0.3 g mL^{−1}.

The **distribution of these elementary charges** on the surface of a molecule of protein seems to be random with little regard for the signs of the elementary charges and no attempt to compensate the charges. Patches of positive charge and patches of negative charge are as common as regions where the charges are evenly divided. Changing these distributions seems to have little effect on the stability of the protein.³¹⁹ The lysozyme from bacteriophage T4 has an excess of nine elementary positive charges over elementary negative charges at neutral pH. When lysines on its surface were changed to glutamates by site-directed mutation to produce a number of single, double, triple, and quadruple mutants in which the net charge number decreased from +9 to +7, +5, +3, and +1, respectively, the mean change in the free energy of folding of the protein was $+3.3 \pm 2.9$ kJ mol^{−1}.³²⁰ Consequently, the protein decreased slightly in stability rather than increasing in stability as its excess charge was neutralized, and the magnitudes of the individual decreases in stability showed no correlation with the magnitude of the decrease in charge. Increases in stability of the same magnitude (−4 to −8 kJ mol^{−1}), however, have been observed upon neutralizing imbalances of charge on the surfaces of ubiquitin³²¹ and the subunit-binding domain of dihydrolipoyllysine-residue acetyltransferase from *Bacillus stearothermophilus*.³²² All of these experiments were performed at an ionic strength of 0.05 M, so the differences in stability observed would

probably have been even smaller at the physiological ionic strength of 0.15 M.

There are experimental results suggesting that the charge of the amino acids on the surface of a protein may electrostatically increase the rate of association^{323,324} or increase the equilibrium constant for association^{325–328} of ligands that bear an opposite charge. These effects, however, are rarely more than a factor of 2 (−1.7 kJ mol^{−1}) at physiological ionic strengths, and often the ionic strength of the solution must be lowered to observe them at all,³²⁶ so they can be of little consequence biologically.

The **acid dissociation constants** of the individual acid–bases of the side chains on the surface of a protein are shifted by the elementary charges on the amino acids in their vicinity. For example, it has been shown that the pK_a of Histidine 64 in subtilisin BPN' decreases by 0.26 unit when Aspartate 99 is mutated to a serine³²⁹ because when the elementary negative charge of Aspartate 99 is no longer in its vicinity, the stability of the histidinium ion decreases relative to that of the neutral histidine. All such pairwise interactions are tautomeric because if one side chain shifts the pK_a of another, then the other side chain must shift the pK_a of the first. If the negative charge of Aspartate 99 shifts the pK_a of Histidine 64, the positive charge of Histidine 64 must shift the pK_a of Aspartate 99. A large constellation of such **tautomeric interactions** determines the individual acid titrations of the side chains on the surface of a protein.

The acid–base **titration curve** of a native protein (Figure 1–11) is the summation of the individual titrations of the accessible acid–bases on its surface. For every 100 amino acids, a normal protein contains about five aspartic acids, six glutamic acids, two histidines, three tyrosines, and six lysines.³³⁰ The aspartic acids ($pK_a = 4.0$) and glutamic acids ($pK_a = 4.4$) account for most of the dissociation of protons between pH 2 and 5.5. The lysines ($pK_a = 10.4$) and tyrosines ($pK_a = 9.8$) account for most of the dissociation of protons between pH 8 and 12. These are the two major features of the titration curve of a protein because these four amino acids account for the majority (90%) of the acid–bases present in the protein. The histidines account for most of the small amount of dissociation that occurs between pH 5.5 and 8.

As the pH is decreased below the isoelectric point, a protein gains net positive charge number as each proton associates, and as the pH is increased above the isoelectric point, the protein gains net negative charge number as each proton dissociates. This change of net charge number with decreases and increases of pH causes the addition of each successive proton or the removal of each successive proton, respectively, to be more difficult. The reason for this is that the gathering of net charge number on a molecule, even one as large as a protein, is an unfavorable reaction relative to dispersing those elementary charges evenly throughout the solution. Because tautomeric interactions are themselves electrostatic, the effect of the resulting charge on the

overall acid dissociation of the molecule of protein is simply the summation of all the tautomeric interactions among all the side chains, the individual titrations of which produce the complete titration curve.

That the **electrostatic work** of creating this charge shifts the observed titration curve is easily demonstrated by changing the ionic strength (Figure 1-11). An increase in ionic strength shrinks the layer of counterions around each individual, charged amino acid in the protein (Equation 1-71) and causes them to exert a decreased effective electrostatic charge in their influence on neighboring acid-bases undergoing titration. This in turn decreases the electrostatic work that must be performed to create charge on the neighboring acid-bases and shifts the titration curve closer to the curve that would have been seen if each acid-base titrated only according to its intrinsic pK_a . This electrostatic shielding due to increased ionic strength produces a steepening of the titration curve for the protein both below and above its isoelectric point (Figure 1-11).

It is possible to correct roughly³³¹ for the electrostatic work involved in creating charge on the molecule of protein by assuming that in a given region of the titration curve, for example, between pH 2 and 5.5, only one type of acid-base is titrating and all of the members of this set have the same intrinsic pK_a , $pK_{a,int}$. Then it is assumed that the charge on the molecule of protein, Q_i , is proportional to the mean net proton charge number, $\bar{Z}_{H,i}$, and that $pK_{a,int}$, which is proportional to a free energy, is shifted arithmetically by the electrostatic work, which is a free energy and which should be proportional to $\bar{Z}_{H,i}$.

The values of the **intrinsic acid dissociation constants** obtained by these corrections for electrostatic work agree with expectation (Table 2-2) to a certain extent. The value of $pK_{a,int}$ for the carboxyl groups in several proteins the titration curves of which between pH 2 and 5.5 have been analyzed in this way³³¹ are between 4.0 and 4.8. The titration of tyrosine side chains in a native protein can be followed independently by using the large difference in ultraviolet absorbance between the phenol and the phenolate anion to calculate f_A and f_{HA} .³³² The values of $pK_{a,int}$, corrected for electrostatic work, for tyrosines in several proteins³³¹ are between 9.4 and 10.8. The contribution of tyrosine to the titration curve between pH 8 and 12 can then be deducted from the overall curve, and values of $pK_{a,int}$ for the lysines in these same proteins can be calculated. They lie between 9.8 and 10.4.

The titration curves of proteins usually fail to meet expectations in one key aspect. There are usually too few acid-bases contributing to the titration.³³¹ The **deficit** is most easily noticed in the case of histidine and tyrosine. The number of moles of protons dissociating from a mole of protein between pH 5.5 and 8 is often less than the moles of histidine in a mole of that protein. This deficit can be explained by assuming that the values of pK_a for one or more of the histidines have been lowered and that their titrations have become buried in those of

the large number of carboxylates. The moles of tyrosine the ultraviolet absorption of which displays the expected shift between pH 9 and 11 upon formation of the phenolate anion is often less than the total moles of tyrosine present in a mole of the protein. For example, only four of the six tyrosines in ribonuclease can be titrated^{331,333} and only two of the four tyrosine side chains of chymotrypsinogen can be titrated³³¹ within accessible ranges of pH. With most proteins, values of pH greater than 11 are inaccessible, so it can be said only that each of these missing tyrosines has a pK_a greater than 11.

Both the decreases in the values of pK_a for the histidines and the increases in the values of pK_a for the tyrosines implied or demonstrated by these results are reasonable. If these shifts in pK_a are due to burying the side chains in the interior of the protein, even though they remain accessible to the solvent and capable of acid-base reactions, their neutral forms should become more stable relative to their charged forms. In most cases, the missing acid-bases in a titration curve are assumed to be buried in the interior of the folded polypeptide. Such **buried acid-bases** can be seen in the crystallographic molecular models of proteins. For example, in the crystallographic molecular model of ribonuclease, Tyrosine 25 is almost completely buried (the solvent accessibility of its phenolic oxygen is only 0.02) and Tyrosine 97 is completely buried.³³⁴ It has been assumed that these are the two tyrosines in the native protein that do not participate in acid-base titrations.

These examples of buried tyrosines or histidines are special cases of the fact that polar amino acids are found in the interior of a protein, even ones that are normally charged. For example, Arginine 30 in the crystallographic molecular model of xylose isomerase from *Anthrobacter* strain B3728 is completely surrounded by both the backbone and mostly carbon-hydrogens of other side chains (Figure 6-41).¹¹² It does not form an ion pair with any anionic side chain. Instead, its five donors form hydrogen bonds with four acyl oxygens from the backbone and a molecule of water. Because one of the hydrogen bonds is to a molecule of water, it cannot be determined whether or not the guanidino group is positively charged. Aspartate 76 is buried in the interior of ribonuclease T₁ of *A. oryzae*³³⁵ and a cluster of three glutamates, two histidines, and an aspartate is buried in the interior of the iron free form of the R2 protein of ribonucleoside-diphosphate reductase.³³⁶

The amino acids that are charged at neutral pH are either the anionic conjugate bases of neutral acids, for example, carboxylates, or the cationic conjugate acids of neutral bases, for example, ammonium cations. It may be the case that when such an amino acid is buried, it is buried as the **neutral acid or the neutral base**, respectively. This conclusion follows from the fact that the farther the pK_a of a normally charged amino acid is from 7.0, the less likely it is to be buried (Table 6-2). Such a buried amino acid usually participates in a set of hydrogen

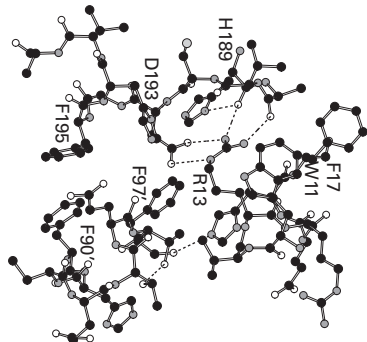
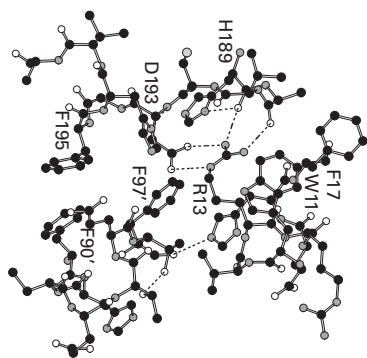
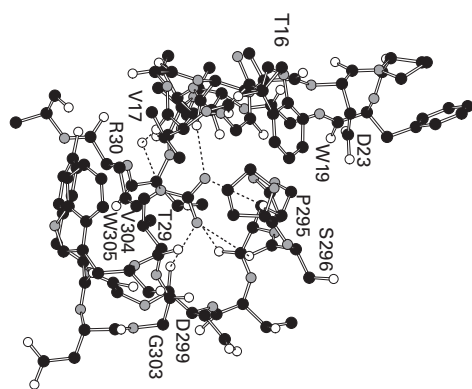
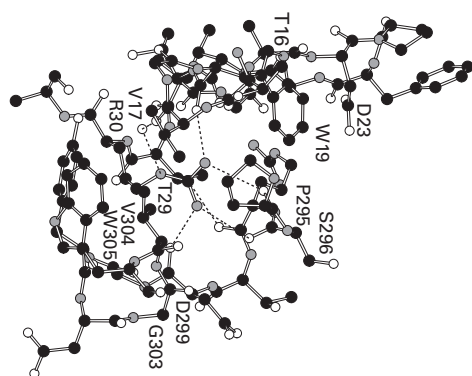


Figure 6-41: Arginine buried in the crystallographic molecular model (Bragg spacing ≥ 0.23 nm) of xylose isomerase from *Anthrobacter* strain B3728.¹¹² The unattached oxygen (open circle) is a fixed location for a molecule of water. Two segments of the folded polypeptide, Threonine 16 to Alanine 31 and Proline 295 to Tryptophan 305, are displayed. This drawing was produced with MolScript.⁵⁷³

bonds that fit its donors or its acceptors (Figure 6-41), so it is neither surprising nor informative that replacing it by site-directed mutation with an amino acid that is isosteric but has a different pattern of donors and acceptors, such as an asparagine for an aspartate, usually produces a less stable protein.³³⁵ From examining closely the constellation of donors and acceptors around such a buried amino acid, however (see Problem 6-8), one often comes to the conclusion that it is buried in its charged form. If this is the case, the protein itself must somehow replace most of the enthalpy of hydration that is lost upon removing the charge from the water. In such situations, it is the large number of complementary donors and acceptors of hydrogen bonds that seems to accomplish this feat in the absence of any compensation of charge.³³⁷ Most of the time, however, when a charged amino acid is found in a crystallographic molecular model at a location removed from the water, it is found as a partner in an ion pair.

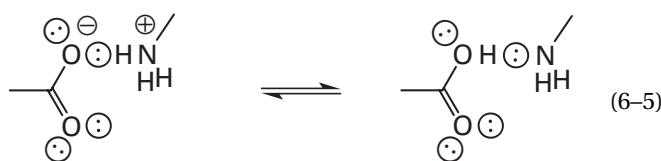
Ion pairs are actually ionized hydrogen bonds. An **ionized hydrogen bond** is a hydrogen bond between a positively charged donor and a negatively charged acceptor. An example would be the hydrogen bonds between an argininium cation and an aspartate anion (Figure 6-42) or an argininium cation and a glutamate anion (Figure 6-58).^{338,339} That almost all ion pairs are ionized hydrogen bonds follows from the fact that the distance between the two heteroatoms, one from the cation and one from the anion, that make the closest contact in an ion pair is usually that of a hydrogen bond (0.3 nm),^{15,24,72,340} the fact that the angles at this point of contact are usually those expected of a hydrogen bond; and the fact that a deuteron can be found between these two heteroatoms in crystallographic molecular models from neutron diffraction of deuterated proteins.³⁴⁰

Although occasionally an ion pair will be as ideal as the one illustrated in Figure 6-42,^{30,97,341} in which the two acceptors of the carboxylate are respectively occupied by two of the donors of the guanidinium ion to form a six-

Figure 6-42: An ionized hydrogen bond between Aspartate 193 and Arginine 13 in the crystallographic molecular model (Bragg spacing ≥ 0.175 nm) of chloramphenicol *O*-acetyltransferase.^{338,339} The hydrogen bonds between donors and acceptors on other side chains, but not those between acyl oxygens and amido nitrogen-hydrogens of the backbone, are drawn. The one apparently unoccupied donor on Arginine 13 is pointing toward the π system of the side chain of Tryptophan 11 and presumably forms a hydrogen bond with it. The segments from one subunit of the protein, Phenylalanine 11 to Phenylalanine 17 and Histidine 189 to Alanine 198, and one segment from an adjacent identical subunit of the protein, Phenylalanine 90' to Phenylalanine 97', are displayed. This drawing was produced with MolScript.⁵⁷³

membered ring, usually it is more peculiar because molecules of protein are the products of evolution by natural selection. For example, a positive side chain and a negative side chain will not be directly hydrogen-bonded to each other but linked through an intermediate, as Arginine 102 and Aspartate 142 are linked by Threonine 107 in α -lytic endopeptidase (Figure 6–43).⁹⁷ As if to illustrate the **irrelevance of placing positive next to negative**, Arginine 8 and Arginine 366 in pepsinogen, albeit each hydrogen-bonded to carboxylates, are nevertheless stacked on top of each other, their π molecular orbital systems parallel to each other, within a buried hydrogen-bonded cluster.³⁴²

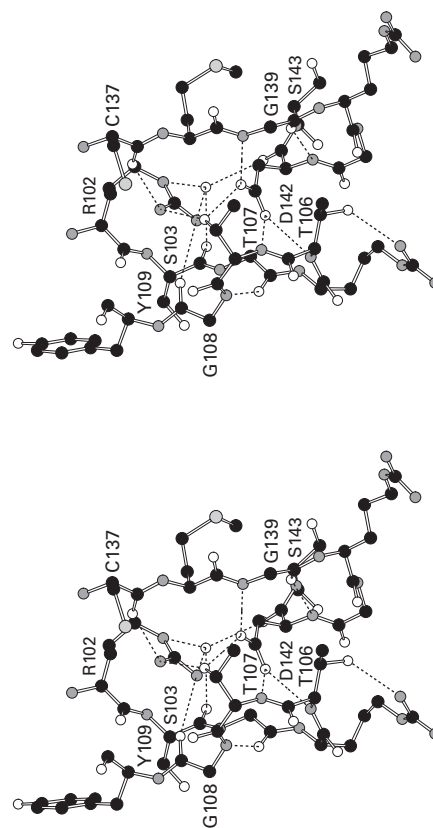
Whether the hydrogen bond between a carboxylate anion on the one hand and a histidinium, ammonium, or guanidinium cation on the other is ionized or not depends formally on whether the proton is on the stronger base, in which case the bond is ionized, or on the weaker base, in which case it is neutral. In almost every instance in which a potentially ionized hydrogen bond is found in a crystallographic molecular model, it is unknown on which atom the proton resides. An estimate of the **effect of the relative permittivity** on the location of the proton in an ionized hydrogen bond has been made,³⁴³ and it was concluded that the relative permittivity of the surroundings would have to be less than that of CCl_4 ($\epsilon_r = 2.2$) before the shift of the proton from an ammonium cation to a carboxylate anion in the ionized hydrogen bond between them



would be favored. If this is the case, such ionized hydrogen bonds are probably ionized even when surrounded by protein because the relative permittivity of protein is thought to be between 3 and 6.³⁴⁴ Certainly, in the crystallographic studies of proteins by neutron diffraction, in which the location of the proton has been observed directly, it usually forms a normal σ bond with the stronger base.

The energy required to transfer an ionized hydrogen bond from water to a region of low relative permittivity is almost as large as the energy required to transfer a monovalent ion,³⁴⁵ so a **buried ion pair** is only marginally less unstable than a buried, uncompensated arginine, lysine, aspartate, or glutamate. Consequently, when one member of an ion pair is replaced by site-directed mutation, the protein becomes less stable, but only by 10–15 kJ mol^{-1} .³⁴⁶ Such observations are, however, ambiguous because the amino acids in the vicinity of the mutation can rearrange to take advantage of alternative compensations. For example, when Aspartate 193 represented in Figure 6–42 is mutated to an asparagine, this

Figure 6–43: Hydrogen-bonded network involving Arginine 102 and Aspartate 142 in the crystallographic molecular model (Bragg spacing ≥ 0.17 nm) of α -lytic endopeptidase from *L. enzymogenes*.⁹⁷ Two strands of polypeptide, from Arginine 102 to Tyrosine 109 and Cysteine 137 to Serine 143, from the interior of the protein are displayed as well as a molecule of water (open circle). The interaction between the two oppositely charged side chains, the arginine and the aspartate, is mediated by their respective hydrogen bonds to Threonine 107. This drawing was produced with MolScript.⁵⁷³



asparagine swings away to form hydrogen bonds with the backbone amido nitrogen–hydrogen of Histidine 196 and the backbone acyl oxygen of Glutamate 95', and Arginine 13 swings away to form a new ionized hydrogen bond with the side chain of Glutamate 95' as well as a hydrogen bond with the acyl oxygen of Threonine 94'.³³⁸

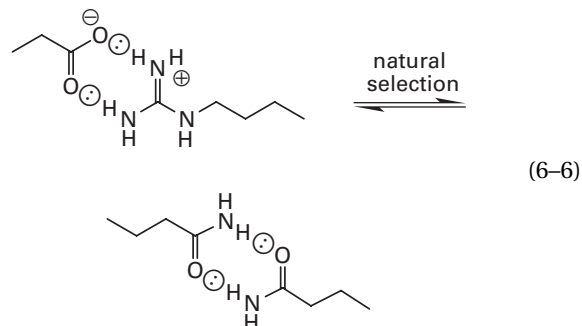
A buried ionized hydrogen bond is less stable than a buried neutral hydrogen bond³⁴⁷ and certainly less stable than an isochoric pair of hydrophobic side chains. When the partially buried ionized hydrogen bonds

among Arginine 31, Glutamate 36, and Arginine 40 in the Arc repressor were replaced with hydrophobic interactions among a methionine, a tyrosine, and a leucine, respectively, the mutant that resulted was -16 kJ mol^{-1} more stable than the wild-type protein.³⁴⁸ Why buried ionized hydrogen bonds uninvolved in the function of the protein have not been eliminated in this way by evolution by natural selection is unknown. Buried, ionized hydrogen bonds, however, are rare; most ionized hydrogen bonds are found on the surfaces of crystallographic molecular models of proteins where they can be stabilized by the solvation of the water. Even then they represent a minority of the ionized hydrogen bonds that potentially could form. Most of the fortuitously juxtaposed, oppositely charged side chains on the surface of a crystallographic molecular model do not participate in hydrogen bonds³¹⁹ “even though in most cases there is no steric reason why they cannot.”³⁴⁹ It is the competition of the donors and acceptors of the water that prevents it.

The **frequency** with which ionized hydrogen bonds are observed in crystallographic molecular models (Table 6–5) is no greater than the probability that they would occur at random. Only hydrogen bonds between two side chains are considered in the tabulation, and the probability that a certain hydrogen bond will form at random is calculated from the frequency with which the amino

acids occur in the usual protein and the number of equivalent donors or equivalent acceptors present on each amino acid. If anything, ionized hydrogen bonds are observed less frequently than predicted by this calculation of probability. This may be due to the fact that both charged donors and charged acceptors will tend to be more exposed to the water and less likely to form hydrogen bonds. A reciprocal argument could be invoked to explain the fact that hydrogen bonds between hydroxyl groups are more frequent than expected (Table 6–5), because amino acids bearing hydroxyl groups are often buried (Table 6–2). Nevertheless, with few exceptions, the frequencies with which each of the particular hydrogen bonds are observed are about those expected from the probability that the respective donor and acceptor would encounter each other at random, regardless of charge.

An example of the interchangeability of charged and uncharged donors and acceptors of hydrogen bonds occurs in phycobiliproteins, where an ionized hydrogen bond between an arginine and an aspartic acid in one species is replaced isochorically by hydrogen bonds between two glutamines in another:³⁵²



This, however, may not be very common because the amino acids surrounding an ionized hydrogen bond have been selected for their ability to solvate the charges and such a tailored environment should resist the neutralization of the bond.³⁵³ Nevertheless, all of these considerations suggest that an ionized hydrogen bond is no different from an un-ionized hydrogen bond except that it should be less stable when exposed to solvent and present greater problems of solvation when it is buried.

Suggested Reading

Gibbs, M.R., Moody, P.C.E., & Leslie, A.G.W. (1990) Crystal structure of the Aspartic acid-199 → Asparagine mutant of chloramphenicol acetyltransferase to 2.35-Å resolution: Structural consequences of disruption of a buried salt bridge. *Biochemistry* 29, 11261–11265.

Timasheff, S.N. (2002) Protein hydration, thermodynamic binding, and preferential hydration. *Biochemistry* 46, 13473–13482.

Problem 6–8: Examine the stereoscopic presentation to the left of the refined map of electron density in the middle of a molecule of protein with the final crystallographic molecular model inserted into it:⁸¹



Problem 6–8

Table 6-5: Frequency of Hydrogen Bonds between Side Chains

hydrogen bond	number ^a	probability ^b (%)	hydrogen bond	number	probability ^b (%)
	2	5		4	4
	10	12		9	7
	6	10		7	4
	0	4		15	8
	5	9		3	5
	0	7		4	6
	12	9		1	7 ^c
	3	18 ^c		17	10

^aFrom tables in refs 8, 22, 97, and 98. ^bProbability that the hydrogen bond would occur at random, calculated only from frequencies of functional groups^{350,351} in proteins and their respective number of donors or acceptors, assuming no preferences for type of hydrogen bond. ^cProbability on the same scale as the others but not included for normalization.

- (A) The side chain of which of the 20 amino acids descends from the top left of the figure into its center?
- (B) Draw the structures of all of the hydrogen bonds made by the donors and acceptors on this side chain. In your drawing include the σ lone pairs of the acceptors and the hydrogens of the donors as in the drawings in Table 6-5. Indicate clearly the chemical identity of each donor and acceptor in your drawing by including enough of its structure that there is no doubt as to what functional group it is and by labeling it.
- (C) Is the side chain charged or neutral? How can you be sure?

- (D) In a solution of protein, are there an excess of donors or acceptors for hydrogen bonds? On the basis of this consideration, why should all donors of hydrogen bonds find a partner? Do all of the donors on the amino acid side chain in the center of the figure find acceptors?
- (E) Because the figure is for a region of electron density from the center of the molecule of protein, what is the most unexpected feature of the arrangement? What seems to permit this unexpected arrangement?

Hydrogen Bonds

Although they are less frequent than the hydrogen bonds between the amido nitrogen–hydrogens and the acyl oxygens of the backbone producing the secondary structure of a protein, hydrogen bonds between the **donors and acceptors on the side chains** of the amino acids are common features of crystallographic molecular models. The **stereochemistry** of such hydrogen bonds is as expected.³⁵⁴ The various acceptors to the nitrogen–hydrogen bonds that are donors on the side chains of glutamine, asparagine, arginine, histidine, and tryptophan are located preferentially at positions to which the sp^2 nitrogen–hydrogen bonds of the donors are pointed. The donors to the oxygens that are acceptors on glutamate, aspartate, asparagine, and glutamine show some preference for the positions at 120° to the carbon–oxygen double bond to which the sp^2 lone pairs of electrons on the oxygens are pointed, but there is much more flexibility to their locations as they pivot around these lone pairs (Figure 5–10D). The distribution of hydrogen-bond donors and acceptors around the hydroxyl oxygens of serines and threonines is even more flexible, but there are noticeable preferences for the two positions at dihedral angles χ_2 of 80° and 280° . The donors and acceptors to the phenolic oxygen of tyrosine, however, have a strong preference to be in the plane of the ring at angles of 120° to the carbon–oxygen bond, as expected from the sp^2 hybridization of the oxygen. The three nitrogen–hydrogen bonds on lysine are almost always occupied by three respective acceptors arranged around the ammonium ion at angles near the 109° expected from its sp^3 hybridization^{15,355,356} but not located at any preferred dihedral angles χ_4 .³⁵⁴

It is fairly common (17% of the tryptophans, 9% of the tyrosines, 6% of the phenylalanines, and 1% of the histidines in crystallographic molecular models from data sets with minimum Bragg spacings less than 0.17 nm) for a nitrogen–hydrogen, an oxygen–hydrogen, or a sulfur–hydrogen bond to be directed towards the **π cloud of an aromatic side chain** with its hydrogen close enough (< 0.3 nm) to conclude that a hydrogen bond has been formed.³⁵⁷ Most frequently, these hydro-

gen bonds are between the π clouds of tyrosine and tryptophan acting as acceptors and the ammonium nitrogen–hydrogens of lysines.^{358,359} When the nitrogen–hydrogen bonds are themselves attached to a π system, however, as with glutamine, asparagine, arginine, and histidine, their π cloud is often stacked on the π cloud of the aromatic side chain. In such instances, the nitrogen–hydrogen bonds point away from the aromatic ring,^{359,360} and none can form a hydrogen bond with it. There are exceptions, however, such as Glutamine 96 in human HLA class I histocompatibility antigen A-2 (Figure 6–22) and Glutamine 399 in ribulose-bisphosphate carboxylase (Figure 6–44B).

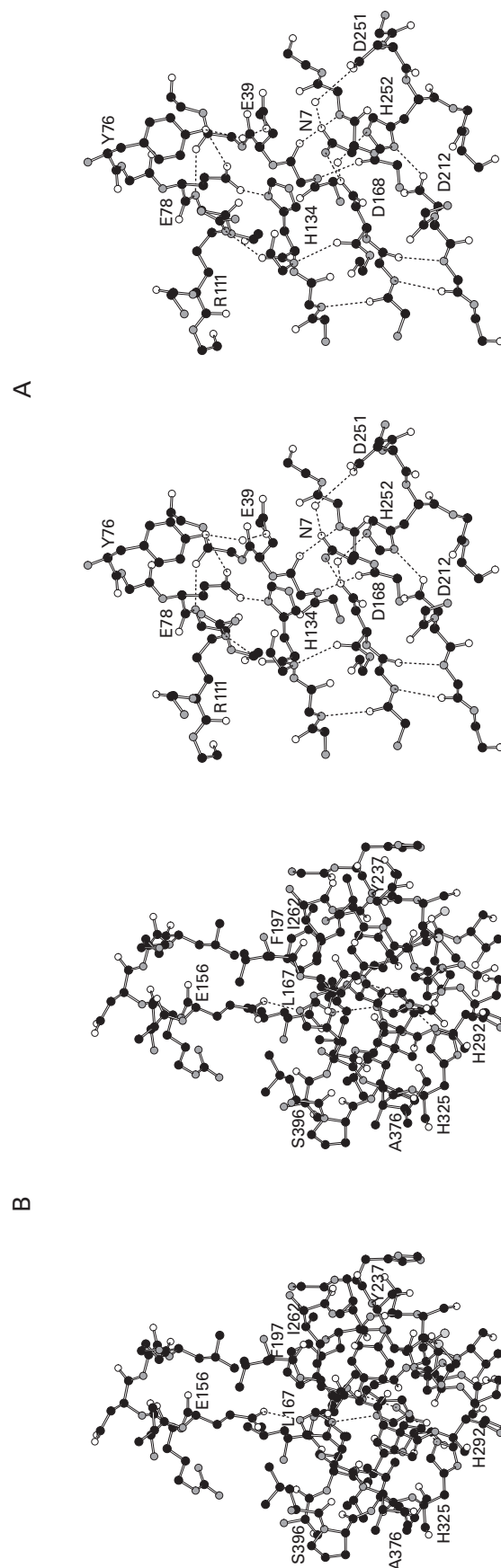
As is the case with the backbone of the polypeptide, when the donors of hydrogen bonds on the side chains of the amino acids are removed from the water and stripped of their hydrogen bonds with the solvent, there would be a considerable increase of enthalpy if they did not find **new partners** in the interior of the protein. Most if not all of them do.

One of the remarkable features of the **buried hydrogen bonds** that result from this energetic imperative is that they tend to be clustered. For example, of the 54 side chains in myoglobin that form hydrogen bonds with atoms in the protein other than bound water, 16 participate in eight closed pairs and nine participate in three closed triplets, but 29 participate in larger clusters.³⁰ These clusters often incorporate buried water. Examples of such clusters occur in deoxyribonuclease I (Figure 6–44A)⁸ and in ribulose-bisphosphate carboxylase (Figure 6–44B).³⁶¹ In these clusters, charged amino acids participate as donors and acceptors of hydrogen bonds as readily as uncharged amino acids and there is no obvious balancing between positive and negative charges (Figure 6–44A).

Clusters of hydrogen bonds serve to orient functionally important amino acids. For example, a “complex network of hydrogen bonds” serves to orient the six histidines responsible for chelating the copper and the zinc in superoxide dismutase.³⁴¹ Histidine 57 in the active site of chymotrypsin is oriented by a hydrogen bond to Aspartate 102, which in turn is oriented by three other hydrogen bonds, one to each of its three remaining acceptors.¹⁷ Histidine 31 in deoxyribonuclease I is functionally important and is held in position by the cluster in which it participates (Figure 6–44A), as is Histidine 325 in ribulose-bisphosphate carboxylase (Figure 6–44B). The hydrogen bond in the case of deoxyribonuclease I forces the dihedral angle χ_2 of Histidine 31 to assume an unfavorable value when it is positioned properly. Carboxylic acids, histidines, and arginines are most susceptible to such pinning because they have donors and acceptors at two or more separate locations on their side chains, and they are rigid structures because of their π molecular orbital systems. These features make them easily immobilized.

Just as an accounting of the concentrations of all of

Figure 6-44: Examples of clusters of hydrogen bonds among side chains. (A) A large cluster of partially buried hydrogen bonds in the active site of the crystallographic molecular model (Bragg spacing ≥ 0.20 nm) of bovine deoxyribonuclease I.⁸ Segments of the folded polypeptide from Phenylalanine 6 to Arginine 9, Isoleucine 59 to Arginine 63, Tyrosine 76 to Tyrosine 80, Serine 110 to Glutamate 112, Alanine 132 to Serine 135, Methionine 166 to Phenylalanine 169, Alanine 232 to Arginine 235, and Serine 250 to Proline 254 are drawn. Only those side chains involved in the hydrogen-bonding are drawn. Locations of molecules of water are drawn as unbonded oxygen atoms (open circles). This drawing was produced with MolScript.⁵⁷³ (B) Completely buried string of hydrogen bonds in the center of the β barrel of parallel consecutive strands in the crystallographic molecular model (Bragg spacing ≥ 0.16 nm) of ribulose-bisphosphate carboxylase.^{361,572} The β barrel has eight β strands and each inserts two amino acids, i and $i + 2$, into the core. The drawing presents the eight consecutively arranged strands of the β barrel from Leucine 167 to Threonine 171, Phenylalanine 197 to *N*-(Carboxy)lysine 199, Glycine 235 to Tyrosine 237, Isoleucine 262 to Aspartate 266, Leucine 288 to Histidine 292, Histidine 323 to Histidine 325, Leucine 373 to Alanine 376, and Serine 396 to Glutamine 399, as well as a segment from Valine 155 to Leucine 160 capping the barrel. Only hydrogen bonds between side chains in the core of the β barrel are drawn. These are a string from Glutamate 156 through Histidine 323 and Histidine 290 to Histidine 325 and hydrogen bonds between the two amido nitrogen-hydrogens on Glutamine 399 and the phenolic oxygen of Tyrosine 237 and the π system of Histidine 290. Consequently, all six of the hydrogen bonds together form a linked cluster. This drawing was produced with MolScript.⁵⁷³



the charged species in a solution, the charge balance, is essential to a quantification of the behavior of acids and bases, an accounting of the individual concentrations of all of the donors and acceptors of hydrogen bonds in a solution, the **hydrogen-bond balance**, is essential to a quantification of their behavior. In a solution of protein, for example, the cytoplasm of a cell, the concentration of acceptors of hydrogen bonds exceeds the concentration of donors. There are 300 acceptors but only 180 donors for every 100 amino acids in a protein.^{350,351} The presence of nucleic acid and carbohydrate only increases this disparity. When a polypeptide is unfolded, its donors and acceptors are freely accessible to the solution and participating in rapidly changing hydrogen bonds with molecules of water, but because at any given instant only one acceptor can pair with each donor, there will be, at all times, a concentration of unoccupied acceptors at least as large as the concentration of this inescapable excess of acceptors over donors. The concentration of unoccupied donors at a given instant, however, will be small if not insignificant.

Whenever the donor of a hydrogen bond is removed from water during the folding of a protein but paired with an acceptor from the protein, the **total number of hydrogen bonds in the solution** does not change. Almost every one of these new hydrogen bonds will have the same enthalpy of formation as the old hydrogen bond between that donor and water because

almost all of the acceptors within a molecule of protein have acid dissociation constants associated with their σ lone pairs of electrons that are not appreciably different from that of the lone pairs of electrons on water ($pK_a = -1.7$). Consequently, the enthalpy of formation of that hydrogen bond will be close to 0 (Equation 5-49). Because both the competition of the water for this donor (Equation 5-45) and the entropies of approximation involved in forming the regular structures of the polypeptide backbone (Figure 5-19) or a hydrogen bond between two side chains are entropic terms, they can be combined into the larger question of the change in standard entropy accompanying the folding of the polypeptide.

If upon folding, however, a **donor for a hydrogen bond**, such as the nitrogen-hydrogen of an amide or the oxygen-hydrogen of a hydroxyl group, finds itself sequestered within the structure without an acceptor, the number of hydrogen bonds in the solution decreases by one. This unsatisfactory sequestration would produce a change in standard enthalpy* of +15 to +20 kJ mol⁻¹ (Table 5-2) and would consequently squander a considerable portion of the net free energy available for folding. Consequently, such a loss must be avoided, and it is likely that every nitrogen-hydrogen bond and oxygen-hydrogen bond of a folded polypeptide participates as a donor in a hydrogen bond, either with water, with an acyl oxygen of a peptide bond, or with a lone pair of electrons on a side chain. It comes as no surprise that there are few^{6,362} if any^{113,363} unoccupied donors of hydrogen bonds in crystallographic molecular models.

If, upon folding, an acceptor such as a lone pair of electrons on the acyl oxygen of an amide or on the oxygen of a phenol or alcohol finds itself without a donor, there is not much of a penalty. For example, if as many as half of the excess of acceptors over donors in the polypeptide were to become sequestered unoccupied, the increase in the free energy of formation of the hydrogen bonds in the solution would be less than $-RT \ln 0.5$ or 1.7 kJ (mol of folded polypeptide)⁻¹. It comes as no surprise that there are quite a few unoccupied acceptors of hydrogen bonds in crystallographic molecular models. The most obvious examples of unoccupied acceptors are the second lone pairs of electrons on the acyl oxygens in a β sheet buried in the center of a molecule of protein (Figure 6-9). The fact that only a fraction of the acyl oxygens on either the backbone or on the side chains end up with two donors (Figure 6-7) is inconsequential because many acceptors were vacant before folding occurred

* This change in standard enthalpy is not to be confused with the dissociation of a hydrogen bond in the reaction described in Equation 5-22. In this situation, in which the dissociated donor and acceptor are not sequestered from the solvent, equienergetic hydrogen bonds are formed between the dissociated donor and the dissociated acceptor with surrounding molecules of water, there is no net decrease in the concentration of hydrogen bonds in the solution, and the change in standard enthalpy is 0.

anyway. All of these considerations should be kept in mind when the standard free energy of formation for a hydrogen bond is being assessed experimentally, because it is often the case that these differences in importance between donor and acceptor affect the results of the experiment.

The necessity that the donor of a hydrogen bond retain an acceptor is particularly relevant when the indole of **tryptophan** is considered. The side chain of tryptophan is remarkably soluble in ethanol,¹⁷⁵ which has twice as many acceptors of hydrogen bonds as donors. Likewise, during partition between water and 1-octanol, the side chain of tryptophan has the greatest preference for 1-octanol of all the amino acids (Figure 5-24). As the indole contains only a donor, a net of one hydrogen bond is created every time it is dissolved in ethanol. When it is transferred from water to 1-octanol, a net of one hydrogen bond is also created because a solution of indole in water has more donors than acceptors and 1-octanol has more acceptors than donors. When indole is transferred, empty donors disappear from water and empty acceptors disappear in the alcohol. Consequently, the side chain of tryptophan is significantly more hydrophilic³⁶⁴ than is indicated by its solubility in ethanol or its transfer between water and 1-octanol, two proposed measurements of its hydrophobicity.^{174,175} Because a solution of protein has, as does ethanol or 1-octanol, more acceptors than donors, similar imbalances of donors and acceptors have a major effect on the distribution of amino acids between the surface and the interior of a molecule of protein or the coupling of the donors and acceptors of hydrogen bonds withdrawn from water during the process of folding the polypeptide.

One example of the strong tendency of tryptophan to retain its hydrogen bond with water occurs in the structure of the Bence-Jones protein Rhe. A tryptophan in the center of the crystallographic molecular model of this protein, though completely buried, is engaged in a hydrogen bond with a buried molecule of water sitting next to its indole nitrogen (Figure 6-39).²⁸⁰ This molecule of water is trapped in the interior during the folding of the polypeptide, and its two donors are hydrogen-bonded in turn to two acyl oxygens from the backbone. In γ -II crystallin, two of the tryptophans are also hydrogen-bonded to buried molecules of water.¹⁶⁸ Usually, however, tryptophan retains the hydrogen bond to the nitrogen-hydrogen bond of its indole in less dramatic ways. For example, all of the donors in the indoles of the tryptophans of chymotrypsin retain hydrogen bonds with the solvent or another acceptor in the interior.¹⁷ The other two tryptophans in γ -II crystallin form hydrogen bonds with acyl oxygens.¹⁶⁸ In deoxyribonuclease I, all of the tryptophans, though mostly buried, retain contact with the solvent at their nitrogen-hydrogen bonds.⁸ The indole nitrogen-hydrogen bond of Tryptophan 21 in the lipoyl domain of dihydrolipoyllysine-residue acetyl-

transferase from *B. stearothermophilus*, which does not fully exchange with $^2\text{H}_2\text{O}$ in the solvent over 3 years, is well buried in the core of the protein but hydrogen-bonded to the acyl oxygen of Proline 61.³⁶⁵

There are also other anecdotal instances in which the **requirement that donors must be occupied** seems to be expressed. Arginine is one of the best examples. When it is partially buried, all of the five hydrogen-bond donors on the guanidinium are provided with acceptors (Figure 6–41).³⁶⁶ In the binding site on trypsin with which both arginine and lysine associate normally, there is a constellation of acceptors that can occupy in turn the five donors on the former and the three donors on the latter even though the dispositions of those donors do not overlap. Consequently, there are empty acceptors in each complex but never empty donors.³⁶⁷ When Tyrosine 385 in 4-hydroxybenzoate 3-monooxygenase is mutated to a phenylalanine, it creates an empty acceptor on Tyrosine 201 and nothing happens, but when Tyrosine 201 is mutated to phenylalanine, it creates an empty donor on Tyrosine 385 and a molecule of water is found in the crystallographic molecular model sitting where the hydroxyl of Tyrosine 201 used to be and occupying that donor.³⁶⁸ When an unoccupied hydrogen-bond donor in the complex between a peptide and penicillopepsin is replaced with a methylene group, the inhibitor binds 400 times more tightly.³⁶⁹ “The pH dependence of chromate binding and the extremely low affinity of phosphate are attributable mainly to the lack of hydrogen bond acceptors in the binding site” of sulfate-binding protein from *Salmonella typhimurium*.³⁷⁰

The difference in the importance of a donor and that of an acceptor affects the magnitude of the free energy of formation of a hydrogen bond in a protein, but so does its location in the structure. One of the unexpected observations resulting from an examination of crystallographic molecular models is the high frequency with which hydrogen bonds between donors and acceptors, each from the protein itself, occur on the surface of the folded polypeptide.¹⁶⁸ Because of the strong hydration of ions or the high relative permittivity of liquid water or both of these factors, ionized hydrogen bonds between monovalent anions such as formates or acetates and monovalent cations such as alkyl ammoniums, imidazoliums, or guanidiniums have negligible standard free energies of formation in aqueous solution.³⁷¹ Consequently, ionized hydrogen bonds on the surface of a molecule of protein should be unstable, but so should neutral hydrogen bonds because the competition for the donors and acceptors by the water surrounding them should prevent them from forming.

Many of the hydrogen bonds found on the surface of a crystallographic molecular model are **artifacts** of the constraints applied during refinement. If potential energies that favor rather than disfavor the formation of an ion pair or a hydrogen bond are incorporated advertently or inadvertently into the procedure for refinement, ionized

and un-ionized hydrogen bonds will form during the refinement rather than in the actual protein. Such fantastical hydrogen bonds on the surface of a crystallographic molecular model tend to appear and disappear as further refinement is performed and as the data set is extended to narrower Bragg spacing. For example, a set of 12 hydrogen bonds on the surface of myoglobin between pairs of amino acid side chains in which both of the partners have been conserved by natural selection throughout all myoglobin sequences had been identified in a refined molecular model of the protein.³⁰ When the Bragg spacing of the data set was decreased and the refinement significantly improved,³⁷² seven of these hydrogen bonds, four of which had been between oppositely charged side chains, were no longer present in the crystallographic molecular model.* All assignments of hydrogen bonds between two amino acid side chains on the surface of a protein should be regarded with skepticism unless properly calculated **omit maps** clearly indicate their existence.†

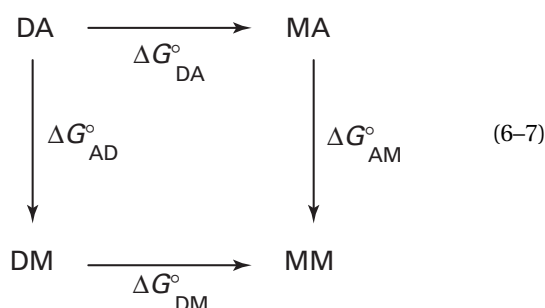
Nevertheless, hydrogen bonds, both ionized and un-ionized, probably do exist on the **surface of a protein**. When they do, there are probably particular reasons for their existence. **Steric effects** of neighboring amino acids and the backbone of the polypeptide can bring a donor and acceptor together in an orientation such that **entropy of approximation** sufficient to overcome solvation of ions and competition by water is realized. It is also possible that these hydrogen bonds are simply the random result of the participation of all of the donors and acceptors on the surface of the molecule of protein in the hydrogen-bonded network of the water surrounding it (Figure 6–38). In this case, these hydrogen bonds would be only the fortuitous outcome of the fact that the positions of these donors and acceptors in the larger lattice happen to be adjacent to each other. This hydrogen-bonded network of waters and donors and acceptors from the protein itself should be a rather fluid structure. The crystallographic molecular model represents only the structure of lowest energy in a constantly fluctuating environment. One observation, however, suggesting that some of these hydrogen bonds on the surface of the crystallographic molecular model are real is that they have negative standard free energies of formation.

The standard free energies of formation for hydrogen bonds seen in the crystallographic molecular models of proteins have been estimated by site-directed mutation. It is not possible to make an accurate estimate of the standard free energy of formation of such a hydrogen bond by mutating only one member of the pair.³⁷³ A single mutation will always have steric, hydrophobic, and electrostatic effects associated with it that are unrelated to the

* C. Chothia and A.M. Lesk, personal communication.

† This is yet another instance in which omit maps must be used to position atoms correctly and eliminate the artifacts inherent in the constraints applied during refinement by simulated annealing.

loss of the hydrogen bond itself but that cannot be separated from the change in standard free energy for only the loss of the hydrogen bond. To correct for these effects, a double-mutant cycle is performed.³⁷⁴ A **double-mutant cycle** is a set of three site-directed mutations: the single mutation of the donor in the hydrogen bond, the single mutation of the acceptor in the hydrogen bond, and the double mutation of both. The four standard free energies of folding of these three mutants and the unmutated wild-type protein are then measured, and they are used to create a linkage relationship:



where DA is the wild-type protein, MA is the single mutant of the donor, DM is the single mutant of the acceptor, and MM is the double mutant. The change in standard free energy $\Delta G_{\text{DA}}^{\circ}$ is the change in free energy of folding when the donor is mutated in the presence of the acceptor; $\Delta G_{\text{DM}}^{\circ}$, when the donor is mutated in the absence of the acceptor; $\Delta G_{\text{AD}}^{\circ}$, when the acceptor is mutated in the presence of the donor; and $\Delta G_{\text{AM}}^{\circ}$, when the acceptor is mutated in the absence of the donor. By definition

$$\Delta G_{\text{DA}}^{\circ} + \Delta G_{\text{AM}}^{\circ} = \Delta G_{\text{AD}}^{\circ} + \Delta G_{\text{DM}}^{\circ} \quad (6-8)$$

Each of these changes in standard free energy is the difference* in the standard free energies of folding between the two proteins connected by the respective arrows defining the mutation. A positive value for the difference states that the mutant version is less stable than the unmutated version. Conveniently, although the actual standard free energy of folding of a protein cannot be estimated accurately, differences in standard free energy of folding can.³⁷⁵

The **free energy of formation of the hydrogen bond** $\Delta G_{\text{AHB}}^{\circ}$ (Table 6-6) should be

$$\Delta G_{\text{AHB}}^{\circ} = \Delta G_{\text{DM}}^{\circ} - \Delta G_{\text{DA}}^{\circ} = \Delta G_{\text{AM}}^{\circ} - \Delta G_{\text{AD}}^{\circ} \quad (6-9)$$

For example, in the crystallographic molecular model of ribonuclease from *Bacillus amyloliquifaciens*, there is an

* As is usually the case in physical chemistry, the difference is the standard free energy of folding for the product of the mutation minus that for the unmutated protein, the reactant.

Table 6-6: Standard Free Energy of Formation of Hydrogen Bonds in Proteins Estimated from Double-Mutant Cycles^a

donor/acceptor	location ^b	$\Delta G_{\text{AHB}}^{\circ c}$ (kJ mol ⁻¹)
lysinium/glutamate ³⁷⁶	surface	-2.3
argininium/aspartate ³⁷⁷	surface	-0.9
argininium/aspartate ³⁷⁷	surface	-2.0
aspartate/argininium/aspartate ^{d 377}	surface	-3.3
amino terminus/glutamate ³⁷⁸	surface	-6.3
lysinium/glutamate ³⁷⁸	surface	0.0
serine/aspartate ³⁷⁹	buried	-5.7
lysine/threonine ³⁸⁰	buried	-6.3 ^e
arginine/glutamate ³⁸⁰	buried	-7.1 ^e
arginine/aspartate ³⁸⁰	buried	-27 ^{ef}
histidine/aspartate ³⁸⁰	buried	-20 ^e

^aWhere available, values were for measurements derived from the standard free energies of folding or standard free energies of association at an ionic strength close to that encountered in cytoplasm (0.15–0.2 M). ^bBased on a crystallographic molecular model of the protein. ^cStandard free energy of formation calculated from double-mutant cycles by Equation 6-9. ^dTriple-mutant cycle; standard free energy is for the mean of the two hydrogen bonds. ^eEstimated from free energies of association rather than free energies of folding. ^fMean of two hydrogen bonds.

ionized hydrogen bond between Arginine 110 and Aspartate 8. The difference in free energy of folding between the protein with both the arginine and the aspartate and a mutant in which Arginine 110 is replaced with alanine, $\Delta G_{\text{DA}}^{\circ}$, is -3.3 kJ mol⁻¹; that between the mutant in which Aspartate 8 is replaced with alanine and the double mutant, $\Delta G_{\text{DM}}^{\circ}$, is -4.2 kJ mol⁻¹; that between the protein with both the arginine and the aspartate and the mutant in which Aspartate 8 is replaced with alanine, $\Delta G_{\text{AD}}^{\circ}$, is +2.3 kJ mol⁻¹; and that between the mutant in which Arginine 110 is replaced with alanine and the double mutant, $\Delta G_{\text{AM}}^{\circ}$, is +1.4 kJ mol⁻¹. Therefore, the free energy of formation of the hydrogen bond³⁷⁷ is -0.9 kJ mol⁻¹. The single mutations of each member of this particular pair would have led to contradictory conclusions concerning the strength of the hydrogen bond; on the one hand that it was endergonic and on the other that it was exergonic.

As expected, the hydrogen bonds between donors and acceptors on the surface of a protein have **marginal stability** (Table 6-6) and are in the range observed for similar hydrogen bonds on an isolated α helix (Table 5-7). The standard free energies of formation for buried hydrogen bonds, however, are significantly more favorable, again as expected. Hydrogen bonds found in buried locations of a molecule of protein are removed from competition with water, found in a region of lower relative permittivity, and fixed more rigidly in their orientations than hydrogen bonds on the surface. All three of these circumstances should significantly increase their stability relative to those on the surface, so it is surprising that the differences observed are as small as they are.

Even the value for the standard free energy of for-

mation of a particular hydrogen bond in a protein obtained by a double-mutant cycle is not completely free of contributions from interactions with neighboring amino acids. When all of the amino acids in a cluster of hydrogen bonds surrounding the hydrogen bond between Arginine 218 (TEM) and Aspartate 49 (BLIP) in the complex between β -lactamase TEM-1 from *E. coli* and its inhibitor protein BLIP were mutated to alanine, the standard free energy of formation of that hydrogen bond increased from -9 to $+1$ kJ mol^{-1} .³⁸¹ Similar increases of 4 – 6 kJ mol^{-1} were observed when the amino acids surrounding three other hydrogen bonds in the same cluster were mutated to alanine. Consequently, the standard free energies of formation listed in Table 6–6 may be only **lower limits** of the value for that hydrogen bond in the absence of assistance from its surroundings.

Approximation is probably the greatest contributor to the stability of a buried hydrogen bond between two side chains. Following formation of the secondary structure and the alignment of secondary structures by packing, the donor and acceptor of a buried hydrogen bond should be efficiently aligned and a considerable amount of entropy of approximation should have been realized, yet the free energies of formation of such buried hydrogen bonds are less than -30 kJ mol^{-1} in the most advantageous circumstances (Table 6–6). The wide variability in the standard free energies of formation could reflect wide differences in the success with which donor and acceptor are aligned given all of the steric problems of the interior of a protein.

There are other experimental observations suggesting that approximation is not so successful as it should be. In a series of tight complexes (dissociation constants less than 750 nM) between thermolysin and a set of ligands that bind to its active site, when the respective nitrogen–hydrogens of the phosphoramidates in the ligands, which each form a hydrogen bond with the acyl oxygen of Alanine 113 in the crystallographic molecular model of the complex,³⁸² were replaced with methylenes, the association constants for the ligands remained the same.³⁸³ When corrected for the removal of the two hydrogen–carbon bonds of the methylene from water, the standard free energy of formation for this buried, rigidly aligned hydrogen bond is -6 kJ mol^{-1} , well within the range of those for buried hydrogen bonds in Table 6–6 but not anywhere near the value predicted from the entropy of approximation that must be realized. Even more surprising is that when the amido nitrogen–hydrogen of a hydrogen bond in the middle of an α helix of T4 lysozyme was replaced with an ester oxygen, the free energy of folding of the protein increased³⁸⁴ by 7 kJ mol^{-1} , but that increase was indistinguishable from the increase expected for the enthalpy of formation of the hydrogen bond between the acyl oxygen of the ester, relative to the acyl oxygen of the unmutated amide, and the nitrogen–hydrogen with which it forms a hydrogen bond.³⁸⁵ In other words, these latter experiments suggest that the

free energy of formation of a hydrogen bond in the middle of an α helix in a protein is indistinguishable from 0 , even though considerable entropy of approximation is realized. All of these results emphasize that it is difficult to form a hydrogen bond in an aqueous solution.

Even though there are hydrogen bonds in a molecule of protein that do have negative standard free energies of formation, it is the structure of the protein that approximates the donor and the acceptor, causing their hydrogen bond to become stable. It is this approximation that overcomes, in many cases meagerly, the otherwise overwhelming competition of the water for the donors and acceptors. The folding of the protein that approximates the donor and acceptor in such a hydrogen bond is driven entirely by the hydrophobic effect. It is only after the hydrophobic effect has collapsed the random coil, withdrawn the donors and acceptors from water, and excluded water from the interior that the hydrogen bonds of the α helices and β structure are able to form. It is only when the hydrophobic effect, expressed as the minimization of the internal volume of the protein, has locked the secondary structures into the tertiary structure, that donors and acceptors of hydrogen bonds between side chains are brought close enough together and are constrained sufficiently that they can form otherwise unfavorable hydrogen bonds. It is only after all of this prelude that the observed hydrogen bond has a lower standard free energy of formation than the hydrated, separated donor and acceptor had in the unfolded polypeptide.

It is the case that such favorable free energy of formation adds to the **stability of the protein**, but this is an illusory contribution. The amino acid sequence of the protein and hence the location and identity of each side chain in its structure is the result of evolution by natural selection. The hydrogen-bonded pair of side chains that currently occupies a particular location in the structure could have been chosen because it was the constellation of atoms that sterically filled that particular location in the structure most effectively relative to all of the other possibilities that were tried, not because it is a hydrogen bond. It has a favorable free energy of formation because the two side chains that were chosen for these other reasons, happened to end up with a donor and an acceptor adjacent to each other. The hydrogen-bonded pair is not necessarily the most energetically favorable pair of side chains that could have occupied that position. In fact, even though it was not so astute a process as evolution by natural selection that determined the choice of the replacements, it is sometimes the case that the double mutant in a double-mutant cycle or even one of the single mutants is as stable as the wild type containing the hydrogen bond.

The relationship between the strength of a hydrogen bond and the **difference in pK_a between donor and acceptor** (Figure 5–14) has been verified in the context of a molecule of protein. As the pK_a of Tyrosine 27 in micro-

coccal nuclease from *Staphylococcus aureus*, which forms a hydrogen bond with Glutamate 10, was lowered by substituting various fluorinated tyrosines, the free energy of folding of the protein, presumably reflecting the decreases in the free energy of formation of the hydrogen bond, decreased³⁸⁶ by 2.0 kJ mol⁻¹ (unit of p*K*_a)⁻¹. This value for the Brønsted coefficient is near that observed (Equation 5–37) for a hydrogen bond in CCl₄ [1.3 kJ mol⁻¹ (unit of p*K*_a)⁻¹]. This increase in strength, as the increase in the acidity of the phenolic side chain matches its p*K*_a more closely with that of the glutamate, suggests that, as in other situations, the hydrogen bond between an acid and its conjugate base should be a strong one.

In crystallographic molecular models there are examples of **hydrogen bonds between an acid and its conjugate base**. In turkey troponin C, Glutamate 57 forms a geometrically ideal hydrogen bond with Glutamate 88 in which it is unknown on which carboxylate the proton resides.²⁴ There is no experimental indication, however, that such hydrogen bonds are unusually stable. The histidinium ion in the hydrogen bond between Histidine 24 and Histidine 119 in sperm whale myoglobin has a p*K*_a of 6.0.³⁸⁷ If this were a particularly stable interaction, the p*K*_a of the acid dissociation that eliminates it should have been much higher (Table 2–2). The hydrogen bond³⁸⁸ between Lysine 206 and Lysine 296 of human transferrin, although necessarily lowering the values of p*K*_a for the lysines participating in it,³⁸⁹ has been shown to destabilize the protein.^{390–392}

No evidence has been presented that hydrogen bonds between acids and their conjugate bases in proteins display properties associated with low-barrier hydrogen bonds, but hydrogen bonds displaying one such property, a low **fractionation factor** (Equation 5–31), have been identified in proteins. The fractionation factor ϕ for a proton in a hydrogen bond in a protein is measured by following the fraction, f_{AHB} , of the hydrogen bond of interest that remains undeuterated, AH⊙B, as a function of the mole fraction $x_{\text{H}_2\text{O}}$ of undeuterated water in a series of mixtures of H₂O and D₂O

$$x_{\text{H}_2\text{O}} = \frac{[\text{H}_2\text{O}]}{[\text{H}_2\text{O}] + [\text{D}_2\text{O}]} \equiv \frac{[\text{L}_2\text{O}\odot\text{HOL}]}{[\text{L}_2\text{O}\odot\text{HOL}] + [\text{L}_2\text{O}\odot\text{DOL}]} \quad (6-10)$$

where L again stands for either H or D.

A physical property that monitors the concentration of the undeuterated hydrogen bond, such as the intensity (i_{AHB}) of the absorption of its proton in a nuclear magnetic resonance spectrum is monitored.^{393,394} Equations 5–31 and 6–10 can be combined to give³⁹⁵

$$\frac{i_{\text{AHB}}}{i_{0,\text{AHB}}} = \frac{[\text{AH}\odot\text{B}]}{[\text{AL}\odot\text{B}]} \equiv f_{\text{AHB}} \equiv \frac{x_{\text{H}_2\text{O}}}{\phi(1 - x_{\text{H}_2\text{O}}) + x_{\text{H}_2\text{O}}} \quad (6-11)$$

where [AL⊙B] is the total concentration of hydrogen bonds, both deuterated and protonated; and $i_{0,\text{AHB}}$ is the intensity of the absorption in H₂O. The normalized intensity of the absorption of the proton in the hydrogen bond as a function of $x_{\text{H}_2\text{O}}$ is fit by nonlinear least squares to Equation 6–11 to obtain ϕ .

In this way, the fractionation factors for the protons within the hydrogen bonds of the secondary structure of a protein can be measured. There are results suggesting that a significant portion of these protons have fractionation factors less than 1. For example, 13 of the 87 amino acids in the phosphocarrier protein HPr from *Bacillus subtilis*³⁹⁴ and 36 of the 231 amino acids in micrococcal nuclease from *S. aureus*³⁹³ have been reported to have amido protons with fractionation factors less than 0.80, and six of the 76 amino acids in ubiquitin have fractionation factors less than 0.90.³⁹⁶ There is some uncertainty to these measurements because it is quite difficult to equilibrate all of the protons in the hydrogen bonds of the secondary structure of a protein with deuterons in the solution,³⁴⁰ and an unequilibrated position would appear artifactually to have a low fractionation factor (Equation 5–31). In more recent studies of the fractionation factors of protons in streptococcal protein G³⁹⁷ and the SH3 domain of proto-oncogene protein-tyrosine kinase from *Gallus gallus*,³⁹⁶ none of the protons in the nitrogen–hydrogens of the backbone had fractionation factors less than 0.9. Nevertheless, it is thought to be the case that some of the protons in the hydrogen bonds of the secondary structure of many proteins have **abnormally low fractionation factors**.³⁹⁶

There also seems to be a correlation between the fractionation factor of a proton and the **length of the hydrogen bond** that it occupies in a crystallographic molecular model of a protein.³⁹⁷ In crystallographic molecular models built from data sets to Bragg spacing of less than 0.1 nm, the maps of electron density are accurate enough that the bond lengths of the hydrogen bonds are of sufficient reliability to identify those that are abnormally short,^{398–400} and there are usually a few abnormally short hydrogen bonds (0.26–0.28 nm) among those between amido nitrogen–hydrogens and acyl oxygens of the backbone.³⁹⁸ It is thought that such shortened hydrogen bonds are the ones that display low fractionation factors and therefore are low-barrier hydrogen bonds.³⁹⁷

It is not possible, however, for these short low-barrier hydrogen bonds to be strong hydrogen bonds³⁹⁶ because the difference in p*K*_a between the nitrogen–hydrogen (p*K*_a = 16) and the oxygen (p*K*_a = –0.5) is so large and any decrease in polarity would only widen the difference. Whenever a polymer as long and heterogeneous as a molecule of protein is folded into a unique conformation, it is hard to believe, in spite of evolution by natural selection, that all of the steric problems can be solved. There must be some places in the structure that are tight fits. When such a tight fit occurs at a hydrogen

bond between an amido nitrogen–hydrogen and acyl oxygen of the backbone, the hydrogen bond shortens to relieve the **strain**, much as the hydrogen bond in hydrogen maleate monoanion shortens in response to the steric compression. This shortened hydrogen bond must be weaker than the unshortened bond because there is repulsion energy in the compressed case that would be relieved on relaxation to the normal distance. This shorter but weaker hydrogen bond has a smaller fractionation factor because this property is determined only by the degree of overlap of the wells of potential energy confining the proton on the donor and the acceptor. It is a low-barrier hydrogen bond not because the strength of the bond has brought donor and acceptor together but because the contraction of the distance is imposed by the rest of the framework. It has also been concluded from studies of complexes between proteins and small ligands that there is no correlation between the length of a hydrogen bond and its strength.⁴⁰¹

Suggested Reading

Horovitz, A., Serrano, L., Avron, B., Bycroft, M., & Fersht, A. (1990) Strength and co-operativity of contributions of surface salt bridges to protein stability, *J. Mol. Biol.* 216, 1031–1044.

Problem 6–9: Aspartate 12 and Arginine 16 are located in an α helix on the surface of a mutant of the ribonuclease from *B. amyloliquifaciens*. The arginine and the aspartate do not form an ionized hydrogen bond in the crystallographic molecular model of the enzyme from the closely related species, *Bacillus intermedius*, even though they are close enough to each other to do so. Three mutants were produced in the ribonuclease from *B. amyloliquifaciens*: Arginine 16 \rightarrow threonine, Aspartate 12 \rightarrow alanine, and the corresponding double mutant. The differences in standard free energies of folding for the three mutants and the original protein were as follows:

	difference in standard free energy of folding (kJ mol ⁻¹)	
	ionic strength = 0.1 M	ionic strength = 0.55 M
ΔG°_{RD}	2.1	1.8
ΔG°_{RA}	1.7	2.0
ΔG°_{DR}	1.8	1.0
ΔG°_{DT}	1.4	1.2

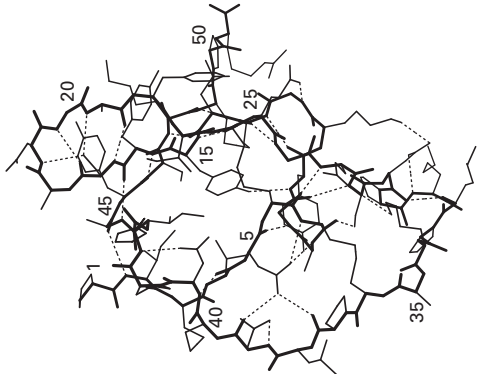
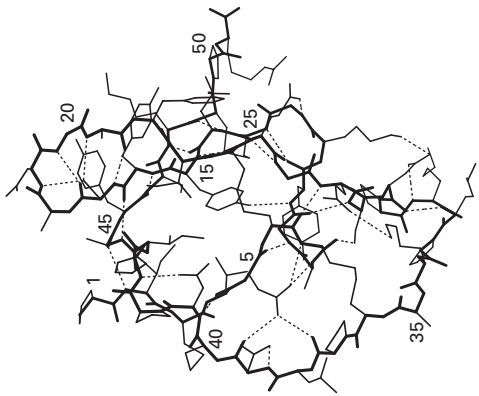
where ΔG°_{RD} is for mutation of Arginine 16 in the presence of aspartate at position 12; ΔG°_{RA} , for mutation of Arginine 16 in the presence of alanine; ΔG°_{DR} , for mutation of Aspartate 12 in the presence of arginine at position 16; and ΔG°_{DT} , for mutation of Aspartate 12 in the presence of threonine.³⁷⁵

- (A) Estimate the interaction between Arginine 16 and Aspartate 12 at the two ionic strengths.

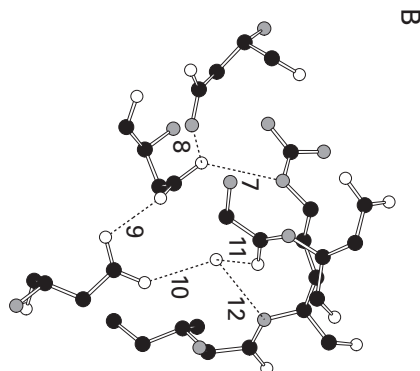
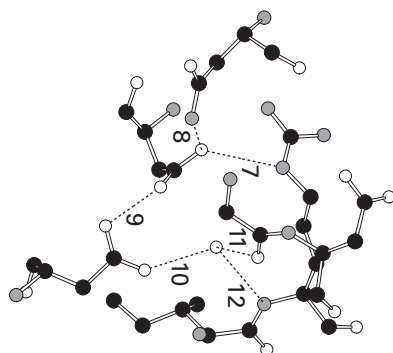
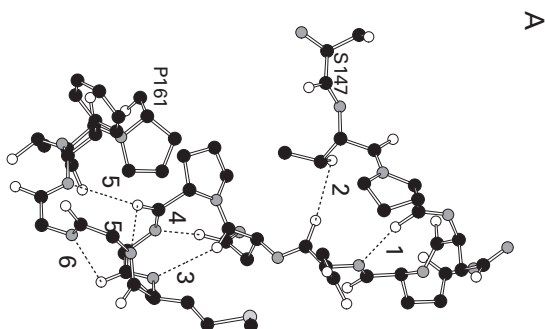
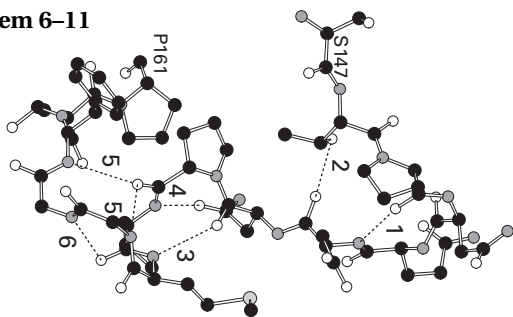
- (B) The uncertainty in your calculated values was estimated by the authors to be ± 0.2 kJ mol⁻¹. Is the electrostatic interaction significantly different from zero at physiological ionic strength? Is this surprising?
- (C) What conclusion would you have reached had only Arginine 16 been mutated?

Problem 6–10:

- (A) Write out the amino acid sequence of the protein in the drawing below of a crystallographic molecular model. This drawing was produced with MolScript.⁵⁷³
- (B) List the pairs of cysteines that participate in the cystines.
- (C) What structural feature of a cystine is illustrated by the model?
- (D) Identify the participants in a small hydrophobic cluster.
- (E) List as a pair the donor and the acceptor of each of the hydrogen bonds in the model by the letter and number of its respective amino acid and by the respective designation defined in Figure 4–14 for the atom participating in the hydrogen bond.
- (F) Which hydrogen bonds are probably artifacts of the procedure used to refine the molecular model?

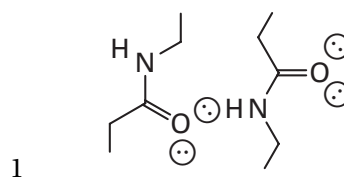


Problem 6-11



Problem 6-11: The two stereo drawings to the right represent portions of the crystallographic molecular models of two proteins.^{274,366} These drawings were produced with MolScript.⁵⁷³

Each of the hydrogen bonds in each stereo drawing is numbered in the figure. Draw the chemical structure of each hydrogen bond in the two stereo drawings. Number each of your drawings with the number for the hydrogen bond assigned in the figure. There are two hydrogen bonds numbered 5 for the same acceptor. Draw all of the lone pairs and all of the hydrogens on each functional group providing the donor and on each functional group providing the acceptor. For example, the correct chemical structure of hydrogen bond number 1 is



Association of Proteins with Nucleic Acid

It is during the association of proteins with nucleic acids that the importance of packing, the existence of fixed positions for molecules of water, the irrelevance of direct compensation of charge, and the stereochemical role of hydrogen bonding are all manifest. A **double helix of DNA** (Figure 3-9)* presents to the protein designed to associate with it a regular structure that can be recognized by its peculiar shape, its pattern of hydration, its high density of negative charge, and its array of donors and acceptors of hydrogen bonds.

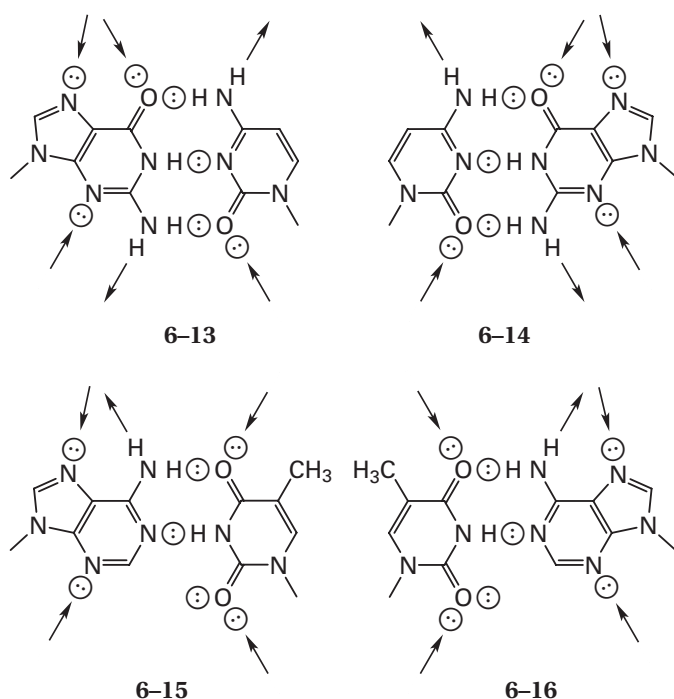
Along each of the two **phosphodiester backbones**, there are regularly spaced pairs of **phosphoryl oxygens** on each phosphorus in each phosphodiester linkage. Within each pair, the two oxygens share between themselves the single negative elementary charge of the phosphoryl group and are directed outward at the sp^3 angle of 109.5° , and each oxygen has the equivalent of two acceptors for hydrogen bonds.

The **pairs of bases** contain their own internal hydrogen bonds, two for the pair between adenine and thymine and three for the pair between guanine and cytosine, and these are located in the core of the structure. Although there are a small number of proteins that can induce a single base to swing out of the stack, thus exposing both itself and its interior donors and acceptors,^{402,403} this conformational change is a difficult one. Most proteins bind to DNA in its normal conformation and never see these hydrogen bonds in the core. The pairs of bases are **stacked** one upon the other, but they

* You should trace the two polynucleotide backbones (O-C-furanose-O-P-O-C-furanose-O-P-...) through the double helix.

do not overlap entirely. Consequently they are a helical staircase but one with narrow treads.

There are two helical grooves, the **major groove** and the **minor groove** (facing the viewer in the upper half and lower half, respectively of Figure 3-9), the former wider than the latter. It is in these grooves that the narrow treads of the stairs are found. Each pair of bases projects a characteristic pattern of donors and acceptors into each groove. The pair of adenine and thymine projects a methyl group, an acceptor, a donor, and an acceptor into the major groove and two acceptors into the minor groove; and the pair of guanine and cytosine projects an acceptor, an acceptor, and a donor into the major groove and an acceptor, a donor, and an acceptor into the minor groove. The order and orientation of these donors and acceptors differs between a guanine-cytosine pair (6-13) and a cytosine-guanine (6-14) pair and between an adenine-thymine pair (6-15) and a thymine-adenine pair (6-16):⁴⁰⁴



so that a sequence containing a G can be distinguished from one containing a C; and a sequence containing an A, from one containing a T.

All of these donors and acceptors of hydrogen bonds provide fixed positions for occupation by molecules of water, but not all of them are firmly occupied. Each segment of DNA has its own characteristic pattern of fixed positions for molecules of water (open circles in Figure 3-9), and it has been observed that when a molecule of protein binds to DNA, some of these positions remain occupied within the complex.⁴⁰⁵ Consequently, the donors and acceptors on these molecules of water that are incorporated into the complex, as well as those provided by the bases in the major groove and the minor

groove and those along the backbone, all provide keys for the recognition of the double-helical DNA by the protein.

There are two **levels of recognition** on which proteins operate in binding to DNA. Certain proteins are required by their function to recognize any segment of double-helical DNA regardless of its sequence. Examples of such proteins are histones that form chromatin from DNA, the RecA protein that catalyzes recombination, helicase and DNA-directed DNA polymerase that are components of the system replicating DNA, and DNA topoisomerase that passes one segment of DNA through another. These proteins recognize only the overall shape of a molecule of DNA and the acceptors along its phosphodiester backbone. Other proteins are required by their function to recognize specific sequences of double-stranded DNA and bind tightly to them. Examples of such proteins are repressors that shut off certain genes, transcription factors that initiate transcription at certain genes, and activators that increase the rates of transcription of certain genes. Many of these latter proteins are able to bind to any segment of a double helix of DNA and then run along the double helix until they reach their targets, and proteins of this type must perform both levels of recognition. Such proteins demonstrate that the ability to recognize specific sequences is a special case of the ability to recognize DNA in general.

One property of the proteins that recognize DNA is that their composition is biased against negatively charged amino acids and in favor of **positively charged amino acids**.⁴⁰⁶ On the open surfaces of these proteins that do not participate in the complexes with DNA, the density of glutamates and aspartates is the same as that on the open surface of any other protein, but in the interfaces between these proteins and DNA, the density of glutamates and aspartates is about 40% of that found in the interfaces between two molecules of protein. On the open surfaces, the density of arginine and lysine in these proteins is 30% greater than that on the open surfaces of other proteins, but at the interfaces between these proteins and DNA, the density of lysines and arginines is 2.5 times that in the interfaces between two molecules of protein. These equivalent biases against electrostatic repulsion and in favor of electrostatic attraction are reasonable responses to the high density of negative charge on the DNA to which these proteins must bind.

These overall biases, however, are distributed evenly over the interface between protein and DNA and not focused only on the **phosphodiester**s within it. Although lysines and arginines do provide donors to the acceptors on the phosphodiester backbone,⁴⁰⁷ they do so no more frequently than other amino acids. For example, in the crystallographic molecular model of the complex of DNA with rat DNA polymerase β , only one of the donors to the phosphodiester is the guanidinium of an arginine; the other nine are nitrogen-hydrogens from the polypeptide backbone and hydroxyls of a tyrosine and a threonine.⁴⁰⁸ In the crystallographic molecular model of the complex

between DNA and the regulatory protein Cro from bacteriophage 434, many of the donors to the phosphoryl oxygens are amido nitrogen–hydrogens from the polypeptide backbone,⁴⁰⁹ while in that between DNA and the regulatory protein Cro from bacteriophage λ , a tyrosine, a threonine, an asparagine, a glutamine, and two amido nitrogen–hydrogens from the backbone provide donors to phosphoryl oxygens (Figure 6–45).⁴¹⁰ In the complex between topoisomerase I and DNA, hydrogen-bond donors to the phosphoryl oxygens are provided by an

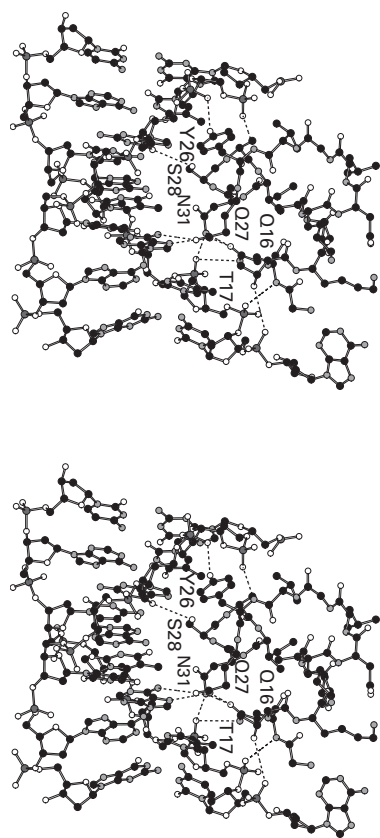


Figure 6–45: Recognition of phosphoryl oxygens on DNA by donors of hydrogen bonds from the amino acids in a protein.⁴¹⁰ The interactions depicted are from the crystallographic molecular model (Bragg spacing ≥ 0.30 nm) of the complex between regulatory protein Cro from bacteriophage λ (66 aa) and the double-helical duplex between d(ACATACACCGGGGTGATAC) and d(TGATACACCGGGGTGATAG). The portion of the model displayed is the double-helical duplex between d(ACATACAC) and d(TGATAG) and the protein from Glycine 15 to Asparagine 31. Tyrosine 26, Asparagine 31, Glutamine 16, Threonine 17, and the amido nitrogen–hydrogens from the backbone at positions 16 and 26 form hydrogen bonds with the phosphoryl oxygens of five bases. Only those hydrogen bonds between donors and acceptors on the protein and acceptors and donors on the DNA are drawn. Note the double hydrogen bond between Glutamine 27 and an adenine. This drawing was produced with MolScript.⁵⁷³

asparagine and a histidine among others.⁴¹¹ And in three successive phosphodiester in the complex between deoxyribonuclease I and DNA, an arginine, two histidines, an aspartic acid, an asparagine, a tyrosine, and a threonine provide the donors.⁴¹²

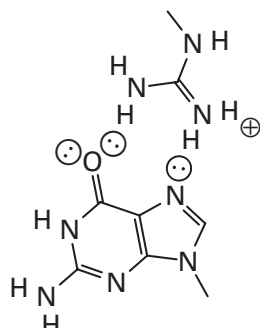
Donors to phosphoryl oxygens are also provided by molecules of water that are incorporated into the complex and form bridges to donors and acceptors on the protein.⁴¹³ In the complex between the repressor from bacteriophage λ and DNA, five amido nitrogen–hydrogens, two lysines, two tyrosines, five asparagines, two glutamines, and 11 waters together provide all of the donors for the 10 phosphodiester in contact with the protein.^{414,415}

Some proteins that recognize only DNA and not specific sequences within it use the regularity of its double-helical structure as a key. For example, each of the octameric complexes of histones around which the DNA winds in chromatin has a surface with a repeating pattern that matches the helical repeat of DNA.⁴¹⁶

Proteins that are required to associate with **specific sequences in DNA**, rather than simply DNA in general, recognize the patterns of donors and acceptors for hydrogen bonds in its grooves in addition to providing donors and acceptors of hydrogen bonds for the phosphodiester backbone and its associated water. The **major groove** in the DNA is the main key used by a protein in recognizing a specific sequence of base pairs. It is in the major groove that the patterns of donors and acceptors projected outward by the pairs of bases (**6–13**, **6–14**, **6–15**, and **6–16**) are the most legible. The protein usually inserts one or two of its segments of polypeptide into the major groove. Often it is an **α helix**. For example, the c-Jun subunit of transcription factor AP-1 (Figure 6–46)⁴¹⁷ and the ETS-domain protein Elk-1⁴¹⁸ insert α helices into the major groove. Such an α helix can participate in hydrogen bonds with donors and acceptors from as many as five⁴¹⁹ or six base pairs⁴¹⁸ in the major groove as well as donors and acceptors on riboses and phosphoryl oxygen from additional base pairs. In the case of the c-Jun subunit of transcription factor AP-1, the α helix inserted into the major groove is the splayed end of one strand of an α -helical coiled coil (Figure 6–29). In each of these complexes the α helix runs along the groove.

Other proteins, such as the *met* repressor from *E. coli*,⁴²⁰ the replication terminator of *E. coli*,⁴²¹ and the *arc* repressor from *E. coli*,⁴²² insert two strands of **β structure** running parallel to the major groove. The α helix or strands of β structure inserted into the major groove provide donors and acceptors to the acceptors and donors projected into the groove by the pairs of bases.

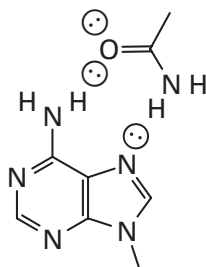
The **hydrogen bonds** formed during the reading of these patterns (Figure 6–47)⁴²³ are as varied as one might expect. A common pair is the double hydrogen bond between a guanine and an arginine (Figures 6–46 and 6–47),^{417,424–428} either at the two η nitrogens of the arginine (Figure 6–46)



6-17

or at one of its η nitrogens and its ϵ nitrogen (Figure 6-47),^{418,423} Arginine also can span two bases, offering a donor to an acceptor on each. Lysine, with its three donors, is often found occupying a single acceptor on a base or spanning two bases.^{429,430} In fact, it is in providing donors to acceptors for the neutral bases, rather than for the negatively charged phosphoryl oxygens, that arginine and lysine seem to be most frequently employed.

Another common hydrogen bond is that between a glutamine or an asparagine and adenine (Figure 6-45)^{414,422,431}



6-18

but often glutamine or asparagine (Figure 6-46)^{417,421,429} or even aspartate (Figure 6-47) bridges a donor on one base and an acceptor on its neighbor. Other donors and acceptors are commonly provided by histidine, serine, and threonine. Aspartate⁴³² can provide an acceptor; tyrosine,⁴²⁶ a donor and an acceptor; tryptophan,⁴³³ a donor; and cysteine,⁴²⁹ a donor and an acceptor. The nitrogen-hydrogen of a cytosine located in the major groove (6-14) makes a hydrogen bond with the π system of a tryptophan in the complex between transcription factor Rob and its cognate DNA.⁴³⁴

Many of the amino acids providing donors and acceptors to the bases in the major groove are themselves **pinned by hydrogen bonds** at one or more of their other donors and acceptors to other amino acids in the protein (Figure 6-47), and some of these other donors and acceptors can make hydrogen bonds to phosphoryl oxygens from the backbone of the nucleic acid to buttress the hydrogen bonds in the center of the groove.^{414,435}

Although there are a few examples in which almost every donor and acceptor in the major groove forms a direct hydrogen bond to an acceptor or donor on the pro-

tein (Figure 6-47),⁴²³ usually a significant fraction of the donors and acceptors from the amino acids on the protein form hydrogen bonds with **waters** that were present at fixed positions in the major groove before the protein bound to it and were subsequently incorporated into the complex.⁴⁰⁵ These waters then bridge donors and acceptors on the protein and donors and acceptors on the DNA. They are as much a key for recognition of the DNA by the protein as the bases themselves. One dramatic indication of this fact is that in crystallographic molecular models of

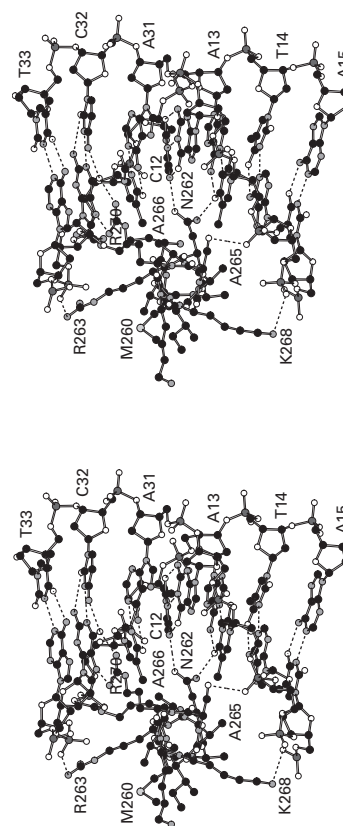
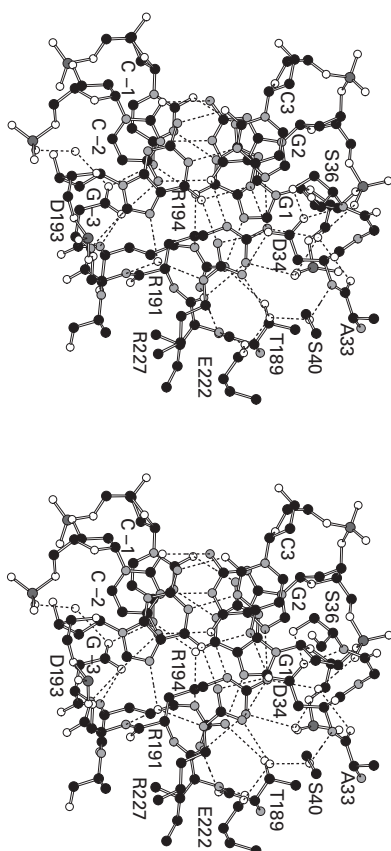


Figure 6-46: α Helix of a protein running along the major groove of a segment of double-helical DNA and reading the donors and acceptors of its bases.⁴¹⁷ The protein in the crystals used to gather the data set for the crystallographic molecular model (Bragg spacing ≥ 0.3 nm) was only the bZIP region (Glutamate 256 to Methionine 313) of the human transcription factor AP-1 expressed in *E. coli*, and the portion displayed in the figure is only the α helix from Arginine 259 to Lysine 279 that fills the major groove of the DNA. Only the portion of the DNA containing the sequence d(AGTCATA) and its complement is displayed. Note the hydrogen bonds between Lysine 268 and Arginines 259 and 263 and phosphoryl oxygens. Four base pairs are recognized as well as phosphodiester on one base pair above and two base pairs below those four. This drawing was produced with MolScript.⁵⁷³

Figure 6-47: Hydrogen bonds formed in the major groove to all of the donors and acceptors on three consecutive base pairs.⁴²³ The crystallographic molecular model is for the complex between *NgbMV* type II site-specific deoxyribonuclease from *Neisseria gonorrhoeae* and the double helix of the self-complementary DNA d(TGGCCGGCGG). Only the portion of the protein—Alanine 33 to Serine 36, Serine 40, Threonine 189 to Alanine 195, Glutamate 222, and Arginine 227—that recognizes specific bases in the DNA and the portion of the DNA that it recognizes, d(GGC) and its complement d(GCC), are drawn. The latter segment of the DNA has been cleaved between its G and the next C by the site-specific deoxyribonuclease in the crystal. All hydrogen bonds are included in the drawing. Three arginines form hydrogen bonds with the three guanines (6-17), and Aspartate 193 spans two cytosines. Note that the carboxylates of Aspartate 193 and Aspartate 34 are both solvated by four nitrogen-hydrogens, three on one oxygen and one on the other of each. The open circles are the locations for oxygens of five molecules of water, one of which screens the carboxylate of Aspartate 193 from an adjacent phosphoryl group. This drawing was produced with MolScript.⁵⁷³



closely related but distinct complexes between related molecules of DNA and related molecules of protein, many of the waters occupy the same locations.⁴³⁶

Waters bridging donors and acceptors on the DNA and donors and acceptors on the protein are common features of the extensive network of hydrogen bonds found at the interface between protein and DNA in the major

groove (Figure 6-48).^{415,425,435} On average, within a crystallographic molecular model between a protein and a molecule of double-helical DNA there are 10 molecules of water that form hydrogen bonds with “both the protein and the DNA simultaneously and thus mediate recognition directly.”⁴³⁷ On average, six of these 10 sit between two acceptors, none between two donors, and four between an acceptor and a donor too distant to form a direct hydrogen bond. These 10 include molecules of water between donors and acceptors on the protein and acceptors on phosphoryl oxygens as well as donors and acceptors on the bases themselves. Many of the waters in such complexes reside between two functional groups with the same charge number and serve to screen the electrostatic repulsion (see for example Aspartate 193 in Figure 6-47).

In Figure 6-47, every donor and acceptor directed into the major groove by the three base pairs is occupied by an acceptor or donor from the protein with the exception of one, which is occupied by a molecule of water. In Figure 6-48, only five of the donors and acceptors directed into the major groove by the three base pairs are occupied, all five by molecules of water bridging DNA and protein. These are the two extremes of a continuously occupied spectrum of hydrogen bonding in the major groove.

The **methyl groups of the thymines** also project into the major groove (Figure 3-9) and are used as keys for recognition. In the crystallographic molecular model of a complex between DNA and a protein that recognizes a particular sequence, the methyl groups on thymine are provided hydrophobic contacts by the protein. For example, the propyl group of a valine,⁴³⁸ the phenyl group of a phenylalanine,⁴²⁹ the butyl group of an isoleucine,⁴³¹ the methylenes of an arginine,⁴²¹ the methyl group of an alanine,⁴¹⁸ or the methyl group of an alanine and the methylene of a serine (Figure 6-46) can cradle the methyl group of a thymine in the major groove.

Although it is narrower and more difficult to enter (Figure 3-9), the **minor groove** is also exploited by proteins recognizing specific sequences in the DNA. Usually, because it is so narrow, only a single loop of polypeptide is inserted into it,^{421,439,440} and a protein that inserts a segment of its polypeptide into the minor groove will also insert a sizeable segment into the adjacent major groove.^{441,442} The less formal arrangement in the minor groove permits even the amido nitrogen-hydrogens of the backbone to occupy acceptors projecting from bases.⁴⁴⁰ Otherwise, the donors and acceptors from the protein that occupy acceptors and donors in the minor groove are the same as those in the major groove. **Lysines**⁴²¹ and particularly **arginines**^{421,439,443,444} are common because their side chains are long, thin, and flexible, but even a short negatively charged aspartate can provide acceptors for donors in the minor groove.^{404,440}

It was originally believed, before crystallographic molecular models of these complexes became available,

that the problem of recognizing a specific sequence would simply require reading enough of the pattern of donors and acceptors in the major groove and minor groove and methyl groups from thymine in the major groove to make an unequivocal identification of the sequence. Although there are a few instances in which side chains on the protein are able to form hydrogen bonds to every donor and acceptor in the major groove (Figure 6-47),⁴²³ and usually many of these features are recognized by the protein either directly or through intervening molecules of water, it is often the case that fewer are recognized than would be necessary to make a positive identification.⁴²⁵ Consequently other strategies must be used to make a positive identification.

The most obvious of these is the use of **packing** to recognize shape, just as in the center of a molecule of protein the dense packing of the side chains of the amino acids is used to position the secondary structures. For the crystallographic molecular models of complexes between a protein and its complementary DNA, calculations of atomic volumes "performed in the presence and absence of water molecules, showed that protein atoms buried at the interface with DNA are on average as closely packed as in the protein interior. Water molecules contribute to the close packing, thereby mediating shape complementarity."⁴⁰⁶ This close packing means that the shape of the surface of the protein fits tightly into the shape of the surface of the DNA and its water, particularly in the major groove.⁴⁰⁵ As the **shape of the surface of the DNA** and water in the major groove represents its sequence, it is the shape of the surface of the protein as much as anything else that reads the sequence of the DNA.

Much of this complementarity in shapes is the networks of hydrogen bonds (Figure 6-47), but the hydrophobic hydrogen-carbon bonds of the protein also contribute to its complementary shape. In fact, there are hydrophobic side chains such as phenylalanines, leucines, and valines^{445,446} that are found in the interface between the protein and the major groove, the functions of which are not just to cradle the methyl groups of the thymines but to form a mold for the DNA. When one such side chain, Leucine 22 in the interface between DNA and transcription factor AREA, was mutated to a valine, the specificity of the transcription factor changed dramatically as it recognized a different set of sequences in the DNA.⁴⁴⁷

The structure of the protein looking for a particular sequence of DNA also recognizes **variation in the overall shape** of a segment of DNA produced by the particular sequence of bases it contains.^{405,425} For example, certain sequences of bases cause the minor groove to become narrower, and this feature of the DNA is recognized by the protein, often by the insertion of an arginine into the minor groove to gauge its width.^{443,448} Overwinding or underwinding of the DNA by particular sequences is recognized,^{420,449} as well as intrinsic curva-

ture of the double helix.⁴²⁵ This strategy of recognizing preexisting, sequence-dependent peculiarities in the structure of the DNA, however, is difficult to separate experimentally from a strategy of recognizing sequence-dependent differences in the resistance of the DNA to distortion by the protein because only rarely⁴⁰⁵ is the preexisting structure of the segment of DNA found in the crystallographic molecular model of the complex known.

When proteins bind specifically to a segment of

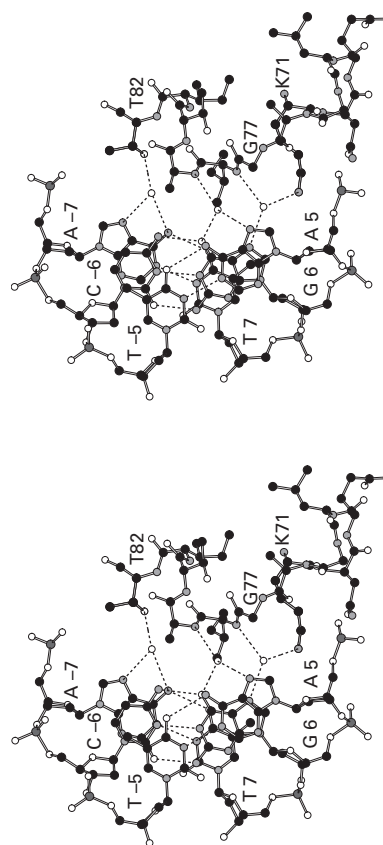
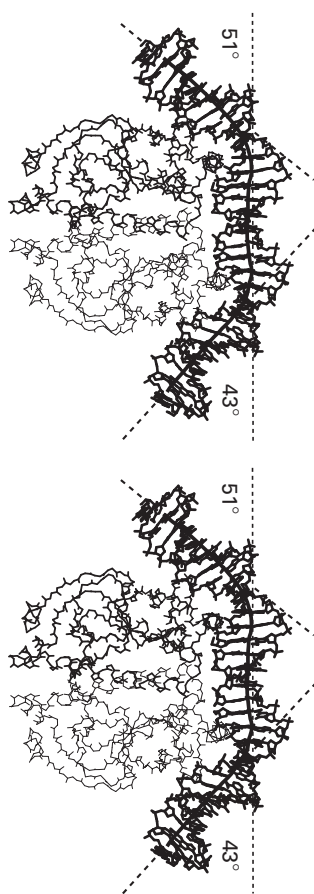


Figure 6-48: Incorporation of molecules of water into the interface between protein and DNA.⁴²⁵ The portion of the interface between protein and double-helical DNA displayed in the figure is from the crystallographic molecular model (Bragg spacing ≥ 0.19 nm) of the complex between the *trp* repressor from *E. coli* and a segment of double-stranded DNA with the self-complementary sequence d(TGTACTAGTAACTAGTAC). The portion displayed contains the second d(AGT) and its complement and Lysine 71 to Threonine 82 from the *trp* repressor. The unattached oxygens (open circles) are locations for molecules of water within the major groove of the DNA that have been incorporated into the complex. Only hydrogen bonds between the bases and hydrogen bonds to the molecules of water are drawn. This drawing was produced with MolScript.⁵⁷³

Figure 6-49: Abrupt kinks produced in double-stranded DNA by the association of a protein.⁴⁵³ The complete crystallographic molecular model (Bragg spacing ≥ 0.30 nm) of the complex between the catabolite gene activator protein from *E. coli* and a segment of double-helical DNA 30 base pairs in length containing the sequences recognized by the protein is displayed. The protein contains two identical folded polypeptides, each forming one of its subunits, and is represented by the two backbones of those two subunits, one drawn with thicker lines than the other. The DNA is drawn with even thicker lines. The two subunits sit side by side joined by a long coiled coil of two α helices, one from each of them. Upon association, each of the two identical subunits of the protein produces an abrupt kink in the DNA. The irregular solid line drawn through the center of the molecule of DNA connects the centers of the consecutive pairs of bases. The dotted lines are drawn arbitrarily to indicate the mean axis in each of the three segments of DNA bounded by the two kinks. The angles are those made by the three dotted lines with each other. This drawing was produced with MolScript.⁵⁷³



DNA with a particular sequence, they often distort its structure. One of the most obvious examples of this fact is the complex between the purine repressor from *E. coli* and its cognate double-helical DNA.⁴⁵⁰ This protein thrusts two of its leucine side chains that are positioned within the minor groove of the DNA into the space between the guanine–cytosine pair and the cytosine–guanine pair at the center of the sequence recognized by the protein. This wedge cants these two pairs of bases and creates an abrupt 45° bend in the DNA centered on this **distortion**. In the complex between TATA-box-binding protein and its cognate DNA, two pairs of phenylalanines from β strands running across the top of the minor groove insert between two pairs of bases, the thymine–adenine and the adenine–thymine at positions 1 and 2 and the thymine–adenine and guanine–cytosine at positions 7 and 8 of the sequence recognized by the protein to cause an overall bend in the DNA of $65\text{--}80^\circ$.^{451,452}

Many proteins **bend the DNA** when they bind to it. Sometimes they form abrupt kinks (Figure 6-49),⁴⁵³ but often the bend induced has a gradual curvature following the curvature of the surface of the globular protein.^{429,454} In the complex with the repressor from bacteriophage 434, the DNA is bent in an irregular arc with a radius of curvature of 6.5 nm closely cleaving to the surface of the globular protein (Figure 6-50).^{449,454} In human DNA-(apurinic or apyrimidinic site) lyase, a rigid, cationic, preformed surface acts as a template upon which the DNA that the protein recognized is bent.⁴⁵⁵ Sometimes two successive segments of DNA that contain sequences recognized by two different proteins, which in turn form a complex with each other, bend smoothly around that pair of proteins.⁴⁵⁶ The ultimate extrapolation of such complexes that smoothly bend DNA is that found in chromatin between the segment of DNA 150 base pairs in length and the complex of eight histones around which the DNA wraps in a smooth superhelix with a radius of curvature of 4.3 nm and two almost complete turns.⁴⁵⁷

Often the bending induced by the protein serves a purpose. The superhelices of double-helical DNA in the complexes with the octamers of histones that constitute chromatin store the DNA compactly. The separate complexes between the two adjacent sites on the DNA and the homeodomain protein MAT α 2 and the MADS-box protein MCM1 bring the two proteins together so they can interact.⁴⁵⁸ It has been proposed, however, that the bending or distortion of the DNA in other instances contributes to the recognition of its particular sequence by the protein.

It is believed that particular sequences of nucleotides are more prone to distortion than others and that the ease with which a double helix of DNA distorts can be recognized by the protein as it bends the DNA during the formation of the complex. The standard free energy required to distort a segment of DNA from the linear B form should be positive and unfavorable. If the standard free energy for

a particular distortion of a particular sequence is significantly less positive than the standard free energies for same distortion of other sequences, then when a complex is formed that requires this distortion, the free energy of formation of the complex will be more negative when the easily distorted sequence is bound.

There are experimental observations indicating that a base pair between adenine and thymine is more flexible than one between guanine and cytosine and that this susceptibility to distortion can be used to recognize this base pair.⁴⁵⁹ In the complex between the repressor protein CI of bacteriophage 434 and its complementary DNA (Figure 6-50), the central six pairs of bases are not part of the sequences on either side that are indispensable for the recognition, but their sequence also determines the magnitude of the dissociation constant between protein and DNA.⁴⁴⁹ When they are **adenine-thymine pairs** rather than guanine-cytosine pairs, the free energy of formation of the complex is more negative. In the crystallographic molecular model of the complex, this region of the DNA is significantly distorted by the protein in a manner that seems as though it should be more readily tolerated by adenine-thymine pairs than it would be by guanine-cytosine pairs.^{449,454} The DNA mismatch repair protein MutS from *E. coli* seems to take advantage of the instability of double-helical DNA at a position where the bases are mismatched to introduce a kink at such a location.^{460,461} The uncomplexed segment of DNA recognized by the *trp* repressor of *E. coli* is already distorted in the direction in which it will be distorted by the complex but is further distorted when the complex forms. It is thought that the partial distortion of the uncomplexed DNA demonstrates the susceptibility of this sequence to the ultimate distortion to which it will be submitted.⁴⁰⁵ It is also thought that the decrease in free energy of formation observed when the N6 anilino group of an adenine in the segment of DNA recognized by *EcoRI* site-specific deoxyribonuclease is deleted results from an increase in the ease with which this segment can be distorted by the protein.⁴⁶²

Just as the DNA is often distorted upon forming a complex with a protein, the protein often has a different conformation in the complex than in solution. Such **conformational changes** are often significant. For example, the carboxy-terminal α helix of *BamHI* site-specific deoxyribonuclease unwinds, and the disordered polypeptide that results turns almost 180° to enter the minor groove of the DNA.⁴⁴⁰ Usually, however, the conformational change of a protein on binding to DNA is the establishment of structure from a disordered segment of the polypeptide⁴⁴⁰ or the tying down of flexible segments of the protein⁴⁶³ by their association with the DNA.

There are several proteins, such as topoisomerase I,⁴⁶⁴ the human Ku heterodimer,⁴⁶⁵ and protein gp 45 from bacteriophage T4,⁴⁶⁶ that have a hole passing through them large enough to contain a double helix of DNA. These are proteins that are required to sur-

round a double-helical molecule of DNA to perform their functions, and they recognize the double helix in part by its fit to the hole. In the empty state, the **ring of protein** around the hole is continuous but always contains at least one interface through which the polypeptide does not pass. It is at such an interface that the ring of protein splits apart to allow the DNA to enter the hole and then closes back around it.^{467,468}

There are also proteins that bind to **single-stranded DNA**. Unlike double-helical DNA, the structure of single-stranded DNA is undefined, but when it is bound by one

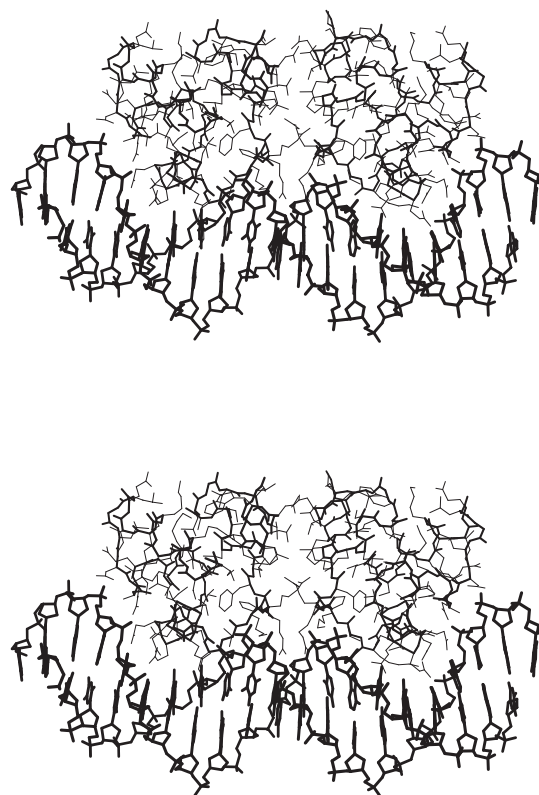
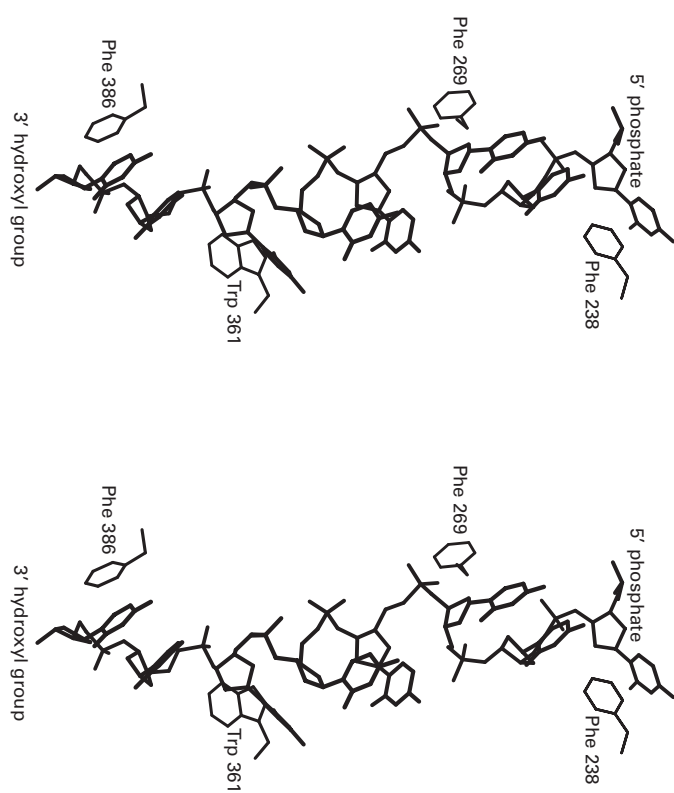


Figure 6-50: Gradual bend produced in DNA as it cleaves to the surface of a globular protein.⁴⁵⁴ The complete crystallographic molecular model (Bragg spacing ≥ 0.25 nm) of the complex between the DNA-binding portion (Serine 1 to Arginine 69) of the repressor protein CI from bacteriophage 434 and a segment of double-helical DNA 20 base pairs in length containing the sequences recognized by the repressor is displayed. The protein is formed from two identical folded polypeptides, each of which participates in an extensive interface with the double-helical DNA that produces the adhesion of the DNA to the surface of the protein. Skeletal structures of the two subunits (above and below each other) are presented with both the backbone (thick lines) and the side chains (thin lines) drawn. Note the arginines inserted into the minor groove and the α helices inserted into the major groove. This drawing was produced with MolScript.³⁷³

Figure 6-51: Stacking of bases in single-stranded DNA between each other and against aromatic amino acids in the protein.⁴⁶⁹ The crystallographic molecular model (Bragg spacing ≥ 0.24 nm) is that of a complex between the DNA-binding portion (Lysine 183 to Glutamine 420) of the DNA-binding subunit (616 aa) of human replication protein A and octadeoxycytidine (drawn with thicker lines). The entire molecule of single-stranded DNA in the complex is displayed as well as the side chains of the four aromatic amino acids (drawn with narrower line segments) participating in the stacks, numbered by their positions in the sequence of the protein. This drawing was produced with MolScript.⁵⁷³



of these proteins, it assumes a structure dictated by that protein. In this sense, the binding of single-stranded DNA is no different from the binding of any other large flexible ligand by a molecule of protein. Replication protein A is a protein that recognizes segments of single-stranded deoxyribonucleic acid with many different sequences. In the crystallographic molecular model of the complex between single-stranded octadeoxycytidine and human replication protein A, the octanucleotide in the complex is stretched into a linear form that is almost fully extended.⁴⁶⁹ The protein forms a number of hydrogen bonds with the phosphoryl oxygens of the backbone similar to those in complexes between double-helical

DNA and proteins that recognize DNA nonspecifically. The hydrogen bonds to the donors and acceptors of the individual bases in the particular complex that was crystallized, however, are thought to arise only from the requirement that these must be occupied somehow to avoid losing hydrogen bonds from the solution upon association of the DNA with the protein. There are sufficient acceptors and donors of sufficient flexibility on the protein in these locations to satisfy any sequence of bases, as replication protein A is required to do.

The novel feature of this complex is the interactions between π systems of amino acid side chains on the protein and the π systems of the bases that are no longer enclosed within the core of a double helix. Phenylalanines 238 and 269 sandwich one stacked pair of cytosines, and Tryptophan 361 and Phenylalanine 386 sandwich another (Figure 6-51).⁴⁶⁹ Both Phenylalanine 238 and Tryptophan 361 have their π aromatic systems normal to those of the cytosines. An almost identical sandwich occurs between a phenylalanine and a tryptophan in single-stranded DNA binding protein from *E. coli* and a stacked pair of cytosines.⁴⁷⁰ Similar sandwiches occur between tyrosines in the telomere end-binding protein of *Oxytricha nova* and pairs of stacked guanines.⁴⁷¹ A different arrangement is found in the same complex between single-stranded binding protein from *E. coli* and single-stranded DNA, in which another tryptophan is surrounded by a cluster of four cytosines.^{470,472-477}

Just as the paradigm of the structure of DNA is the double helix (Figure 3-9), the paradigm of the structure of RNA is a molecule of **transfer RNA** (Figure 6-52).⁴⁷²⁻⁴⁷⁷ There are several novel structural features of RNA that are not encountered in DNA.

Although there are two double helices in the crystallographic molecular model in Figure 6-52, a horizontal one at the top of the structure containing 13 pairs of bases and a vertical one at the bottom containing 5 pairs of bases, neither is formed from two separate strands of RNA because the entire molecule is formed from only one strand of RNA.* The vertical double helix of 5 pairs of bases is formed from the two uninterrupted lines of a **double-helical hairpin of RNA**. At the bottom of the hairpin there is a **loop** in this RNA of 7 bases (2'-O-Methylcytosine 32† to Adenine 38). In transfer RNA, this loop displays the anticodon, but in most double-helical hairpins, this loop has only a structural function. One of the most common sequences in such a loop in which the chain reverses is UUCG, which produces a structure referred to as a **tetraloop** containing five hydrogen bonds that efficiently change the direction of the RNA.⁴⁸⁰ A

* You should trace the polynucleotide backbone (O-P-O-C-furanose-O-P-O-C-furanose-O-...) through the whole molecule of transfer RNA.

† Just as some proteins are posttranslationally modified on their side chains, all transfer RNAs are posttranscriptionally modified on many of their bases.^{478,479}

double-helical hairpin and its loop is one of the basic structures formed by RNA.

Double-helical hairpins of RNA can be as long as 50 or more pairs of bases, but they are usually interrupted one or more times with **bulges** at which there is a mismatch of the bases on the two strands that face each other. The mismatch causes an interruption in the double helix. A bulge can be as small as one or two extra unmatched bases that protrude out of the double helix on one of the strands while the strand on the other side contains no mismatched base. Uracil 59 and Cytosine 60, found between the 12th (Guanine 53 and Cytosine 61) and the 13th (5-Methyluracil 54 and 1-Methyladenine 58) pairs of bases of the horizontal double-helical hairpin in Figure 6-52, form such a small bulge immediately before the loop of three bases (Pseudouracil 55 to Guanine 57) following the 13th base pair. Bulges can also occur across from each other on both strands of double-helical RNA. The number of bases on one strand of such a bulge can be the same as or different from the number of bases on the other strand, and the two strands are usually independent of each other until they rejoin in the double helix at the other end of the bulge. An inconsequential bulge of one mismatched pair of bases occurs at Guanine 4 and Uracil 69 in the horizontal double-helical hairpin in Figure 6-52.

The most interesting bulges, however, are the larger ones. For example, the entire lower portion (Uracil 8 to Cytosine 48) of the transfer RNA in Figure 6-52 is a bulge out of the horizontal double-helical hairpin. It protrudes between the seventh pair (Uracil 7 and Adenine 66) and the eighth pair (5-Methylcytosine 49 and Guanine 65) of bases while the opposite strand of the horizontal double helix does not skip a beat. The returning strand of this bulge picks up the beat at the eighth pair of bases that was dropped by the departing strand at the seventh pair.

In the central region of a molecule of transfer RNA (Uracil 8 to N^2,N^2 -Dimethylguanine 26 and Adenine 44 to Cytosine 48 in Figure 6-52), the polynucleotide strand **meanders randomly** through the region, forming a complex **tertiary structure** that rigidifies the molecule and holds the two double helices perpendicular to each other. In this region, there are numerous intramolecular hydrogen bonds orienting the strand of RNA as it passes through. It is in this central region that the RNA becomes almost reminiscent of a molecule of protein.

There is, however, one unique characteristic of this central region in which the structure is unmistakably nucleic acid. Even though the structure has become random meander, most of the bases are still **stacked** one upon the other as they are in the double-helical regions. Even the last two bases to the right, beyond the end of the horizontal double-helical hairpin, are stacked upon themselves and the last base of the double helix.

These novel features—double-helical hairpins, bulges, rigid random meander, and stacking of nonhelical bases—are regularly found in molecules of RNA.

Unlike DNA, which is rarely unassociated with protein, there are species of RNA such as transfer RNA and messenger RNA that spend at least a part of their lives free in solution. Unlike DNA, in which the proteins with which it is associated change dramatically as it is transferred from storage, to transcription, to replication, and to recombination; complexes between RNA and protein, such as ribosomes and the small nuclear ribonucleoprotein particles that form spliceosomes, often have fixed structures that remain essentially unchanged during their lifetimes. Such **ribonucleoproteins** are biologically distinguished from proteins only by the fact that they almost always operate on other molecules of RNA.

Many of the atomic details of the association

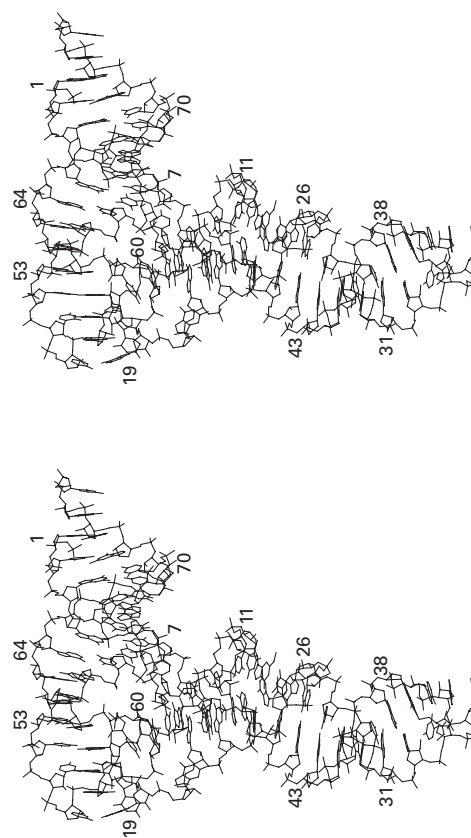


Figure 6-52: Crystallographic molecular model (Bragg spacing ≥ 0.19 nm) of phenylalanyl transfer RNA from *S. cerevisiae*.^{472,477} The structure is formed from a single, folded molecule of RNA 76 bases long. The last two bases, which are fully extended to the right of the upper double helix, have been omitted. Fourteen of the bases in this particular transfer RNA have been posttranslationally modified by reduction, methylation, and isomerization,^{478,479} but the modifications are difficult to spot at this magnification. This drawing was produced with MolScript.⁵⁷³

between proteins and RNA are indistinguishable from those for the association of proteins and DNA. There are **hydrogen bonds** formed to the acceptors on the phosphoryl oxygens and the donors and acceptors on the bases, and molecules of water are participants in these networks of hydrogen bonds.^{481–485} The main difference is that many of the bases are not in pairs, and in those bases all of their donors and acceptors of hydrogen bonds are available for recognition by acceptors and donors on both the side chains and the polypeptide backbone of the protein. **Hydrophobic contacts** are also made, often with the exposed π systems of the bases, which are accessible in the regions of the RNA that are not double-helical.⁴⁸⁶ There are even instances of α helices lying in grooves of the RNA.⁴⁸⁶

When a large globular molecule of RNA such as a transfer RNA is bound in a transient complex by a protein such as an aminoacyl-tRNA ligase, the complex is reminiscent of one between a protein and double-helical DNA in that the surface of the protein and the surface of the RNA in the interface fit together as cast in mold.^{487,488} In the more **permanent complexes**, however, the RNA and protein are more intimately intertwined. For example, in the U1 small nuclear ribonucleoprotein particle, a representative component of the spliceosome, 10 separate proteins form a complex with one molecule of RNA in which some individual proteins and multimeric complexes of other proteins associate with different segments of the RNA.⁴⁸⁹ The RNA contains four double-helical hairpins in an open, extended structure, and the proteins bind to the ends of individual hairpins⁴⁹⁰ or to the double-helical portions emerging from the center of the molecule of RNA.⁴⁸⁹ In such a small nuclear ribonucleoprotein particle, only 20% of the mass is RNA, and the RNA is a loose scaffold that ties together the proteins, which are responsible for most of the structure of the particle.

The ultimate complex between protein and RNA is a **ribosome**. A ribosome is a ribonucleoprotein that is by mass about two-thirds RNA and one-third protein. It contains three different molecules of RNA, about 3000, 1500, and 120 bases long, respectively, and about 50 different molecules of protein, totalling about 7200 aa, the largest about 350 aa long, the smallest about 50 aa long.* There are two different subunits comprising a ribosome. The 50S subunit contains the largest and smallest molecules of RNA and 30 of the proteins; the 30S subunit contains the RNA 1500 bases long and 20 of the proteins. Yonath and her associates have obtained crystals of the 50S subunit from *Haloarcula marismortui*^{491,492} and the 30S subunit from *Thermus thermophilus*,⁴⁹³ and Yusupov and his associates have obtained crystals of the intact ribosome from *T. thermophilus*^{494–496} and the 30S subunit from *T. thermophilus*.^{495,497} All of these crystals have

proven satisfactory for crystallographic studies, and crystallographic molecular models have been obtained from data sets gathered from them.^{498–503} These crystallographic molecular models provide the atomic details of the structure of a ribosome as well as insight into its ability to translate messenger RNA into protein.⁵⁰⁴

As the distribution of mass suggests, the basic structural element of a ribosome is the RNA. The 4600 bases of the three molecules of RNA form a **globular structure** with which the proteins associate. The RNA, although it is 60 times larger, is reminiscent of a molecule of tRNA and displays all of its characteristic features: double-helical hairpins, loops, bulges, and random meander. One of the few novel features is that many of the hairpins are so long that they form smoothly curved double helices that wrap around other curved double-helical hairpins in structures reminiscent of coiled coils.

For the most part, the various **proteins** are found associated with the outer surface of the much larger globular RNA. Many of the proteins are entirely globular, but some of them have long segments of polypeptide, either interior loops or segments at their carboxy-terminal or amino-terminal ends, emerging from their globular portions and meandering widely through the RNA. Some of the globular portions of these proteins sit upon the surface of the RNA, others are buried within it, but all are subordinate to it both structurally and functionally.

The RNA is responsible for the ability of a ribosome to translate messenger RNA into protein. The RNA of the 30S subunit aligns the codon of the messenger RNA with the anticodon of the transfer RNA,⁵⁰⁵ and the RNA of the 50S subunit appears to catalyze⁵⁰⁶ the formation of the peptide bond from aminoacyl transfer RNA and the peptidyl tRNA.⁵⁰⁷

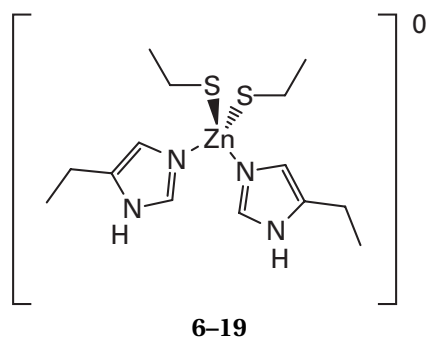
There is a set of small modules of protein known as **zinc fingers** that recognize specific sequences of double-helical DNA mainly by forming bonds within the major groove (Figure 6–53).⁵⁰⁸ Each of the many different zinc fingers is capable of recognizing the specific sequence of a segment of DNA 3–4 bases in length, and each recognizes a different sequence.^{509,510} Sets of these **modules** are strung together within the same polypeptide and together recognize longer specific sequences in DNA. Four of the five zinc fingers in the zinc finger protein GLI1 from *Homo sapiens* (Figure 6–53) together recognize a segment of DNA 14 bp long by binding consecutively to segments 3–4 bp in length.⁵⁰⁸

Transcription factor IIIA has nine successive zinc fingers sequentially joined together within a segment of its overall sequence.^{511,512} These nine zinc fingers together associate with a segment of DNA 55 bp long but directly recognize sequences only in a segment 11 bp long beginning 8 bp from one end of the overall segment, a segment 10 bp long beginning 9 bp from the other end, and a segment 3 bp long in the middle. The first three zinc fingers each associate with overlapping sequences 4 bp long in the segment 10 bp long, the fifth zinc finger

* The uncertainty reflects both experimental ambiguity and differences among species.

associates with the sequence 3 bp long in the middle, and the last three zinc fingers associate with the segment 11 bp long.⁵¹³ Side chains from the various fingers associate with the phosphoribosyl backbone outside of the three segments the sequences of which are recognized. The fourth and the sixth zinc fingers do not enter the major groove and consequently do not recognize and bind to sequences of base pairs.

Each zinc finger is a segment of polypeptide about 30 amino acids long. Ordinarily a segment of polypeptide this short would be unable to form a specific structure because the small size would prevent the folded protein from removing a sufficient number of hydrogen-carbon bonds from contact with the water to provide enough of a hydrophobic effect to overcome the change in standard entropy required for folding.⁵¹⁴ The most common solution to this problem is that a small protein or small module of protein will contain several cysteines in its core, the cross-links of which provide sufficient rigidity to the structure to overcome this deficit in standard free energy. The zinc finger solves the problem in a similar way, but instead of using cysteines, it uses a Zn^{2+} cation that forms four **covalent bonds** with two cysteines and two histidines in the sequence of the module (Figure 6-53), cross-linking the four amino acids together:



Consequently, a zinc finger is an interesting example of a metalloprotein.

Suggested Reading

Shakke, Z., Guzikevich-Guerstein, G., Frolow, F., Rabinovich, D., Joachimiak, A., & Sigler, P.B. (1994) Determinants of repressor/operator recognition from the structure of the *trp* operator binding site, *Nature* 368, 469-473.

Ban, N., Nissen, P., Hansen, J., Moore, P.B., & Steitz, T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution, *Science* 289, 905-920.

Problem 6-12: What is the sequence of the segment of DNA in Figure 3-9?

Problem 6-13: List the hydrogen bonds between the amino acids of the zinc finger in Figure 6-53 and the bases in the DNA with which it is associated. Identify the amino acids and bases by their respective positions

in the sequences and the functional groups of each by their names and their numbers (2-9, 2-10, 2-11, and 2-12 and Figure 4-14).

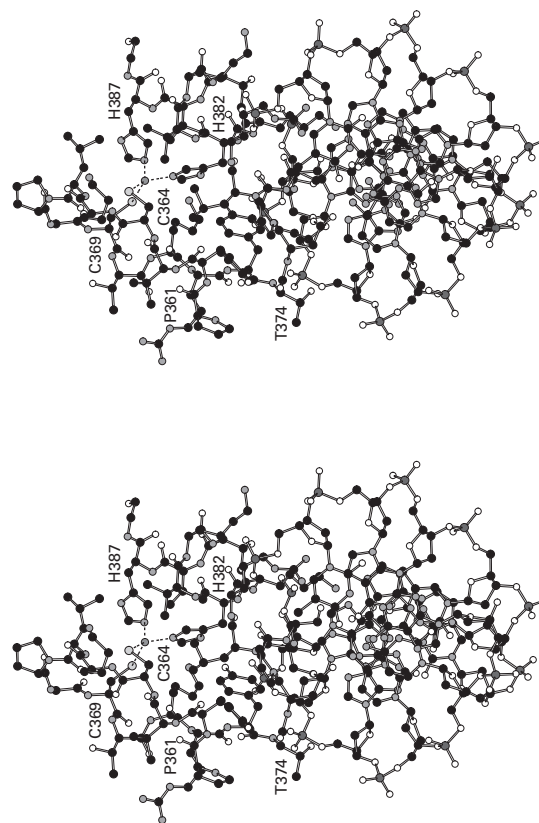
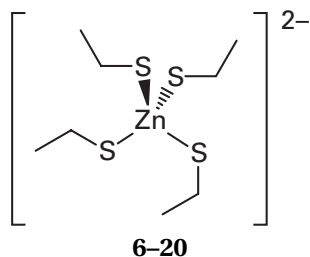


Figure 6-53: A zinc finger bound to the major groove of the double-helical DNA that it recognizes.⁵⁰⁸ The crystallographic molecular model (Bragg spacing ≥ 0.26 nm) is that of the complex between the five consecutive zinc fingers (Valine 232 to Alanine 391) of the human zinc finger protein GLI1 expressed as a separate protein and a segment of double-helical DNA 21 base pairs in length containing sequences recognized by those five fingers. Only the fifth zinc finger (Proline 361 to Glycine 388) and the four pairs of bases [d(GACC) paired with d(GGTC)] that it recognizes in addition to one pair of bases on each side (dG-dC and dA-dT, respectively) are included in the figure. The view is down the axis of the B conformation of the DNA. The DNA is in the bottom of the figure and the zinc finger is in the top. The polypeptide is numbered according to the complete amino acid sequence of the zinc finger protein GLI1. Side chains from the protein read the donors and acceptors projecting into the major groove of the DNA, but none of the responsible hydrogen bonds has been drawn. In the protein crystallized, the Zn^{2+} had been replaced by Co^{2+} . The Co^{2+} is the gray sphere near the top of the finger forming four tetrahedrally arranged covalent bonds (dashed lines) with Cysteines 364 and 369 and Histidines 382 and 387. Although its van der Waals radius is about 10% shorter than that of Co^{2+} and it is a somewhat softer acid, Zn^{2+} would have formed covalent bonds with the same four ligands to produce the identical tetrahedral geometry. This drawing was produced with MolScript.⁵⁷³

Metalloproteins

As does a zinc finger, many proteins incorporate one or more **metallic cations** into their structure. Aside from cations such as lithium, sodium, potassium, rubidium, magnesium, and calcium that are dissolved in the cytoplasm or the extracellular solution and bind adventitiously and randomly over the surface of a protein, there is a set of metallic cations that participate as specific and necessary structural and functional elements of **metalloproteins**. These are the cations of sodium, potassium, magnesium, calcium, vanadium, manganese, iron, cobalt, nickel, copper, zinc, molybdenum, and tungsten. The nontransition metals, sodium, potassium, magnesium, and calcium, and zinc, a transition metal inactive in oxidation–reduction, occur exclusively in their most common oxidation states: Na^+ , K^+ , Mg^{2+} , Ca^{2+} , and Zn^{2+} , respectively. Because of the availability of two or more readily accessible oxidation levels, the other transition metals, for example, iron or copper, are often used as one-electron carriers, and in this role alternate between oxidation levels, for example, Fe^{2+} and Fe^{3+} or Cu^+ and Cu^{2+} . In other situations transition metals, such as the Ni^{2+} in urease or the Fe^{2+} in myoglobin (Figure 4–18), fill roles in the active sites of enzymes in which no changes in oxidation level are required and in fact are to be avoided.

Eventually, the role of a metallic cation in **maintaining the structure** of a protein must be distinguished from its role as a **catalytic functional group** in its active site. Aspartate carbamoyltransferase from *E. coli* is a protein constructed from two different folded polypeptides, the regulatory subunit ($n_{\text{aa}} = 152$) and the catalytic subunit ($n_{\text{aa}} = 310$). In the crystallographic molecular model of aspartate carbamoyltransferase, a Zn^{2+} is tetrahedrally coordinated to the four sulfurs of Cysteines 109, 114, 137, and 140 in the regulatory subunit.⁵¹⁵ This Zn^{2+} forms a tetrahedral, covalent complex with the structure



that resembles closely structures observed in polynuclear clusters that form between Zn^{2+} and small mercaptans. When the zinc is displaced from the thiols by organic mercurials, the two subunits separate from each other but can be reassociated^{516,517} by reincorporating the Zn^{2+} . In the crystallographic molecular model, the Zn^{2+} is located adjacent to the boundary between the regulatory subunit and its neighboring catalytic subunit but distant from both the active site of the enzyme located on

the catalytic subunit and sites for binding ligands on the regulatory subunits. Therefore, the role of the Zn^{2+} is entirely structural. Its complexation with the thiols creates and stabilizes the proper structure at the surface of a regulatory subunit. Only when this stable structure is formed can the properly folded regulatory subunit associate with a complementary structure on the surface of a catalytic subunit, just as only when the proper structure of a zinc finger is formed by the binding of the Zn^{2+} can it associate with the proper site on DNA (Figure 6–53). A metallic cation fulfills such a structural role because the bonds it forms either covalently or ionically with lone pairs of electrons on bases within the protein are strong ones, especially when the protein itself assists in orienting the bases advantageously.

In the case of aspartate carbamoyltransferase, removal of the Zn^{2+} from the protein produces catalytic subunits with full enzymatic activity. In most instances, however, the removal of a metallic cation from a protein leads to loss of function, and separating the effect of the cation on the structure of a protein, which itself is responsible for that function, from a direct effect at an active site, in which a metallic cation is often a catalytic group, is difficult. For example, when mammalian liver arginase, which is formed from four identical folded polypeptides, is treated with the chelating agent *N,N,N',N'*-tetracarboxymethyl-1,2-diaminoethane (**5–1**), it loses all of its enzymatic activity.⁵¹⁸ At the same time, however, it dissociates into individual folded polypeptides. When Mn^{2+} is added to the inactive protein, enzymatic activity is regained, but the individual folded polypeptides reassociate. It was possible that the dissociation of the tetramer was responsible for the inactivation and that the Mn^{2+} required for activity was necessary to retain the proper structure of the protein rather than as a catalytic group in the active site. If, however, a crystallographic molecular model of the protein is available it is possible to determine, as was the case with both a zinc finger and aspartate carbamoyltransferase, whether or not the metal is distant from sites involved in the function of the protein and consequently performs purely a structural role. In the case of arginase, for example, it has been shown crystallographically that Mn^{2+} cations form a binuclear cluster within the active site intimately involved in the catalysis performed by the enzyme.⁵¹⁹

The metallic cations incorporated into proteins in aqueous solution, because they are themselves **Lewis acids**, are at all times surrounded by **Lewis bases**. The strongest Lewis bases present in biological fluids are the lone pairs of electrons on oxygens, nitrogens, and sulfurs and the chloride ion. The proton is also a Lewis acid, and in biological fluids every acidic proton is usually surrounded by lone pairs of electrons on oxygens, nitrogens, or sulfurs. The proton is so small, however, that it can accommodate directly only two Lewis bases at a time in one hydrogen bond. Because metallic cations have core electrons, they are larger than a proton and can accom-

moderate more Lewis bases simultaneously. The metallic cations incorporated into proteins are always surrounded by pairs of electrons from oxygens, nitrogens, sulfurs, or halides. The atoms providing the lone pairs of electrons surrounding a metallic cation are its **ligands**, and the number of ligands surrounding the cation is its **coordination number**. The complexes formed between metallic cations and proteins are tetra-coordinate, penta-coordinate, hexa-coordinate, hepta-coordinate, octa-coordinate, and nona-coordinate.

The preferences of a metallic cation for a particular type of lone pair of electrons are usually discussed in terms of the **hardness or softness** of the Lewis acid and the Lewis base.⁵²⁰ The rule is that hard acids prefer hard bases and soft acids prefer soft bases. For the divalent metal ions of importance to the structure of proteins, the series of hardness is $\text{Mg}^{2+} > \text{Ca}^{2+} > \text{Mn}^{2+} > \text{Fe}^{2+} > \text{Co}^{2+} > \text{Ni}^{2+} > \text{Cu}^{2+}, \text{Zn}^{2+}$. For the commonly encountered bases, the series of hardness is lone pairs on oxygen > lone pairs on chloride > lone pairs on nitrogen > lone pairs on sulfur. These rankings, for example, are consistent with the fact that calcium ion has a strong preference for oxygen ligands while zinc ion has a preference for thiol ligands. It also explains why the ligands on metallothionein, a protein responsible for chelating and thus removing from solution soft, toxic heavy metal cations, are entirely the thiols of cysteine side chains in the protein.

The Lewis bases surrounding a metallic cation in solution are attached to it by bonds the characteristics of which span the **spectrum between ionic and covalent**. The bonds between hard cationic Lewis acids and hard Lewis bases are usually ionic, and those between soft cationic Lewis acids and soft Lewis bases are usually covalent. The calcium dication is an example of a hard, purely ionic metallic cation. In biological solutions, its ligands are invariably oxygen atoms,⁵²¹ the hardest of bases, and those oxygens are held by ionic bonds. The zinc dication in the crystallographic molecular model of aspartate carbamoyltransferase (6–20) is a good example of a metallic cation participating in covalent bonds. In this arrangement, a soft metallic cation, Zn^{2+} , is bonding covalently to four soft bases, $(\text{RS}^-)_4$. Soft bases such as sulfur or even nitrogen are rarely found as ligands to hard metallic cations such as Na^+ , K^+ , Mg^{2+} , and Ca^{2+} , but one way that a protein reconciles the steric difficulty of arranging ligands precisely enough to form unhindered covalent bonds with soft metallic cations such as Fe^{2+} ,

Cu^{2+} , and Zn^{2+} is to use hard bases such as oxygen or nitrogen as ligands. The bonds formed by these harder ligands, because they have more ionic character, are more flexible in their angles. Regardless of whether the bonds are ionic or covalent, their lengths are usually governed by the size of the ion, which is reflected in its **ionic radius** (Table 6–7).

The major structural difference between ionic bonds and covalent bonds is the directional properties of the arrangements of the ligands. **Ionic bonds** are created by the electrostatic forces between the metallic cation and an anion or a dipole on a ligand. If the forces holding the ligands are entirely ionic, the number and orientation of the Lewis bases around the cation are determined solely by steric considerations. In the case of Ca^{2+} , the number and orientation of the oxygens surrounding the dication depend entirely on the size and shape of the functional groups that provide them.⁵²³ When ligands are bonded ionically, the larger the cation or the smaller the bases, the more ligands will be gathered. **Covalent bonds** result from the overlap of atomic orbitals to form bonding molecular orbitals. Because the degree of overlap determines the strength of the bond and because the degree of overlap depends on the bond lengths and bond angles, covalent bonds are characterized by specific bond lengths and bond angles. Because zinc is a d_{10} transition metal, its $2d$ shell is filled. As a result, the covalent bonds between Zn^{2+} and sulfur in the regulatory subunit of aspartate carbamoyltransferase are formed from sp^3 hybrid orbitals on the zinc, and this produces the usual tetrahedral arrangement. A similar tetrahedral disposition is assumed by the four Lewis bases around the Zn^{2+} in a zinc finger (Figure 6–53). Covalent bonds position the participating atoms in strict geometric orientations while ionic bonds are completely malleable, resembling pigs at a trough.

The fact that covalent bonds involving metals are so rigid is reflected in the practice during crystallographic refinement of considering them as fixed geometrically as the bonds involving carbon, nitrogen, and oxygen. For example, in the initial crystallographic molecular model of aspartate carbamoyltransferase built directly from the unrefined map of electron density, the arrangement of the four sulfurs around the Zn^{2+} was restrained to a tetrahedral geometry, just as an sp^3 carbon would have been, and this geometry was retained in all the subsequent refinement.⁵²⁴ This practice can be dangerous, however,

Table 6–7: Ionic Radii and Lengths of Bonds to Ligands of Metallic Cations Found as Structural Elements in Proteins

	Na^+	K^+	Mg^{2+}	Ca^{2+}	Mn^{2+}	Fe^{2+}	Ni^{2+}	Cu^{2+}	Zn^{2+}
ionic radius ^a (nm)	0.102	0.138	0.072	0.100	0.083	0.061	0.069	0.073	0.074
bond length ^b (nm)		0.22–0.34	0.20–0.21	0.22–0.26	0.20–0.23	0.19–0.23	0.20–0.23		0.20–0.23

^aIonic radii for the hexacoordinated metallic cation⁵²² and for the dications of transition metals to permit direct comparison. ^bValues for the metallic cations in crystallographic molecular models of proteins from the references cited in the text.

particularly if the ligands to the Zn^{2+} are harder, less covalent bases.⁵²⁵ In such intermediate cases, various mixtures of ionic and covalent behavior are observed. The main indication of such deviations from covalent behavior is the loss of directional ligation.

The monovalent cations of **sodium** and **potassium** are hard Lewis acids and are almost always surrounded by lone pairs from oxygen in any situation, as they are when they are dissolved in water. There is, however, one crystallographic molecular model in which a Na^+ has the π system of a tryptophan as one of its ligands.^{526,527} Because cytoplasm has a high concentration of K^+ and a

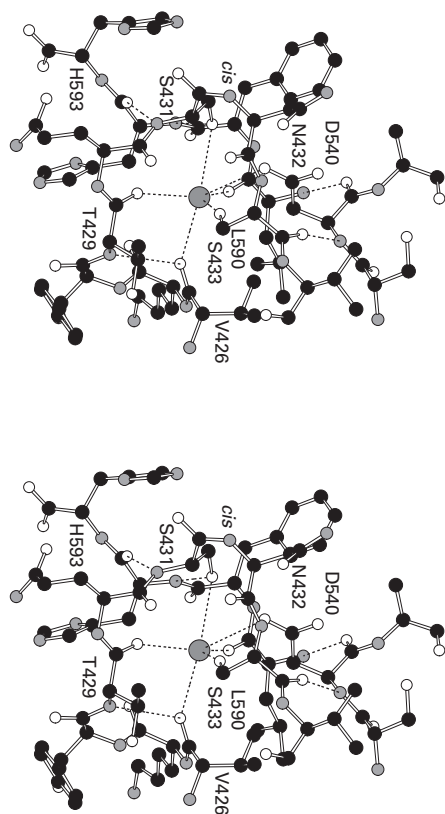


Figure 6-54: A site at which a structural K^+ is bound in the crystallographic molecular model (Bragg spacing ≥ 0.175 nm) of the bifunctional protein containing phosphoribosylaminoimidazolecarboxamide formyltransferase and IMP cyclohydrolase from *G. gallus*. Segments of the folded polypeptide from Valine 426 to Valine 434, Serine 539 to Alanine 541, and Leucine 590 to Histidine 593 are drawn. All of the ligands to K^+ are provided by the protein. Histidine 593 is the carboxy terminus of the protein. The binding site for K^+ is in the enzymatic domain responsible for the activity of phosphoribosylaminoimidazolecarboxamide formyltransferase (Lysine 200 to Histidine 593). This drawing was produced with MolScript.⁵⁷³

low concentration of Na^+ , it is K^+ that is almost invariably incorporated as a structural metallic cation in cytoplasmic proteins. There are examples, however, of cytoplasmic proteins incorporating Na^+ during their crystallization, in one case at a site formed by five acyl oxygens from the backbone and the oxygen of a carboxylate⁵²⁸ and in another at a site formed from two acyl oxygens from the backbone and a molecule of water forming three hydrogen bonds with groups on the protein.⁵²⁹ Whether or not these sites are occupied by Na^+ when the proteins are in the cytoplasm is unknown. Extracytoplasmic proteins, such as α -thrombin, however, do seem to incorporate Na^+ .⁵³⁰

In structural sites for K^+ ,⁵³¹⁻⁵³⁴ the complex can be anywhere from tetracoordinate to heptacoordinate. The ion is large enough (Table 6-7) to support seven oxygens easily, but the final number of ligands is dictated by the structure of the protein. The ligands to a structural K^+ in a crystallographic molecular model of a protein are contributed by as many as three acyl oxygens of the peptide backbone, as many as three waters, sometimes a hydroxyl from a serine or the acyl oxygen of a glutamine or asparagine, and often, but not always, one of the oxygens of a carboxylate from an aspartate or glutamate. For example, in the hexacoordinate structural site for K^+ in phosphoribosylaminoimidazolecarboxamide formyltransferase from *G. gallus*, the cation is liganded by the hydroxyl groups of two serines, the carboxylate oxygen of one aspartate, and three acyl oxygens from the backbone (Figure 6-54).⁵³⁵ The eight ligands from the protein to the two potassium cations bound within the selectivity filter of the KcsA potassium channel, however, are all acyl oxygens from the polypeptide backbone,⁵³⁶ and at the entry to the selectivity filter a potassium can be observed liganded by four acyl oxygens from the backbone and four molecules of water, presumably in the act of being dehydrated.⁵³⁶ In keeping with the size of the potassium cation and the hardness of both the cation and the oxygens, the bonding is ionic, the geometry of the ligands around the cation is unpredictable, and unlike those of most other metallic cations, the bond lengths span a large range (Table 6-7).

Calcium ion, a hard Lewis acid like K^+ , also participates in purely ionic bonds with no particular geometric requirements. Its smaller ionic radius (Table 6-7) is more than compensated by its increased charge, and it associates with as many as nine ligands. Calcium ion has a low affinity for nitrogen,⁵²¹ and it can be distinguished from Mg^{2+} by this characteristic. It is probably the dication that is most often bound by proteins, and its role is usually entirely structural.

When bound to proteins, Ca^{2+} serves in its structural role by gathering around itself oxygens from the backbone of the polypeptide, its side chains, and molecules of water. This **exclusive preference for oxygen** is entirely consistent with its hardness. The octacoordinate site on the surface of endopeptidase K (Figure 6-55)⁵³⁷

that binds Ca^{2+} with a dissociation constant⁵³⁸ of 8×10^{-8} M is typical of a complex between a Ca^{2+} and protein. It is representative of such complexes because the Ca^{2+} is surrounded by molecules of water, one of which is positioned by the molecule of protein, by acyl oxygens from the backbone and by the carboxylate of an aspartate that provides simultaneously two oxygens as a bidentate ligand. The charge number on the Ca^{2+} in this site is not matched by that of its ligands, as is also the case in the heptacoordinate site for Ca^{2+} in α -lactalbumin from *Papio cynocephalus*, in which the ligands are the carboxylates of three aspartates, each a monodentate ligand to the Ca^{2+} , as well as two molecules of water and two acyl oxygens from the backbone.⁵³⁹ There are also structural Ca^{2+} that are more completely surrounded by the protein, such as the heptacoordinate site in thermolysin,⁵⁴⁰ which is surrounded by three acyl oxygens from the backbone, an acyl oxygen of the side chain of an asparagine, a single oxygen of the side chain of an aspartate, and the two carboxylate oxygens of another aspartate acting as a bidentate ligand.

In these complexes, the distance between the Ca^{2+} and the heteroatoms of the ligands is between 0.22 and 0.26 nm (Table 6–7). The position taken by the Ca^{2+} relative to an acyl oxygen of the backbone is remarkably similar to that of an amido nitrogen–hydrogen in a hydrogen bond to such an acyl oxygen (Figure 5–11) with a broad distribution of angle b from 140° to 180° and a strong tendency to lie in the plane of the peptide.⁵⁴¹

These complexes are usually specific for Ca^{2+} . The specificity is provided by the distribution of the oxygens within the protein and the donors and acceptors of hydrogen bonds between the protein and molecules of water (Figure 6–55) retained by the Ca^{2+} as it enters the complex. The number of ligands provided by the protein and the conformation to which they are confined by the rest of its structure⁵⁴² permits them to recognize both the two units of charge number and the radius of the Ca^{2+} . For example, Aspartate 200 in the complex between Ca^{2+} and endopeptidase K (Figure 6–55)⁵³⁷ is positioned properly by resting in a groove formed by the backbone and Cysteine 178/249 as well as by a hydrogen bond to an amido nitrogen–hydrogen of the backbone.

A **magnesium** cation, although also a hard, alkaline earth metallic cation, displays directional bonding reminiscent of a covalent-coordinate metallic cation not because of covalence but because of the intensity of its electrostatic field and the steric effects associated with its significantly smaller ionic radius (Table 6–7). Its complexes are almost always hexacoordinate with the ligands in an octahedral arrangement enforced by the severe steric effects, and the interatomic distances between the metal and the ligands are shorter and much less variable (0.20–0.21 nm) than those for calcium (0.22–0.26 nm). In a structural role, Mg^{2+} is often associated with phosphoryl oxygens, as for example those on nucleic acid. When bound to protein or nucleic acid, Mg^{2+} usually retains

several of its waters, often all of them.⁴⁸⁵ For example, in the complex between Mg^{2+} and inorganic diphosphatase from *E. coli*, the Mg^{2+} retains all six of its octahedrally arrayed molecules of water, each of which forms one to three hydrogen bonds with donors and acceptors on the protein.⁵⁴³

It appears from crystallographic and spectral⁵⁴⁴ observations that Mg^{2+} , when bound to a protein, prefers oxygen ligands exclusively, in particular the oxygens of phosphates and carboxylates, but the dication of **man-**

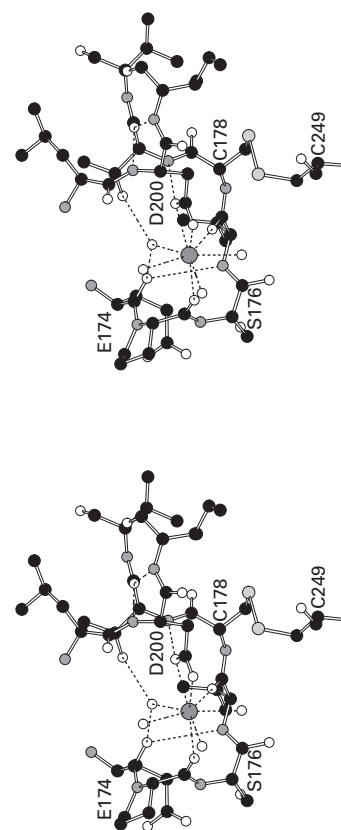


Figure 6–55: A site at which a structural Ca^{2+} is bound on the surface of the crystallographic molecular model (Bragg spacing ≥ 0.15 nm) of endopeptidase K from *Tritirachium album* Limber.⁵³⁷ Segments of the folded polypeptide from Glutamate 174 to Valine 180 and Leucine 199 to Isoleucine 201 are drawn. Of the ligands to the Ca^{2+} , the two acyl oxygens are provided by the second and fourth amino acids of a β turn of type I (Glutamate 174 to Valine 177) that has been stretched out of its normal conformation as it cradles the cation, and the bidentate carboxylate is provided by an aspartate from elsewhere in the amino acid sequence. The unattached oxygens (open circles) are locations for molecules of water. The Ca^{2+} is the gray sphere in the center of the cluster of ligands. The interior of the protein is beyond the valines, the leucine, the isoleucine, and the cysteine to the back and right of the drawing and the water is to the front and left. This drawing was produced with MolScript.⁵⁷³

manganese, Mn^{2+} , a softer metallic ion that is about the same size as Mg^{2+} and that gathers its ligands just as tightly (Table 6–7), forms complexes with both nitrogen and oxygen bases such as ammonia, imidazole, 1,2-diaminoethane, water, alcohols, carboxylates, the carbonyl oxygens of ketones and aldehydes, and the acyl oxygens of amides. Consistent with its degree of hardness, oxygen bases and nitrogen bases are roughly equivalent in their affinities for Mn^{2+} . In aqueous solution, the hexammine complex is observed only at concentrations of ammonia greater than 2 M, but hexaimidazole salts can

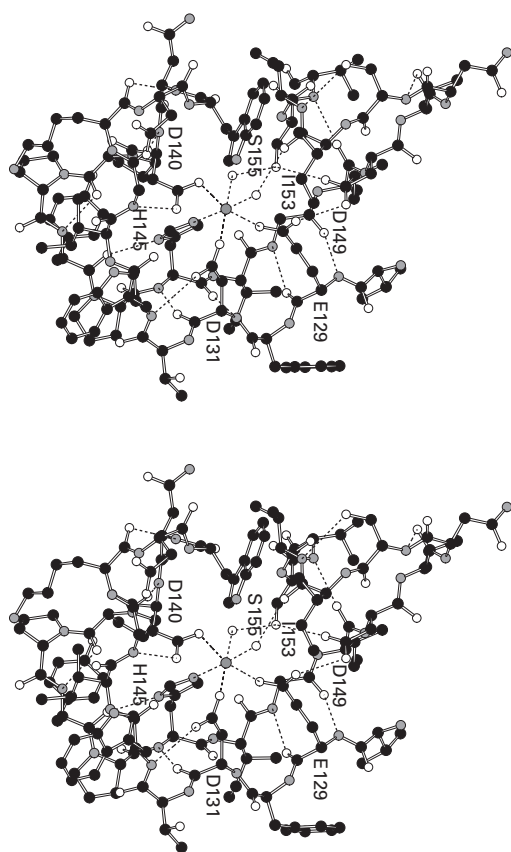
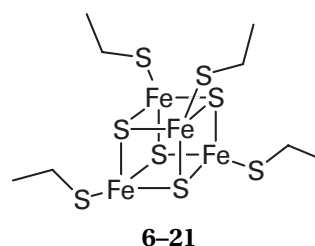


Figure 6-56: Site for the structural Mn^{2+} in the crystallographic molecular model (Bragg spacing ≥ 0.20 nm) of lectin IV from *Griffonia simplicifolia*.² Only one intact segment of the folded polypeptide (Valine 128 to Valine 156) is drawn. It surrounds the Mn^{2+} and provides all of the ligands directly to the metal and to its two bound molecules of water (open circles). The side chain of Asparagine 135 has been erased to provide a better view of the metal and its ligands. The Mn^{2+} is the gray sphere in the center of the cluster of ligands. This drawing was produced with MolScript.⁵⁷³

be crystallized from anhydrous ethanol. Because of its small size and degree of hardness, all of the complexes between Mn^{2+} and such unhindered hard and intermediate bases are hexacoordinate and octahedral, reminiscent of directional covalent bonds; and in these complexes, mixtures of various ligands around manganese can occur. For example, each of the species $[\text{Mn}(\text{OH}_2)_n(\text{NH}_3)_{6-n}]^{2+}$ with $0 < n < 6$ is observed in mixtures of ammonia and water. When Mn^{2+} is bound to a protein, it is complexed octahedrally by Lewis bases both from amino acids on the protein and from molecules of water (Figure 6–56).²

Iron, when it is found in a protein, is almost always in a coenzyme such as a heme (Figure 4–18) or an iron–sulfur cluster:

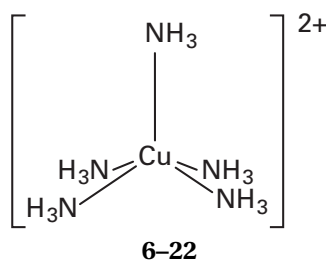


In most of these instances, the iron acts either as an electron carrier because of its ability to convert between $\text{Fe}(\text{II})$ and $\text{Fe}(\text{III})$ or as a catalytic group that can bind and activate oxygen, but the iron–sulfur cluster in DNA-(apurinic or apyrimidinic site) lyase of *E. coli*,⁵⁴⁵ as well as the one in glycosylase MutY from *E. coli*,⁵⁴⁶ performs a structural role. In a heme the Fe^{2+} is hexacoordinate, but in an iron–sulfur cluster it is tetracoordinate, consistent with the softness of the thiolates. There is also a pentacoordinate site for a structural Fe^{2+} in UTP-hexose-1-phosphate uridylyltransferase in which the ligands are the nitrogens from three histidines and the two carboxyl oxygens of a glutamate acting as a bidentate ligand.⁵⁴⁷ The irregular arrangement of the ligands in this case is permitted by their hardness, which causes the ligation to be more ionic. One of the most peculiar sites for an iron cation is that on nitrile hydratase of *Rhodococcus*, in which the Fe^{3+} is liganded by two amido nitrogens from the backbone and the sulfurs of three cysteines, one of which is oxidized to a sulfenic acid and another to a sulfinic acid (Figure 2–8).⁵⁴⁸

Cobalt is incorporated into proteins as the metal in the center of coenzyme B_{12} . **Nickel** is used as a Lewis acid in the active site of urease⁵⁴⁹ and as the electrochemically active component in the active site of a few enzymes catalyzing oxidation–reduction.^{550–552} In at least one of these latter enzymes, it is found coordinated within a tetrapyrrole that resembles a porphyrin.⁵⁵³ **Molybdenum** and **vanadium** are found in nitrogenases and **molybdenum** and **tungsten** in other enzymes catalyzing oxidation–reduction such as nitrate reductase, aldehyde:ferredoxin oxidoreductase, and xanthine oxidase. In each of these proteins, the molybdenum, tungsten, or vanadium is enclosed within a pterin coenzyme that provides the

thiols coordinating the metallic cation and holding it within the protein.⁵⁵⁴

Copper exists in biochemical situations as the kinetically stable cations Cu^+ and Cu^{2+} . It is usually used as a one-electron carrier, often in reactions involving oxygen activation such as those catalyzed by monooxygenases. Although it is a soft metallic cation, Cu^{2+} can form a number of coordination complexes with lone pairs from oxygen, nitrogen, and sulfur. The variety of these complexes defies categorization. They range from dicoordinate to octacoordinate. Even in the more common tetracoordinate and hexacoordinate stereochemistries, the terms used to describe the variations, such as square planar, compressed tetrahedral, elongated tetragonal octahedral, and trigonal octahedral, indicate that the arrangement of many of the Lewis bases around copper, as with calcium, is governed mainly by steric effects among the ligands, rather than by covalent bonding. Examples of complexes between Cu^{2+} and simple biochemical ligands are $[\text{Cu}(\text{NH}_3)_5]^{2+}$



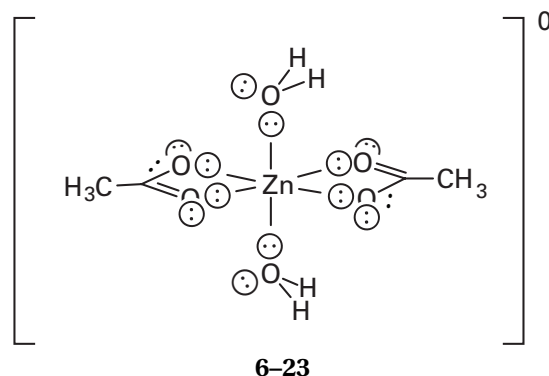
in which four of the nitrogens are equivalent and the fifth forms a longer bond to Cu^{2+} , and $[\text{Cu}(\text{imidazole})_4(\text{OH}_2)_2]^{2+}$ and $[\text{Cu}(\text{formate})_2(\text{OH}_2)_2]^{2+}$, which are both elongated tetragonal octahedral structures. Simple thiols such as mercaptans reduce Cu^{2+} to Cu^+ and form complex polymeric structures with Cu^+ in which the coppers are multidentate and the thiols are bidentate.

The azurins and the plastocyanins are related metalloproteins involved in one-electron transfers in which the single copper passes reversibly from Cu^{2+} to Cu^+ to carry the electron. In crystallographic molecular models of these proteins,^{555,556} the copper is coordinated to two histidines, a methionine, and a cysteine with no particular geometric regularity. In the apoprotein,* the two nitrogens and the two sulfurs that surround the copper in the holoprotein assume the same orientations even though the copper is not present.⁵⁵⁷ Consequently, unlike the situation in a zinc finger in which the Zn^{2+} dictates the structure assumed by the protein, azurins and plastocyanins are large enough proteins that they dictate the stereochemistry of the ligation.

The most versatile metallic dication performing

* The metal-free form of a metalloprotein is the apoprotein, and the form of the metalloprotein when it contains the metal is the holoprotein.

structural roles in proteins is that of **zinc**. Its versatility in this role arises from its ability to form both tetracoordinate and pentacoordinate complexes with Lewis bases and its ability, even though it is one of the softest metallic cations, to form complexes with lone pairs from **oxygen and nitrogen, as well as sulfur**. Often a mixture of two or three of these rather different bases forms the site on a metalloprotein for the Zn^{2+} . When it is bound by four sulfurs, which are soft bases complementary to the soft Zn^{2+} , the bonds are covalent and tetrahedral. The complex between Zn^{2+} and the harder base ammonia is tetracoordinate $[\text{Zn}(\text{NH}_3)_4]^{2+}$ and tetrahedral, probably because it is also covalent. This tetrahedral covalent form of Zn^{2+} is the most common and is observed when Zn^{2+} forms complexes with 1,2-diaminoethane, cyclic lactams, and imidazole. As the ligands become harder, however, geometries become more variable. For example, the complex $[\text{Zn}(\text{OH}_2)_6]^{2+}$ between Zn^{2+} and water, a hard base, is hexacoordinate and octahedral; but as protons are removed, it eventually decreases its coordination to four, as $[\text{Zn}(\text{OH})_3(\text{OH}_2)]^-$, as a result of electrostatic repulsion. An ionic, octahedral complex forms with carboxylates:

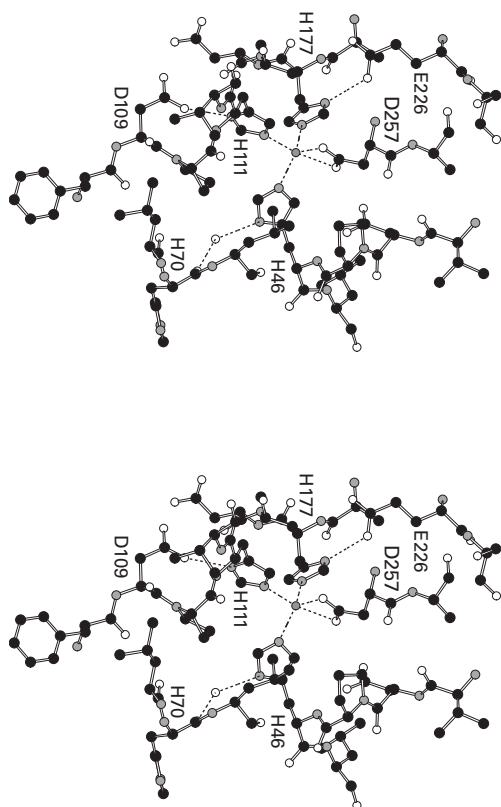


Zinc dication forms a pentacoordinate complex with, among other ligands, 8-aminoquinazoline,⁵⁵⁸ in which the four nitrogens from two aminoquinazolines and a molecule of water are the five Lewis bases that generate the complex $[\text{Zn}(\text{N}_2\text{C}_9\text{H}_8)_2(\text{OH}_2)]^{2+}$.

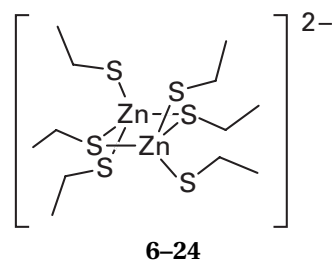
A **pentacoordinate complex** is formed by the structural Zn^{2+} in the periplasmic zinc-binding protein TroA of *Treponema pallidum*. In this complex with the protein, the Zn^{2+} is surrounded by the two oxygens of the carboxylate of Aspartate 257 as a bidentate ligand and three imidazolyl nitrogens from Histidine 46, Histidine 111, and Histidine 177 in an irregular arrangement (Figure 6-57).⁵⁵⁹ Most structural sites for zinc, however, are tetracoordinate, resembling the one in a zinc finger (Figure 6-53). For example, in the structural sites for zinc in UTP-hexose-1-phosphate uridylyltransferase from *E. coli* and alanine-tRNA ligase from *E. coli*, the cations are also surrounded by two histidines and two cysteines in a tetrahedral array.^{547,560}

There are a number of proteins that contain mod-

Figure 6-57: Site for the structural Zn^{2+} in the crystallographic molecular model (Bragg spacing ≥ 0.18 nm) of periplasmic zinc-binding protein TroA from *Treponema pallidum*.⁵⁵⁹ In this case, unlike that of the Mn^{2+} binding site in lectin IV (Figure 6-56), the four ligands to the Zn^{2+} are provided by side chains distant from each other in the amino acid sequence. Consequently, four short segments of the folded polypeptide (Valine 43 to Valine 47, Phenylalanine 108 to Valine 112, Alanine 176 to Alanine 179, and Aspartate 257 to Alanine 258) are drawn as well as two others (Glutamate 226 to Serine 227 and Leucine 69 to Leucine 71) that provide acceptors for hydrogen bonds buttressing the ligands to the metal. The Zn^{2+} is the gray sphere in the center of the cluster of ligands. A molecule of water, not a ligand to the metal, is drawn as an open circle. The amino acid sequence is numbered according to that of the posttranslationally modified version of the protein. This drawing was produced with MolScript.⁵⁷³



ules resembling zinc fingers in that they bind to specific sequences in DNA and are also zinc metalloproteins. Some of them have complexes that resemble the one in the regulatory subunit of aspartate carbamoyltransferase (**4-48**) because the zinc forms covalent bonds with four cysteines.^{438,561} Others have a site formed from three cysteines and only one histidine.⁵⁶² Others contain clusters formed from two Zn^{2+} and the sulfurs from six cysteines



that resemble iron-sulfur clusters (**6-21**) and in which each Zn^{2+} forms covalent bonds to four sulfurs.⁵⁶³⁻⁵⁶⁵ In all of these modules, as in the zinc fingers, the cross-linking of the polypeptide performed by the respective complex is essential for maintenance of the proper structure.

Because the distances (Table 6-7) at which the ligands are held by Mn^{2+} , Fe^{2+} , Ni^{2+} , Co^{2+} , and Zn^{2+} in a crystallographic molecular model are so similar, it is not surprising that these metallic cations are often **interchangeable in their structural roles**. For example, even though Zn^{2+} is the only cation found in aspartate carbamoyltransferase when it is purified from *E. coli*, Ni^{2+} and Co^{2+} are capable of promoting the proper reassembly of the apoprotein.⁵¹⁷ The metal site within the active site of arylalkylphosphatase⁵⁶⁶ is as happy with Ni^{2+} or Mn^{2+} as it is with the Zn^{2+} it naturally contains. The diphtheria toxin repressor from *Corynebacterium diphtheriae* under normal conditions of growth incorporates an Fe^{2+} into its structure,⁵⁶⁷ which is required for it to fold properly. The structural role of this Fe^{2+} can be played just as well by Ni^{2+} or Mn^{2+} .⁵⁶⁸ In the properly folded form of this protein, the metallic cation is octahedrally coordinated by the carboxylate oxygen of an aspartate, the carboxylate oxygen of a glutamate, the imidazolyl nitrogen of a histidine, the sulfur of a methionine, an acyl oxygen from the backbone, and a molecule of water, making this site one of the most eclectic (Figure 6-58).⁵⁶⁷

Sometimes the site for a structural metallic cation associates with its metallic cation so weakly that the cation is lost during its purification. If several metallic cations are as effective at producing its proper conformation, it is difficult to say with any certainty which is used in vivo.

Suggested Reading

Lee, Y.H., Deka, R.K., Norgard, M.V., Radolf, J.D., & Hasemann, C.A. (1999) *Treponema pallidum* TroA is a periplasmic zinc-binding protein with a helical backbone, *Nat. Struct. Biol.* 6, 628-633.

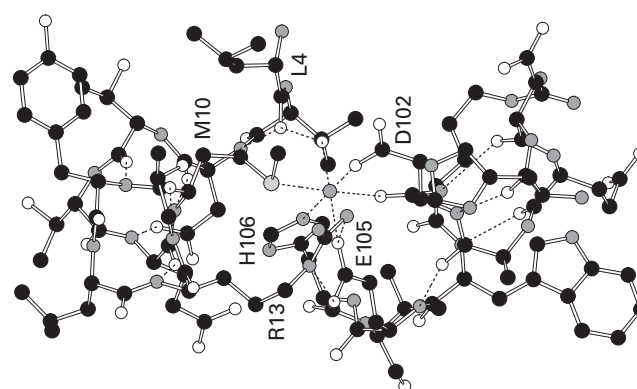
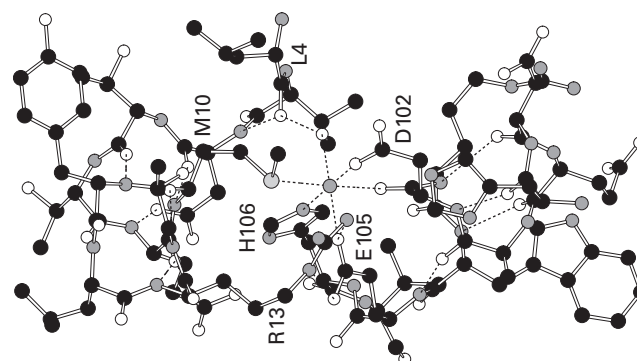
Problem 6-14: The drawing in the figure on the next page is of a site for the binding of Mg^{2+} within the crystallographic molecular model of inorganic pyrophosphatase from *E. coli*.⁵⁴³ This drawing was produced with MolScript.⁵⁷³

Identify the donors and acceptor for each of the hydrogen bonds.

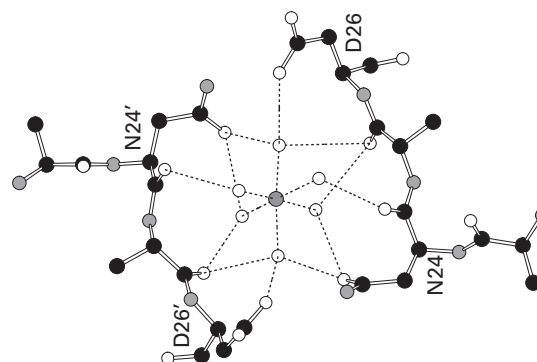
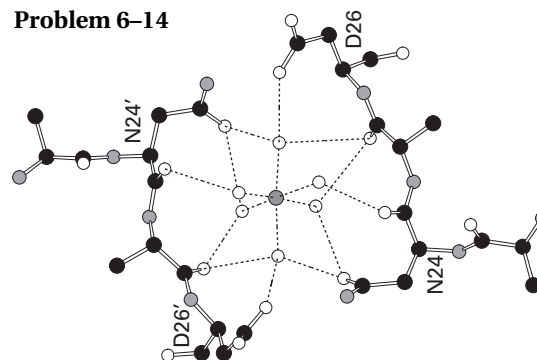
Figure 6-58: Site for the structural Fe^{2+} in the crystallographic molecular model (Bragg spacing ≥ 0.24 nm) of diphtheria toxin repressor from *Corynebacterium diphtheriae*.⁵⁶⁷ The protein purified from the bacterium contains an Fe^{2+} at this site, but that Fe^{2+} was replaced with Ni^{2+} for the crystallographic study. The Ni^{2+} is the gray sphere in the center of the cluster of ligands. Presumably, an Fe^{2+} would gather the same ligands (Table 6-7). The ligands are provided by side chains and backbone from two α helices. The amino-terminal portion of the upper α helix and a segment amino-terminal to it (Leucine 4 to Threonine 14) and the carboxy-terminal portion of the lower α helix and a segment carboxy-terminal to it (Histidine 98 to Valine 107) are drawn. Together they provide all of the ligands to the metal and donors and acceptors for hydrogen bonds buttressing the ligands. Note the pair of hydrogen bonds between Arginine 13 and Glutamate 105 and the hydrogen bond between Aspartate 6 and the open amido nitrogen-hydrogen at position 9 at the amino-terminal end of the upper α helix. This drawing was produced with MolScript.⁵⁷³

References

1. Pal, D., & Chakrabarti, P. (1999) *J. Mol. Biol.* 294, 271–288.
2. Delbaere, L.T., Vondonselaar, M., Prasad, L., Quail, J.W., Wilson, K.S., & Dauter, Z. (1993) *J. Mol. Biol.* 230, 950–965.
3. Brandts, J.F., Halvorson, H.R., & Brennan, M. (1975) *Biochemistry* 14, 4953–4963.
4. Herzberg, O., & Moulton, J. (1991) *Proteins: Struct., Funct., Genet.* 11, 223–229.
5. Blake, C.C., Ghosh, M., Harlos, K., Avezoux, A., & Anthony, C. (1994) *Nat. Struct. Biol.* 1, 102–105.
6. Wilson, K.S., Butterworth, S., Dauter, Z., Lamzin, V.S., Walsh, M., Wodak, S., Pontius, J., Richelle, J., Vaguine, A., Sander, C., Hooft, R.W.W., Vriend, G., Thornton, J.M., Laskowski, R.A., MacArthur, M.W., Dodson, E.J., Murshudov, G., Oldfield, T.J., Kaptien, R., & Rullmann, J.A.C. (1998) *J. Mol. Biol.* 276, 417–436.
7. Edison, A.S. (2001) *Nat. Struct. Biol.* 8, 201–202.
8. Oefner, C., & Suck, D. (1986) *J. Mol. Biol.* 192, 605–632.
9. Kirby, A.J., Komarov, I.V., & Feeder, N. (1998) *J. Am. Chem. Soc.* 120, 7101–7102.
10. Ramachandran, G.N., Ramakrishnan, C., & Sasisekharan, V. (1963) *J. Mol. Biol.* 7, 95–99.
11. Adzhubei, A.A., & Sternberg, M.J. (1993) *J. Mol. Biol.* 229, 472–493.
12. Mandel, N., Mandel, G., Trus, B.L., Rosenberg, J., Carlson, G., & Dickerson, R.E. (1977) *J. Biol. Chem.* 252, 4619–4636.
13. Peti, W., Hennig, M., Smith, L.J., & Schwalbe, H. (2000) *J. Am. Chem. Soc.* 122, 12017–12018.
14. Smith, L.J., Bolin, K.A., Schwalbe, H., MacArthur, M.W., Thornton, J.M., & Dobson, C.M. (1996) *J. Mol. Biol.* 255, 494–506.
15. Vrielink, A., Lloyd, L.F., & Blow, D.M. (1991) *J. Mol. Biol.* 219, 533–554.
16. Strynadka, N.C., & James, M.N. (1991) *J. Mol. Biol.* 220, 401–424.
17. Birktoft, J.J., & Blow, D.M. (1972) *J. Mol. Biol.* 68, 187–240.
18. Esposito, L., Vitagliano, L., Sica, F., Sorrentino, G.,



Problem 6-14



- Zagari, A., & Mazzarella, L. (2000) *J. Mol. Biol.* 297, 713–732.
19. Takano, T. (1984) in *Methods and Applications in Crystallographic Computing: Papers presented at the International Summer School on Crystallographic Computing, held at Kyoto, Japan* (Ashida, T., & Hall, S. R., Eds.) p 262, Clarendon Press, Oxford, U.K.
20. Matthews, B.W., Weaver, L.H., & Kester, W.R. (1974) *J. Biol. Chem.* 249, 8030–8044.
21. Pauling, L., Corey, R.B., & Branson, H.R. (1951) *Proc. Natl. Acad. Sci. U.S.A.* 37, 205–211.
22. Artymiuk, P.J., & Blake, C.C. (1981) *J. Mol. Biol.* 152, 737–762.
23. Barlow, D.J., & Thornton, J.M. (1988) *J. Mol. Biol.* 201, 601–619.
24. Herzberg, O., & James, M.N. (1988) *J. Mol. Biol.* 203, 761–779.
25. Watson, H.C. (1969) *Prog. Stereochem.* 4, 299–333.
26. Gray, T.M., & Matthews, B.W. (1984) *J. Mol. Biol.* 175, 75–81.
27. Bolin, J.T., Filman, D.J., Matthews, D.A., Hamlin, R.C., & Kraut, J. (1982) *J. Biol. Chem.* 257, 13650–13662.
28. Manas, E.S., Getahun, Z., Wright, W.W., DeGrado, W.F., & Vanderkooi, J.M. (2000) *J. Am. Chem. Soc.* 122, 9883–9890.
29. Blundell, T., Barlow, D., Borkakoti, N., & Thornton, J. (1983) *Nature* 306, 281–283.
30. Takano, T. (1977) *J. Mol. Biol.* 110, 537–568.
31. Remington, S., Wiegand, G., & Huber, R. (1982) *J. Mol. Biol.* 158, 111–152.
32. Sauer, U.H., San, D.P., & Matthews, B.W. (1992) *J. Biol. Chem.* 267, 2393–2399.
33. Pflugrath, J.W., & Quiocho, F.A. (1988) *J. Mol. Biol.* 200, 163–180.
34. Hasemann, C.A., Ravichandran, K.G., Peterson, J.A., & Deisenhofer, J. (1994) *J. Mol. Biol.* 236, 1169–1185.
35. Heinz, D.W., Baase, W.A., Zhang, X.J., Blaber, M., Dahlquist, F.W., & Matthews, B.W. (1994) *J. Mol. Biol.* 236, 869–886.
36. Keefe, L.J., Sondek, J., Shortle, D., & Lattman, E.E. (1993) *Proc. Natl. Acad. Sci. U.S.A.* 90, 3275–3279.
37. Richardson, J.S., & Richardson, D.C. (1988) *Science* 240, 1648–1652.
38. Presta, L.G., & Rose, G.D. (1988) *Science* 240, 1632–1641.
39. Harper, E.T., & Rose, G.D. (1993) *Biochemistry* 32, 7605–7609.
40. Bordo, D., & Argos, P. (1994) *J. Mol. Biol.* 243, 504–519.
41. Serrano, L., Sancho, J., Hirshberg, M., & Fersht, A.R. (1992) *J. Mol. Biol.* 227, 544–559.
42. Bell, J.A., Becktel, W.J., Sauer, U., Baase, W.A., & Matthews, B.W. (1992) *Biochemistry* 31, 3590–3596.
43. Schellman, C. (1980) in *Protein Folding: Proceedings of the 28th Conference of the German Biochemical Society* (Jaenicke, R., Ed.) p 53, Elsevier/North-Holland Biomedical Press, Amsterdam.
44. Hol, W.G., van Duijnen, P.T., & Berendsen, H.J. (1978) *Nature* 273, 443–446.
45. Wishart, D.S., Sykes, B.D., & Richards, F.M. (1991) *J. Mol. Biol.* 222, 311–333.
46. Aqvist, J., Luecke, H., Quiocho, F.A., & Warshel, A. (1991) *Proc. Natl. Acad. Sci. U.S.A.* 88, 2026–2030.
47. Lockhart, D.J., & Kim, P.S. (1993) *Science* 260, 198–202.
48. Rogers, N.K., & Sternberg, M.J. (1984) *J. Mol. Biol.* 174, 527–542.
49. Nicholson, H., Becktel, W.J., & Matthews, B.W. (1988) *Nature* 336, 651–656.
50. McLachlan, A.D., & Stewart, M. (1975) *J. Mol. Biol.* 98, 293–304.
51. Chatton, B., Walter, P., Ebel, J.P., Lacroute, F., & Fasiolo, F. (1988) *J. Biol. Chem.* 263, 52–57.
52. Zhou, N.E., Zhu, B.Y., Sykes, B.D., & Hodges, R.S. (1992) *J. Am. Chem. Soc.* 114, 4320–4326.
53. Chou, P.Y., & Fasman, G.D. (1978) *Adv. Enzymol. Relat. Areas Mol. Biol.* 47, 45–148.
54. Fasman, G.D. (1967) *Poly- α -Amino Acids; Protein Models for Conformational Studies*, Marcel Dekker, New York.
55. Myers, J.K., Pace, C.N., & Scholtz, J.M. (1997) *Biochemistry* 36, 10923–10929.
56. Serrano, L., Neira, J.L., Sancho, J., & Fersht, A.R. (1992) *Nature* 356, 453–455.
57. Blaber, M., Zhang, X.J., Lindstrom, J.D., Pepiot, S.D., Baase, W.A., & Matthews, B.W. (1994) *J. Mol. Biol.* 235, 600–624.
58. Myers, J.K., Pace, C.N., & Scholtz, J.M. (1997) *Proc. Natl. Acad. Sci. U.S.A.* 94, 2833–2837.
59. O'Neil, K.T., & DeGrado, W.F. (1990) *Science* 250, 646–651.
60. Chakrabartty, A., Schellman, J.A., & Baldwin, R.L. (1991) *Nature* 351, 586–588.
61. Yang, D.S., Sax, M., Chakrabartty, A., & Hew, C.L. (1988) *Nature* 333, 232–237.
62. Spek, E.J., Wu, H.C., & Kallenbach, N.R. (1997) *J. Am. Chem. Soc.* 119, 5053–5054.
63. Blaber, M., Zhang, X.J., & Matthews, B.W. (1993) *Science* 260, 1637–1640.
64. Hasson, M.S., Muscate, A., McLeish, M.J., Polovnikova, L.S., Gerlt, J.A., Kenyon, G.L., Petsko, G.A., & Ringe, D. (1998) *Biochemistry* 37, 9918–9930.
65. Kelly, M.A., Chellgren, B.W., Rucker, A.L., Troutman, J.M., Fried, M.G., Miller, A.F., & Creamer, T.P. (2001) *Biochemistry* 40, 14376–14383.
66. Romao, M.J., Turk, D., Gomis-Rueth, F.X., Huber, R., Schumacher, G., Moellering, H., & Ruessmann, L. (1992) *J. Mol. Biol.* 226, 1111–1130.
67. Ma, K., Kan, L.-s., & Wang, K. (2001) *Biochemistry* 40, 3427–3438.
68. Boyington, J.C., Gaffney, B.J., & Amzel, L.M. (1993) *Science* 260, 1482–1486.
69. Dickinson, C.D., Veerapandian, B., Dai, X.P., Hamlin, R.C., Xuong, N.H., Ruoslahti, E., & Ely, K.R. (1994) *J. Mol. Biol.* 236, 1079–1092.
70. Eklund, H., Nordstrom, B., Zeppezauer, E., Seoderlund, G., Ohlsson, I., Boiwe, T., Seoderberg, B.O., Tapia, O., Breandaen, C.I., & Akesson, A. (1976) *J. Mol. Biol.* 102, 27–59.
71. Chothia, C. (1973) *J. Mol. Biol.* 75, 295–302.
72. Herzberg, O. (1991) *J. Mol. Biol.* 217, 701–719.
73. Richardson, J.S., Getzoff, E.D., & Richardson, D.C. (1978) *Proc. Natl. Acad. Sci. U.S.A.* 75, 2574–2578.
74. Scapin, G., Gordon, J.L., & Sacchettini, J.C. (1992) *J. Biol. Chem.* 267, 4253–4269.

75. Sacchettini, J.C., Gordon, J.I., & Banaszak, L.J. (1989) *J. Mol. Biol.* 208, 327–339.
76. Wilmanns, M., Priestle, J.P., Niermann, T., & Jansonius, J.N. (1992) *J. Mol. Biol.* 223, 477–507.
77. Murzin, A.G., Lesk, A.M., & Chothia, C. (1994) *J. Mol. Biol.* 236, 1382–1400.
78. Yoder, M.D., Keen, N.T., & Jurnak, F. (1993) *Science* 260, 1503–1507.
79. Beaman, T.W., Binder, D.A., Blanchard, J.S., & Roderick, S.L. (1997) *Biochemistry* 36, 489–494.
80. Xia, Z., Dai, W., Zhang, Y., White, S.A., Boyd, G.D., & Mathews, F.S. (1996) *J. Mol. Biol.* 259, 480–501.
81. Varghese, J.N., & Colman, P.M. (1991) *J. Mol. Biol.* 221, 473–486.
82. Crennell, S.J., Garman, E.F., Philippon, C., Vasella, A., Laver, W.G., Vimr, E.R., & Taylor, G.L. (1996) *J. Mol. Biol.* 259, 264–280.
83. Ormo, M., Cubitt, A.B., Kallio, K., Gross, L.A., Tsien, R.Y., & Remington, S.J. (1996) *Science* 273, 1392–1395.
84. Hopf, M., Gohring, W., Ries, A., Timpl, R., & Hohenester, E. (2001) *Nat. Struct. Biol.* 8, 634–640.
85. Novotny, J., Bruccoleri, R.E., & Newell, J. (1984) *J. Mol. Biol.* 177, 567–573.
86. Lasters, I., Wodak, S.J., Alard, P., & van Cutsem, E. (1988) *Proc. Natl. Acad. Sci. U.S.A.* 85, 3338–3342.
87. Musil, D., Bode, W., Huber, R., Laskowski, M., Jr., Lin, T.Y., & Ardelt, W. (1991) *J. Mol. Biol.* 220, 739–755.
88. Arnold, E., & Rossmann, M.G. (1990) *J. Mol. Biol.* 211, 763–801.
89. Koronakis, V., Sharff, A., Koronakis, E., Luisi, B., & Hughes, C. (2000) *Nature* 405, 914–919.
90. Uhlin, U., & Eklund, H. (1996) *J. Mol. Biol.* 262, 358–369.
91. Baumann, U. (1994) *J. Mol. Biol.* 242, 244–251.
92. Steinbacher, S., Seckler, R., Miller, S., Steipe, B., Huber, R., & Reinemer, P. (1994) *Science* 265, 383–386.
93. Wu, H., Maciejewski, M.W., Marintchev, A., Benashski, S.E., Mullen, G.P., & King, S.M. (2000) *Nat. Struct. Biol.* 7, 575–579.
94. Teeter, M.M., Roe, S.M., & Heo, N.H. (1993) *J. Mol. Biol.* 230, 292–311.
95. Venkatachalam, C.M. (1968) *Biopolymers* 6, 1425–1436.
96. Wilmot, C.M., & Thornton, J.M. (1988) *J. Mol. Biol.* 203, 221–232.
97. Fujinaga, M., Delbaere, L.T., Brayer, G.D., & James, M.N. (1985) *J. Mol. Biol.* 184, 479–502.
98. James, M.N., & Sielecki, A.R. (1983) *J. Mol. Biol.* 163, 299–361.
99. Hamada, K., Bethge, P.H., & Mathews, F.S. (1995) *J. Mol. Biol.* 247, 947–962.
100. Louie, G.V., & Brayer, G.D. (1990) *J. Mol. Biol.* 214, 527–555.
101. Bragg, L., Kendrew, J.C., & Perutz, M.F. (1950) *Proc. R. Soc. London, A* 203, 321–357.
102. Taylor, H.S. (1941) *Proc. Am. Philos. Soc.* 85, 1–12.
103. Richardson, J.S. (1981) *Adv. Protein Chem.* 34, 167–339.
104. Pavone, V., Di Blasio, B., Santini, A., Benedetti, E., Pedone, C., Toniolo, C., & Crisma, M. (1990) *J. Mol. Biol.* 214, 633–635.
105. Dijkstra, B.W., Kalk, K.H., Hol, W.G., & Drenth, J. (1981) *J. Mol. Biol.* 147, 97–123.
106. Sibanda, B.L., & Thornton, J.M. (1985) *Nature* 316, 170–174.
107. Efimov, A.V. (1986) *Mol. Biol. (USSR)* 20, 250–260.
108. Volz, K., & Matsumura, P. (1991) *J. Biol. Chem.* 266, 15511–15519.
109. Chou, P.Y., & Fasman, G.D. (1977) *J. Mol. Biol.* 115, 135–175.
110. Moore, S.A., & James, M.N. (1995) *J. Mol. Biol.* 249, 195–214.
111. Milner-White, E., Ross, B.M., Ismail, R., Belhadj-Mostefa, K., & Poet, R. (1988) *J. Mol. Biol.* 204, 777–782.
112. Henrick, K., Collyer, C.A., & Blow, D.M. (1989) *J. Mol. Biol.* 208, 129–157.
113. Baker, E.N., & Hubbard, R.E. (1984) *Prog. Biophys. Mol. Biol.* 44, 97–179.
114. Merritt, E.A., Kuhn, P., Sarfaty, S., Erbe, J.L., Holmes, R.K., & Hol, W.G. (1998) *J. Mol. Biol.* 282, 1043–1059.
115. Martinez-Oyanedel, J., Choe, H.W., Heinemann, U., & Saenger, W. (1991) *J. Mol. Biol.* 222, 335–352.
116. Karplus, P.A., & Schulz, G.E. (1987) *J. Mol. Biol.* 195, 701–729.
117. Borgstahl, G.E., Rogers, P.H., & Arnone, A. (1994) *J. Mol. Biol.* 236, 817–830.
118. Goddette, D.W., Paech, C., Yang, S.S., Mielenz, J.R., Bystroff, C., Wilke, M.E., & Fletterick, R.J. (1992) *J. Mol. Biol.* 228, 580–595.
119. Wlodawer, A., Svensson, L.A., Sjolín, L., & Gilliland, G.L. (1988) *Biochemistry* 27, 2705–2717.
120. Ponder, J.W., & Richards, F.M. (1987) *J. Mol. Biol.* 193, 775–791.
121. Schrauber, H., Eisenhaber, F., & Argos, P. (1993) *J. Mol. Biol.* 230, 592–612.
122. Lovell, S.C., Word, J.M., Richardson, J.S., & Richardson, D.C. (2000) *Proteins: Struct., Funct., Genet.* 40, 389–408.
123. Dunbrack, R.L., Jr. (2002) *Curr. Opin. Struct. Biol.* 12, 431–440.
124. Wilson, M.A., & Brunger, A.T. (2000) *J. Mol. Biol.* 301, 1237–1256.
125. Roberts, J.D., & Caserio, M.C. (1977) *Basic Principles of Organic Chemistry*, 2nd ed., p 457, W.A. Benjamin, Menlo Park, CA.
126. Janin, J., & Wodak, S. (1978) *J. Mol. Biol.* 125, 357–386.
127. Russell, R.J., Ferguson, J.M., Hough, D.W., Danson, M.J., & Taylor, G.L. (1997) *Biochemistry* 36, 9983–9994.
128. Kyte, J. (1995) *Structure in Protein Chemistry*, 1st ed., p 212, Garland Publishing, Inc., New York.
129. Wang, J.F., Hinck, A.P., Loh, S.N., & Markley, J.L. (1990) *Biochemistry* 29, 4242–4253.
130. Kossiakoff, A.A., & Shteyn, S. (1984) *Nature* 311, 582–583.
131. Word, J.M., Lovell, S.C., LaBean, T.H., Taylor, H.C., Zalis, M.E., Presley, B.K., Richardson, J.S., & Richardson, D.C. (1999) *J. Mol. Biol.* 285, 1711–1733.
132. Kossiakoff, A., Shpungin, J., & Shteyn, M. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87, 4468–4472.
133. Lovell, S.C., Word, J.M., Richardson, J.S., & Richardson, D.C. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 400–405.
134. Gellman, S.H. (1991) *Biochemistry* 30, 6633–6636.
135. Nemethy, G., Gibson, K.D., Palmer, K.A., Yoon, C.N., Paterlini, G., Zagari, A., Rumsey, S., & Scheraga, H.A. (1992) *J. Phys. Chem.* 96, 6472–6484.

136. Kelly, J.A., & Kuzin, A.P. (1995) *J. Mol. Biol.* 254, 223–236.
137. Sorensen, S.B., Raaschou-Nielsen, M., Mortensen, U.H., Remington, S.J., & Breddam, K. (1995) *J. Am. Chem. Soc.* 117, 5944–5950.
138. Endrizzi, J.A., Breddam, K., & Remington, S.J. (1994) *Biochemistry* 33, 11106–11120.
139. Nelsen, S.F., Teasley, M.F., Bloodworth, A.J., & Eggelte, H.J. (1985) *J. Org. Chem.* 50, 3299–3302.
140. Morris, A.L., MacArthur, M.W., Hutchinson, E.G., & Thornton, J.M. (1992) *Proteins: Struct., Funct., Genet.* 12, 345–364.
141. Czapinska, H., Otlewski, J., Krzywda, S., Sheldrick, G.M., & Jaskolski, M. (2000) *J. Mol. Biol.* 295, 1237–1249.
142. Webba da Silva, M., Sham, S., Gorst, C.M., Calzolari, L., Brereton, P.S., Adams, M.W.W., & La Mar, G.N. (2001) *Biochemistry* 40, 12575–12583.
143. Kuwajima, K., Ikeguchi, M., Sugawara, T., Hiraoka, Y., & Sugai, S. (1990) *Biochemistry* 29, 8240–8249.
144. Lee, B., & Richards, F.M. (1971) *J. Mol. Biol.* 55, 379–400.
145. Janin, J. (1979) *Nature* 277, 491–492.
146. Miller, S., Janin, J., Lesk, A.M., & Chothia, C. (1987) *J. Mol. Biol.* 196, 641–656.
147. Chothia, C. (1974) *Nature* 248, 338–339.
148. Chothia, C. (1976) *J. Mol. Biol.* 105, 1–12.
149. Wolfenden, R.V., Cullis, P.M., & Southgate, C.C. (1979) *Science* 206, 575–577.
150. Tavares, G.A., Beguin, P., & Alzari, P.M. (1997) *J. Mol. Biol.* 273, 701–713.
151. van Raaij, M.J., Schoehn, G., Burda, M.R., & Miller, S. (2001) *J. Mol. Biol.* 314, 1137–1146.
152. Eriksson, A.E., Baase, W.A., Zhang, X.J., Heinz, D.W., Blaber, M., Baldwin, E.P., & Matthews, B.W. (1992) *Science* 255, 178–183.
153. Otzen, D.E., Rheinhecker, M., & Fersht, A.R. (1995) *Biochemistry* 34, 13051–13058.
154. Pace, C.N. (1992) *J. Mol. Biol.* 226, 29–35.
155. Pace, C.N. (2001) *Biochemistry* 40, 310–313.
156. Holder, J.B., Bennett, A.F., Chen, J., Spencer, D.S., Byrne, M.P., & Stites, W.E. (2001) *Biochemistry* 40, 13998–14003.
157. Mendel, D., Ellman, J.A., Chang, Z., Veenstra, D.L., Kollman, P.A., & Schultz, P.G. (1992) *Science* 256, 1798–1802.
158. Shortle, D., Stites, W.E., & Meeker, A.K. (1990) *Biochemistry* 29, 8033–8041.
159. Kellis, J.T., Jr., Nyberg, K., & Fersht, A.R. (1989) *Biochemistry* 28, 4914–4922.
160. Kellis, J.T., Jr., Nyberg, K., Sali, D., & Fersht, A.R. (1988) *Nature* 333, 784–786.
161. Jackson, S.E., Moracci, M., elMasry, N., Johnson, C.M., & Fersht, A.R. (1993) *Biochemistry* 32, 11259–11269.
162. Southall, N.T., Dill, K.A., & Haymet, A.D.J. (2002) *J. Phys. Chem., B* 106, 521–533.
163. Dao-pin, S., Anderson, D.E., Baase, W.A., Dahlquist, F.W., & Matthews, B.W. (1991) *Biochemistry* 30, 11521–11529.
164. De Vos, S., Backmann, J., Prevost, M., Steyaert, J., & Loris, R. (2001) *Biochemistry* 40, 10140–10149.
165. Takano, K., Yamagata, Y., & Yutani, K. (2001) *Biochemistry* 40, 4853–4858.
166. Schwehm, J.M., Kristyenne, E.S., Biggers, C.C., & Stites, W.E. (1998) *Biochemistry* 37, 6939–6948.
167. Cornish, V.W., Kaplan, M.I., Veenstra, D.L., Kollman, P.A., & Schultz, P.G. (1994) *Biochemistry* 33, 12022–12031.
168. Wistow, G., Turnell, B., Summers, L., Slingsby, C., Moss, D., Miller, L., Lindley, P., & Blundell, T. (1983) *J. Mol. Biol.* 170, 175–202.
169. Burley, S.K., & Petsko, G.A. (1985) *Science* 229, 23–28.
170. Jorgensen, W.L., & Severance, D.L. (1990) *J. Am. Chem. Soc.* 112, 4768–4774.
171. Pai, E.F., Kabsch, W., Krengel, U., Holmes, K.C., John, J., & Wittinghofer, A. (1989) *Nature* 341, 209–214.
172. Hine, J., & Mookerjee, P.K. (1975) *J. Org. Chem.* 40, 292–298.
173. Eisenberg, D., & McLachlan, A.D. (1986) *Nature* 319, 199–203.
174. Fauchere, J.L., & Pliska, V. (1983) *Eur. J. Med. Chem. Chim. Ther.* 18, 369–375.
175. Nozaki, Y., & Tanford, C. (1971) *J. Biol. Chem.* 246, 2211–2217.
176. McPhalen, C.A., & James, M.N. (1987) *Biochemistry* 26, 261–269.
177. Hyde, C.C., Ahmed, S.A., Padlan, E.A., Miles, E.W., & Davies, D.R. (1988) *J. Biol. Chem.* 263, 17857–17871.
178. Choinowski, T., Hauser, H., & Piontek, K. (2000) *Biochemistry* 39, 1897–1902.
179. Bondi, A. (1964) *J. Phys. Chem.* 68, 441–451.
180. Richards, F.M. (1974) *J. Mol. Biol.* 82, 1–14.
181. Chothia, C. (1975) *Nature* 254, 304–308.
182. Li, A.J., & Nussinov, R. (1998) *Proteins: Struct., Funct., Genet.* 32, 111–127.
183. Tsai, J., Taylor, R., Chothia, C., & Gerstein, M. (1999) *J. Mol. Biol.* 290, 253–266.
184. Chalikian, T.V., Totrov, M., Abagyan, R., & Breslauer, K.J. (1996) *J. Mol. Biol.* 260, 588–603.
185. Jung, K., Jung, H., Colacurcio, P., & Kaback, H.R. (1995) *Biochemistry* 34, 1030–1039.
186. Klapper, M.H. (1971) *Biochim. Biophys. Acta* 229, 557–566.
187. Prieve, A., Almagor, A., Yedgar, S., & Gavish, B. (1996) *Biochemistry* 35, 2061–2066.
188. Cohn, E.J., McMeekin, T.L., Edsall, J.T., & Blanchard, M.H. (1934) *J. Am. Chem. Soc.* 56, 784–794.
189. Cohn, E.J., & Edsall, J.T. (1943) *Proteins, Amino Acids and Peptides as Ions and Dipolar Ions*, Reinhold, New York.
190. Traube, J. (1899) *Samml. Chem. Chem. Tech. Vortr.* 4, 255–331.
191. Takusagawa, F., & Kamitori, S. (1996) *J. Am. Chem. Soc.* 118, 8945–8946.
192. Taylor, W.R. (2000) *Nature* 406, 916–919.
193. Saper, M.A., Bjorkman, P.J., & Wiley, D.C. (1991) *J. Mol. Biol.* 219, 277–319.
194. Stewart-Jones, G.B., McMichael, A.J., Bell, J.I., Stuart, D.I., & Jones, E.Y. (2003) *Nat. Immunol.* 4, 657–663.
195. Chothia, C., Levitt, M., & Richardson, D. (1977) *Proc. Natl. Acad. Sci. U.S.A.* 74, 4130–4134.
196. Klug, A., Crick, F.H.C., & Wyckoff, H.W. (1958) *Acta Crystallogr.* 11, 199–213.
197. Chothia, C., Levitt, M., & Richardson, D. (1981) *J. Mol. Biol.* 145, 215–250.

198. Betts, L., Xiang, S., Short, S.A., Wolfenden, R., & Carter, C.W., Jr. (1994) *J. Mol. Biol.* 235, 635–656.
199. Chothia, C., & Finkelstein, A.V. (1990) *Annu. Rev. Biochem.* 59, 1007–1039.
200. Doolittle, R.F., Goldbaum, D.M., & Doolittle, L.R. (1978) *J. Mol. Biol.* 120, 311–325.
201. Crick, F.H.C. (1953) *Acta Crystallogr.* 6, 689–697.
202. Whitby, F.G., Kent, H., Stewart, F., Stewart, M., Xie, X., Hatch, V., Cohen, C., & Phillips, G.N., Jr. (1992) *J. Mol. Biol.* 227, 441–452.
203. Brown, J.H., Kim, K.H., Jun, G., Greenfield, N.J., Dominguez, R., Volkmann, N., Hitchcock-DeGregori, S.E., & Cohen, C. (2001) *Proc. Natl. Acad. Sci. U.S.A.* 98, 8496–8501.
204. O'Shea, E.K., Klemm, J.D., Kim, P.S., & Alber, T. (1991) *Science* 254, 539–544.
205. Harbury, P.B., Zhang, T., Kim, P.S., & Alber, T. (1993) *Science* 262, 1401–1407.
206. Phillips, G.N., Jr. (1992) *Proteins: Struct., Funct., Genet.* 14, 425–429.
207. Chen, L., Glover, J.N., Hogan, P.G., Rao, A., & Harrison, S.C. (1998) *Nature* 392, 42–48.
208. Cusack, S., Berthet-Colominas, C., Hartlein, M., Nassar, N., & Leberman, R. (1990) *Nature* 347, 249–255.
209. Harbury, P.B., Kim, P.S., & Alber, T. (1994) *Nature* 371, 80–83.
210. Shu, W., Liu, J., Ji, H., & Lu, M. (2000) *J. Mol. Biol.* 299, 1101–1112.
211. Lovejoy, B., Choe, S., Cascio, D., McRorie, D.K., DeGrado, W.F., & Eisenberg, D. (1993) *Science* 259, 1288–1293.
212. Pascual, J., Pfuhl, M., Walther, D., Saraste, M., & Nilges, M. (1997) *J. Mol. Biol.* 273, 740–751.
213. Tarbouriech, N., Curran, J., Ruigrok, R.W., & Burmeister, W.P. (2000) *Nat. Struct. Biol.* 7, 777–781.
214. Bowman, G.D., Nodelman, I.M., Levy, O., Lin, S.L., Tian, P., Zamb, T.J., Udem, S.A., Venkataraghavan, B., & Schutt, C.E. (2000) *J. Mol. Biol.* 304, 861–871.
215. Ultsch, M.H., Somers, W., Kossiakoff, A.A., & de Vos, A.M. (1994) *J. Mol. Biol.* 236, 286–299.
216. Vlassi, M., Steif, C., Weber, P., Tsernoglou, D., Wilson, K.S., Hinz, H.J., & Kokkinidis, M. (1994) *Nat. Struct. Biol.* 1, 706–716.
217. Malashkevich, V.N., Kammerer, R.A., Efimov, V.P., Schulthess, T., & Engel, J. (1996) *Science* 274, 761–765.
218. Weaver, T.M., Levitt, D.G., Donnelly, M.I., Stevens, P.P., & Banaszak, L.J. (1995) *Nat. Struct. Biol.* 2, 654–662.
219. Calladine, C.R., Sharff, A., & Luisi, B. (2001) *J. Mol. Biol.* 305, 603–618.
220. Milburn, M.V., Prive, G.G., Milligan, D.L., Scott, W.G., Yeh, J., Jancarik, J., Koshland, D.E., Jr., & Kim, S.H. (1991) *Science* 254, 1342–1347.
221. Stetefeld, J., Jenny, M., Schulthess, T., Landwehr, R., Engel, J., & Kammerer, R.A. (2000) *Nat. Struct. Biol.* 7, 772–776.
222. Bailey, K., Astbury, W.T., & Ruddall, K.M. (1943) *Nature* 151, 716–717.
223. Heimbürg, T., Schunemann, J., Weber, K., & Geisler, N. (1999) *Biochemistry* 38, 12727–12734.
224. Hecht, M.H., Richardson, J.S., Richardson, D.C., & Ogden, R.C. (1990) *Science* 249, 884–891.
225. Monera, O.D., Kay, C.M., & Hodges, R.S. (1994) *Biochemistry* 33, 3862–3871.
226. Varughese, K.I., Skinner, M.M., Whiteley, J.M., Matthews, D.A., & Xuong, N.H. (1992) *Proc. Natl. Acad. Sci. U.S.A.* 89, 6080–6084.
227. Nordlund, P., & Eklund, H. (1993) *J. Mol. Biol.* 232, 123–164.
228. Li, J.D., Carroll, J., & Ellar, D.J. (1991) *Nature* 353, 815–821.
229. Li, H., Dunn, J.J., Luft, B.J., & Lawson, C.L. (1997) *Proc. Natl. Acad. Sci. U.S.A.* 94, 3584–3589.
230. Huang, X., Nakagawa, T., Tamura, A., Link, K., Koide, A., & Koide, S. (2001) *J. Mol. Biol.* 308, 367–375.
231. Spadaccini, R., Crescenzi, O., Tancredi, T., De Casamassimi, N., Saviano, G., Scognamiglio, R., Di Donato, A., & Temussi, P.A. (2001) *J. Mol. Biol.* 305, 505–514.
232. Koide, S., Huang, X., Link, K., Koide, A., Bu, Z., & Engelman, D.M. (2000) *Nature* 403, 456–460.
233. Cohen, F.E., Sternberg, M.J.E., & Taylor, W.R. (1981) *J. Mol. Biol.* 148, 253–272.
234. Liao, D.I., Kapadia, G., Reddy, P., Saier, M.H., Jr., Reizer, J., & Herzberg, O. (1991) *Biochemistry* 30, 9583–9594.
235. Chothia, C., & Janin, J. (1981) *Proc. Natl. Acad. Sci. U.S.A.* 78, 4146–4150.
236. Matthews, D.A., Appelt, K., & Oatley, S.J. (1989) *J. Mol. Biol.* 205, 449–454.
237. Hrabal, R., Chen, Z., James, S., Bennett, H.P., & Ni, F. (1996) *Nat. Struct. Biol.* 3, 747–752.
238. Chothia, C., & Janin, J. (1982) *Biochemistry* 21, 3955–3965.
239. Xu, Z., Bernlohr, D.A., & Banaszak, L.J. (1992) *Biochemistry* 31, 3484–3492.
240. Jones, T.A., Bergfors, T., Sedzik, J., & Unge, T. (1988) *EMBO J.* 7, 1597–1604.
241. Cowan, S.W., Newcomer, M.E., & Jones, T.A. (1990) *Proteins: Struct., Funct., Genet.* 8, 44–61.
242. Yarbrough, D., Wachter, R.M., Kallio, K., Matz, M.V., & Remington, S.J. (2001) *Proc. Natl. Acad. Sci. U.S.A.* 98, 462–467.
243. Janin, J., & Chothia, C. (1980) *J. Mol. Biol.* 143, 95–128.
244. Rabijns, A., De Bondt, H.L., & De Ranter, C. (1997) *Nat. Struct. Biol.* 4, 357–360.
245. Beamer, L.J., Carroll, S.F., & Eisenberg, D. (1997) *Science* 276, 1861–1864.
246. Uhlin, U., & Eklund, H. (1994) *Nature* 370, 533–539.
247. Van Petegem, F., Contreras, H., Contreras, R., & Van Beeumen, J. (2001) *J. Mol. Biol.* 312, 157–165.
248. Gassner, N.C., Baase, W.A., & Matthews, B.W. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 12155–12158.
249. Liu, R., Baase, W.A., & Matthews, B.W. (2000) *J. Mol. Biol.* 295, 127–145.
250. Baldwin, E., Xu, J., Hajiseyedi, O., Baase, W.A., & Matthews, B.W. (1996) *J. Mol. Biol.* 259, 542–559.
251. Hamilton, J.A., Steinrauf, L.K., Braden, B.C., Liepnieks, J., Benson, M.D., Holmgren, G., Sandgren, O., & Steen, L. (1993) *J. Biol. Chem.* 268, 2416–2424.
252. Lim, W.A., Farruggio, D.C., & Sauer, R.T. (1992) *Biochemistry* 31, 4324–4333.
253. Martensson, L.G., Jonsson, B.H., Andersson, M.,

- Kihlgren, A., Bergenhem, N., & Carlsson, U. (1992) *Biochim. Biophys. Acta* 1118, 179–186.
254. Collyer, C.A., Guss, J.M., Sugimura, Y., Yoshizaki, F., & Freeman, H.C. (1990) *J. Mol. Biol.* 211, 617–632.
 255. Connolly, M.L. (1983) *Science* 221, 709–713.
 256. Xu, J., Baase, W.A., Baldwin, E., & Matthews, B.W. (1998) *Protein Sci.* 7, 158–177.
 257. Bruns, C.M., & Karplus, P.A. (1995) *J. Mol. Biol.* 247, 125–145.
 258. Morton, A., Baase, W.A., & Matthews, B.W. (1995) *Biochemistry* 34, 8564–8575.
 259. Buckle, A.M., Cramer, P., & Fersht, A.R. (1996) *Biochemistry* 35, 4298–4305.
 260. Varadarajan, R., & Richards, F.M. (1992) *Biochemistry* 31, 12315–12327.
 261. McRee, D.E., Redford, S.M., Getzoff, E.D., Lepock, J.R., Halleswell, R.A., & Tainer, J.A. (1990) *J. Biol. Chem.* 265, 14234–14241.
 262. Lim, W.A., & Sauer, R.T. (1989) *Nature* 339, 31–36.
 263. Behe, M.J., Lattman, E.E., & Rose, G.D. (1991) *Proc. Natl. Acad. Sci. U.S.A.* 88, 4195–4199.
 264. Hanukoglu, I., & Fuchs, E. (1983) *Cell* 33, 915–924.
 265. Contreras-Martel, C., Martinez-Oyanedel, J., Bunster, M., Legrand, P., Piras, C., Vernede, X., & Fontecilla-Camps, J.C. (2001) *Acta Crystallogr., D* 57, 52–60.
 266. Charron, C., Kadri, A., Robert, M.C., Giege, R., & Lorber, B. (2002) *Acta Crystallogr., D* 58, 2060–2065.
 267. Bjorkman, A.J., Binnie, R.A., Zhang, H., Cole, L.B., Hermodson, M.A., & Mowbray, S.L. (1994) *J. Biol. Chem.* 269, 30206–30211.
 268. Matthews, B.W. (1977) in *The Proteins*, 3rd Ed. (Neurath, H., & Hill, R.L., Eds.) Vol. III, pp 404–590, Academic Press, New York.
 269. Blake, C.C., Pulford, W.C., & Artymiuk, P.J. (1983) *J. Mol. Biol.* 167, 693–723.
 270. Noguchi, S., Satow, Y., Uchida, T., Sasaki, C., & Matsuzaki, T. (1995) *Biochemistry* 34, 15583–15591.
 271. Levitt, M., & Park, B.H. (1993) *Structure* 1, 223–226.
 272. Tanaka, N., Arai, J., Inokuchi, N., Koyama, T., Ohgi, K., Irie, M., & Nakamura, K.T. (2000) *J. Mol. Biol.* 298, 859–873.
 273. Katti, S.K., LeMaster, D.M., & Eklund, H. (1990) *J. Mol. Biol.* 212, 167–184.
 274. Knox, J.R., & Moews, P.C. (1991) *J. Mol. Biol.* 220, 435–455.
 275. Malin, R., Zielenkiewicz, P., & Saenger, W. (1991) *J. Biol. Chem.* 266, 4848–4852.
 276. Carrell, C.J., Schlarb, B.G., Bendall, D.S., Howe, C.J., Cramer, W.A., & Smith, J.L. (1999) *Biochemistry* 38, 9590–9599.
 277. Lu, G., Lindqvist, Y., Schneider, G., Dwivedi, U., & Campbell, W. (1995) *J. Mol. Biol.* 248, 931–948.
 278. Otting, G., Liepinsh, E., & Wuthrich, K. (1991) *Science* 254, 974–980.
 279. Denisov, V.P., & Halle, B. (1995) *J. Mol. Biol.* 245, 682–697.
 280. Furey, W., Wang, B.C., Yoo, C.S., & Sax, M. (1983) *J. Mol. Biol.* 167, 661–692.
 281. Ramaswamy, S., Eklund, H., & Plapp, B.V. (1994) *Biochemistry* 33, 5230–5237.
 282. Benning, M.M., Smith, A.F., Wells, M.A., & Holden, H.M. (1992) *J. Mol. Biol.* 228, 208–219.
 283. van Aalten, D.M., Synstad, B., Brurberg, M.B., Hough, E., Riise, B.W., Eijsink, V.G., & Wierenga, R.K. (2000) *Proc. Natl. Acad. Sci. U.S.A.* 97, 5842–5847.
 284. Denisov, V.P., Peters, J., Horlein, H.D., & Halle, B. (1996) *Nat. Struct. Biol.* 3, 505–509.
 285. Leslie, A.G., Moody, P.C., & Shaw, W.V. (1988) *Proc. Natl. Acad. Sci. U.S.A.* 85, 4133–4137.
 286. Housset, D., Habersetzer-Rochat, C., Astier, J.P., & Fontecilla-Camps, J.C. (1994) *J. Mol. Biol.* 238, 88–103.
 287. Fontecilla-Camps, J.C., Habersetzer-Rochat, C., & Rochat, H. (1988) *Proc. Natl. Acad. Sci. U.S.A.* 85, 7443–7447.
 288. Yu, B., Blaber, M., Gronenborn, A.M., Clore, G.M., & Caspar, D.L. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 103–108.
 289. Ernst, J.A., Clubb, R.T., Zhou, H.X., Gronenborn, A.M., & Clore, G.M. (1995) *Science* 267, 1813–1817.
 290. Schrag, J.D., & Cygler, M. (1993) *J. Mol. Biol.* 230, 575–591.
 291. Nakasako, M. (1999) *J. Mol. Biol.* 289, 547–564.
 292. Teeter, M.M. (1984) *Proc. Natl. Acad. Sci. U.S.A.* 81, 6014–6018.
 293. Kielkopf, C.L., Ding, S., Kuhn, P., & Rees, D.C. (2000) *J. Mol. Biol.* 296, 787–801.
 294. Watenpaugh, K.D., Margulis, T.N., Sieker, L.C., & Jensen, L.H. (1978) *J. Mol. Biol.* 122, 175–190.
 295. Thanki, N., Thornton, J.M., & Goodfellow, J.M. (1988) *J. Mol. Biol.* 202, 637–657.
 296. Jiang, J.S., & Brunger, A.T. (1994) *J. Mol. Biol.* 243, 100–115.
 297. Svergun, D.I., Richard, S., Koch, M.H., Sayers, Z., Kuprin, S., & Zaccai, G. (1998) *Proc. Natl. Acad. Sci. U.S.A.* 95, 2267–2272.
 298. Denisov, V.P., & Halle, B. (1994) *J. Am. Chem. Soc.* 116, 10324–10325.
 299. Scanlon, W.J., & Eisenberg, D. (1975) *J. Mol. Biol.* 98, 485–502.
 300. Fisher, H.F. (1965) *Biochim. Biophys. Acta* 109, 544–550.
 301. Grant, E.H., Mitton, B.G., South, G.P., & Sheppard, R.J. (1974) *Biochem. J.* 139, 375–380.
 302. McMeekin, T.L., Groves, M.L., & Hipp, N.J. (1954) *J. Am. Chem. Soc.* 76, 407–413.
 303. Bull, H.B., & Breese, K. (1968) *Arch. Biochem. Biophys.* 128, 488–496.
 304. Arakawa, T., & Timasheff, S.N. (1982) *Biochemistry* 21, 6536–6544.
 305. Lee, J.C., & Timasheff, S.N. (1981) *J. Biol. Chem.* 256, 7193–7201.
 306. Mc Clure, R.J., & Craven, B.M. (1974) *J. Mol. Biol.* 83, 551–555.
 307. Kuntz, I.D., Brassfield, T.S., Law, G.D., & Purcell, G.V. (1969) *Science* 163, 1329–1331.
 308. Tanford, C. (1961) *Physical Chemistry of Macromolecules*, John Wiley, New York.
 309. Kumosinski, T.F., & Pessen, H. (1982) *Arch. Biochem. Biophys.* 219, 89–100.
 310. Wang, J.H. (1954) *J. Am. Chem. Soc.* 76, 4755–4763.
 311. Oncley, J.L. (1943) in *Proteins, Amino Acids, and Peptides* (Cohn, E.J., & Edsall, J.T., Eds.) pp 543–568, Reinhold, New York.

312. Usha, M.G., & Wittebort, R.J. (1989) *J. Mol. Biol.* 208, 669–678.
313. Bull, H.B., & Breese, K. (1968) *Arch. Biochem. Biophys.* 128, 497–502.
314. Miller, S., Lesk, A.M., Janin, J., & Chothia, C. (1987) *Nature* 328, 834–836.
315. Vossen, K.M., Wolz, R., Daugherty, M.A., & Fried, M.G. (1997) *Biochemistry* 36, 11640–11647.
316. Dzingeleski, G.D., & Wolfenden, R. (1993) *Biochemistry* 32, 9143–9147.
317. Colombo, M.F., Rau, D.C., & Parsegian, V.A. (1992) *Science* 256, 655–659.
318. Rand, R.P., Fuller, N.L., Butko, P., Francis, G., & Nicholls, P. (1993) *Biochemistry* 32, 5925–5929.
319. Sun, D.P., Sauer, U., Nicholson, H., & Matthews, B.W. (1991) *Biochemistry* 30, 7142–7153.
320. Sun, D.P., Soderlind, E., Baase, W.A., Wozniak, J.A., Sauer, U., & Matthews, B.W. (1991) *J. Mol. Biol.* 221, 873–887.
321. Loladze, V.V., Ibarra-Molero, B., Sanchez-Ruiz, J.M., & Makhatadze, G.I. (1999) *Biochemistry* 38, 16419–16423.
322. Spector, S., Wang, M., Carp, S.A., Robblee, J., Hendsch, Z.S., Fairman, R., Tidor, B., & Raleigh, D.P. (2000) *Biochemistry* 39, 872–879.
323. Escobar, L., Root, M.J., & MacKinnon, R. (1993) *Biochemistry* 32, 6982–6987.
324. Imoto, K., Busch, C., Sakmann, B., Mishina, M., Konno, T., Nakai, J., Bujo, H., Mori, Y., Fukuda, K., & Numa, S. (1988) *Nature* 335, 645–648.
325. Stocker, M., & Miller, C. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 9509–9513.
326. Rodgers, K.K., Pochapsky, T.C., & Sligar, S.G. (1988) *Science* 240, 1657–1659.
327. Stayton, P.S., & Sligar, S.G. (1990) *Biochemistry* 29, 7381–7386.
328. Fujimori, K., Sorenson, M., Herzberg, O., Moulton, J., & Reinach, F.C. (1990) *Nature* 345, 182–184.
329. Russell, A.J., Thomas, P.G., & Fersht, A.R. (1987) *J. Mol. Biol.* 193, 803–813.
330. Doolittle, R.F. (1981) *Science* 214, 149–159.
331. Tanford, C. (1962) *Adv. Protein Chem.* 17, 69–165.
332. Crammer, J.L., & Neuberger, A. (1943) *Biochem. J.* 37, 302–310.
333. Lenstra, J.A., Bolscher, B.G., Beintema, J.J., & Kaptein, R. (1979) *Eur. J. Biochem.* 98, 385–397.
334. Matthew, J.B., & Richards, F.M. (1982) *Biochemistry* 21, 4989–4999.
335. Giletto, A., & Pace, C.N. (1999) *Biochemistry* 38, 13379–13384.
336. Aberg, A., Nordlund, P., & Eklund, H. (1993) *Nature* 361, 276–278.
337. Hecht, H.J., Kalisz, H.M., Hendle, J., Schmid, R.D., & Schomburg, D. (1993) *J. Mol. Biol.* 229, 153–172.
338. Gibbs, M.R., Moody, P.C., & Leslie, A.G. (1990) *Biochemistry* 29, 11261–11265.
339. Leslie, A.G. (1990) *J. Mol. Biol.* 213, 167–186.
340. Cheng, X.D., & Schoenborn, B.P. (1991) *J. Mol. Biol.* 220, 381–399.
341. Tainer, J.A., Getzoff, E.D., Beem, K.M., Richardson, J.S., & Richardson, D.C. (1982) *J. Mol. Biol.* 160, 181–217.
342. James, M.N., & Sielecki, A.R. (1986) *Nature* 319, 33–38.
343. Zheng, Y.J., & Ornstein, R.L. (1996) *J. Am. Chem. Soc.* 118, 11237–11243.
344. Simonson, T., & Brooks, C.L. (1996) *J. Am. Chem. Soc.* 118, 8452–8458.
345. Parsegian, A. (1969) *Nature* 221, 844–846.
346. Tissot, A.C., Vuilleumier, S., & Fersht, A.R. (1996) *Biochemistry* 35, 6786–6794.
347. Hendsch, Z.S., & Tidor, B. (1994) *Protein Sci.* 3, 211–226.
348. Waldburger, C.D., Jonsson, T., & Sauer, R.T. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 2629–2634.
349. Baker, E.N. (1988) *J. Mol. Biol.* 203, 1071–1095.
350. Baud, F., & Karlin, S. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 12494–12499.
351. Johnson, M.S., & Overington, J.P. (1993) *J. Mol. Biol.* 233, 716–738.
352. Schirmer, T., Huber, R., Schneider, M., Bode, W., Miller, M., & Hackert, M.L. (1986) *J. Mol. Biol.* 188, 651–676.
353. Warshel, A. (1987) *Nature* 330, 15–16.
354. Ippolito, J.A., Alexander, R.S., & Christianson, D.W. (1990) *J. Mol. Biol.* 215, 457–471.
355. Yao, N., Trakhanov, S., & Quioco, F.A. (1994) *Biochemistry* 33, 4769–4779.
356. Murray-Rust, J., Leiper, J., McAlister, M., Phelan, J., Tilley, S., Santa Maria, J., Vallance, P., & McDonald, N. (2001) *Nat. Struct. Biol.* 8, 679–683.
357. Steiner, T., & Koellner, G. (2001) *J. Mol. Biol.* 305, 535–557.
358. Burley, S.K., & Petsko, G.A. (1986) *FEBS Lett.* 203, 139–143.
359. Gallivan, J.P., & Dougherty, D.A. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 9459–9464.
360. Mitchell, J.B., Nandi, C.L., McDonald, I.K., Thornton, J.M., & Price, S.L. (1994) *J. Mol. Biol.* 239, 315–331.
361. Knight, S., Andersson, I., & Branden, C.I. (1990) *J. Mol. Biol.* 215, 113–160.
362. McDonald, I.K., & Thornton, J.M. (1994) *J. Mol. Biol.* 238, 777–793.
363. Chattopadhyaya, R., Meador, W.E., Means, A.R., & Quioco, F.A. (1992) *J. Mol. Biol.* 228, 1177–1192.
364. Kyte, J., & Doolittle, R.F. (1982) *J. Mol. Biol.* 157, 105–132.
365. Dardel, F., Davis, A.L., Laue, E.D., & Perham, R.N. (1993) *J. Mol. Biol.* 229, 1037–1048.
366. Ludwig, M.L., Metzger, A.L., Patridge, K.A., & Stallings, W.C. (1991) *J. Mol. Biol.* 219, 335–358.
367. Bode, W., Walter, J., Huber, R., Wenzel, H.R., & Tschesche, H. (1984) *Eur. J. Biochem.* 144, 185–190.
368. Lah, M.S., Palfey, B.A., Schreuder, H.A., & Ludwig, M.L. (1994) *Biochemistry* 33, 1555–1564.
369. Khan, A.R., Parrish, J.C., Fraser, M.E., Smith, W.W., Bartlett, P.A., & James, M.N. (1998) *Biochemistry* 37, 16839–16845.
370. Jacobson, B.L., & Quioco, F.A. (1988) *J. Mol. Biol.* 204, 783–787.
371. Tanford, C. (1954) *J. Am. Chem. Soc.* 76, 945–946.
372. Phillips, S.E. (1980) *J. Mol. Biol.* 142, 531–554.
373. Myers, J.K., & Pace, C.N. (1996) *Biophys. J.* 71, 2033–2039.
374. Carter, P.J., Winter, G., Wilkinson, A.J., & Fersht, A.R. (1984) *Cell* 38, 835–840.

375. Serrano, L., Horovitz, A., Avron, B., Bycroft, M., & Fersht, A.R. (1990) *Biochemistry* 29, 9343–9352.
376. Heinz, D.W., Baase, W.A., & Matthews, B.W. (1992) *Proc. Natl. Acad. Sci. U.S.A.* 89, 3751–3755.
377. Horovitz, A., Serrano, L., Avron, B., Bycroft, M., & Fersht, A.R. (1990) *J. Mol. Biol.* 216, 1031–1044.
378. Strop, P., & Mayo, S.L. (2000) *Biochemistry* 39, 1251–1255.
379. Myers, J.K., & Oas, T.G. (1999) *Biochemistry* 38, 6761–6768.
380. Schreiber, G., & Fersht, A.R. (1995) *J. Mol. Biol.* 248, 478–486.
381. Albeck, S., Unger, R., & Schreiber, G. (2000) *J. Mol. Biol.* 298, 503–520.
382. Tronrud, D.E., Holden, H.M., & Matthews, B.W. (1987) *Science* 235, 571–574.
383. Morgan, B.P., Scholtz, J.M., Ballinger, M.D., Zipkin, I.D., & Bartlett, P.A. (1991) *J. Am. Chem. Soc.* 113, 297–307.
384. Koh, J.T., Cornish, V.W., & Schultz, P.G. (1997) *Biochemistry* 36, 11314–11322.
385. Chapman, E., Thorson, J.S., & Schultz, P.G. (1997) *J. Am. Chem. Soc.* 119, 7151–7152.
386. Thorson, J.S., Chapman, E., Murphy, E.C., Schultz, P.G., & Judice, J.K. (1995) *J. Am. Chem. Soc.* 117, 1157–1158.
387. Hennig, M., & Geierstanger, B.H. (1999) *J. Am. Chem. Soc.* 121, 5123–5126.
388. Kurokawa, H., Mikami, B., & Hirose, M. (1995) *J. Mol. Biol.* 254, 196–207.
389. Dewan, J.C., Mikami, B., Hirose, M., & Sacchettini, J.C. (1993) *Biochemistry* 32, 11963–11968.
390. He, Q.Y., Mason, A.B., Tam, B.M., MacGillivray, R.T., & Woodworth, R.C. (1999) *Biochemistry* 38, 9704–9711.
391. MacGillivray, R.T., Bewley, M.C., Smith, C.A., He, Q.Y., Mason, A.B., Woodworth, R.C., & Baker, E.N. (2000) *Biochemistry* 39, 1211–1216.
392. Nurizzo, D., Baker, H.M., He, Q.Y., MacGillivray, R.T., Mason, A.B., Woodworth, R.C., & Baker, E.N. (2001) *Biochemistry* 40, 1616–1623.
393. Loh, S.N., & Markley, J.L. (1994) *Biochemistry* 33, 1029–1036.
394. Bowers, P.M., & Klevit, R.E. (1996) *Nat. Struct. Biol.* 3, 522–531.
395. Loh, S.N., & Markley, J.L. (1993) in *Techniques in Protein Chemistry IV* (Angeletti, R. H., Ed.) pp 517–524, Academic Press, San Diego, CA.
396. Bowers, P.M., & Klevit, R.E. (2000) *J. Am. Chem. Soc.* 122, 1030–1033.
397. Khare, D., Alexander, P., & Orban, J. (1999) *Biochemistry* 38, 3918–3925.
398. Dauter, Z., Wilson, K.S., Sieker, L.C., Meyer, J., & Moulis, J.M. (1997) *Biochemistry* 36, 16065–16073.
399. Kuhn, P., Knapp, M., Soltis, S.M., Ganshaw, G., Thoenes, M., & Bott, R. (1998) *Biochemistry* 37, 13446–13452.
400. Morales, R., Chron, M.H., Hudry-Clergeon, G., Petillot, Y., Norager, S., Medina, M., & Frey, M. (1999) *Biochemistry* 38, 15764–15773.
401. Schwartz, B., Drueckhammer, D.G., Usher, K.C., & Remington, S.J. (1995) *Biochemistry* 34, 15459–15466.
402. Klimasauskas, S., Kumar, S., Roberts, R.J., & Cheng, X. (1994) *Cell* 76, 357–369.
403. Slupphaug, G., Mol, C.D., Kavli, B., Arvai, A.S., Krokan, H.E., & Tainer, J.A. (1996) *Nature* 384, 87–92.
404. Horton, J.R., Nastri, H.G., Riggs, P.D., & Cheng, X. (1998) *J. Mol. Biol.* 284, 1491–1504.
405. Shakked, Z., Guzikevich-Guerstein, G., Frolov, F., Rabinovich, D., Joachimiak, A., & Sigler, P.B. (1994) *Nature* 368, 469–473.
406. Nadassy, K., Wodak, S.J., & Janin, J. (1999) *Biochemistry* 38, 1999–2017.
407. Huai, Q., Colandene, J.D., Topal, M.D., & Ke, H. (2001) *Nat. Struct. Biol.* 8, 665–669.
408. Pelletier, H., Sawaya, M.R., Kumar, A., Wilson, S.H., & Kraut, J. (1994) *Science* 264, 1891–1903.
409. Mondragon, A., & Harrison, S.C. (1991) *J. Mol. Biol.* 219, 321–334.
410. Albright, R.A., & Matthews, B.W. (1998) *J. Mol. Biol.* 280, 137–151.
411. Redinbo, M.R., Stewart, L., Kuhn, P., Champoux, J.J., & Hol, W.G. (1998) *Science* 279, 1504–1513.
412. Weston, S.A., Lahm, A., & Suck, D. (1992) *J. Mol. Biol.* 226, 1237–1256.
413. Kostrewa, D., & Winkler, F.K. (1995) *Biochemistry* 34, 683–696.
414. Jordan, S.R., & Pabo, C.O. (1988) *Science* 242, 893–899.
415. Beamer, L.J., & Pabo, C.O. (1992) *J. Mol. Biol.* 227, 177–196.
416. Arents, G., & Moudrianakis, E.N. (1993) *Proc. Natl. Acad. Sci. U.S.A.* 90, 10489–10493.
417. Glover, J.N., & Harrison, S.C. (1995) *Nature* 373, 257–261.
418. Mo, Y., Vaessen, B., Johnston, K., & Marmorstein, R. (2000) *Nat. Struct. Biol.* 7, 292–297.
419. Mo, Y., Ho, W., Johnston, K., & Marmorstein, R. (2001) *J. Mol. Biol.* 314, 495–506.
420. Somers, W.S., & Phillips, S.E. (1992) *Nature* 359, 387–393.
421. Kamada, K., Horiuchi, T., Ohsumi, K., Shimamoto, N., & Morikawa, K. (1996) *Nature* 383, 598–603.
422. Raumann, B.E., Rould, M.A., Pabo, C.O., & Sauer, R.T. (1994) *Nature* 367, 754–757.
423. Deibert, M., Grazulis, S., Sasnauskas, G., Siksnys, V., & Huber, R. (2000) *Nat. Struct. Biol.* 7, 792–799.
424. Pavletich, N.P., & Pabo, C.O. (1991) *Science* 252, 809–817.
425. Otwinowski, Z., Schevitz, R.W., Zhang, R.G., Lawson, C.L., Joachimiak, A., Marmorstein, R.Q., Luisi, B.F., & Sigler, P.B. (1988) *Nature* 335, 321–329.
426. Batchelor, A.H., Piper, D.E., de la Brousse, F.C., McKnight, S.L., & Wolberger, C. (1998) *Science* 279, 1037–1041.
427. Muller, C.W., Rey, F.A., Sodeoka, M., Verdine, G.L., & Harrison, S.C. (1995) *Nature* 373, 311–317.
428. Ghosh, G., van Duyne, G., Ghosh, S., & Sigler, P.B. (1995) *Nature* 373, 303–310.
429. Hegde, R.S., Grossman, S.R., Laimins, L.A., & Sigler, P.B. (1992) *Nature* 359, 505–512.
430. Clarke, N.D., Beamer, L.J., Goldberg, H.R., Berkower, C., & Pabo, C.O. (1991) *Science* 254, 267–270.
431. Kissinger, C.R., Liu, B.S., Martin-Blanco, E., Kornberg, T.B., & Pabo, C.O. (1990) *Cell* 63, 579–590.
432. Fairall, L., Schwabe, J.W., Chapman, L., Finch, J.T., & Rhodes, D. (1993) *Nature* 366, 483–487.

433. Wuttke, D.S., Foster, M.P., Case, D.A., Gottesfeld, J.M., & Wright, P.E. (1997) *J. Mol. Biol.* 273, 183–206.
434. Kwon, H.J., Bennik, M.H., Demple, B., & Ellenberger, T. (2000) *Nat. Struct. Biol.* 7, 424–430.
435. Keller, W., Konig, P., & Richmond, T.J. (1995) *J. Mol. Biol.* 254, 657–667.
436. Hovde, S., Abate-Shen, C., & Geiger, J.H. (2001) *Biochemistry* 40, 12013–12021.
437. Reddy, C.K., Das, A., & Jayaram, B. (2001) *J. Mol. Biol.* 314, 619–632.
438. Luisi, B.F., Xu, W.X., Otwinowski, Z., Freedman, L.P., Yamamoto, K.R., & Sigler, P.B. (1991) *Nature* 352, 497–505.
439. Meinke, G., & Sigler, P.B. (1999) *Nat. Struct. Biol.* 6, 471–477.
440. Newman, M., Strzelecka, T., Dorner, L.F., Schildkraut, I., & Aggarwal, A.K. (1995) *Science* 269, 656–663.
441. Bell, C.E., & Lewis, M. (2000) *Nat. Struct. Biol.* 7, 209–214.
442. Wojciak, J.M., Iwahara, J., & Clubb, R.T. (2001) *Nat. Struct. Biol.* 8, 84–90.
443. Pellegrini, L., Tan, S., & Richmond, T.J. (1995) *Nature* 376, 490–498.
444. Santelli, E., & Richmond, T.J. (2000) *J. Mol. Biol.* 297, 437–449.
445. Kosa, P.F., Ghosh, G., DeDecker, B.S., & Sigler, P.B. (1997) *Proc. Natl. Acad. Sci. U.S.A.* 94, 6042–6047.
446. Kim, J.L., & Burley, S.K. (1994) *Nat. Struct. Biol.* 1, 638–653.
447. Starich, M.R., Wikstrom, M., Schumacher, S., Arst, H.N., Jr., Gronenborn, A.M., & Clore, G.M. (1998) *J. Mol. Biol.* 277, 621–634.
448. Cho, Y., Gorina, S., Jeffrey, P.D., & Pavletich, N.P. (1994) *Science* 265, 346–355.
449. Aggarwal, A.K., Rodgers, D.W., Drott, M., Ptashne, M., & Harrison, S.C. (1988) *Science* 242, 899–907.
450. Schumacher, M.A., Choi, K.Y., Zalkin, H., & Brennan, R.G. (1994) *Science* 266, 763–770.
451. Kim, J.L., Nikolov, D.B., & Burley, S.K. (1993) *Nature* 365, 520–527.
452. Kim, Y., Geiger, J.H., Hahn, S., & Sigler, P.B. (1993) *Nature* 365, 512–520.
453. Schultz, S.C., Shields, G.C., & Steitz, T.A. (1991) *Science* 253, 1001–1007.
454. Shimon, L.J., & Harrison, S.C. (1993) *J. Mol. Biol.* 232, 826–838.
455. Mol, C.D., Izumi, T., Mitra, S., & Tainer, J.A. (2000) *Nature* 403, 451–456.
456. Li, T., Stark, M.R., Johnson, A.D., & Wolberger, C. (1995) *Science* 270, 262–269.
457. Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F., & Richmond, T.J. (1997) *Nature* 389, 251–260.
458. Tan, S., & Richmond, T.J. (1998) *Nature* 391, 660–666.
459. Sierk, M.L., Zhao, Q., & Rastinejad, F. (2001) *Biochemistry* 40, 12833–12843.
460. Obmolova, G., Ban, C., Hsieh, P., & Yang, W. (2000) *Nature* 407, 703–710.
461. Lamers, M.H., Perrakis, A., Enzlin, J.H., Winterwerp, H.H., de Wind, N., & Sixma, T.K. (2000) *Nature* 407, 711–717.
462. Lesser, D.R., Kurpiewski, M.R., Waters, T., Connolly, B.A., & Jen-Jacobson, L. (1993) *Proc. Natl. Acad. Sci. U.S.A.* 90, 7548–7552.
463. King, D.A., Zhang, L., Guarente, L., & Marmorstein, R. (1999) *Nat. Struct. Biol.* 6, 64–71.
464. Lima, C.D., Wang, J.C., & Mondragon, A. (1994) *Nature* 367, 138–146.
465. Walker, J.R., Corpina, R.A., & Goldberg, J. (2001) *Nature* 412, 607–614.
466. Moarefi, I., Jeruzalmi, D., Turner, J., O'Donnell, M., & Kuriyan, J. (2000) *J. Mol. Biol.* 296, 1215–1223.
467. Soumillion, P., Sexton, D.J., & Benkovic, S.J. (1998) *Biochemistry* 37, 1819–1827.
468. Latham, G.J., Dong, F., Pietroni, P., Dozono, J.M., Bacheller, D.J., & von Hippel, P.H. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 12448–12453.
469. Bochkarev, A., Pfuetzner, R.A., Edwards, A.M., & Frappier, L. (1997) *Nature* 385, 176–181.
470. Raghunathan, S., Kozlov, A.G., Lohman, T.M., & Waksman, G. (2000) *Nat. Struct. Biol.* 7, 648–652.
471. Classen, S., Ruggles, J.A., & Schultz, S.C. (2001) *J. Mol. Biol.* 314, 1113–1125.
472. Shi, H., & Moore, P.B. (2000) *RNA* 6, 1091–1105.
473. Quigley, G.J., & Rich, A. (1976) *Science* 194, 796–806.
474. Quigley, G.J., Wang, A.H., Seeman, N.C., Suddath, F.L., Rich, A., Sussman, J.L., & Kim, S.H. (1975) *Proc. Natl. Acad. Sci. U.S.A.* 72, 4866–4870.
475. Kim, S.H., Suddath, F.L., Quigley, G.J., McPherson, A., Sussman, J.L., Wang, A.H., Seeman, N.C., & Rich, A. (1974) *Science* 185, 435–440.
476. Kim, S.H., Sussman, J.L., Suddath, F.L., Quigley, G.J., McPherson, A., Wang, A.H., Seeman, N.C., & Rich, A. (1974) *Proc. Natl. Acad. Sci. U. S. A.* 71, 4970–4974.
477. Suddath, F.L., Quigley, G.J., McPherson, A., Sneden, D., Kim, J.J., Kim, S.H., & Rich, A. (1974) *Nature* 248, 20–24.
478. Bjork, G.R. (1995) in *tRNA: Structure, Biosynthesis, and Function* (Söll, D., & RajBhandary, U.L., Eds.), ASM Press, Washington, DC.
479. Bjork, G.R., Ericson, J.U., Gustafsson, C.E., Hagervall, T.G., Jonsson, Y.H., & Wikstrom, P.M. (1987) *Annu. Rev. Biochem.* 56, 263–287.
480. Ennifar, E., Nikulin, A., Tishchenko, S., Serganov, A., Nevskaya, N., Garber, M., Ehresmann, B., Ehresmann, C., Nikonov, S., & Dumas, P. (2000) *J. Mol. Biol.* 304, 35–42.
481. Oubridge, C., Ito, N., Evans, P.R., Teo, C.H., & Nagai, K. (1994) *Nature* 372, 432–438.
482. Rould, M.A., Perona, J.J., & Steitz, T.A. (1991) *Nature* 352, 213–218.
483. Antson, A.A., Dodson, E.J., Dodson, G., Greaves, R.B., Chen, X., & Gollnick, P. (1999) *Nature* 401, 235–242.
484. Valegard, K., Murray, J.B., Stonehouse, N.J., van den Worm, S., Stockley, P.G., & Liljas, L. (1997) *J. Mol. Biol.* 270, 724–738.
485. Batey, R.T., Rambo, R.P., Lucast, L., Rha, B., & Doudna, J.A. (2000) *Science* 287, 1232–1239.
486. Biou, V., Yaremchuk, A., Tukalo, M., & Cusack, S. (1994) *Science* 263, 1404–1410.
487. Frugier, M., Soll, D., Giege, R., & Florentz, C. (1994) *Biochemistry* 33, 9912–9921.
488. Rould, M.A., Perona, J.J., Soll, D., & Steitz, T.A. (1989) *Science* 246, 1135–1142.

489. Stark, H., Dube, P., Luhrmann, R., & Kastner, B. (2001) *Nature* 409, 539–542.
490. Price, S.R., Evans, P.R., & Nagai, K. (1998) *Nature* 394, 645–650.
491. Shevack, A., Gewitz, H.S., Hennemann, B., Yonath, A., & Wittmann, H.G. (1985) *FEBS Lett.* 184, 68–71.
492. von Bohlen, K., Makowski, I., Hansen, H.A., Bartels, H., Berkovitch-Yellin, Z., Zaytzev-Bashan, A., Meyer, S., Paulke, C., Franceschi, F., & Yonath, A. (1991) *J. Mol. Biol.* 222, 11–15.
493. Yonath, A., Glotz, C., Gewitz, H.S., Bartels, K.S., von Bohlen, K., Makowski, I., & Wittmann, H.G. (1988) *J. Mol. Biol.* 203, 831–834.
494. Karpova, E.A., Serdiuk, I.N., Tarkhovskii, I.S., Orlova, E.V., & Boroviagin, V.L. (1986) *Dokl. Akad. Nauk SSSR* 289, 1263–1266.
495. Trakhanov, S.D., Yusupov, M.M., Agalarov, S.C., Garber, M.B., Ryazantsev, S.N., Tischenko, S.V., & Shirokov, V.A. (1987) *FEBS Lett.* 220, 319–322.
496. Trakhanov, S., Yusupov, M., Shirokov, V., Garber, M., Mitschler, A., Ruff, M., Thierry, J.C., & Moras, D. (1989) *J. Mol. Biol.* 209, 327–328.
497. Yusupov, M.M., Trakhanov, S.D., Barynin, V.V., Boroviagin, V.L., Garber, M.B., Sedelnikova, S.E., Selivanova, O.M., Tishchenko, S.V., Shirokov, V.A., & Edintsov, I.M. (1987) *Dokl. Akad. Nauk SSSR* 292, 1271–1274.
498. Wimberly, B.T., Brodersen, D.E., Clemons, W.M., Jr., Morgan-Warren, R.J., Carter, A.P., Vonnrhein, C., Hartsch, T., & Ramakrishnan, V. (2000) *Nature* 407, 327–339.
499. Yusupov, M.M., Yusupova, G.Z., Baucom, A., Lieberman, K., Earnest, T.N., Cate, J.H., & Noller, H.F. (2001) *Science* 292, 883–896.
500. Ban, N., Nissen, P., Hansen, J., Moore, P.B., & Steitz, T.A. (2000) *Science* 289, 905–920.
501. Harms, J., Schlunzen, F., Zarivach, R., Bashan, A., Gat, S., Agmon, I., Bartels, H., Franceschi, F., & Yonath, A. (2001) *Cell* 107, 679–688.
502. Schlunzen, F., Tocilj, A., Zarivach, R., Harms, J., Gluehmann, M., Janell, D., Bashan, A., Bartels, H., Agmon, I., Franceschi, F., & Yonath, A. (2000) *Cell* 102, 615–623.
503. Pioletti, M., Schlunzen, F., Harms, J., Zarivach, R., Gluehmann, M., Avila, H., Bashan, A., Bartels, H., Auerbach, T., Jacobi, C., Hartsch, T., Yonath, A., & Franceschi, F. (2001) *EMBO J.* 20, 1829–1839.
504. Ramakrishnan, V. (2002) *Cell* 108, 557–572.
505. Ogle, J.M., Brodersen, D.E., Clemons, W.M., Jr., Tarry, M.J., Carter, A.P., & Ramakrishnan, V. (2001) *Science* 292, 897–902.
506. Thompson, J., Kim, D.F., O'Connor, M., Lieberman, K.R., Bayfield, M.A., Gregory, S.T., Green, R., Noller, H.F., & Dahlberg, A.E. (2001) *Proc. Natl. Acad. Sci. U.S.A.* 98, 9002–9007.
507. Nissen, P., Hansen, J., Ban, N., Moore, P.B., & Steitz, T.A. (2000) *Science* 289, 920–930.
508. Pavletich, N.P., & Pabo, C.O. (1993) *Science* 261, 1701–1707.
509. Choo, Y., & Klug, A. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 11168–11172.
510. Choo, Y., & Klug, A. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 11163–11167.
511. Miller, J., McLachlan, A.D., & Klug, A. (1985) *EMBO J.* 4, 1609–1614.
512. Brown, R.S., Sander, C., & Argos, P. (1985) *FEBS Lett.* 186, 271–274.
513. Nolte, R.T., Conlin, R.M., Harrison, S.C., & Brown, R.S. (1998) *Proc. Natl. Acad. Sci. U.S.A.* 95, 2938–2943.
514. Dill, K.A., Alonso, D.O., & Hutchinson, K. (1989) *Biochemistry* 28, 5439–5449.
515. Honzatko, R.B., Crawford, J.L., Monaco, H.L., Ladner, J.E., Edwards, B.F.P., Evans, D.R., Warren, S.G., Wiley, D.C., Ladner, R.C., & Lipscomb, W.N. (1982) *J. Mol. Biol.* 160, 219–263.
516. Rosenbusch, J.P., & Weber, K. (1971) *Proc. Natl. Acad. Sci. U.S.A.* 68, 1019–1023.
517. Nelbach, M.E., Pigiet, V.P., Jr., Gerhart, J.C., & Schachman, H.K. (1972) *Biochemistry* 11, 315–327.
518. Carvajal, N., Venegas, A., Oestreicher, G., & Plaza, M. (1971) *Biochim. Biophys. Acta* 250, 437–442.
519. Kim, N.N., Cox, J.D., Baggio, R.F., Emig, F.A., Mistry, S.K., Harper, S.L., Speicher, D.W., Morris, S.M., Jr., Ash, D.E., Traish, A., & Christianson, D.W. (2001) *Biochemistry* 40, 2678–2688.
520. Pearson, R.G. (1966) *Science* 151, 1721–1727.
521. Martin, R.B. (1984) in *Metal Ions in Biological Systems: Volume 17, Calcium and Its Role in Biology* (Sigel, H., Ed.) pp 1–50, Marcel Dekker, New York.
522. Lide, D.R. (1998) *CRC Handbook of Chemistry and Physics*, 79th Ed., CRC Press, Boca Raton, FL.
523. Einspahr, H., & Bugg, C.E. (1984) in *Metal Ions in Biological Systems: Volume 17, Calcium and Its Role in Biology* (Sigel, H., Ed.) pp 51–97, Marcel Dekker, New York.
524. Kim, K.H., Pan, Z., Honzatko, R.B., Ke, H.M., & Lipscomb, W.N. (1987) *J. Mol. Biol.* 196, 853–875.
525. Holmes, M.A., & Matthews, B.W. (1981) *Biochemistry* 20, 6912–6920.
526. Wouters, J. (1998) *Protein Sci.* 7, 2472–2475.
527. De Wall, S.L., Meadows, E.S., Barbour, L.J., & Gokel, G.W. (2000) *Proc. Natl. Acad. Sci. U.S.A.* 97, 6271–6276.
528. DeLaBarre, B., Thompson, P.R., Wright, G.D., & Berghuis, A.M. (2000) *Nat. Struct. Biol.* 7, 238–244.
529. Mueller, U., Perl, D., Schmid, F.X., & Heinemann, U. (2000) *J. Mol. Biol.* 297, 975–988.
530. Wells, C.M., & Di Cera, E. (1992) *Biochemistry* 31, 11721–11730.
531. Rhee, S., Parris, K.D., Ahmed, S.A., Miles, E.W., & Davies, D.R. (1996) *Biochemistry* 35, 4211–4221.
532. Toney, M.D., Hohenester, E., Cowan, S.W., & Jansonius, J.N. (1993) *Science* 261, 756–759.
533. Yusupov, M.N., Antson, A.A., Dodson, E.J., Dodson, G.G., Dementieva, I.S., Zakomirdina, L.N., Wilson, K.S., Dauter, Z., Lebedev, A.A., & Harutyunyan, E.H. (1998) *J. Mol. Biol.* 276, 603–623.
534. Larsen, T.M., Laughlin, L.T., Holden, H.M., Rayment, I., & Reed, G.H. (1994) *Biochemistry* 33, 6301–6309.
535. Greasley, S.E., Horton, P., Ramcharan, J., Beardsley, G.P., Benkovic, S.J., & Wilson, I.A. (2001) *Nat. Struct. Biol.* 8, 402–406.
536. Zhou, Y., Morais-Cabral, J.H., Kaufman, A., & MacKinnon, R. (2001) *Nature* 414, 43–48.
537. Betzel, C., Pal, G.P., & Saenger, W. (1988) *Eur. J. Biochem.* 178, 155–171.

538. Bajorath, J., Hinrichs, W., & Saenger, W. (1988) *Eur. J. Biochem.* 176, 441–447.
539. Acharya, K.R., Stuart, D.I., Walker, N.P., Lewis, M., & Phillips, D.C. (1989) *J. Mol. Biol.* 208, 99–127.
540. Teplyakov, A.V., Kuranova, I.P., Harutyunyan, E.H., Vainshtein, B.K., Frommel, C., Hohne, W.E., & Wilson, K.S. (1990) *J. Mol. Biol.* 214, 261–279.
541. Chakrabarti, P. (1990) *Biochemistry* 29, 651–658.
542. Snyder, E.E., Buoscio, B.W., & Falke, J.J. (1990) *Biochemistry* 29, 3937–3943.
543. Kankare, J., Salminen, T., Lahti, R., Cooperman, B.S., Baykov, A.A., & Goldman, A. (1996) *Biochemistry* 35, 4670–4677.
544. Ray, W.J., Jr., & Multani, J.S. (1972) *Biochemistry* 11, 2805–2812.
545. Kuo, C.F., McRee, D.E., Fisher, C.L., O'Handley, S.F., Cunningham, R.P., & Tainer, J.A. (1992) *Science* 258, 434–440.
546. Guan, Y., Manuel, R.C., Arvai, A.S., Parikh, S.S., Mol, C.D., Miller, J.H., Lloyd, S., & Tainer, J.A. (1998) *Nat. Struct. Biol.* 5, 1058–1064.
547. Wedekind, J.E., Frey, P.A., & Rayment, I. (1995) *Biochemistry* 34, 11049–11061.
548. Nagashima, S., Nakasako, M., Dohmae, N., Tsujimura, M., Takio, K., Odaka, M., Yohda, M., Kamiya, N., & Endo, I. (1998) *Nat. Struct. Biol.* 5, 347–351.
549. Jabri, E., Carr, M.B., Hausinger, R.P., & Karplus, P.A. (1995) *Science* 268, 998–1004.
550. Teixeira, M., Moura, I., Xavier, A.V., Dervartanian, D.V., Legall, J., Peck, H.D., Jr., Huynh, B.H., & Moura, J.J.G. (1983) *Eur. J. Biochem.* 130, 481–484.
551. Ragsdale, S.W., Clark, J.E., Ljungdahl, L.G., Lundie, L.L., & Drake, H.L. (1983) *J. Biol. Chem.* 258, 2364–2369.
552. Ellefson, W.L., Whitman, W.B., & Wolfe, R.S. (1982) *Proc. Natl. Acad. Sci. U.S.A.* 79, 3707–3710.
553. Hamilton, C.L., Scott, R.A., & Johnson, M.K. (1989) *J. Biol. Chem.* 264, 11605–11613.
554. Chan, M.K., Mukund, S., Kletzin, A., Adams, M.W., & Rees, D.C. (1995) *Science* 267, 1463–1469.
555. Guss, J.M., & Freeman, H.C. (1983) *J. Mol. Biol.* 169, 521–563.
556. Norris, G.E., Anderson, B.F., & Baker, E.N. (1983) *J. Mol. Biol.* 165, 501–521.
557. Garrett, T.P.J., Clingeffer, D.J., Guss, J.M., Rogers, S.J., & Freeman, H.C. (1984) *J. Biol. Chem.* 259, 2822–2825.
558. Kerr, M.C., Preston, H.S., Ammon, H.L., Huheey, J.E., & Stewart, J.M. (1981) *J. Coord. Chem.* 11, 111–115.
559. Lee, Y.H., Deka, R.K., Norgard, M.V., Radolf, J.D., & Hasemann, C.A. (1999) *Nat. Struct. Biol.* 6, 628–633.
560. Miller, W.T., Hill, K.A., & Schimmel, P. (1991) *Biochemistry* 30, 6970–6976.
561. Qian, X., Gozani, S.N., Yoon, H., Jeon, C.J., Agarwal, K., & Weiss, M.A. (1993) *Biochemistry* 32, 9944–9959.
562. Green, L.M., & Berg, J.M. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 4047–4051.
563. Kraulis, P.J., Raine, A.R., Gadhave, P.L., & Laue, E.D. (1992) *Nature* 356, 448–450.
564. Pan, T., & Coleman, J.E. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87, 2077–2081.
565. Swaminathan, K., Flynn, P., Reece, R.J., & Marmorstein, R. (1997) *Nat. Struct. Biol.* 4, 751–759.
566. Benning, M.M., Kuo, J.M., Raushel, F.M., & Holden, H.M. (1995) *Biochemistry* 34, 7973–7978.
567. White, A., Ding, X., vanderSpek, J.C., Murphy, J.R., & Ringe, D. (1998) *Nature* 394, 502–506.
568. Tao, X., & Murphy, J.R. (1992) *J. Biol. Chem.* 267, 21761–21764.
569. Zheng, Y.J., Xia, Z., Chen, Z., Mathews, F.S., & Bruice, T.C. (2001) *Proc. Natl. Acad. Sci. U.S.A.* 98, 432–434.
570. Skarzynski, T., Moody, P.C., & Wonacott, A.J. (1987) *J. Mol. Biol.* 193, 171–187.
571. Nicholson, H., Anderson, D.E., Dao-pin, S., & Matthews, B.W. (1991) *Biochemistry* 30, 9816–9828.
572. Andersson, I. (1996) *J. Mol. Biol.* 259, 160–174.
573. Kraulis, P.J. (1991) *J. Applied Crystallog.* 24, 946–950.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Chapter 7

Evolution

Although it is mutations in the DNA that produce the diversity upon which natural selection operates, it is within the proteins encoded by that DNA that most of the diversity is expressed. Consequently, natural selection accepts or rejects mutated proteins, not mutated genes. The two genes encoding the two calmodulins in *Arbacia punctulata*, which arose from the duplication of a single gene, differ in nucleotide sequence from each other at 45 out of 393 positions, but in the two calmodulins themselves, which unlike the genes have been continuously scrutinized by natural selection, only two of those 45 differences have been permitted to change the amino acid sequence.¹

It is within the proteins existing today that the history of evolution by natural selection can be read. The later episodes of this history are read by comparing the amino acid sequences of the same protein from different species. Two new species arise from one ancestral species as soon as subpopulations of that ancestral species become so different from each other that two individuals of different sex, one from each of the subpopulations, are no longer able to breed successfully. Even when two closely related species that have only recently diverged from their common ancestor are compared, the amino acid sequences of the same respective proteins from each species will often differ at one or more positions. For example, myoglobin from domestic sheep differs in amino acid sequence at three of its 143 positions from myoglobin of domestic goats. Even the amino acid sequences of the myoglobins from human and chimpanzee differ at one of their 153 positions.

The reason for this divergence of amino acid sequence is that once speciation has occurred and interbreeding becomes impossible, two versions of the same protein are established. These two versions begin to evolve in isolation from each other, and mutations occur at random in the respective genes encoding each version. Once in a while one of these mutations in one version that produces an acceptable change in amino acid sequence of the protein is fixed by genetic drift or natural selection independent of any fixation occurring in the other version, and slowly the amino acid sequences of the encoded proteins become different, one position at a time. Because the geologic instant at which the two species were established from one common ancestral species coincides with the instant at which the two versions of the same protein began to evolve separately,

amino acid sequences retain the history of the speciation of organisms. This history can be reconstructed by comparing the amino acid sequences of the same protein from an array of different species.

Even in the most advantageous instances in which amino acid sequences are compared to each other, connections can usually be made only as far back as the common ancestors of prokaryotes and eukaryotes. What has been found, however, is that the tertiary structure of a particular protein, when viewed in crystallographic molecular models from distantly related species, changes less rapidly than its amino acid sequence during evolution by natural selection. Because of this, comparisons of crystallographic molecular models permit one to look back in evolutionary history to the time at which the individual proteins themselves were diverging from common ancestors: to the time, for example, when L-lactate dehydrogenase and glyceraldehyde-3-phosphate dehydrogenase or triose-phosphate isomerase and indole-3-glycerol-phosphate synthase diverged from their common ancestor. Through such comparisons, the speciation of proteins can be traced. Because amino acid sequences change more rapidly than tertiary structures, only a few of the pedigrees of proteins, those that diverged recently in geologic time or those in which mutations are fixed slowly, can be traced by comparing amino acid sequences. Most of our insight into the speciation of organisms has come from comparisons of amino acid sequences, but most of our insight into the speciation of proteins has come from comparisons of tertiary structures.

From the comparisons that can be made among the tertiary structures that are now available, it has become clear that the larger proteins often, if not always, have arisen during evolution by the chance fusion of two genes encoding smaller proteins, each of which could fold independently and each of which usually had an independent function prior to the fusion. As a consequence of such fusions, larger and larger proteins appeared. If a particular fusion produced a protein that was not impaired functionally, the new gene for the larger protein may have been fixed in the population by genetic drift; or, if the fusion produced a protein with advantageous features, the new gene for the larger protein may have been fixed in the population by natural selection. The history of these fusions can be observed in the existing domains from which these larger proteins

are constructed. The domains of a protein are discrete regions in the tertiary structure of that protein which arose from separate, previously independent proteins that were fused together, one after the other, to produce the present protein. Because a polypeptide shorter than about 70 amino acids usually cannot fold spontaneously to form a tertiary structure, domains, when defined in this way, are usually larger than this. They appear in the crystallographic molecular model as independently folded regions. Because they are the fundamental units in the evolution of proteins, domains must be identified by a set of conservative, objective criteria, if our description of the evolutionary history of a set of proteins is to be accurate.

It may be possible, by examining enough crystallographic molecular models, to trace the ancestry of the proteins that presently exist, in a sense to derive a molecular phylogeny of the proteins. Because most of the existing proteins were produced by fusion of smaller units, this molecular phylogeny of the proteins must be based on a reconstruction of two processes. First, the family trees of the individual, ancestrally related domains from different proteins must be reconstructed. In almost every instance these family trees must be based on patterns in which the secondary structures are arranged to form the tertiary structures of the domains being compared because similarity in amino acid sequence has been completely lost. Second, the separate events that produced each of the fusions of the independent domains to produce the larger chimeric proteins must also be reconstructed.

Although the most interesting question may be how the large array of existing proteins arose from a much smaller array of smaller proteins present in the distant past, it should be stressed that new proteins are continuously being made by this process of fusion of different pieces. We know this because, in some instances, domains that have homologous amino acid sequences can be found in otherwise completely different proteins. Because similarity in amino acid sequence usually disappears quite rapidly over geologic time, these domains must have been separately incorporated into their respective proteins fairly recently; the greater the similarity of amino acid sequence among them, the more recently the separate fusions must have occurred.

Molecular Phylogeny from Amino Acid Sequence

The amino acid sequences of a set of related polypeptides retain a record of the history of their evolution by natural selection. That record provides information about the speciation of organisms, the specialization of tissues, and the conversion of older proteins into newer ones. This evolutionary history is read from aligned amino acid sequences.

As the amino acid sequences of the same protein from different species have become available, it has usually been found that they are similar enough to be readily aligned with each other. An **alignment** of two or more amino acid sequences is a display in which positions that are thought to be directly related to each other from the respective sequences are aligned directly above and below each other. The decision that the aligned positions are related is based on the fact that they are occupied by the same amino acid or the fact that they are each surrounded by similar sequences of amino acids. The cytochromes *c* from human, corn, and yeast can be readily aligned (Figure 7–1A).^{2,3} There is no uncertainty about the alignment even though the three proteins are from distantly related species.

The fact that, in most instances, the amino acid sequences of the three respective proteins responsible for a particular function in humans, yeast, and corn can be readily aligned, as can the three cytochromes *c*, is the strongest evidence for the fact that these three species share a **common ancestor**. Consequently, each of the proteins that are responsible for a particular function and the amino acid sequences of which can be aligned also share a common ancestor. Any two proteins that have descended from a common ancestor are **homologues** of each other.

In the distant past, when only the common ancestral species was present, all of the individuals in the population of that ancestral species contained, for all practical purposes, a cytochrome *c* the amino acid sequence of which was the same, just as all individuals of an extant species contain a cytochrome *c* with the same sequence. As natural selection operated upon the genetic

Figure 7–1: Alignment of amino acid sequences of cytochromes *c* and replacements observed at each of the positions in the common sequence. (A) Alignment of ungapped amino acid sequences. The three amino acid sequences below the numerical scale are the aligned amino acid sequences of the cytochromes *c* from *Homo sapiens*, *Zea mays*, and *Saccharomyces cerevisiae*. The amino acid sequence of cytochrome *c* from *Thunnus alalunga* is immediately above the numerical scale, which is based on this latter sequence. Above each position in this top sequence is a list of the other amino acids found in this position in a collection of 40 cytochromes *c* from various eukaryotes.^{2,3} Letters below the horizontal lines in each of the columns are variations found among cytochromes *c* of animals, and letters above the horizontal lines are the additional variations found in fungi and plants, more distantly related eukaryotes. (B) Insertion of gaps for the purpose of alignment. The two aligned amino acid sequences are those for cytochromes *c* from *T. alalunga* and *Paracoccus dinitrificans*. Each set of dashes represents a gap that must be made in one of the sequences to align it reasonably with the other sequence. You should convince yourself that the gaps are inescapable. When the two sequences are aligned in this way, the size of each gap is determined by the number of extra amino acids in the sequence that does not have a gap. (C) Gaps visualized as insertions. Instead of introducing gaps to permit the alignment shown in panel B, the extra amino acids in each intervening segment are shown as loops. This presents a more realistic picture of the situation but is a significantly more awkward method for displaying an alignment.

tuna

human
corn
yeast

[illegible]

tuna

human
corn
yeast

C

10 20 30 40 50
 GDVAKGKKTFFVQKCAQCHTV-----ENGGKHKVGPNLWGLFGRKTGQAEGYSYTDANKS-----KGIV
 QDGDAAKGEKEF-NKCKACHMIQAPDGTDIKGGKTGPNLYGVVGRKIASEEGFKYGEIGILEVAEKNPDLT
 60 70 80 90 100
 WNNDTLMEYLENPKKYI-----PGTKMIFAGIKKKGERQDLVAYLKSATS
 WTEADLIEYVTDPKPWLVKMTDDKGAKTKMTFKMGK---NOADVVAFLAQNSPDAGGDGEAA

GDVAKGKKT^VOKCAQCHTVENG^QGKHKVGP^NNLWGLFGRKTGQAE^GYSYTDANKSKGIV^D
 QDGDAAKGEKEFNKCKACH^IIKGGKTGPNLYGVVGRKIAS^EE^GFKYGE^IGI^DLT^N
 W^NNDTLMEYLENPKKYIPG^TTKMIFAG^IQDLVAYLK^SSATS^T
 WTEADLIEYVTDPKP^NTKMTFKMGKNQADVVAF^LAQNSPDAGGDGEAA^A

variation present within the population of that ancestral species, varieties arose that occupied different ecological niches. These varieties eventually diverged sufficiently to become separate species. At that point, the genes for cytochrome *c* in these two new species became disconnected, and the amino acid sequences encoded by those genes from that time forth were altered independently and continuously by mutation, genetic drift, and natural selection. As a result of a long series of such disconnections, the distantly related species *Homo sapiens*, *Zea mays*, and *Saccharomyces cerevisiae* eventually appeared. The differences and similarities between the three extant amino acid sequences for the three respective cytochromes *c* are the accumulated result of the individual steps in this process of speciation. An underlying assumption of this description is that a function performed only by the protein encoded by a certain gene remains the exclusive property of the product of that gene as it passes from species to species. Although this is usually the case, there are isolated examples in which the amino acid sequence of a protein from one species seems to be unrelated to that of the protein performing the same role in another species and must be the result of convergent rather than divergent evolution.⁴

Evolution by natural selection is usually viewed from its optimistic side. Natural selection operates on the variation inherent in any large population of a given species of organisms to shift the distribution of its assembled abilities gradually in a direction that makes that species or its descendant species more successful. Beneficial traits are patiently nurtured and multiplied. The major portion of the variation upon which natural selection operates to achieve this progress is variation in the sequences of the proteins within the population of a given species.

It is unlikely, however, that more than a small number of the differences seen when two aligned amino acid sequences are compared (Figure 7-1A) reflect improvements in the ability of the individuals of that species to survive relative to that of individuals of other species or their common ancestors. There is little evidence that the cytochrome *c* from either *H. sapiens* or *Z. mays* is an improved version of the cytochrome *c* that was used by their common ancestor or that any of the proteins the amino acid sequences of which are being presently compared are improved versions. The majority of the differences that accumulate in the sequences of the same protein in two lineages, following their divergence from their common ancestor, are neutral replacements.^{5,6} A **neutral replacement** is a change of one amino acid for another that is harmless enough that the biological function of the protein does not deteriorate sufficiently to cause the elimination of the replacement by natural selection. These neutral replacements arise from mutations in the DNA encoding the protein. Each that is now in existence began as a mutation in the genome of one individual and then spread through the

population of its species, or became **fixed**, by **genetic drift**. When one views aligned sequences of the same protein from different species, one is examining the record of this gradual increase in entropy.

This increase in entropy, however, is biased. From examining aligned amino acid sequences of the same protein from many species, it is clear that each position in the underlying sequence that gives the protein its unique character is under a different degree of negative selective pressure. Mutations can occur with equal frequency at any position in the sequence of the DNA encoding for the sequence of a protein. Each of these individual mutations is assessed by natural selection, and the majority⁵ disappear almost immediately because they adversely affect the function of the protein or are otherwise deleterious. For example, in the human population, there are many mutant forms of hemoglobin that bind oxygen improperly or are unstable proteins.⁷ These represent deleterious mutations that survive for a limited time before disappearing from the population. These mutant forms can be contrasted with fetal hemoglobin that has been fixed in the human population because it is a stable protein and has beneficial properties. The most **deleterious mutation** is one that kills the individual in which it arises before that individual has had an opportunity to mate or otherwise reproduce after the mutation occurs. The more critical a particular amino acid in the sequence of the protein is to its function, the less prone will that position be to substitution over time. For this reason, the aligned amino acid sequences of the same protein from an array of species evaluate the scope of the intolerance to variation expressed at each position in the sequence of the protein.

This **record of intolerance** can be read from examining consecutively each position in the aligned sequences of a large collection of the same proteins from different species. Above the numerical scale in Figure 7-1A, the sequence of cytochrome *c* from *Thunnus alalunga* is presented, and above each of its positions in a column of letters are tallied the amino acids found there in the cytochromes *c* from 40 other eukaryotes.^{2,3} The horizontal lines in each of these columns of letters separate amino acids found in the sequences of cytochromes *c* from animals, of which far more are available, from amino acids found in the sequences of cytochromes *c* from fungi and plants, which represent more distant relationships. A similar record of intolerance is observed when the amino acid sequence of a particular protein is mutated at random and the resulting mutants are selected for their ability to function properly.⁸

This intolerance to substitution is most strongly manifested at an invariant position. An **invariant position** in a protein is a position at which no replacement has been made over the history encompassed by the aligned amino acid sequences. A few of the positions in the aligned sequences of the cytochromes *c* have remained absolutely invariant, for example, Cysteine 14, Cysteine 17, Histidine 18, and Methionine 80 because

these are functionally irreplaceable and consequently define a cytochrome *c*. Some positions such as those occupied by Glycine 6, Glycine 34, Glycine 41, Glycine 77, and Glycine 84 are invariant among the eukaryotes but are replaced in bacterial cytochromes *c*. Several of these glycines are examples of the fact that glycines with angles ϕ and ψ outside the boundaries on a Ramachandran plot (Figure 6-4B) are difficult to replace.⁹ Nevertheless, their eventual replacement demonstrates that a designation of invariant is always provisional. As more and more amino acid sequences of the same protein from different species become available, the number of invariant positions usually decreases.¹⁰

The fact that any designation of invariant is necessarily based on a limited set of amino acid sequences may explain why site-directed mutation of amino acids at apparently invariant positions often has little effect on the function of a protein.¹¹⁻¹⁵ For example, site-directed mutation of five of the 15 highly conserved amino acids in lathosterol oxidase had little effect on its function.¹⁶ Consequently, the intuition that an invariant position must be structurally or functionally important is unreliable. The situation is even more confusing when a position known to be functionally critical nevertheless displays several replacements even among closely related species.¹⁷

Many of the changes accumulating over time seem to be conservative replacements. A **conservative replacement** is a replacement at a position in which only similar amino acids, either in size or in chemical properties, can be tolerated. For example, only valine, isoleucine, phenylalanine, and leucine, each of the side chains of which is a hydrocarbon, seem to occur in position 35 of eukaryotic cytochrome *c* (Figure 7-1A). Either glycine or alanine, the side chains of which are small, seems to be necessary in position 29. Either serine, threonine, or glutamine, the side chains of which are polar but uncharged, seems to be necessary in position 42.

It was once thought that each position in the amino acid sequence of a protein could be assigned unambiguously to one of a few categories, for example, invariant, conservative, physicochemically constant, and variable.¹⁸ When it became possible, however, to compare the amino acid sequences of the same protein from a large number of distantly related species, the majority of the replacements observed could not be easily explained. This fact suggests that even the specific designations just presented may themselves be rationalizations of more subtle processes that are not understood. A close examination of the actual results, however, sequence position by sequence position (Figure 7-1), does produce an intuitive feeling for the play of evolution.

In addition to the capacity of a particular amino acid to be tolerated at a particular position in the sequence of a protein, the nature of the **genetic code** itself also affects the patterns in which replacements in the sequence

occur during evolution. Because there are three bases coding for each amino acid and mutation occurs one base at a time, replacements requiring one base change should be more common than those requiring two or more.⁵ There are, however, some interesting apparent exceptions to this generalization that occur even in the comparisons of the various eukaryotic cytochromes *c* (Figure 7-1). For example, at position 31, only asparagine and alanine are found; at position 72, only lysine and serine; at position 45, only lysine and glycine; and at position 19, only glycine and threonine. Each of these four replacements would require that two bases of the respective codon be mutated consecutively. Although these are unlikely events, the constraints on the occupation of these positions seem to have been severe enough to confine the replacements among the eukaryotes to these choices. In the short term, however, the difficulty of changing more than one base to effect a replacement is more acute. When the detailed history of the mutational events that have occurred during the recent evolution of artiodactyl fibrinopeptides⁵ was examined, it was observed that replacements requiring the mutation of only one base were far more frequent than those requiring two consecutive mutations. It is altogether likely that, in circumstances where two consecutive mutations seem to have occurred, the amino acid sequence of the protein displaying the intermediate single mutation, although it exists, has not yet been determined.

One approach to examining quantitatively the progress of natural selection has been to calculate a mutation probability for every pair of possible replacements.² This was accomplished by reconstructing a probable sequence of events in the evolution of 10 different groups of closely related sequences. Sequences for common ancestors were predicted from alignments, and all of the replacements that should have occurred following the divergence of the progeny from that ancestor were tabulated to provide the basis for the calculation of probabilities. The results of this study were presented as mutation probabilities. A **mutation probability** is the probability that a certain replacement will occur during a time long enough for a particular number of replacements to accumulate for every 100 amino acids of the sequence.

Values for mutation probabilities over a period of time long enough for two replacements for every 100 amino acids (Table 7-1) register changes that occur over the **short term**. Almost all of the replacements with the highest mutation probabilities over this period require only one base change to occur and are also remarkably conservative. For example, the 12 most frequent replacements do not involve any change in charge number or even polarity. Replacements involving **alanine** are the most frequent by a considerable margin, an observation suggesting that a truncation to the β carbon is the most readily tolerated change. A large number of replacements are not tolerated well at all (mutational probab-

Table 7-1: Mutation Probabilities for Various Pairs of Amino Acid Replacements

pair ^a	base changes ^b	mutation probability ^c (%)	pairs with mutation probability less than 0.005%		
V/I	1	2.3	W/A	Y/A	R/A
E/D	1	1.9	R/W	R/Y	R/D
S/A	1	1.6	W/N	Y/N	R/C
S/T	1	1.5	W/D	Y/D	R/E
Y/F	1	1.3	W/V	Y/V	R/G
S/N	1	1.2	W/Q	Y/Q	R/I
L/M	1	1.1	W/E	Y/E	R/L
K/R	1	0.9	W/G	Y/G	R/P
V/M	1	0.9	W/I	Y/I	R/T
G/A	1	0.8	W/L	Y/L	R/V
P/A	1	0.8	W/K	Y/M	C/N
T/A	1	0.8	W/M	Y/S	C/D
N/D	1	0.8	C/W	C/Y	C/Q
S/G	1	0.7	W/S	Y/T	C/E
I/L	1	0.7	W/T	P/Y	C/H
N/A	2	0.6	P/W	F/N	C/L
E/A	1	0.6	P/H	F/D	C/K
V/A	1	0.6	P/L	F/Q	F/C
N/K	1	0.6	P/M	F/E	P/C
Q/E	1	0.6	H/M	F/G	P/H
Q/A	2	0.5	G/M	F/K	
V/L	1	0.5	G/I	F/P	
H/N	1	0.5	D/I		
D/S	2	0.5			

^aPairs of amino acids occupying the same position in pairs of aligned sequences.

^bMinimum number of changes required to change a codon from one member of the pair into a codon for the other. ^cProbability that the given pair will occur at the same position in two aligned sequences that have only two replacements for every 100 positions.² Only pairs with probabilities greater than or equal to 0.5% are tabulated (24 of the 190 possible pairs).

ity $\leq 0.005\%$) over the short term (Table 7-1). Many of these require two or three consecutive base changes. The frequency at which other amino acids are turned into tryptophan or methionine or into cysteine, tyrosine, or phenylalanine will also be decreased by the fact that these amino acids have only one or two codons, respectively.⁶ The amino acids most intolerant to promiscuous replacement are tryptophan, tyrosine, arginine, and cysteine (mutation probabilities are less than 0.005% for 11 or more of the 19 possible replacements), in part because they along with glycine and proline are so peculiar. The amino acids that are the most promiscuously replaced are alanine, serine, glutamine, threonine, valine, methionine, lysine, and asparagine (mutation probabilities are greater than 0.12% for 11 or more replacements).

There are several ways in which the DNA encoding a protein can be altered over time. The most common is by **point mutation**, which is the ultimate source of the exchanges of one amino acid for another that are observed in the alignments of the four eukaryotic cytochromes *c* (Figure 7-1A). It is also possible for a **start site** or a **stop site** for translation to be mutated and another start site or stop site, either already present or

arising by mutation, to take over, causing the protein to become longer or shorter at one or the other of its ends (Figure 7-1A). Because the amino-terminal and carboxy-terminal segments of a protein are seldom involved in its function, this is usually an inconsequential change.

Eukaryotic genes contain **introns**. These are segments of DNA, often quite long, inserted at several locations within the coding sequence of the genomic DNA. Introns are removed by **splicing** at the level of the messenger RNA. Often a cell will contain two or more¹⁹ versions of the same protein, one in which all the splices were successful and one or more in which one or more of the splices has failed. For example, there are two crystallins found in the lenses of the eyes of *Zapus hudsonius*, the longer containing an unremoved insert between positions 63 and 64 in the amino acid sequence of the shorter.²⁰ Errors in splicing can also lead to two forms of a protein that differ in their amino-terminal sequences because amino-terminal segments under the control of two different promoters, respectively, have been alternatively spliced to the same coding sequence for the remainder of the protein.²¹ Likewise, there are two different versions of subunit β of isocitrate dehydrogenase (NAD⁺) in *Bos taurus* that differ only in their carboxy-terminal sequences. Both are encoded by the same genomic DNA. One ends with a sequence of 28 aa; the other, with a completely different sequence of 26 aa even though their sequences are exactly the same for the first 357 aa. The former results from a messenger RNA in which the final exon in the genomic DNA is properly spliced to those that go before; the latter, from a messenger RNA in which this exon is skipped and the following exon is spliced to those that go before.²² Each of these types of **alternative splicing** has the potential to produce different versions of the same proteins.

One of the common changes that occurs over evolutionary time is the **insertion** into or the **deletion** from a protein of a short segment of amino acids. For example, most forms of adenylate kinase have an additional 25 aa in their amino acid sequence between the valine and the aspartate in the sequence –GRVDDN– found near position 140 in the amino acid sequences of isoform 1 of adenylate kinase from mammalian cytoplasm.²³ This additional segment of 25 aa is present in the adenylate kinases from bacteria, fungi, and mitochondria, so it can be concluded that it must have been deleted during one of the genetic events leading to the appearance of mammalian isoform 1. Usually, however, it is difficult to tell whether a deletion or an insertion has occurred. Because such insertions or deletions appear as frequently in proteins from prokaryotes as they do in proteins from eukaryotes, they must arise from processes independent of alternative splicing. When the sequences of two proteins that differ by a deletion or an insertion are aligned, a gap is included in the shorter to permit the alignment of the sequences on the two sides of the aberration.

A **gap** is a series of blank spaces inserted into one

amino acid sequence that is missing a segment of amino acids present in the other amino acid sequence with which the first is being aligned. For example, it is necessary to insert three gaps of 6, 5, and 8 spaces in length into the amino acid sequence of cytochrome *c* from *T. alalunga* and two gaps of 1 and 3 spaces in length into the amino acid sequence of cytochrome *c*-550 from *Paracoccus denitrificans* in order to achieve the most reasonable alignment of these two proteins (Figure 7-1B). On either side of each gap, the alignments are convincing enough to justify the insertions of the gaps required to bring those alignments into register. It must be kept in mind that in the actual polypeptide there is no gap; rather it is the other polypeptide, the one with the ungapped sequence, that has additional amino acids at that point (Figure 7-1C). The use of a gap is simply a convenient method for displaying the alignment of the sequences.

When two sequences are aligned, their similarity is usually quantified by stating their percentage of identity and the gap percentage. The **percentage of identity** is the percentage of the average number of positions in the two aligned sequences that are occupied by the same amino acid. The **gap percentage** is the number of gaps that had to be inserted for every 100 amino acids in the alignment. For example, in the alignment of the cytochromes *c* from *T. alalunga* and *P. denitrificans* (Figure 7-1B), an average of 110.5 positions from the two sequences are aligned, there are 38 identities for a percentage of identity of 34% identity, and there are 5 gaps for a gap percentage of 4.5 gap percent.*

The alignments of the cytochromes *c* in Figure 7-1 are so obvious that they can be performed unassisted. Even for the cytochromes *c* from *T. alalunga* and *P. denitrificans*, with only 34% identity and 4.5 gap percent, the sequences are easily aligned by eye. The amino acid sequence of each protein changes, however, at a different rate during evolution. Although there are proteins that change more slowly than cytochrome *c*, such as histone H4 (20 times more slowly), calmodulin (4 times more slowly), α tubulin (2 times more slowly), ubiquitin (2 times more slowly), and protein phosphatase 2A (2 times more slowly),²⁴ most proteins change more rapidly than cytochrome *c*. As more and more time has passed following the divergence of the amino acid

sequences of two proteins from that of their common ancestor, the percentage of identity decreases and the gap percentage increases until it becomes difficult to align them. Appropriately programmed digital computers are used to align such distantly related sequences.²⁵

The **computational alignment** of two distantly related amino acid sequences is accomplished by constructing a **matrix**.²⁶ If one sequence A has *p* amino acids, arranged in the order $a_1 a_2 a_3 \dots a_p$, and the other sequence B has *q* amino acids, arranged in the order $b_1 b_2 b_3 \dots b_q$, the product of these two vectors is a matrix C, the coefficients of which, c_{ij} , are equal to $a_i \times b_j$, where a_i and b_j are particular amino acids. For example, in the alignment of the cytochromes *c* from *T. alalunga* and *P. denitrificans* (Figure 7-1B), $a_9 \times b_{11}$ would be Thr \times Glu. The numerical value assigned to a particular position c_{ij} in the matrix representing $a_i \times b_j$ depends on the schemes chosen to weight the comparisons.

The simplest scheme is to decide that when $a_i = b_j$, $c_{ij} = a_i \times b_j = 1$, and when $a_i \neq b_j$, $c_{ij} = a_i \times b_j = 0$. This produces a matrix the coefficients of which, c_{ij} , are either 1 or 0. When the amino acid in position a_i in the first sequence is the same as the amino acid in position b_j in the second sequence, $c_{ij} = 1$; when they are different, regardless of the difference, $c_{ij} = 0$. Such a matrix, spread upon a two-dimensional field, can be represented diagrammatically by placing a dot on every position with a score of 1 (Figure 7-2).²⁷ In such a **dot matrix**, the alignment is represented by diagonal strings of dots. In the dot matrices comparing the amino acid sequence of the cytochrome *c* of human with those of monkey and fish in Figure 7-2, the diagonals are obvious and unbroken. In the matrix comparing those of human and bacterium, the alignment is a set of at least three diagonal segments that can be picked out by eye if the figure is tilted and viewed along the diagonal direction. The offsets between the diagonal segments are the gaps in the alignments.

There are, however, 231 different outcomes* for $a_i \times b_j$ if one assumes symmetry, namely, that Glu \times Thr = Thr \times Glu, if one treats cysteine and cystine as separate amino acids, and if one treats each of the 21 types of identity as a unique result. It has always seemed that some of these 231 outcomes are more probable and that recognition of this probability with the proper weighting scheme might enhance the ability to align distantly related sequences. Each of the more than 18 available **weighting schemes**²⁸ is a table of numbers assigned to the 231 possible identities and replacements. Each of these entries reflects the author's view of the probability that such an outcome is the result of evolution by natural selection.

The ultimate goal in aligning two amino acid sequences is to decide whether position a_i in sequence A and position b_j in sequence B arose from the same position in the sequence of a common ancestor or position a_i

* There is no agreement as to the length of amino acid sequence to be used in calculating the percentage of identity. The most common choice is the length of the shorter sequence. The justification for this choice is that the inserts in the longer protein cannot be compared to anything and should therefore be discounted. This choice, however, in a self-contradiction, ignores the inserts in the smaller protein. Probably, the best choice would be to use the length of the common amino acid sequence in which a gap appears in neither protein. The problem with this choice is that it would inflate both percentages and would be misleading in the absence of universal agreement. The choice made in the present calculations was to use the mean length of the two sequences being compared, which produces somewhat smaller percentages than any of the other choices.

* $[(21 \times 20)/2] + 21$.

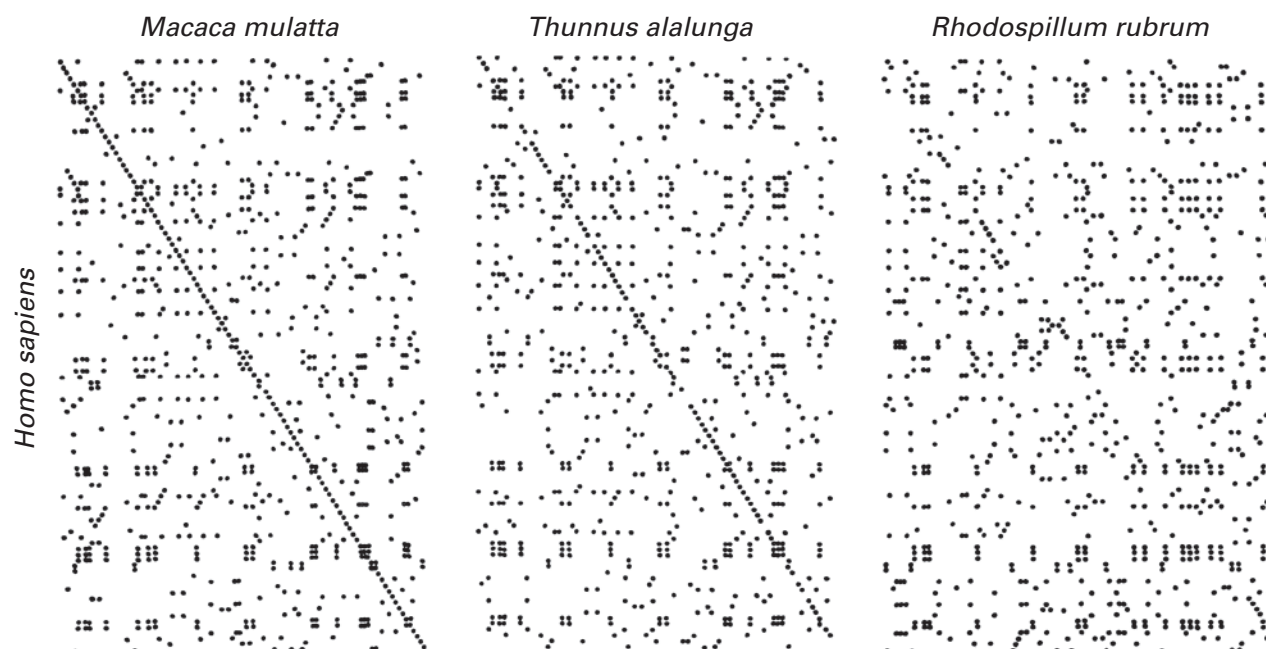


Figure 7-2: Dot matrices²⁷ for the amino acid sequences of the cytochromes *c* from *Macaca mulatta*, *T. alalunga*, and *Rhodospirillum rubrum*, each compared to the amino acid sequence of human cytochrome *c*. The sequence of human cytochrome *c* is the vertical vector (top to bottom, amino to carboxy terminus), and the respective sequences with which it is compared are the horizontal vectors (left to right, amino to carboxy terminus). A dot is placed in the matrix when the amino acids at those two positions, horizontal and vertical, are the same. Reprinted with permission from ref 27. Copyright 1970 Springer-Verlag.

in sequence A and position b_j in sequence B are unrelated to each other either because protein A and protein B do not share a sufficiently recent common ancestor that they can be aligned or because the two sequences are misaligned. If the two amino acid sequences are unrelated, $a_i \times b_j$ is governed solely by chance. If the respective positions are descended from the same position in an ancestral sequence, then $a_i \times b_j$ should retain some of the biases enforced by natural selection, and these biases, if they can be quantified, should be considered while a decision is reached. If a particular replacement has a higher probability of occurring as a result of evolutionary change than it does of occurring as a result of random change, then whenever that particular replacement is encountered, those two positions have a higher probability of being evolutionarily related than of being unrelated. For example, every $a_i \times b_j$ where a_i and b_j are interconvertible by only one base change should have a higher probability of being evolutionarily related than those $a_i \times b_j$ where a_i and b_j are interconvertible only by two or three base changes.²⁹ Every $a_i \times b_j$ where a_i and b_j are similar in size or chemical properties should have a higher probability of being evolutionarily related than those in which they are dissimilar.³⁰ The mutation probability (Table 7-1)² can also be used to weight $a_i \times b_j$.

The net effect of any one of the different weighting schemes, or some combination of them, is to assign a number to every coefficient c_{ij} of the matrix. The magnitude of this number is thought to quantify either the effect of natural selection relative to chance on the par-

ticular replacement $a_i \times b_j$ or the probability that the particular replacement $a_i \times b_j$ would arise as the result of evolution rather than chance. Consequently, logarithms of these probabilities are used as entries in the matrix so that the summation to be performed will represent products of probabilities.² It is also possible to incorporate weights into a dot matrix by assigning a dot to any c_{ij} the weight of which exceeds a certain threshold.³¹

When the matrix has been constructed to the taste of the practitioner, the alignment can be performed.²⁶ Any alignment of the two sequences A and B can be represented as a set of consecutive diagonal segments running through the matrix, for example, the three diagonal segments in the dot matrix comparing the cytochromes *c* of humans and *R. rubrum* (Figure 7-2). To be included in the alignment, the end of one of these diagonal segments must be in line with, above, or to the left of the beginning of the next diagonal segment when the diagonals run from top left to bottom right as in Figure 7-2. Each discontinuity requiring a negative vertical shift or a positive horizontal shift to connect the previous diagonal to the next diagonal represents a gap in one of the sequences being aligned. Associated with each individual alignment is an **alignment score**

$$AS = \sum_i \sum_j c_{ij} - \sum_k P_k \quad (7-1)$$

where the respective sums are over all c_{ij} intersected by the diagonal segments and over all gaps k that must be

inserted and P_k is a penalty assessed for creating the gap k . A computer can be programmed to find the **path of diagonal segments** through the matrix that has the largest alignment score,²⁶ and this path produces the most appropriate alignment of the two sequences dictated by the choice of weighting scheme and gap penalty.

The penalty assessed for each gap is an estimate of the logarithm of the probability of a gap the length of gap k appearing during evolution by natural selection on the same numerical scale used to assign the logarithms of the probabilities to each c_{ij} . It is possible to optimize such a gap penalty for the particular weighting scheme being used.²⁸ For example, it has been shown that when values of 1 are assigned to identities and values of 0 are assigned to nonidentities, the appropriate **gap penalty** is

$$P = 1.2 + 0.23l \quad (7-2)$$

where l is the length of the gap.

The most important responsibility of an investigator who performs such a computation and produces an alignment with the maximum alignment score is to provide an assessment of its **statistical significance**. The accepted criterion for this assessment is a statistical evaluation of a set of alignments produced from randomly jumbled sequences of the same length and amino acid composition as the actual amino acid sequences.² First the two actual sequences are aligned, and a maximum alignment score for the optimum alignment is calculated. Then each of the two actual sequences is randomly jumbled a number of times to produce for each a set of nonsense sequences that have the same amino acid composition and length as the actual amino acid sequence from which they were generated. This produces two sets of randomly **jumbled amino acid sequences**, one derived from each of the two actual sequences. All of the different combinations of one jumbled sequence from one of these two sets and one jumbled sequence from the other set are aligned by the same algorithm that was used to align the two actual, unjumbled sequences, and a large number of maximum alignment scores for the nonsense sequences is gathered in this way. The mean and standard deviation of the alignment scores of this collection of randomly jumbled nonsense sequences are calculated by the usual statistical formulas.

The number of standard deviations that the alignment score for the two actual amino acid sequences lies above the mean for the maximum alignment scores for the jumbled sequences is a measure of the confidence that can be assigned to the decision that the two actual sequences share a common ancestor and to the decision that the alignment has juxtaposed positions in the sequence that have evolved independently from the same position in the ancestral sequence. For example, when human $\beta 2$ microglobulin was aligned with the κ -constant region of human immunoglobulin, the maxi-

mum alignment score was 3.5 standard deviations larger than the mean of the alignments of the jumbled sequences (Table 7-2). Unfortunately, there is no accepted level of statistical significance above which an alignment is judged to be real. Consequently, each person is left to make her own decision. Two sequences of amino acids are considered to be homologous to each other once the decision has been made that their alignment is statistically significant.

As the data bases from which candidates for alignment are drawn become larger and larger, the risk that the alignment of the amino acid sequences of two unrelated proteins will nevertheless be judged to be statistically significant becomes greater. For example, if the lengths of the sequences are disregarded, in a data base containing 100,000 amino acid sequences, each of the amino acid sequences should be able to be aligned with two or three other unrelated sequences in the data base with alignment scores that are at least 4 standard deviations greater than the means of the jumbles.

There is a frequently encountered sleight of hand that is practiced in the alignment of amino acid sequences and that violates the rules of statistics. This trick is to align two sequences and then select only the regions in which there is a higher frequency of coincidence for the statistical test. Because the sample has been preselected, it usually shows a higher frequency of coincidences than occurs when jumbled sequences of the same small regions are compared. Ordinarily, statistical evaluation of an alignment of two amino acid sequences shorter than those of complete, naturally occurring, and logically defensible domains within the native protein should not be accepted without the closest scrutiny.

At the present time, statistically significant alignments can be made only between two amino acid sequences that have a **percentage of identity of 15% or greater** upon alignment.^{28,29,32} If a set of three or four amino acid sequences can be assembled, however, that are from a set of proteins that share some structural or functional feature, it is often possible to demonstrate with high statistical confidence that the members of this set all share the same common ancestor even when pairwise comparisons between the members of the set fail to demonstrate convincing homology.^{33,34} In these instances, the statistical significance only becomes convincing when the whole set is aligned together. Such methods for multiple alignment can detect with statistical significance many more correct relationships between distantly related proteins than can pairwise alignments.³⁵

Although they also identify statistically significant^{36,37} relationships among proteins, computational procedures that rapidly **search large banks of amino acid sequences** should be distinguished from the computational procedure for aligning two sequences by using a complete matrix. Banks of the currently available

amino acid sequences, such as the Swiss-Prot Sequence Database (www.expasy.ch) and the Protein Sequence Database of the Protein Information Resource (<http://pir.georgetown.edu>), contain the sequences of hundreds of thousands of proteins. When the sequence of a new protein becomes available, it is not possible to attempt a complete computational alignment between it and each of the proteins in such large collections. Consequently, other strategies have been developed to search such a bank and rapidly find as many candidates for a relationship as possible. Each of these candidates can then be aligned by the standard matrix method with the new amino acid sequence to validate statistically significant relationships.

The methods that are used to search the banks take advantage of the fact that evolution operates unevenly over a sequence of amino acids. Because segments of the sequence in the core of the structure of a protein or in functionally important locations change far less rapidly than regions on the surface;³⁸ in distantly related amino acid sequences, identities and conservative replacements tend to be clustered. For example, in the amino acid sequences of the group of proteins containing the ATP-binding cassette, the sequence –SGCGKST–, or limited variations of it in which the first serine is replaced by a proline or a threonine, the cysteine is replaced by a serine, the second serine is replaced by a threonine or a glycine, or the threonine is replaced by a serine or a glutamine, appears in all of the members even though the rest of the amino acid sequences show low percentages of identity.³⁹ Consequently, searching a bank for **short segments of amino acid sequence** that have a high degree of similarity with short segments in the new amino acid sequence has a significant probability of locating relatives and has the advantage that it can be done rapidly.

The bank of amino acid sequences is searched for short segments of sequence that are similar to any of the segments of a certain length in the newly sequenced protein. The similarity is quantified by summing the weights given to each identity and each replacement in the aligned segments by use of the weighting scheme preferred by the investigator. If the score for the match is above a certain predetermined threshold, the amino acid sequence in the bank containing this segment is judged to be a candidate for a relationship.

The three most widely used **algorithms** for searching banks of amino acid sequences differ only in how they find the segments. In the BLAST algorithm,⁴⁰ every sequence four amino acids in length that appears in the complete sequence of the protein is tabulated. The amino acid sequences in the bank are then searched for segments identical or highly similar to one of the tabulated segments. When such a match is found, the alignment is extended in both directions to find a longer segment that has a high degree of similarity to the corresponding segment of the sequence being matched. It is

the score for this longer segment, based on the identities and replacements it contains, that must exceed the final threshold. In the FASTA algorithm,⁴¹ regions within the new amino acid sequence and regions within an amino acid sequence in the bank with the highest density of identities are located. Ten of these regions are then trimmed until the portion of each giving the highest score is identified. All of these regions with scores above a threshold are joined, and the gaps resulting from the joining are penalized. It is the score for this rough alignment of a portion of the protein that must exceed the final threshold. The SSEARCH algorithm^{32,42} searches directly each sequence in the bank for the segment that has the highest score when aligned with a segment of the new sequence. Because each of these procedures focuses only on short segments of the sequences being searched, each misses some statistically significant matches, but the advantage gained is that they are rapid enough that such searches of the large extant banks of sequences can be performed in a reasonable amount of time.

Computer-assisted searches of banks and complete computational alignments have permitted the relatives of a **newly sequenced protein** to be located so that it can be joined with a known group. The most obvious successes occur when the new amino acid sequence is identical to one that already is known, because this identification often demonstrates that the same protein has two different, unsuspected, and unconnected functions.⁴³ For example, it has been discovered that the protein responsible for the function of neuroleukin, autocrine motility factor, maturation factor, and myofibril-bound serine endopeptidase inhibitor is glucose-6-phosphate isomerase.⁴⁴ It also happens that the amino acid sequence of a protein of known function can be matched with known amino acid sequences of proteins with unknown function and such an alignment gives a strong indication of the identity of that hitherto unknown function.^{45,46} When a new genome is sequenced, one of the banks is searched for matches to the new amino acid sequences of each of the previously unidentified proteins it contains. For example, when the genome of *Archaeoglobus fulgidus* had been sequenced, it was found to encode 1797 proteins that could be matched with amino acid sequences already known while it encoded only 639 proteins that could not be matched.⁴⁷

The aligned amino acid sequences of polypeptides have been used to provide information about the **speciation of organisms**. The sequences of the same protein from a set of different species, for example, the sequences of the cytochromes *c* from different eukaryotic species, serve as the data on which such studies are based. The goal of the exercise is to construct a **phylogenetic tree** (Figure 7–3)⁴⁸ that displays the evolutionary history of the species bearing that protein. The lengths of the limbs in the tree are estimates of the evolutionary distances between any two present-day species and their common ancestor, represented by the node at which the

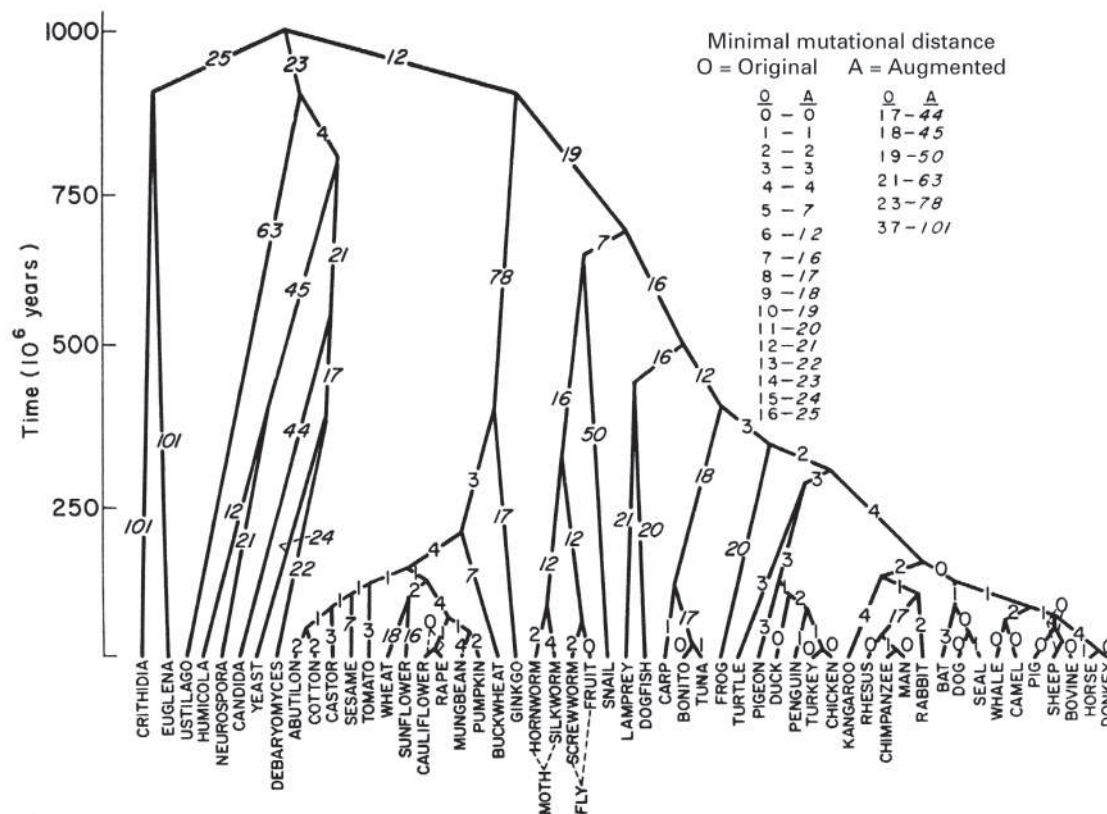


Figure 7-3: Phylogenetic tree⁴⁸ for the cytochromes *c* from 53 species of eukaryotes. Each of the possible 1378 pairs of sequences was aligned and the minimal mutational distance between each pair was tabulated. These numerical values were then adjusted statistically for mutations that would have not left a trace to obtain estimates of evolutionary distances. The magnitudes of these corrections are indicated in the upper right corner. These estimates of evolutionary distance were used to construct the phylogenetic tree. If one passes along the branches of the tree between any two species, the numbers on the branches that are passed through sum to give the estimated evolutionary distance. For example, the evolutionary distance between carp and dogfish is 67. The noted length of the branches, in evolutionary distance, and the positions of the nodes are determined by the most parsimonious sequence of events that satisfy the requirement that the evolutionary distances from the alignments of the sequences of amino acids equal the distances along the branches. In this figure, the nodes connecting marsupials (kangaroo) with the eutheria (the rest of the mammals), reptiles and birds with mammals, amphibians (frog) with amniotes (reptiles, birds, and mammals), fish (tuna, bonito, and carp) with tetrapods (amphibians, reptiles, birds, and mammals), and cartilaginous fish (lamprey and dogfish) with bony fish were placed on the scale of geologic time (millions of years) by using the dates from the fossil record at which the respective divergence from a common ancestor took place. Adapted with permission from ref 48. Copyright 1976 Academic Press.

branches to those two species join. The **evolutionary distance** between the amino acid sequences of two proteins is the value of any quantity that is thought to be directly proportional to the time that has elapsed since those proteins shared a common ancestor. The first step in constructing a phylogenetic tree is to estimate the evolutionary distances between each of the $(n^2 - n)/2$ pairs of sequences in the set of n sequences.

There are a number of **problems** involved in transforming the differences in the aligned sequences of two proteins into an evolutionary distance.^{5,49} First, the number of replacements observed in the two aligned sequences is always an underestimate of the number of replacements that have actually occurred since the two proteins diverged from a common ancestor because several successive unregistered replacements at the same site have often taken place. Second, each position in the two aligned sequences varies at a different rate (Figure 7-1A), each type of amino acid has a characteristic sus-

ceptibility to replacement (Table 7-1), and each protein has its own characteristic rate of change.²⁴ Third, there are examples of accelerated changes occurring along only one branch of a tree containing species that seem indistinguishable from each other. For example, rat ribonuclease seems to have accumulated replacements at 4 times the rate of its close relatives the ribonucleases from mice, muskrats, and hamsters.⁵⁰ Such accelerations in the replacement of amino acids are also observed in those members of a related set of proteins that happen to be involved in the battle between a species and one of its pathogens because the weapons of such a battle are often rapid changes in amino acid sequence either on the part of the pathogen to avoid the defenses or on the part of the attacked to reinstate defenses.⁵¹ Fourth, the size of the population of a given species or its generation time may affect the rate at which mutations become fixed. This may explain why the sequences of the cytochromes *c* from three closely related strains of the

bacterial genus *Pseudomonas* show as much variation in their amino acid sequences (78–61% identity)⁵² as is shown between mammals and amphibians (82% identity) or between mammals and insects (67% identity). These problems have been addressed with varying success by the two methods used to estimate evolutionary distance. One method relies on the percentage of identity between the two aligned sequences; the other, on the minimal mutational distance between them.

It has been shown, both theoretically⁵³ and by simulation,⁵⁴ that

$$q = \frac{\ln(1 + D)}{2D} \quad (7-3)$$

where q is the fraction of the positions in the alignment occupied by identical, unreplaced amino acids and D is the evolutionary distance. This equation corrects for the variations in rate of replacement both among the different positions in the two aligned sequences and among the different types of amino acids and provides an estimate of evolutionary distance from the percentage of identity.

The minimal mutational distance, however, focuses on the changes that have occurred rather than on the positions that have remained unchanged. In theory, there should be more information in these replacements of one amino acid for another because they are progressive rather than static, but the corrections required to account for the unrecorded changes that have occurred over time are inaccurate enough that the advantages of the greater information are significantly diminished.

To calculate its minimal mutational distance, a pair of aligned amino acid sequences in the set is compared position by position, and the minimum number of mutations that had to be fixed to accomplish each replacement is scored. Because of the redundancy of the genetic code, these individual minimum numbers of mutations are most accurately assessed if the actual codons used for each amino acid are known from the nucleic acid sequence. These individual minimum numbers of mutations are added together to obtain the minimum total number of mutations that had to be fixed to convert either of the two sequences into the other. This sum is the **minimal mutational distance** between the two proteins.⁵⁵

The actual number of mutations that were fixed in the two lineages diverging from the common ancestor represented by each of the comparisons between two species is almost always greater than the number calculated, even when the nucleic acid sequences are known, because mutations fixed in the past but then replaced by mutations fixed at the same position at a later date cannot be scored. If the nucleic acid sequence encoding either protein is unknown, mutations to an alternative codon for the same amino acid are also missed. The minimal mutational distances must be corrected statisti-

cally^{48,56} for all of these missing mutations to obtain estimates of evolutionary distances (Figure 7-3).

The major contributors to the minimal mutational distances calculated for each pair of aligned amino acid sequences are the regions of the protein that have experienced the greatest change over time. Unfortunately, these are also the most difficult segments of the amino acid sequences to align convincingly. As a result, the choice of the method used to align the sequences can have a significant effect on the structure of the final tree. With this in mind, a method of **progressive alignment of amino acid sequences** has been developed to provide the most suitable and internally consistent alignments of a large collection of sequences of the same protein from different species.⁵⁷ The basis of this method is the assumption from the beginning that all of the amino acid sequences to be aligned share a common ancestor and have diverged from that common ancestor along their own unique lineages. The most closely related sequences are aligned first, and the gaps in these more certain alignments are retained as the more distant alignments are made. This is advantageous because it is the uncertainty in the precise location of the gaps that must be inserted to align distantly related sequences that creates the greatest uncertainty in the final value for the minimal mutational distance. An example of the product of this method is the progressive alignment of the amino acid sequences of 11 globins (Figure 7-4).⁵⁷ The important feature of these alignments is that the gaps are confined to specific locations rather than being more randomly distributed as would result from simple pairwise alignments.

The tabulated values of evolutionary distances are used to construct a tree the branches of which connect the species being compared (Figure 7-3).⁴⁸ The tree is arranged so that the connections made produce the **most parsimonious sequence** of events that can reproduce the observed evolutionary distances. The overall length of the line segments connecting any two present day species is equal to the evolutionary distance between the two aligned sequences of the proteins from each of them. The branching order in such a tree conveys a historical sequence of the relationships among the species represented, and these historical sequences seem to be reasonable, based on the fossil record and anatomical resemblances.

Usually the phylogenetic trees that are built from the amino acid sequences of only one protein, for example, those of the cytochromes *c* (Figure 7-3), are unsatisfactory. Often there are sequences of a particular protein available for only a limited number of species. Often the phylogenetic tree based on the amino acid sequences of one protein disagrees with the phylogenetic tree based on the sequences of another.⁵⁸ There are a number of solutions to this problem. For example, a more comprehensive and detailed phylogenetic tree of the eukaryotes than the one displayed in Figure 7-3 has been built by

hghu	GHFTEEDKATI	TSLW	GKV	NVEDAGGETLGRLLVVYPWTQRFFDSFGNLSASAIMGNPK	VKAHGKKVLTSLG
hbhu	VHLTPEEKSAV	TALW	GKV	NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDVAMGNPK	VKAHGKKVLGAFS
hahu	VLSPADKTNV	KAAW	GKVGAGAGEYGAEALERMFLSFPTTKTYFPHF	DLSH GSAQ	VKGHGKKVADALT
heha	PITDHGQPPTLSEGDKKAI	RESW	PQIYKNFEQNSLAVLLEFLKKFPKAQDSFPKFSAKKS	HLEQDPA	VKLQAEVIINAVN
hbrl	PIVDSGSVAPLSAAEKTKI	RSWA	APVYSNYETSGVDILVKFFTSTPAAQEFPPKFKGMTSADQLKKSAD		VRWHAERIINAVN
myhu	GLSDGEWQLV	LNWV	GKVEADIPGHGQEVLIIRLFKGHPETLEKFDKFKHLKSEDEMKASED		LKKHGATVLTALG
mycr	SLQPASKSAL	ASSWKTAKDAATIQNNGATLFSLLFKQFPDTRNYTFHFGNM	SDAEMKTTGV		GKAHSMVAVFAGIG
haew	KKQCGVLEGLVKVSEWGRAYGSGHDREAFSQAIWRATFAQVRESRSLFKR		VHGDHTSDPA		FIAHAERVLGGLD
hety	TDCGILQRIKVKQQAQVYSVSGESRTDFAIDVFNNFFRTNPD	RSLFNR	VNGDNVYSPE		FKAHMRVVFAGFD
gpfb	GAFTEKQEALVNSSW	EAFK	GNIPQYSVVFYTSILEKAPAAKNLFSF	LANGVDPNPK	LTAAHESLFGFLVR
hbvs	MLDQQTINIIKATV	PVLK	EHGVTITTTFYKNLFAKHPEVRPLFD	MGRQESLEQPKALAMTVLAAAQNI	

hghu	DAIKHLD	DLKGTFAQLSELHCDKLHVDPENFKLLGNVLVTVLAIHFGKEFTPEVQASWQKMV	TGVASALSSRYH
hbhu	DGLAHL	NLKGTFATLSELHCDKLHVDPENFRLLGNVLVLCVLAHFGKEFTPPVQAAYQKVV	AGVANALAHKYH
hahu	NAVAHVD	DMPNALSALSDLHAHKLVRDPVNFKLLSHCLLVTLAAHLPAAFTPAVHASLDKFL	ASVSTVLTSKYR
heha	HTIGLMDKEAMKKYLDLSTKHSTEFQVNPDMFKELSAVFVSTM	GGKAAEYKLF	SIATLLRSTYDA
hbrl	DAVASMDDTEKMSMKLRDLGSKHAKSQVDPQYFKVLAIAVIADTV	AAGDAGFEKLM	SMICILLRSAY
myhu	GILKKKGHHE	AEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGMANKAL	ELFRKDMASNYKE LGFQG
mycr	SMDSMDADCMNGLALKLSRNHIQKIGASRFGE	MRQVFPNFDLDEALGGGASGDVKGAWDALL	AYLQDNKQA QA L
haew	IAISTLDQPATLKEELDHLQVQHEGRKIPDNYFDA	FKTAILHVAAQLGERCYSNNEEIHDAIACDGFARVLPQVLERG	IKGHH
hety	ILISVLDDKPVLDQALAHYAAFH	LQFGTIPFKA FGQTMFQTIAEHI	HGADIGAWRAC YA EQIVT G ITA
gpfb	DSAAQLRANGAVVAD	AALGSIHSQKGVSNQDLV	VKEALLKTLQAV YDELAAGI KKAYA
hbvs	NLPAILPAVKKIIVKHCQACVAAAHYPIVGQELLGAIKEVLGDAATDDI	LDWAGKAYGVIADV	FIQVEADLYAQAVE

Figure 7-4: Multiple alignment of 11 globins by the progressive method.⁵⁷ The amino acid sequences aligned were the γ polypeptide of human hemoglobin (hghu), the β polypeptide of human hemoglobin (hbhu), the α polypeptide of human hemoglobin (hahu), globin III from *Myxine glutinosa* (heha), globin I from *Petromyzon marinus* (hbrl), human myoglobin (myhu), myoglobin from *Cerithidea rhizophorarum* (mycr), globin II from *Lumbricus terrestris* (haew), globin I from *Tylorhynchus heterochaetus* (hety), leghemoglobin from *Rhaseolus vulgaris* (gpfb), and bacterial hemoglobin from *Vitreoscilla stercoraria* (hbvs). Reprinted with permission from ref 57. Copyright 1987 Springer-Verlag.

combining alignments of the amino acid sequences of α -tubulins, β -tubulins, actins, and elongation factors 1 α .⁵⁹ The conflicts between three phylogenetic trees for gnathosomes were resolved by considering the positions in the sequences of gaps, the patterns of alternative splicing, and the distributions of introns in the genomic DNA.⁵⁸ Nevertheless, disagreements on the branching order of phylogenetic trees, especially the most ancient, persist.⁶⁰

Because the rate of replacement varies dramatically among the positions in the sequence of a protein (Figure 7-1A), minimal mutational distance changes more rapidly than percentage of identity over the short term, and corrections of minimal mutational distance are also less significant over the short term (Figure 7-3). Consequently, historical sequences based on minimal mutational distance are preferred for examinations of **recent speciation**. For example, a detailed phylogenetic tree for the order of artiodactyls covering the last 50 million years has been constructed from considerations of minimal mutational distances for aligned fibrinopeptides⁶¹ and pancreatic ribonucleases.⁶² In fact, ribonucleases with the amino acid sequences predicted for common ancestors at the nodes on that tree were produced by site-directed mutation and shown to display the functional traits characteristic of artiodactyl ribonucleases.

The phylogenetic tree, however, in addition to the historical sequence of events, conveys estimates of the evolutionary distances from existing species to common ancestors. These evolutionary distances can be calibrated (Figures 7-3 and 7-5)⁴⁹ with estimates from the fossil record of the time at which divergence occurred. Eutheria and marsupials diverged from a common

ancestor 130 million years ago (mya); mammals and either reptiles or birds, 300 mya; amniotes and amphibians, 365 mya; and tetrapods and fish, 405 mya.⁴⁹ The respective nodes on the tree should fall at these dates (Figure 7-3). Once the distances are calibrated, the times at which divergences unavailable in the fossil record have occurred can be estimated by extrapolation. To overcome the problems of the different rates of change from one protein to the other and rapid rate of change of a particular protein within a particular branch of the tree, these calibrations are usually performed with sets of amino acid sequences for as many proteins as possible (Figure 7-5).⁴⁹ In this way, the final factor converting evolutionary distance to time should be as reliable as possible to permit a realistic **extrapolation beyond the fossil record** to be performed.

Because the corrections required to convert minimal mutational distances to estimates of evolutionary distance become more significant and less reliable over the long term (Figure 7-3), estimates of evolutionary distance by percentage of identity (Equation 7-3) are preferred for assigning a date to distant common ancestors. For example, it has been estimated from percentage of identity that eukaryotes and archaebacteria diverged from a common ancestor 2.3 billion years ago.⁴⁹ Estimates, however, from both percentage of identity⁴⁹ and minimal mutational distance⁵⁶ agree that deuterostomes and protostomes diverged from a common ancestor 0.7 billion years ago.

At the point at which the lineages of two presently existing species diverged from their common ancestor, the gene for a particular protein carried by the common ancestor became two separate and disconnected genes,

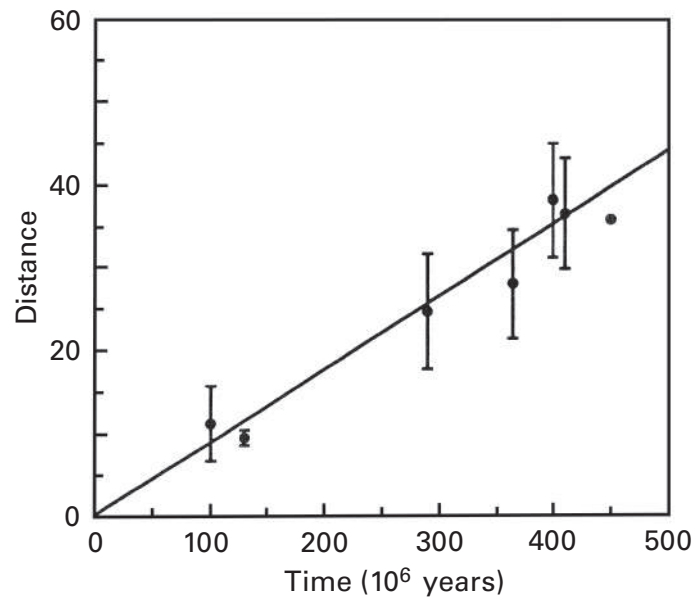


Figure 7-5: Calibrating evolutionary distance with dates from the fossil record.⁴⁹ Sets containing amino acid sequences of the same protein from different species were assembled. Within each set, the amino acid sequences were aligned and the percentage of identity of all pairs of aligned sequences were converted into evolutionary distances with Equation 7-3. Within each set, the evolutionary distances between pairs of sequences in which the two members were from two different groups were sorted into five categories based on which two groups they respectively represented (eutheria and marsupial, mammal and either bird or reptile, amniote and amphibian, tetrapod and fish, and gnathostome and lamprey). The evolutionary distances for pairs of sequences that diverged at these nodes were averaged over each set and then over all of the sets, and the means and standard deviations of these averages are plotted against the date (million years ago, mya) at which the groups in each set diverged from a common ancestor, as determined by the fossil record: 130, 300, 365, 405, and 450 mya, respectively. Averages are also plotted for comparisons of major mammalian groups that diverged from a common ancestor 100 mya. The sets of amino acid sequences did not each have representatives of all groups being compared. The number of sets of amino acid sequences of the same protein from different species that could be used were 48 for mammal/mammal, 3 for eutheria/marsupial, 16 for mammal/bird or reptile, 11 for amniote/amphibian, 15 for tetrapod/fish, and only 1 for gnathostome/lamprey.

one carried by each of the new, independent ancestral species. At that time, natural selection began to operate on these two genes independently, and the differences now observed in the sequences of the same protein from the two existing species began to accumulate. A similar disconnection of two genes for the same protein can occur within a single genome by gene duplication.⁵

Gene duplication is the result of a mistake in recombination causing the DNA in an individual suddenly to contain two copies of the same gene where before there was only one. If this duplication spreads through the population by genetic drift and becomes fixed,* the

* Gene duplication is a fairly common event, its spread, however, over the entire population of a species is a rare event.

genome of the affected species will now contain two copies of the same gene. Both copies will usually continue to produce their respective proteins; but, because of the disconnection, the amino acid sequences of these two proteins have become capable of independent variation to produce isoforms of a given protein or isoenzymes of a given enzyme. **Isoforms** of the same polypeptide are polypeptides found in the same organisms that are encoded by different genes and have different amino acid sequences but nevertheless share a common ancestor and, when properly folded and assembled, perform the same function.

Two proteins are homologues of each other if they both descended from a common ancestor, usually as a result of the speciation of organisms but often as a result of gene duplication. Two proteins are **orthologues** of each other if they both descended in direct lineage from a common ancestor and neither lineage contains a point of gene duplication. Two proteins are **paralogues** of each other if the gene encoding one of them descended in direct lineage from one member of a pair of duplicated genes in the genome of a common ancestor and the gene encoding the other descended in direct lineage from the other member of the same pair of duplicated genes. Isoforms are always paralogues of each other. The A isoform of L-lactate dehydrogenase from dogfish muscle and the A isoform of L-lactate dehydrogenase from pig muscle are orthologous to each other, but the B isoform of L-lactate dehydrogenase from pig heart and the A isoform of L-lactate dehydrogenase from pig muscle are paralogous to each other.

The advantage of having separate isoforms to both the ancestral species and those that diverged from it was that these isoforms could gradually specialize to meet separate demands. These demands are often expressed at the level of individual tissues within the organism, and the sequences of the two isoenzymes or isoforms diverge, in part, in response to the particular demands of sets of tissues. For example, one set of tissues may require that an enzyme respond to changes in levels of its substrates in a different range from the range in which it should respond to them in another set of tissues, and the respective isoenzymes can be tailored to the separate sets of requirements. **Positive selection** causes advantageous mutations that produce changes in biological properties causing the isoform of a protein to be more suitable to a particular set of tissues to be preferentially fixed in the populations at the expense of the parental types, and it is in the adaptation of isoforms of the same protein that positive selection of amino acid sequences is perhaps most readily detected.

It is alignments of the amino acid sequences of the isoforms of a given protein in one species with those of the isoforms of the same protein in another species that permit them to be identified and classified^{63,64} and the history of their divergence to be described.⁶⁵ For example, among the tissues of mammals and birds, at least

have evolutionary histories that are independent of the histories of the organisms carrying them.

The genome of a particular species can encode two or more paralogues of a given protein if the appropriate gene duplications have occurred. It is the potential to **evolve into a new protein** that distinguishes such a set of paralogues from a single, unduplicated orthologue. As long as there is only one gene for a given protein in the genome, the protein that it encodes is required to perform its designated function. That protein evolves along its lineage as its gene diverges into separate species, but it must remain the same protein. If there are two or more paralogues in each individual of a species, one of those paralogues has the opportunity to become a new protein with a new function. Even though paralogues usually retain the same function and specialize to handle different situations, as in the case of the isoenzymes of L-lactate dehydrogenase, often one of them changes sufficiently to perform another function as in the case of chymotrypsinogen and haptoglobin (Table 7–2).

One way to demonstrate that a paralogue, freed from the necessity of performing its function by the existence of a sibling, is able to change sufficiently to perform an entirely new function is by alignment of amino acid sequences. In addition to alignments of the same protein from different species and alignments of different isoforms of the same protein, it has been possible to perform statistically significant **alignments of the amino acid sequences of different proteins** (Table 7-2).²⁵ Many of these connections make sense from a functional standpoint. For example, parvalbumin can be successfully aligned with troponin *c* (Table 7-2); hemoglobin, with myoglobin (Table 7-2); the coiled coil of human vimentin, with the coiled coil of human lamin A (36%

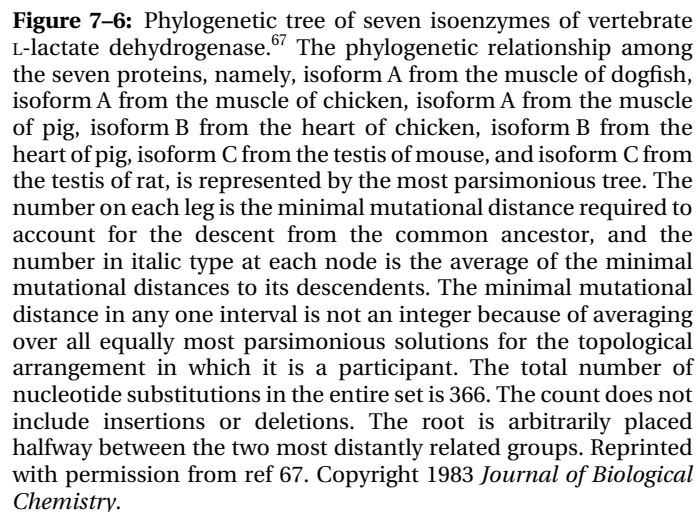


Table 7-2: Examples of Pairs of Proteins Thought to Share a Common Ancestor on the Basis of an Alignment of Their Amino Acid Sequences^a

protein I ^b	protein II ^b	% identity ^c	gaps ^d	gap percent	standard deviations ^e
chymotrypsinogen <i>a</i> , bovine (245)	chymotrypsinogen <i>b</i> , bovine (245)	79.2	0	0	65.8
hemoglobin β , human (146)	hemoglobin γ , human (146)	73.3	0	0	40.7
carbonic anhydrase <i>b</i> , human (260)	carbonic anhydrase <i>c</i> , human (259)	60.6	1	0.4	56.8
chymotrypsinogen <i>a</i> , bovine (245)	trypsinogen, bovine (229)	46.1	6	2.6	22.6
lysozyme (egg white), chicken (129)	lactalbumin (milk), human (123)	38.2	3	2.4	10.7
viral coat protein, PF1 (46)	viral coat protein, Xf(44)	40.5	1	2.3	5.0
hemoglobin α , human (141)	myoglobin, human (153)	27.0	1	0.7	9.3
ovalbumin, chicken (386)	antithrombin III, human (423)	28.1	6	1.6	14.3
parvalbumin, carp (108)	troponin <i>c</i> , bovine (161)	27.0	2	2.0	6.1
cytochrome <i>c</i> , pig (104)	cytochrome <i>f</i> , <i>Spirulina maxima</i> (89)	27.0	3	2.9	7.2
β_2 -microglobulin, human (100)	immunoglobulin κ -constant region, human (102)	18.8	1	1.0	3.5
plastocyanin, spinach (99)	azurin, <i>Pseudomonas</i> (128)	27.3	4	4.0	2.9
histocompatibility antigen, mouse (173)	immunoglobulin SH λ chain ^f (183)	16.7	3	1.7	4.1
leghemoglobin, yellow lupin (153)	invertebrate hemoglobin, midge (151)	22.0	3	2.0	2.6
chymotrypsinogen <i>b</i> , bovine (245)	haptoglobin <i>b</i> , human (245)	19.4	4	1.6	5.4

^aAlignment was performed on a matrix where $a_i \times b_j = 1$ when $a_i = b_j$ and $a_i \times b_j = 0$ when $a_i \neq b_j$ and the gap penalty was 2.5. When $a_i = b_j = \text{cysteine}$, at $a_i \times b_j = 2.0$. Reproduced from Doolittle.²⁵ ^bTwo proteins the amino acid sequences of which were aligned. Number of amino acids in each protein is shown in parentheses. ^cPercentage of the positions in the aligned sequences at which the same amino acid was found in both sequences, based on the length of the shortest. ^dTotal number of gaps that had to be introduced to get the best alignment. ^eDistance in standard deviations that the alignment score for the actual sequences was above the mean of the alignment scores of 36 comparisons of jumbled sequences. Represents only a portion of the entire sequence.

identity, 1.2 gap percent);⁷⁰ vacuolar H⁺-transporting two-sector ATPase from *Daucus carota*, with the β subunit of H⁺-transporting two-sector ATPase from *Spinacia oleracea* (15 standard deviations above the mean of the jumbles);⁷¹ dihydrolipoyllysine acetyltransferase from *Escherichia coli*, with dihydrolipoyllysine succinyltransferase of *E. coli* (30% identity, 1.7 gap percent);⁷² tripeptidyl-peptidase II from *H. sapiens*, with subtilisin from *Bacillus subtilis* (34% identity, 5.3 gap percent);⁷³ acetolactate synthase III from *E. coli*, with tartronate-semialdehyde synthase from *E. coli* (34% identity, 0.7 gap percent);⁷⁴ and 4-hydroxy-2-oxoglutarate aldolase from *E. coli*, with 2-dehydro-3-deoxy-phosphogluconate aldolase from *E. coli* (45% identity with no gaps).⁷⁵

The identification of a set of paralogues in the same species by alignment of their sequences often provides clues as to the functions of those members of the set that have not yet been studied.⁷⁶ For example, the fact that three **proteins of unknown function** in *E. coli* were paralogues of methylmalonyl-CoA mutase and acetate CoA-transferase allowed them to be identified as a methylmalonyl-CoA decarboxylase, a succinate-propionate CoA-transferase, and another methylmalonyl-CoA mutase.⁷⁷

The more interesting though rarer connections, however, are those between **functionally unrelated proteins**. For example, ovalbumin can be successfully aligned with antithrombin III (Table 7-2); bovine angio-

genin, with bovine ribonuclease (33% identity, 3.2 gap percent);⁷⁸ chicken $\delta 2$ crystallin, with human argininosuccinate lyase (69% identity, 0.2 gap percent);⁷⁹ and glucarate dehydratase from *Pseudomonas putida*, with mandelate racemase from *P. putida* (23% identity, 5.6 gap percent).⁸⁰

All of these alignments demonstrate that the evolution of duplicated proteins is completely analogous to the evolution of species. The fixation of the two forms of a duplicated gene within a population produces two paralogues of the ancestral protein. As the paralogues of the protein evolve independently, they drift slowly from isoforms with the same function, to proteins with similar but different functions, just as daughter species drift apart from each other, creating separate genera. Occasionally, a dramatic leap occurs, for example, the one turning argininosuccinate lyase into $\delta 2$ crystallin, a change resembling on a small scale the appearance of chordates. Usually, however, the process is one of slow, continuous divergence. The alignment of amino acid sequences gives only hints of the evolving pedigrees. A more complete picture is seen only when the crystallographic molecular models of proteins are superposed.

Suggested Reading

Feng, D.F., Johnson, M.S., & Doolittle, R.F. (1985) Aligning amino acid sequences: comparison of commonly used methods, *J. Mol. Evol.* 21, 112–125.

Vogt, G., Etzold, T., & Argos, P. (1995) An assessment of amino acid exchange matrixes in aligning protein sequences: the twilight zone revisited, *J. Mol. Biol.* 249, 816–831.

Problem 7-1: Calculate alignment scores (Equation 7-1) for the five cytochromes *c* aligned in Figure 7-9 based on the rules that when $a_i = b_j$, $a_i \times b_j = 1$; when $a_i \neq b_j$, $a_i \times b_j = 0$; and $P = 1.2 + 0.23 l$.

Problem 7-2: This exercise will illustrate the method for assessing the validity of a particular alignment of two sequences. Pick a number between 1 and 80 at random and write it on a piece of paper. Turn to Figure 7-1 and the alignment of the two amino acid sequences of the cytochromes *c* from *T. alalunga* and *P. denitrificans*, respectively. Start at the amino acid in the sequence of the cytochrome *c* from *T. alalunga* corresponding to the number you picked at random, and write the next 20 amino acids in that sequence across the page. Below this sequence write the corresponding amino acids of the aligned sequence from the cytochrome *c* of *P. denitrificans*, as in the figure. Calculate an alignment score by the rules in Problem 7-1 for these two segments of aligned sequences. Take 20 playing cards and to each of them assign one of the 20 amino acids in the amino acid sequence from the cytochrome *c* of the amino acid sequence with the least number of gaps. Shuffle the cards well and deal them into a row. Copy out the jumbled sequence dictated by

this shuffle on a piece of paper. Align this jumbled sequence with the segment of real amino acid sequence from the other cytochrome *c* by shifting and gapping until you think the alignment will give the highest alignment score. Record that score. Repeat the process five times. How do the alignment scores of the jumbled sequences compare to the alignment score of the one unjumbled sequence?

Problem 7-3: On the basis of their locations in the crystallographic molecular models of proteins, their structural roles, and their chemical properties, the amino acids can be divided into three categories: hydrophobic, neutral, and hydrophilic. The hydrophobic amino acids are isoleucine (I), valine (V), leucine (L), phenylalanine (F), cystine (C–C), methionine (M), and alanine (A). The neutral amino acids are glycine (G), cysteine (C), threonine (T), tryptophan (W), serine (S), tyrosine (Y), and proline (P). The hydrophilic amino acids are histidine (H), glutamate (E), glutamine (Q), aspartate (D), asparagine (N), lysine (K), and arginine (R).

The following alignment is from Figure 7-1B:

```
T F V Q K C A Q C H T V - - - - - E N G G
E F - N K C K A C H M I Q A P D G T D I I K
```

Because cytochrome *c* is a cytoplasmic protein, none of the cysteines participates in a cystine.

- (A) Construct a 16×21 matrix on a sheet of graph paper for the two segments of sequence involved in this alignment using the following rules:
 - (1) $a_i \times b_j = 1$ for an identity
 - (2) $a_i \times b_j = 0.6$ for hydrophobic \times hydrophobic
 - (3) $a_i \times b_j = 0.6$ for apathetic \times apathetic
 - (4) $a_i \times b_j = 0.6$ for hydrophilic \times hydrophilic
 - (5) $a_i \times b_j = 0.2$ for hydrophobic \times apathetic
 - (6) $a_i \times b_j = 0.2$ for apathetic \times hydrophilic
 - (7) $a_i \times b_j = 0.0$ for hydrophobic \times hydrophilic
- (B) Trace the alignment presented above through the matrix.
- (C) Calculate an alignment score for that trajectory with the gap penalty of Equation 7-2.
- (D) What is the most serious difficulty with the rules?

Problem 7-4: From the genetic code, calculate the minimum number of base changes between the amino acid sequences of the γ and β polypeptides of human hemoglobin as they are aligned in Figure 7-4. To do this, make a list containing all of the replacements between the two sequences, find the minimum number of base changes required for each, and add up the individual minimum base changes to obtain the total.

Problem 7-5: The sequences of the fibrinopeptides A and B from a series of primates are given in the table

below.⁸¹ Construct a tree of minimal mutational distances. Treat a gap as if it were two base changes.

primate	fibrinopeptide A	fibrinopeptide B
green monkey	ADTGE ^a GDFLAEGGGVR	PCA ^a -GVNGNEEGLFGGR
human	ADSGEGDFLAEGGGVR	PCA-GVNDNEEGFFSAR
drill	ADTGDGDFITEGGGVR	PCA-GVNGNEEGLFGGR
macaque	ADTGE ^a GDFLAEGGGVR	NEESLFSGR
chimpanzee	ADSGEGDFLAEGGGVR	PCA-GVNDNEEGFFSAR

^aPyrrolidone-5-carboxylic acid, a cyclized form of glutamine.

Molecular Phylogeny from Tertiary Structure

Just as the amino acid sequences either of two different proteins or of the same protein from two different species can be aligned, so can their tertiary structures be superposed.⁸² The crystallographic molecular models of two proteins that have tertiary structures so similar to each other that they are thought to share a common ancestor are chosen for comparison. Those pairs of respective α carbon atoms that unambiguously occupy equivalent positions in equivalent strands of secondary structure in the two crystallographic models are identified statistically⁸³ and designated as forming the **cores** of the structures to be superposed. To **superpose** these two crystallographic molecular models is to translate and rotate one of them relative to the other until the sum of the squares of the distances between these pairs of equivalent α carbon atoms in the cores of the two structures is minimized. Because two different crystallographic molecular models are being superposed, the structures never coincide exactly, even if they are of the same protein. The two proteins the crystallographic molecular models of which are being compared are considered to be homologous once a decision has been made by the investigator that the superposition of the two crystallographic molecular models is significant enough to demonstrate that they share a common ancestor.

An example of such a superposition is that between porcine pancreatic elastase and trypsin from rat mast cells (Figure 7-7A).^{83,84} These proteins are both serine endopeptidases sharing a common enzymatic mechanism, their amino acid sequences are readily aligned (33% identity, 2.5 gap percent), and the superposition confirms the fact that they share a common ancestor. A more distant relationship was validated by the superposition of a portion (amino acids 157-403) of creatinase from *P. putida* and methionyl aminopeptidase from *E. coli* (Figure 7-7B).⁸⁵ Although both of these enzymes catalyze the hydrolysis of an amide and their amino acid sequences can be aligned computationally, their respective substrates are quite different from each other and the alignment is marginal (17% identity, 1.9 gap percent).

The degree to which two superposed structures

coincide is usually quantified by the root mean square deviation. The **root mean square deviation** is the square root of the mean of the values for the squares of the distances between only those pairs of α carbon atoms designated as belonging to the cores. Consequently, both the root mean square deviation and the percentage of the total number of α carbons that were included in the cores from the two crystallographic molecular models must be noted (Table 7-3). For example, in the superposition of porcine pancreatic elastase and trypsin from rat mast cells (Figure 7-7A), 65% of the α carbons were included in the two cores and they aligned with a root mean square deviation of 0.07 nm. When the crystallographic molecular model of 4- α -glucanotransferase from *Thermus aquaticus* was superposed in turn upon the crystallographic molecular models of nine amylases from bacteria, fungi, and mammals, 270-320 aa (41-65% of the entire amino acid sequences of these proteins) was designated as belonging to the cores, and those amino acids in the cores aligned with root mean square deviations between 0.29 and 0.35 nm.⁹³

An alternative method of quantifying the superposition of two crystallographic molecular models is to note the percentage of the total number of α carbons in the superposition that lie less than a certain distance from their equivalent partners. For example, in the superposition of creatinase from *P. putida* and methionyl aminopeptidase from *E. coli* (Figure 7-7B), 86% of the α carbons lie less than 0.25 nm from their partners.

As two proteins diverge steadily from a common ancestor, first as orthologues in different species, then, following gene duplication, as paralogues of the same protein, then as paralogues with related functions such as adenosine kinase and ribokinase, and finally as paralogues with unrelated functions such as phosphoribosylamine-glycine ligase and glutathione synthase, their structures drift apart from each other by greater and greater deviations (Table 7-3).

Whenever the crystallographic molecular models of two proteins, the amino acid sequences of which can be aligned computationally with statistical significance, have been compared, they could always be unambiguously superposed.⁹⁴⁻¹⁰⁰ The root mean square deviation of a pair of superposed crystallographic molecular models can be plotted as a function of the percentage of identity between the respective aligned amino acid sequences (Figure 7-8).^{101,102} When the percentage of identity in just the segments chosen as the core reaches around 15%, the range in which statistically meaningful alignments of complete amino acid sequences can no longer be made, the root mean square deviation of the α carbons of the amino acids in the core (50% or greater of the total α carbons) was only 0.2 nm, and the topological similarity between the structures being compared was still unmistakable.

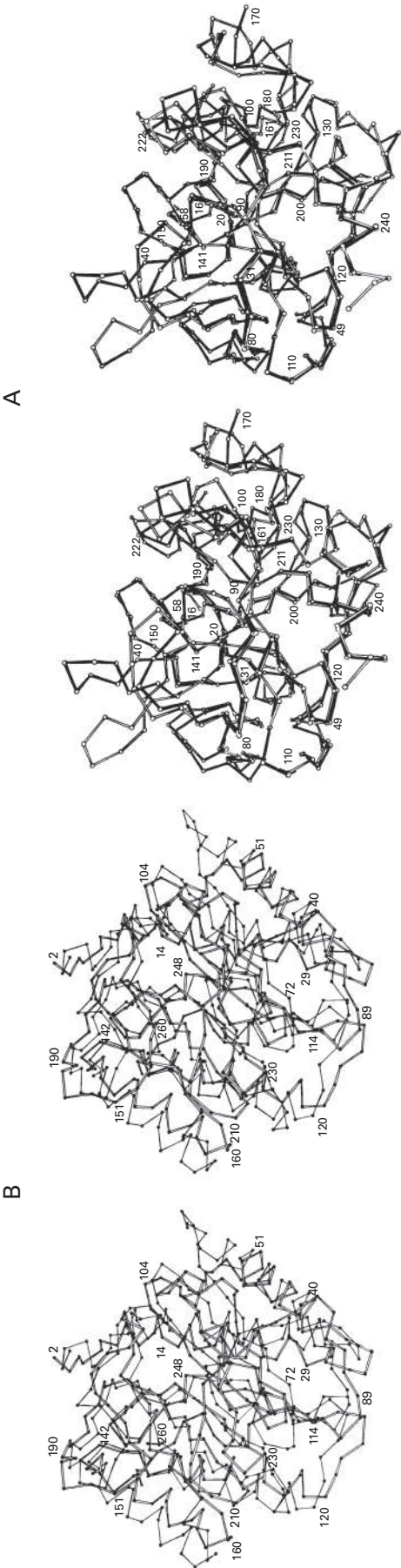
Consequently, structural superpositions are able to establish more **distant evolutionary relationships** than

Figure 7-7: Superposition of the α carbons of crystallographic molecular models. (A) Superposition of the α -carbon diagrams of crystallographic molecular models of pancreatic elastase from *Sus scrofa* (open lines) and trypsin from mast cells of *Rattus norvegicus* (solid lines).⁸⁴ The α carbons of 152 amino acids in each crystallographic molecular model were determined statistically⁸³ to be respectively equivalent to each other. The two models were rotated and translated until the root mean square deviation for these 152 pairs of α -carbons was minimized (0.07 nm). The resulting superposition is numbered according to the amino acid sequence of trypsin. There was no electron density for amino acid 171 of trypsin. (B) Superposition of the α -carbon diagrams of the enzymatic portion of creatinase (amino acids 157–403) from *P. putida* (solid lines) and methionyl aminopeptidase from *E. coli* (open lines).⁸⁵ The α carbons of 218 amino acids from each crystallographic molecular model were determined to be respectively equivalent and the two models were rotated and translated until the root mean square deviation for these 218 pairs was minimized. The numbering is for the methionyl aminopeptidase.

Table 7-3: Levels of Coincidence for the Superposition of Crystallographic Molecular Models^a

proteins superposed	rms ^b (nm)	α carbons ^c (%)
superoxide dismutase <i>Thermus thermophilus</i> superoxide dismutase <i>E. coli</i> ⁸⁶	0.11	90
hexokinase <i>H. sapiens</i> hexokinase <i>S. cerevisiae</i> ⁸⁷	0.13	80
lysozyme <i>Gallus gallus</i> lysozyme bacteriophage λ ⁸⁸	0.20	44
adenosine kinase <i>H. sapiens</i> ribokinase <i>E. coli</i> ⁸⁹	0.24	91
chaperone protein PapD <i>E. coli</i> immunoglobulin G CH2 <i>H. sapiens</i> ⁹⁰	0.23	50
thiamin pyridinylase <i>Bacillus thiaminolyticus</i> D-maltodextrin binding protein <i>E. coli</i> ⁹¹	0.36	84
phosphoribosylamine-glycine ligase <i>E. coli</i> glutathione synthase <i>E. coli</i> ⁹²	0.38	64
phosphoribosylamine-glycine ligase <i>E. coli</i> biotin carboxylase <i>E. coli</i> ⁹²	0.50	57

^aThe two crystallographic molecular models were superposed by use of the respective coordinates of equivalent α carbons in the cores of the two structures. ^bRoot mean square deviation between atoms in the respective cores that were considered to be equivalent during the superposition. ^cPercent of the total number of α carbons in the two crystallographic molecular models that were designated as being in the core and that were aligned in pairs during the superposition.



those established by alignment of amino acid sequences. For example, by superposing crystallographic molecular models it could be shown that adenylyl-sulfate kinase, adenylate kinase, guanylate kinase, and 6-phosphofructo-2-kinase are all members of a large group of kinases that share a common ancestor.¹⁰³ This ability to recognize distant relationships has permitted the history of the evolution of proteins to be traced just as the alignment of amino acid sequences has permitted the history of the evolution of species to be traced.

From Figure 7-8, it can be concluded that whenever two proteins have amino acid sequences that can be aligned with statistical significance, they will also have superposable tertiary structures. For example, the fact that the amino acid sequences of the three enzymes Ca^{2+} -transporting ATPase, Na^+/K^+ -exchanging ATPase, and H^+/K^+ -exchanging ATPase can be aligned¹⁰⁴ demonstrates that their tertiary structures are superposable.¹⁰⁵ This rule is important because far more sequences are available than tertiary structures. If the amino acid sequence of a protein the crystallographic molecular model of which is unavailable can be related to a protein for which a crystallographic molecular model is available

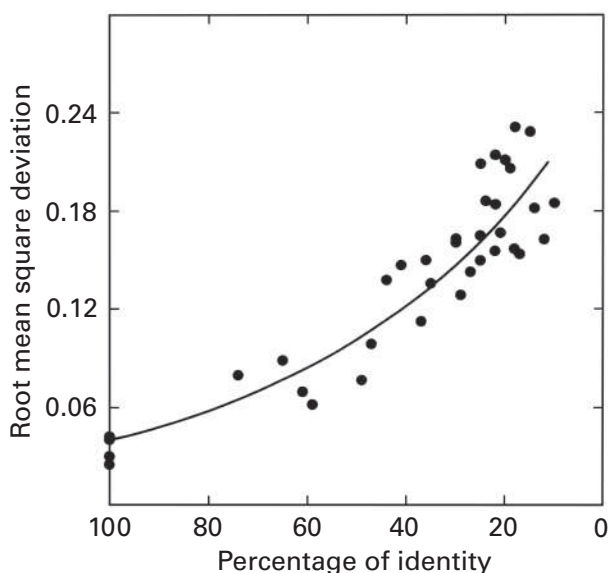


Figure 7-8: Relationship between superposition of crystallographic molecular models and the alignment of amino acid sequence.¹⁰¹ Thirty-two pairs of homologous proteins were chosen for which a crystallographic molecular model of each member of the pair was available. Structural cores were defined for each superposed pair of molecular models by including all α carbon atoms in the polypeptide backbone the distance between which was less than 0.3 nm. The cores for the pairs that were most distantly related included only 50% of the amino acids in the sequence. The root mean square distance between pairs of aligned α carbon atoms of the cores were calculated and plotted against the frequency at which the same amino acid was found at the equivalent positions in the two cores, expressed as a percentage. By this procedure, the percentage of identity is considerably higher than it would be if the entire sequences had been aligned. Reprinted with permission from ref 101. Copyright 1986 IRL Press.

through a valid alignment of the two amino acid sequences, reliable conclusions can be drawn about the unknown tertiary structure by a comparison with the known tertiary structure.

Cytochromes *c* are present in all organisms, they are small proteins, and their structures have changed at a rate such that comparisons between them illustrate many of the facts that can be learned from superposition of tertiary structures. The eukaryotic cytochromes *c* are indistinguishable from each other in tertiary structure,¹⁰⁶ if those from rice and tuna are assumed to represent the evolutionary extremes. Consequently, the eukaryotes can be represented by the crystallographic molecular model of the protein from the tuna (structure C in Figure 7-9).^{107,108} The other four of the five α carbon drawings of the crystallographic molecular models in Figure 7-9 are those of four bacterial cytochromes *c*: the one from *Chlorobium thiosulfatophilum* (Figure 7-9A), the one from *Pseudomonas aeruginosa* (Figure 7-9B), the one from *Rhodospirillum rubrum* (Figure 7-9D), and the one from *P. denitrificans* (Figure 7-9E). A similar comparison of the crystallographic molecular models of four other bacterial cytochromes *c* with that of a eukaryotic cytochrome *c* is also available.¹⁰⁹

When the drawings of the crystallographic molecular models of the cytochromes *c* (Figure 7-9) are compared, the **reason for the gaps** in their aligned sequences (lower part of Figure 7-9) is immediately apparent. They represent **loops** in the longer protein that are smaller or are missing entirely from the shorter protein. These loops (darkened in the figure) can appear or disappear because positions in the sequence at their bases are near enough to each other to be connected without disrupting the structure in a major way. It has usually been observed that the insertions and deletions that are found in superposed tertiary structures of proteins and that are responsible for the gaps in the aligned amino acid sequences occur in regions such as these that are peripheral to the central elements of the structure. The resulting loops usually extend out into the solvent, for example, the loop between positions 20 and 30 in the cytochromes *c* (Figure 7-9), or extend across the outer surface of the protein, for example, the loop between positions 210 and 230 in methionyl aminopeptidase (Figure 7-7B). An interesting spreading apart of two flexible flaps, however, to accommodate the large loop present in the cytochrome *c* from tuna but missing in the two smaller bacterial cytochromes illustrates a more disruptive outcome.

These loops that come and go can be of **various lengths**. An insertion of one amino acid often causes almost no change in the structure of the protein except for a small bulge sufficient to accommodate two amino acids where there is only one in the shorter cousin;¹¹⁰ the two amino acids to the sides of the bulge remain in the same positions. The extra two amino acids in elastase between positions 60 and 64 (Figure 7-7A) cause a bulge that is unmistakable but completely confined to this

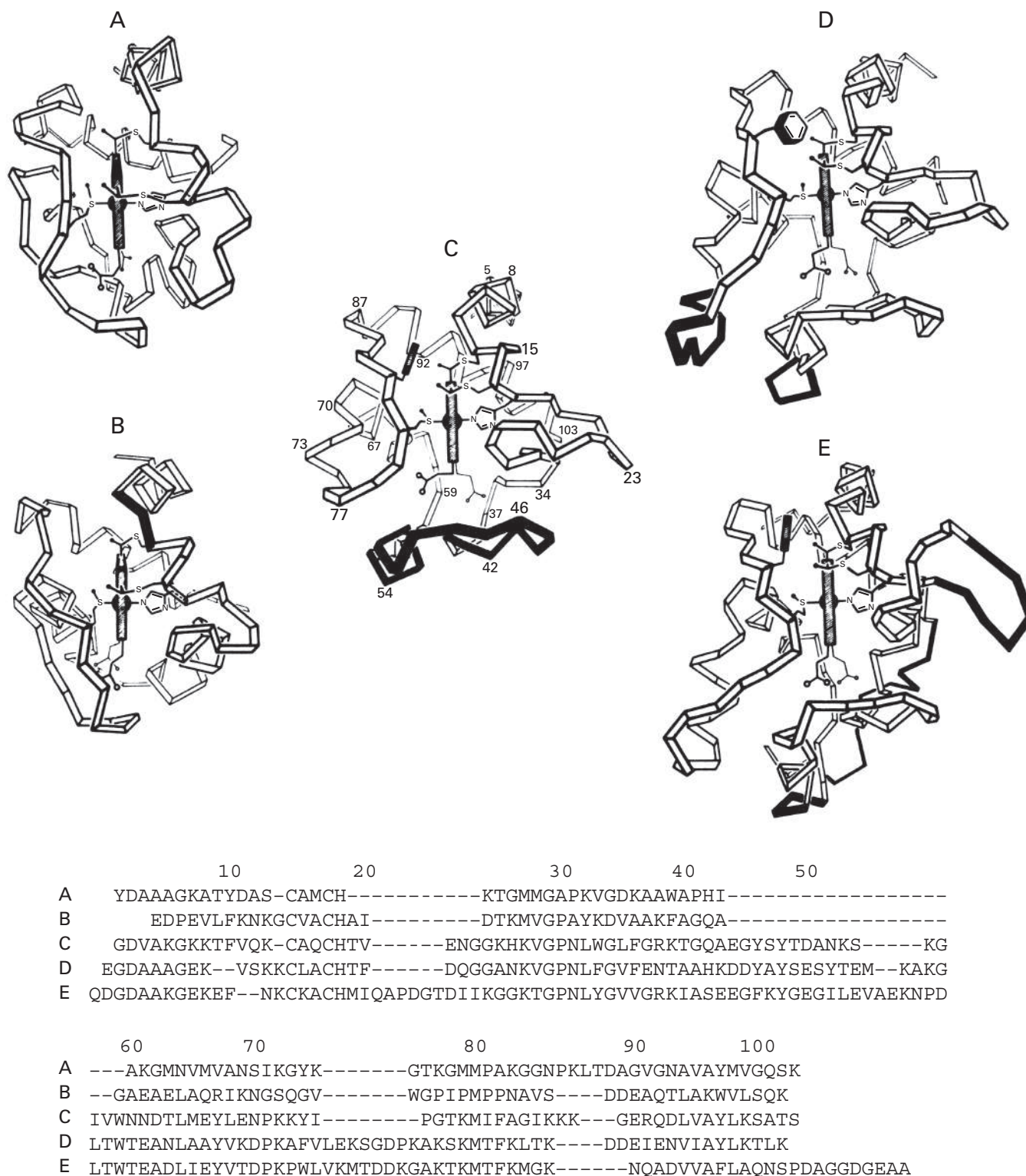


Figure 7-9: Tertiary structures of five cytochromes *c*.^{107,108} Ribbon diagrams with creases at each α carbon were made from the crystallographic molecular models of (A) cytochrome *c*-555 from *Chlorobium thiosulfatophilum*, (B) cytochrome *c*-551 of *Pseudomonas aeruginosa*, (C) cytochrome *c* of tuna mitochondria, (D) cytochrome *c*₂ of *R. rubrum*, and (E) cytochrome *c*-550 of *P. denitrificans*. The hemes, the iron ions of which are each liganded by a methionine and a histidine, and a conserved phenylalanine are also drawn. The sequences of these five cytochromes *c* are structurally aligned at the bottom of the figure (identified by the respective letters). The alignment of *R. rubrum* is represented by the dot matrix in Figure 7-2. Filled portions of the ribbon diagrams highlight loops representing insertions into the basic structure. The central structure for tuna cytochrome *c* is numbered with the same numbers as those used in the alignments. Reprinted with permission from ref 107. Copyright 1982 Academic Press.

short segment, but the one extra amino acid in tryptase between positions 180 and 188 seems to be more disruptive. If the loops become long enough they can assume structure of their own. The extra amino acids between positions 203 and 266 in methionyl aminopeptidase from *Pyrococcus furiosus*, relative to methionyl aminopeptidase from *E. coli*, form an almost spherical knob containing three α helices and some random meander sitting on the surface of the common structure.¹¹¹ The extra amino acids between positions 180 and 317 in serine-type carboxypeptidase from *S. cerevisiae* relative to the structures of other members of the group of related hydrolases form a compact globular structure, but a long loop of polypeptide protrudes from it and wanders over the surface of the common structure.¹¹² The events that produced such insertions can be mimicked by inserting segments of synthetic DNA into the gene encoding a protein and expressing the elongated version. When these inserts are placed at a location where they can loop out of the original structure, they cause little alteration in that structure.¹¹³

The structural alignment of the five amino acid sequences presented in Figure 7–9 is based on superpositions^{114,115} of the five crystallographic molecular models. A **structural alignment** of two amino acid sequences is an alignment in which two respective amino acids that occupy the same positions upon superposition of the two respective crystallographic molecular models also occupy the same position in the alignment.

A structural alignment of amino acid sequences often differs significantly from a computational alignment of the same sequences. For example, a previously performed computational alignment of human basic fibroblast growth factor and human interleukin 1 β agreed with the subsequent structural alignment from positions 90 to 144 but was out of register with the structural alignment by at least seven amino acids between positions 20 and 89.¹¹⁶ It has been argued that when they are available, structural alignments of amino acid sequences are more reliable than computational alignments. This is an interesting argument because it assumes that no relative movement of β strands relative to each other or advancement of the screw of an α helix has occurred during the divergence from a common ancestor. There is some support for this assumption.¹¹⁷

Because the tertiary structures of proteins change more slowly than their amino acid sequences, a structural alignment based on a superposition can be performed between the amino acid sequences of two proteins that diverged from their common ancestor so long ago that a computational alignment would be insignificant. For example, the amino acid sequence of phosphopyruvate hydratase can be structurally aligned with that of mandelate racemase,¹¹⁸ that of P1 nuclease from *Penicillium citrinum*, with that of phospholipase C from *Bacillus cereus*,¹¹⁹ and that of aspartate-semialdehyde dehydrogenase, with that of glyceraldehyde-

3-phosphate dehydrogenase.¹²⁰ In such structural alignments between distantly related sequences, there are significantly more gaps than would have been inserted during a computational alignment because gaps are more readily recognized. For example, the structural alignment of coat protein VP1 from Mengo virus and coat protein from human rhinovirus 14 (13% identity) has a 4.9 gap percent, and eight of the 14 gaps are only one amino acid in length.¹²¹ When crystallographic molecular models are available for many of the members of a large group of related proteins, structural alignments of the amino acid sequences of those members for which molecular models are available can be combined with multiple computational alignments of the sequences of the rest of the members of the group to produce a statistical template that can be used to search databases for additional as yet unrecognized members of the group.¹²²

It is the ability to perform structural alignments of amino acid sequences based on superpositions that allows the frequencies with which one amino acid is substituted for another in more **distantly related proteins** to be ascertained.¹²³ In one study, crystallographic molecular models of 235 proteins from 65 different groups were chosen for superposition.¹²⁴ All of the proteins within each family were superposed, and their amino acid sequences were structurally aligned. For each type of amino acid in the alignments, the number of times that it shared that position with itself or with each of the other amino acids was tabulated. From this tabulation, the **frequency of substitution** for each amino acid could be calculated (Table 7–4).

The table of the frequency of substitution in the more distantly related proteins (Table 7–4) complements the tabulation of mutation probabilities for closely related proteins (Table 7–1). Over the longer term, there seems to be a stronger preference for maintaining the **hydropathy** of a position. This enhanced preference probably arises from the fact that enough time has elapsed that significant substitutions have accumulated in the more intolerant sites. In addition to its hydropathy, the **size** of the amino acids has a significant effect on the frequencies at which specific replacements occur. Small amino acids are usually replaced by small amino acids; the large aromatic amino acids are most often replaced by other aromatic amino acids.

Each amino acid displays a characteristic **tolerance to replacement** by any amino acid other than itself. The most peculiar amino acids, cystine, tryptophan, and glycine, are the most intolerant of replacement. Cystine is so intolerant that it is deleted more than twice as often as it is replaced by alanine, the most common substitution. Proline is also deleted more often than it is replaced by any other amino acid.

The observed frequencies of substitution listed in Table 7–4 do not reflect the full effects of the steric and chemical properties of each side chain on its capacity to replace another or its tolerance to replacement because

Table 7–4: Frequency with Which Substitutions Occur in Distantly Related Proteins^a

ccy ^b (0.9) ^c		trp 1 (1.6)		gly 4 (8.8)		cys ^g (1.2)		pro 4 (4.5)		tyr 2 (3.8)		leu 6 (7.6)		phe 2 (3.9)		val 4 (7.3)		his 2 (2.2)		asp 2 (6.0)	
ccy	88	trp	55	gly	53	cys	46	pro	44	tyr	43	leu	41	phe	40	val	37	his	36	asp	35
gap	3.0	phe	8.2	ala	6.8	ala	9.6	gap	7.3	phe	8.1	val	10.9	leu	11.5	ile	11.8	asn	5.7	ser	8.3
ala	1.4	tyr	6.6	gap	6.7	val	7.9	ser	6.8	gap	4.9	ile	8.8	tyr	8.6	leu	11.6	lys	5.5	gap	7.2
phe	1.0	leu	6.2	ser	6.5	thr	4.7	ala	6.0	val	4.7	phe	5.6	ile	5.4	ala	7.1	gln	5.2	asn	6.9
glu	0.9	gap ^e	2.9	asp	3.1	ser	4.0	gly	4.7	leu	4.4	met	4.4	val	5.4	thr	4.4	gap	4.8	glu	6.9
—	— ^d	val	2.3	asn	2.8	gly	3.8	lys	4.5	ser	4.0	—	—	ala	3.7	ser	3.3	ser	4.8	gly	5.2
val	0.3	ile	2.2	—	—	gap	2.8	—	—	thr	3.5	gln	1.4	—	—	gap	3.2	asp	4.3	ala	5.2
asp	0.2	ser	2.2	tyr	1.1	leu	2.7	tyr	0.8	—	—	arg	1.3	asp	1.0	—	—	—	—	—	—
his	0.1	ala	2.1	his	0.9	—	—	his	0.7	met	1.2	asn	1.3	glu	0.9	arg	1.0	pro	1.6	leu	1.2
asn	0.1	gly	2.1	phe	0.5	ccy	1.0	trp	0.3	gln	1.2	glu	1.2	gln	0.9	his	0.8	ile	1.5	phe	0.7
trp	0.1	—	—	met	0.5	his	0.7	met	0.2	pro	0.9	asp	0.8	arg	0.8	cys	0.6	met	0.9	met	0.4
tyr	0.1	gln	0.6	trp	0.4	glu	0.6	cys	0.2	cys	0.3	his	0.7	ccy	0.4	trp	0.5	trp	0.6	trp	0.4
pro	0.1	cyx ^f	0.2	cyx	0.3	trp	0.4	ccy	0.0	ccy	0.0	cyx	0.3	cys	0.3	ccy	0.1	cyx	0.3	cyx	0.2
ser 6 (7.4)		thr 4 (6.3)		ile 3 (5.2)		arg 6 (3.7)		lys 2 (5.9)		gln 2 (3.6)		ala 4 (8.4)		glu 2 (5.1)		asn 2 (4.7)		met 1 (1.9)			
ser	33	thr	32	ile	31	arg	31	lys	30	gln	30	ala	30	glu	29	asn	24	leu	20.9		
thr	10.2	ser	14.1	val	18.1	lys	11.0	gap	6.9	glu	7.3	ser	8.7	asp	8.6	ser	10.9	met	20.6		
ala	7.6	ala	6.5	leu	14.2	ser	7.7	arg	6.3	lys	6.9	gly	7.6	gap	6.9	asp	9.1	ile	8.7		
gap	6.8	lys	5.1	ala	4.8	gln	6.1	ala	6.1	ser	6.8	gap	6.4	lys	6.7	gap	7.3	val	8.4		
gly	6.3	gap	4.7	phe	4.2	ala	5.4	ser	6.1	ala	6.7	val	6.3	ala	6.5	thr	6.6	ala	5.5		
asn	4.8	val	4.6	thr	3.2	gap	5.2	thr	6.1	thr	6.0	thr	5.3	gln	6.5	gly	6.1	gap	4.6		
asp	4.8	asn	4.0	—	—	—	—	glu	5.2	arg	5.1	—	—	thr	5.7	ala	5.5	lys	3.5		
—	—	—	—	pro	1.2	pro	1.8	gln	4.8	—	—	tyr	1.5	—	—	—	—	—	—		
phe	1.2	his	1.0	arg	0.9	ile	1.3	—	—	tyr	1.2	met	1.1	his	1.1	phe	1.5	arg	1.2		
his	1.0	met	0.8	gln	0.8	phe	0.9	met	1.0	ile	0.9	his	0.9	phe	0.8	met	0.7	his	1.1		
met	0.5	cys	0.4	trp	0.8	trp	0.7	phe	1.0	phe	0.8	cys	0.6	met	0.8	trp	0.4	pro	0.4		
trp	0.4	trp	0.2	his	0.6	met	0.6	trp	0.5	cyx	0.3	trp	0.4	ccx	0.4	cys	0.3	ccy	0.4		
cyx	0.3	ccy	0.1	cyx	0.3	cyx	0.5	cyx	0.2	trp	0.2	ccy	0.2	trp	0.3	ccy	0.0	cys	0.3		

^aStructural alignments were performed of the amino acid sequences within each of 65 groups of proteins. Each of the 65 groups contained a different set of superposable proteins. For each of the 21 amino acids, the frequencies with which it was paired with itself or with each of the other 20 amino acids over all of the structural alignments were calculated.¹²⁴ The number to the right of each amino acid is the frequency (in percent) with which it was paired with the amino acid at the top of the column. ^bCystine. ^cThe number in boldface type at the head of each column is the number of codons that encode that amino acid, and the number in parentheses is the percentage in which it occurs in the amino acid composition of the complete data set (208,000 amino acids). ^dThe horizontal lines divide the amino acids with the highest frequencies of replacement from those with the lowest. All of the amino acids that are not listed in a given column have frequencies between the highest of the lowest group and the lowest of the highest group. ^eFrequency with which a gap occupied the aligned position. ^fFrequency of cysteine plus that of cystine. ^gCysteine; there are two codons for cysteine and cystine combined. The decision between cysteine and cystine is made post translationally.

they are biased. The most important of these **biases** is that of the number of codons for each of the amino acids (boldface numbers next to their abbreviations at the head of each column in Table 7-4). It has been shown⁶ that the frequency with which an amino acid occurs in the overall population of proteins (number in parentheses next to the number of codons in Table 7-4) is roughly proportional to the number of its codons (compare the numbers in boldface type with those in parentheses). Consequently, the number of codons for an amino acid must significantly affect its frequency of substitution over the long term. For example, if the frequencies of substitution for leucine in Table 7-4 were corrected only for number of codons of each of the amino acids, conservation would still have the highest frequency but by much less of a margin, and the preferred substitution, by almost a factor of 2, would become methionine, with valine, isoleucine, and phenylalanine tied for second place. This result would make sense because methionine has the hydrophobic side chain that is the most similar to that of leucine. In addition to the number of codons, however, other factors such as the number of base changes for a given substitution and the frequency with which the codons are used by a given species are also factors affecting the frequency of substitution. Because these effects have yet to be quantitatively assessed, no corrections were applied to the directly observed frequencies of substitution in Table 7-4.

Structural alignments also allow the success of computational alignments and the procedures for searching banks of amino acid sequences to be evaluated. To perform such an evaluation, a bank of amino acid sequences of only those proteins for which crystallographic molecular models are available is assembled. The amino acid sequences in the bank can then be distributed into groups within which each member can be superposed on every other member. Consequently, all of the members of one of these groups share a common ancestor. How many of these established relationships can be detected by a particular algorithm operating only on the amino acid sequences?

When **computational alignments** were evaluated,²⁸ the success with which they aligned two sequences known to share a common ancestor was quantified as the percentage of the positions aligned structurally that were also aligned statistically. It was found that a greater percentage of the positions was correctly matched when weighting schemes were used that assigned values to the $a_i \times b_j$ other than just 1 and 0, but the improvement at best was only 1.25 times (from a 51% rate of success to a 64% rate of success). It made no difference whether the weighting scheme was based on frequencies of identity and replacement from entirely computational alignments (Table 7-1) or from entirely structural alignments, but weighting schemes based on frequencies drawn from larger numbers of alignment performed somewhat better than those drawn from

smaller numbers of alignments. Weighting schemes that were not based on frequencies of replacement and identity, however, did almost as well as those that were. Even the best weighting scheme was unable to align sequences (<30% correctly matched positions) when the percentage of identity fell below 15%.

When the **procedures used to search databanks** were evaluated,³² the success with which they found matches was assessed by comparing coverage to error rate. **Coverage** is the fraction of the known relatives that had scores above the chosen **threshold**. **Error rate** is the percentage of the unrelated amino acid sequences in the bank that had scores above the chosen threshold. As the threshold was raised, the error rate decreased, but so did the coverage. In plots of error rate against coverage, the three currently used algorithms, WU-BLAST2, FASTA, and SSEARCH, all performed equivalently. When a bank containing sets of sequences that were known to be related by superposition but in which none of the members had a percentage of identity greater than 40% was chosen for the searches, at a threshold producing an error rate of 1%, the coverage was only 0.18. In other words, from a bank containing 100,000 amino acid sequences, the search would give 1000 false matches but miss 82% of the real relationships with percentage of identities less than 40%.

There are several different **levels of agreement** within the superposition of two crystallographic molecular models. In the core of the structure, where changes occur less rapidly, the superposition is usually acceptable (Figure 7-7). When the two proteins are functionally related, the segments of the polypeptide in the core that participate in this common function usually superpose the most precisely.¹²⁵ For example, the segments between positions 51 and 56, 102 and 108, and 195 and 200 in the superposition of tryptase and elastase (Figure 7-7A) contain the histidine, the aspartate, and the serine, respectively, that participate in their common mechanism. As more and more replacements of amino acids accumulate, the steric effects of these changes cause the backbone to shift to accommodate them. For example, the respective replacement of a serine and a valine in rice cytochrome *c* with a threonine and an isoleucine, both larger side chains, in tuna cytochrome *c* causes a displacement further into the solvent of the polypeptide to the exterior of this substitution.¹²⁶ The significant shifts of secondary structure in the polypeptide between creatinase and methionyl aminopeptidase (Figure 7-7B) within the core of the common structure result from the accumulation of such steric effects. Flexible loops such as the one between positions 32 and 42 in the superposition of tryptase and elastase (Figure 7-7A) often differ dramatically in their disposition, but such differences may reflect only the effects of crystal packing that pins down an otherwise fluctuating structure.¹²⁷

As one traces the polypeptides through the superposed α carbon diagrams (Figure 7-7), the distance

between the backbones fluctuates as one moves through the core, out through the loops, and back into the core. These fluctuations can be represented graphically in a plot of the distance between the paired α carbons as a function of their position in the amino acid sequence.¹²⁸

The **globins** are a group of the same proteins and their isoforms from different species, for which many sequences are available. They include myoglobins, hemoglobins, erythrocrurins, and leghemoglobins. The **details of the variations** that occur in the tertiary structure of a protein as amino acids are slowly replaced at the toleration of natural selection have been examined by superposing the nine available crystallographic molecular models of different globins¹¹⁷ and using these superpositions to align their amino acid sequences.¹¹⁷ Each globin is formed from eight α helices stacked one upon the others as a bundle of sticks would be in a fire. As the sequences of the globins have varied, the interdigitations of the amino acid side chains situated between the α helices has adjusted to accommodate changes in their size, and this has caused the helices to shift as rigid bodies with respect to each other. These adjustments are necessary because the amino acid side chains between the α helices are tightly packed together and many atomic contacts occur. As the shifting proceeds, accommodating changes in size, the individual pairs of atomic contacts persist between two amino acids at different positions in the amino acid sequence but next to each other in the tertiary structure even though the identities of the amino acids themselves change.

About 60 amino acids out of the 140 in the polypeptide of a typical globin remain in equivalent locations in the nine superposed crystallographic molecular models and account for the core of the native structure. Only half of these are buried; the ones on the surface remain fixed because they are within α helices that are themselves rigid structures. The regions in which the greatest variation in sequence and tertiary structure occurs are in the seven loops connecting the eight helices. This is due to their almost exclusive location at the surfaces of the molecular models but may also reflect the changes in the end to end distances between the α helices that were required to accommodate the slow shifts of the helices relative to each other as the packing among them has been altered by the substitutions.

These observations suggest that the degree of conservation that is displayed by a position in the sequence of a protein may provide an indication of its location in the tertiary structure. Positions showing the least tolerance to replacement are often located on the interior of the protein and those displaying the greatest tolerance tend to be located on flexible surface loops, but the tendency is not overwhelming.

A crystallographic molecular model of myoglobin from the sperm whale has been prepared,⁹⁴ and the structural roles of the 82 **invariant amino acids** among the 24 myoglobins that had been sequenced at the time

were tabulated (Table 7-5). This list represents a combination that has been retained since the time that all of these myoglobins shared a common ancestor. Positions marked (Hb) in Table 7-5 are those invariant among mammalian hemoglobins and myoglobins, and they represent amino acids that have been retained for an even longer period of time. Finally, the amino acids that appear at these 82 positions in the nine globins aligned by superposition have been entered into the tabulation. An examination of Table 7-5 reinforces several features of the atomic structure of molecules of protein.

Positions in the sequence that are buried in **hydrophobic clusters** are the most conserved. Usually three or four members of the group isoleucine, valine, phenylalanine, leucine, methionine, and alanine (Table 7-4) will substitute among themselves in this role, but occasionally only one or two are suitable. For example, only leucine is found at position 2NA and only valine or isoleucine at position 11E in the globins. These two preferences presumably reflect the constraints of the intricate, interlocking stereochemistry in the interior. In two locations, positions 1CD and 4CD, only phenylalanine is found among all the globins, and presumably in this location the flat disk of the phenyl ring is essential to maintain the structure. The phenylalanine at position 1CD is stacked upon the heme.

There are usually a number of locations in the structure of a protein where difficulties resulting from the packing of the backbone of the polypeptide arise. At position 2C in the globins, a proline seems essential to enforce a sharp turn. When two strands of polypeptide are forced too closely together, these **tight locations**, such as positions 6B, 8E, 5F, and 7H in the globins, are occupied by glycine, proline, alanine, serine, or threonine (Table 7-5). Both serine and threonine, by forming hydrogen bonds to acyl oxygens (Figure 6-7A), are able to hug the polypeptide. Tight fits can also result from the juxtaposition of a large and bulky amino acid. The amino acid at position 16E is crowded by the tryptophan at position 12A in both hemoglobin and myoglobin.

There are several instances in which side chains cap one end or the other of an α helix; for example, Serine 1A, Threonine 4C, Serine 1E, or Tyrosine 23H. It is often stated (Table 7-5) that this arrangement has the effect of initiating the α helix. The fact that at position 4C other globins lack an amino acid capable of forming a hydrogen bond and still contain the α helix suggests that the assignment of such a purpose in this case is an overstatement. Remarkably, four pairs of participants in **ionized hydrogen bonds** between side chains on the surface of myoglobin are invariant in the short term. When these particular interactions are examined, however, over all nine of the globins, which represent a much longer history of evolution, all of these hydrogen bonds are found to be dispensable (Table 7-6).

A deeply buried position in the sequence of a folded

polypeptide remains invariably hydrophobic, but a buried location near the surface will occasionally erupt toward the water. For example, at position 65 in cytochrome *c* (Figure 7–1) an arginine appears at a location usually occupied by hydrophobic amino acids. Presumably the alkane portion of the arginine traverses the hydrophobic region and the guanidinium can push through the surface into the solvent.

Often a hydrophilic location on the exterior is occupied by a hydrophobic amino acid. For example, position 9A in the globins (Table 7–5) is on the exterior of the protein and is usually occupied by hydrophilic amino acids, but in the myoglobins it is occupied by leucine. Because such a substitution has no effect on the free energy of folding for myoglobin compared to the other globins because the leucine is solvated equivalently in both the unfolded and folded polypeptide, such exchanges are common during evolution. There is, however, a price to be paid for such an exchange because a hydrophobic amino acid that replaces a hydrophilic amino acid on the surface of a protein makes it less soluble. The helical polymers formed by human deoxyhemoglobin S, in which a glutamate on the surface at position 4A has been replaced by a valine, are an example of such a problem.

The globins also provide a particularly informative example of the **focused constraints** that natural selection places on the gradual shifts in position among segments of secondary structure during evolution. The invariant feature of both the structure and the function of a globin is the heme (Figure 4–18). The only functions of a globin are to provide a fifth ligand to the iron, to make its heme soluble in water, and to prevent its heme from approaching another heme too closely. Through all of the alterations encountered during evolution, the amino acids responsible for surrounding the heme and supporting it within the protein were required by natural selection to maintain these roles. The record of this series of accommodations can be inferred from superposing crystallographic molecular models of present globins so that their hemes are made to coincide rather than their polypeptides. The situation is most graphically illustrated when the α subunit from equine hemoglobin is superposed in this way on leghemoglobin from *Lupinus luteus* (Figure 7–10).^{117,130} Over this long period of evolution, the amino acids supporting the heme have shifted their positions relative to it by only small distances. At the same time, however, the ends most distant from the heme of the two α helices in which these functionally critical amino acids reside (E and G in Figure 7–10) have

Table 7–5: Role and Location of Invariant Residues in Myoglobin^a

structural location ^b	amino acid ^c	location ^d	role ^e
2NA	LEU	I	in contact with helix H: L ^f
1A	SER	E	hydrogen bond to GLU 4A NH to initiate helix A: TS
4A	GLU	S	hydrogen bond to SER 1A NH to stabilize LEU 2NA: DQE
5A	TRP	S	between LEU 2NA and LYS 2EF: KWRIA
8A	VAL	I	in contact with helix H; in bottom hydrophobic cluster: IV
9A	LEU	E	protruding into solvent: KTLRAE
12A	TRP	S	hydrogen bond to GLU 16A, which in turn is bound to LYS 20E to hold helices A and E together (Hb) ^h :
WF			
14A	LYS	E	hydrogen bond to asp (glu, gln, asn) 4GH to stabilize GH-corner (Hb): KPDE
15A	VAL	I	in bottom hydrophobic cluster: VIF
16A	GLU	E	hydrogen bond to TRP 12A: G-EYAKN
IB	ASP	E	hydrogen bond to gly 4B NH to stabilize AB-corner: HND
5B	HIS	I	hydrogen bond to HIS 1GH to stabilize GH-corner: YVHSD
6B	GLY	I	in close contact with GLY 8E: GPT
10B	LEU	I	in hydrophobic cluster on HIS 7E side (Hb): LF
11B	ILE	S	blocking an opening between helices B and D: EGIIVY
13B	LEU	I	in hydrophobic cluster on HIS 7E side: MLFHV
14B	PHE	I	in hydrophobic cluster on HIS 7E side: FL
1C	HIS	S	in close contact with phe (leu) 7G: FYHTDA
2C	PRO	E	sharp turn from helix B to helix C (Hb): P
3C	GLU	E	hydrogen bond to GLU 3C NH: TWEAS
4C	THR	I	van der Waals contact with heme, hydrogen bond to HIS 1C CO to initiate helix C (Hb): TAMI
5C	LEU	S	blocking an opening formed by helix C and CD-corner: KQLEAMK
6C	GLU	E	in contact with LYS 8CD through a water molecule: TAERD
1CD	PHE	I	van der Waals contact with heme parallel to heme plane; in hydrophobic cluster on HIS 7E side (Hb): F
2CD	ASP	E	hydrogen bond to LYS 5CD to stabilize CD-corner: PEDGTS
4CD	PHE	I	in hydrophobic cluster on HIS 7E side of heme (Hb): F
5CD	LYS	E	hydrogen bond to ASP 2CD to stabilize CD-corner: G-KSAL
6CD	HIS	E	in contact with ASP 2CD CO through a water molecule to stabilize CD-corner: DHGK
7CD	LEU	S	in hydrophobic cluster on HIS 7E side (Hb): L-G
8CD	LYS	E	in contact with ASP 2CD CO through a water molecule to stabilize CD-corner: SKTG

Table 7-5: Role and Location of Invariant Residues in Myoglobin^a – continued

structural location ^b	amino acid ^c	location ^d	role ^e
5D	MET	I	blocking an opening formed by CD-corner and helix D: VMLIP
1E	SER	S	hydrogen bond to LEU 4E CO to initiate helix E: NSDT
2E	GLU	E	in contact with lys (arg) 5E through a water molecule: APE
4E	LEU	I	in hydrophobic cluster on HIS 7E side: VLF
6E	LYS	E	hydrogen bond to neighboring molecule: KRAEQ
7E	HIS	I	hydrogen bond to the sixth ligand of the heme; van der Waals contact with heme (Hb): HL
8E	GLY	I	in close contact with GLY 6B (Hb): GA
11E	VAL	I	van der Waals contact with heme; in hydrophobic cluster on HIS 7E side: VI
12E	LEU	I	in bottom hydrophobic cluster: ALGIVF
14E	ALA	S	van der Waals contact with heme: ASEFL
15E	LEU	I	in bottom hydrophobic cluster. van der Waals contact with heme vinyl group: LFVI
16E	GLY	S	in contact with TRP 12A: TSGDY
18E	ILE	I	in hydrophobic cluster on HIS 8F side: AGI
19E	LEU	I	in bottom hydrophobic cluster: VLIA
20E	LYS	E	hydrogen bond to GLU 16A to keep helices A and E stable: AGHKSI
1EF	LYS	E	hydrogen bond to neighboring molecule: HKSEQ
2EF	LYS	E	hydrogen bond to glu (asp) 2A to stabilize amino terminus of helix A: DKG–
3EF	GLY	E	in contact with solvent: –TGV
5EF	HIS	S	hydrogen bond to ASP 18H to stabilize EF-corner and helix H: MLHKIS
7EF	ALA	E	in contact with solvent: NGAS
4F	LEU	I	in hydrophobic cluster on HIS 8F side; in contact with heme (Hb): LVF
5F	ALA	S	in close contact with helix H: SAGV
7F	SER	S	van der Waals contact with heme; hydrogen bond to pro (his) 3F CO or LEU 4F CO or HIS 8F N: LSKRV
8F	HIS	I	the fifth ligand to heme iron (Hb): H
9F	ALA	S	in close contact with helix H: AKV
1FG	LYS	E	no electron density (Hb): KSYR
2FG	HIS	S	van der Waals contact with heme; hydrogen bond to propionic acid residue to stabilize heme and FG-corner: LHFG
3FG	LYS	E	no electron density: RHKEV
1G	PRO	E	sharp turn from FG-corner to helix G: DPKTA
5G	LEU	I	van der Waals contact with heme; in hydrophobic cluster on HIS 8F side: FL
6G	GLU	E	hydrogen bond to ARG 16H to stabilize helices G and H: KRENP
8G	ILE	I	van der Waals contact with heme: LIFV
9G	SER	S	hydrogen bond to LEU 5G CO: SGARK
12G	ILE	I	in bottom hydrophobic cluster: LIF
1GH	HIS	S	hydrogen bond to HIS 5B to stabilize GH-corner: LFHITV
5GH	PHE	I	in bottom hydrophobic cluster: FMW
3H	ASP	E	hydrogen bond to ala (val) 4H NH: APED
7H	ALA	S	in close contact with helix A: SAG
8H	MET	I	in bottom hydrophobic cluster: LYMFV
10H	LYS	E	hydrogen bond to GLU 4A to stabilize amino-terminal end of helix A (Hb): KAI
11H	ALA	I	in bottom hydrophobic cluster: FVALT
12H	LEU	I	in bottom hydrophobic cluster: LVY
13H	GLU	E	in contact with asn 9H through a water molecule: ASERD
14H	LEU	E	protruding into solvent: SGLMDTE
15H	PHE	I	van der Waals contact with heme; in hydrophobic cluster on HIS 8F side: VFIL
16H	ARG	S	hydrogen bond to GLU 6G to stabilize helices G and H: SARF
18H	ASP	E	hydrogen bond to HIS 5EF to stabilize EF-corner and helix H: VADFM
20H	ALA	E	in contact with solvent: TARIFK
23H	TYR	I	hydrogen bond to ile (val) 4FG CO to cap helix H (Hb): YLM
24H	LYS	E	hydrogen bond to the carboxy terminus: RHKED
1HC	GLY	E	
4HC	GLY	E	

^aAdapted from Takano.⁹⁴ The amino acids listed are those that are invariant in all myoglobins, and the structural roles assigned are those in the crystallographic molecular model (Bragg spacing ≥ 0.2 nm) of myoglobin from *Physeter catodon*. ^bPosition in the common crystallographic molecular model of the globins. Capital letters (A–H) indicate which α helix, from amino- to carboxy-terminal, and the numbers indicate the position in the α helix. Double letters refer to turns between the respective helices. The globins are all bundles of eight α helices (Figure 4–18). ^cAmino acids that are invariant over all myoglobins. ^dLocation in the crystallographic molecular model of myoglobin: I, internal; E, external; S, surface crevice. ^eThree-letter amino acid abbreviations given in uppercase letters represent invariant residues in myoglobin; those given in lowercase letters are not invariant. *Amino acids appearing at each of these positions in nine superposed globins¹¹⁷ are noted in one-letter code. Dash indicates deletion. ⁸Amino acids noted with (Hb) are invariant in all mammalian hemoglobins.

Table 7-6: Evolutionary Variation of Ionized Hydrogen Bonds^a

structural location ^b	pairs of amino acids
2CD	P E P D D P G T S ^c
5CD	— G — G K K S A L
5EF	M L L L H K M I S
18H	V A V A D L A M V
IB	H N H N D D N — N
19G	H H H H R G R H V
2EF	D D D D K D D — G
2A	P P A G E A A A E

^aFour invariant ionized hydrogen bonds that were present in an earlier refined crystallographic molecular model (Table 7-5)⁹⁴ and a later refined crystallographic molecular model (Bragg spacing ≥ 0.16 nm)¹²⁹ of myoglobin were chosen for examination. Each of the four pairs is between the amino acids in boldface type above and below each other in the central positions of the four paired strings of letters. The two amino acids forming each of these four hydrogen bonds in the two crystallographic molecular models, aspartate and lysine, histidine and aspartate, aspartate and arginine, and lysine and glutamate, were conserved among all of the myoglobins. The amino acids occupying each of these eight positions in a structural alignment of eight other globins¹¹⁷ are listed to the right and left of the pair occupying each of the eight positions in myoglobin. ^bCode assigned to the positions in the common crystallographic molecular model of the globin class (Table 7-5). ^cThe amino acid at the respective position in each of the globins is aligned above or below the amino acid at the other position in the same globin.

shifted significantly in their position, and another α helix that provides no amino acids in contact with the heme (B in Figure 7-10) has shifted even more.

In any protein, a few amino acids that embody its function can be identified. Over evolution, natural selection maintains the relative separations and orientations of these amino acids because if it did not, the protein could no longer be what it is. An extreme example of this fact is found in the group of related enzymes to which phosphopyruvate hydratase, mandelate racemase, galactonate dehydratase, glucarate dehydratase, muconate cycloisomerase, and methylaspartate ammonia-lyase belong. Although each of these enzymes has diverged widely from its distant common ancestor, the positions of the **functional groups in the active sites** of these proteins that are responsible for the abstraction of the proton α to the respective carboxylate, a function common to the mechanism of each of them, have been conserved.¹¹⁸ The more distant a location within the protein is from such invariant points of reference, however, the more likely its position will drift as mutations accumulate that shift the orientations of the segments of secondary structure within the overall molecular structure of the protein.

An exception to this rule that functional groups are usually the most invariant features of a protein can be seen in a comparison of the crystallographic molecular models of the phospholipase A₂ from cobra venom and the phospholipase from bee venom.¹³¹ From aligned amino acid sequences and superposed crystallographic molecular

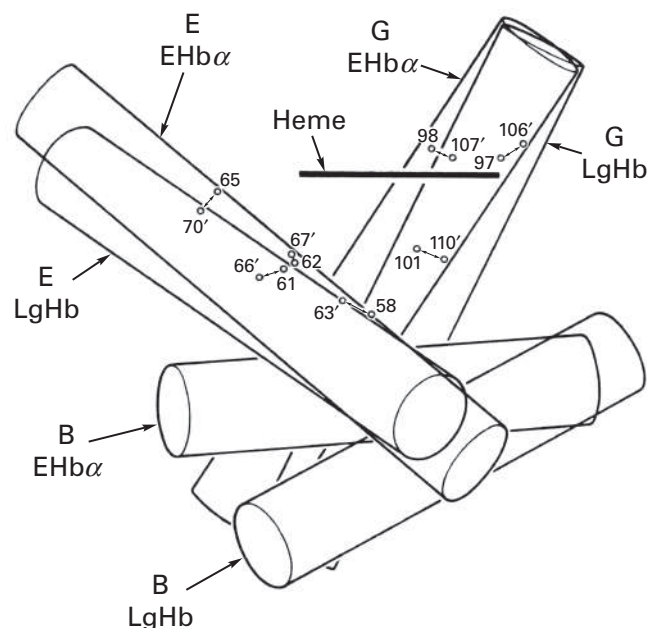


Figure 7-10: Arrangement of the α helices and contact side chains that form part of the heme pocket in the α subunit of equine hemoglobin (EHb α) and in leghemoglobin II from *Lupinus luteus* (LgHb).¹³⁰ The hemes in the two proteins are superposed. The three α helices are designated as B, E, and G, in order of their appearance in the globin molecule. The positions of homologous pairs of side chains (sequence positions in leghemoglobin are primed) that are in contact with the heme are indicated by open circles joined by arrows. The coupling of the shifts at the E-B and B-G helix interfaces keeps the side chains that form the heme pocket in the same relative positions. Reprinted with permission from ref 130, originally from ref 117. Copyright 1980 Academic Press.

models, there is no doubt that these two proteins share a common ancestor. The α helix in the protein from cobra venom that binds to the surface of a biological membrane, in which the reactants for the enzyme are found, is formed by the first 20 amino acids of the polypeptide, but these 20 amino acids are missing from the protein from the bee. The missing tertiary structure necessary for adhering the protein to the membrane is supplied by an additional α helix at the carboxy-terminal end of the polypeptide from the bee that takes the place of the α helix from the amino-terminal end of the polypeptide from the cobra. It was the ability to superpose the crystallographic molecular models of these two proteins that permitted the substitution of one segment of the polypeptide for another in a functional role to be demonstrated.

The inability to superpose two crystallographic molecular models of two proteins can demonstrate an example of convergent evolution. For proteins, **convergent evolution** is the assumption of the same function by two proteins that do not share a common ancestor. For example, as had been predicted by aligning sequences,⁴ chorismate mutase from *S. cerevisiae* cannot be superposed on chorismate mutase from *B. subtilis*.¹³² Consequently, it can be concluded that these two unrelated proteins nevertheless evolved so

that they each were able to perform the same function. Other examples of such convergent evolution are the 3-dehydroquinate dehydratases from *Salmonella typhimurium* and *Mycobacterium tuberculosis*¹³³ and the Cu,Zn-superoxide dismutase of *B. taurus*¹³⁴ and the Fe-superoxide dismutase of *M. tuberculosis*.¹³⁵ A particularly interesting example of convergent evolution that was elucidated by superposing crystallographic molecular models is that of the [2Fe-2S] ferredoxin from *Clostridium pasteurianum*. This protein is completely unrelated to the other [2Fe-2S] ferredoxins and turns out to be a thioredoxin that has been converted into a ferredoxin.¹³⁶ The more common examples of convergent evolution, however, are those in which two unrelated proteins catalyze similar but not identical functions. For example, although they share the same mechanism of activating molecular oxygen for insertion into a carbon-hydrogen bond and both use a heme to do so, crystallographic molecular models of nitric-oxide synthase and cytochrome P-450 have different, unrelated structures.¹³⁷

Often the catalytic amino acids in the active sites of examples of convergent evolution are arranged similarly in the two proteins even though the overall structures are completely different. The serine, histidine, and aspartate responsible for the nucleophilic catalysis in the active site of subtilisin are superposable on the serine, histidine, and aspartate in the active site of chymotrypsin even though the two proteins themselves cannot be superposed,¹³⁸ and the functional groups in the active site of alanine dehydrogenase are similarly arranged to those in L-lactate dehydrogenase.¹³⁹ The catalytic amino acids, however, around the flavin adenine dinucleotide in the active sites of the two flavoenzymes L-lactate dehydrogenase (cytochrome) and D-amino-acid oxidase are arranged in patterns that are mirror images of each other.¹⁴⁰

As the number of crystallographic molecular models has increased, instances have become more common in which two proteins that display no similarity in amino acid sequence nevertheless have segments of their tertiary structure that can be superposed. An example of such a **segment of recurring structure** is found in the crystallographic molecular models of L-lactate dehydrogenase,⁸² alcohol dehydrogenase,¹⁴¹ phosphoglycerate kinase,¹⁴² and phosphorylase.¹⁴³ This common segment is 140–200 aa in length and occurs at different locations in the overall sequences of these proteins. It is formed from the amino acids in the sequence between Asparagine 21 and Glycine 162 in isoform A of L-lactate dehydrogenase from *Squalus acanthius*, between Phenylalanine 207 and Serine 392 in equine phosphoglycerate kinase, between Serine 193 and Phenylalanine 319 in isoform E of equine alcohol dehydrogenase, and between Asparagine 559 and Arginine 713 in the isoform of glycogen phosphorylase from muscle of *Oryctolagus cuniculus*. All four structures can be superposed.^{82,143} The superposition of these

regions from phosphorylase and L-lactate dehydrogenase is presented in Figure 7–11A.^{82,143}

This particular topological pattern of secondary structures can be identified as a **doubly wound, parallel β sheet**. It is a sheet of six parallel β strands flanked on both sides by α helices. The basic rhythm of the recurring theme is β strand, α helix, β strand, α helix, β strand, random meander, β strand, α helix, β strand, α helix, β strand. The β strands numbered from amino terminus to carboxy terminus occur in the order 321456 across the sheet (Figure 7–20). You should trace this pattern in Figure 7–11A. The six β strands all run parallel to each other to form a pleated sheet, and the helices arch above or below the sheet to connect the end of one β strand to the next. The complete and concise theme is developed in L-lactate dehydrogenase (Figure 7–11), and there are variations on this theme in the other proteins. For example, in phosphoglycerate kinase, there are two α helices after the second β strand and a long additional loop after the third β strand, and in alcohol dehydrogenase the last α helix is replaced by an additional antiparallel β strand. An interesting variation occurs after the first β strand in the structure from phosphorylase, where a large bulge has appeared in the first β strand that pushes up the loop between the third and fourth strands of β structure (Figure 7–11A).

Flavodoxin shares this pattern but with a more significant variation. In this protein, the second α helix and the third β strand have been deleted (Figure 7–11B). This deletion seems to have been very similar to those seen in cytochrome *c* (Figure 7–9) in that the loop containing the α helix and the β strand has simply been pinched off from the open end of the common structure. Nevertheless, the superposition of flavodoxin upon the corresponding region from L-lactate dehydrogenase is quite close, even though the sequences of these two superposed polypeptides appear to be completely unrelated. When the sequences are aligned, even with the assistance of the superposition, they have identical amino acids in only 9% of their aligned positions.

The conclusion that has been drawn from these superpositions is that all of these regions from these very different proteins together share a common ancestor. As these structures represent only a portion of each of the presently existing proteins, and as the other portions of the proteins bear no resemblance to each other, this common ancestor must have been a small primordial protein that was combined covalently with other small primordial proteins by gene fusion to produce respectively these larger chimeric proteins. **Gene fusion** is a process in which genomic DNA is recombined incorrectly so that segments of different genes become fused together rather than, as in the usual process of recombination, allelic segments of the same gene being interchanged in precise alignment. Like gene duplication, gene fusion occurs frequently, but only infrequently will

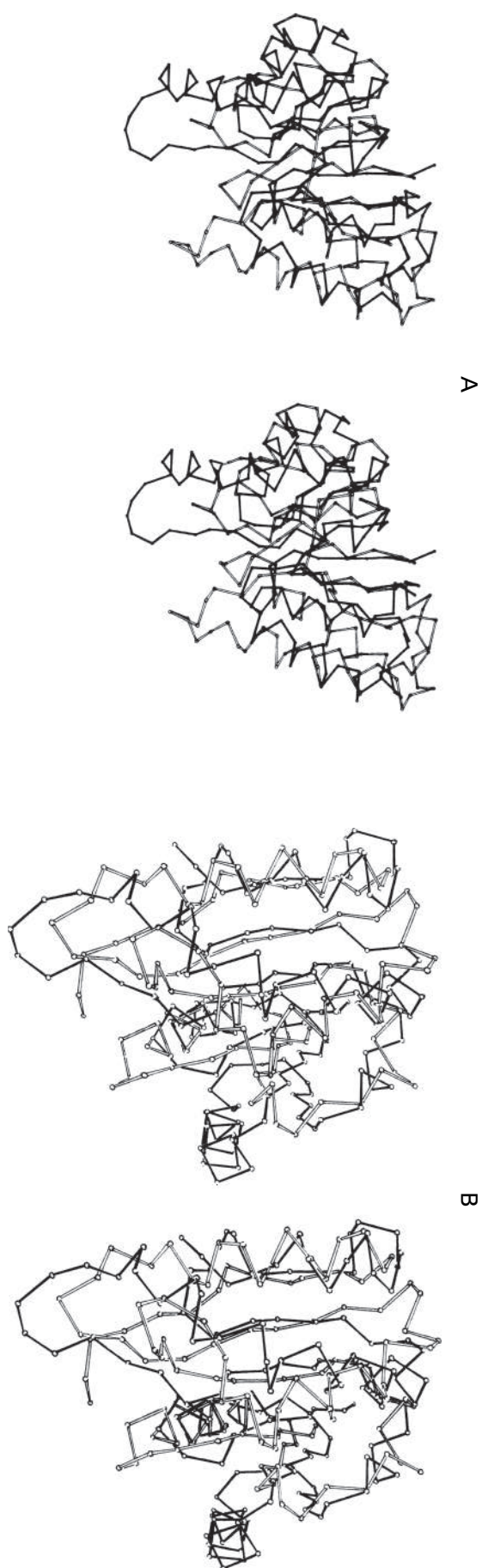


Figure 7-11: Doubly wound, parallel β sheets found in three crystallographic molecular models. (A) An α -carbon diagram of the doubly wound, parallel β sheet found in isoform A of L-lactate dehydrogenase from *Squalus acanthias* (solid lines) superposed on the α -carbon diagram of this structure found in glycogen phosphorylase from muscle of *Oryzotilus cinctus* (open lines).¹⁴³ Reprinted with permission from ref 143. Copyright 1976 *Journal of Biological Chemistry*. (B) An α -carbon diagram of the doubly wound, parallel β sheet found in L-lactate dehydrogenase (solid lines) superposed on the α -carbon diagram of this structure found in flavodoxin from *Clostridium beijerinckii* (open lines).⁸² In both panels the amino termini of the α -carbon diagrams are the upper ends of the middle β strands of the sheets; the carboxy termini are the ends of the strands at the side of each of the sheets. Reprinted with permission from ref 82. Copyright 1974 *Nature*.

the gene that results from the fusion spread over the entire population of a species by genetic drift and become fixed in its genome. The evidence of early gene fusions that did spread over a primordial population can still be observed in the progeny, not only in their structures but also in the distribution of methionines that are the fossilized remains of the initiation sites of the smaller ancient proteins that were fused together.¹⁴⁴ The reason that each of the primordial proteins now resides at a different location in the sequences of the present proteins is that each was combined with different proteins in different orders during these gene fusions. This description of evolutionary history requires that at one time in the distant past each of the segments of these polypeptides now folding to produce each of these superposable regions was not attached to the remainder of the polypeptide to which it is now joined. If this is the case, then each of the doubly wound β -pleated sheets and the other regions in each of these proteins to which they are now attached were at one time separate proteins, and a significant part of the evolution of proteins is a history of the joining together of smaller proteins to produce ever larger proteins.

Suggested Reading

Rossmann, M.G., Moras, D., & Olsen, K.W. (1974) Chemical and biological evolution of a nucleotide-binding protein, *Nature* 250, 194–199.

Problem 7-6: The following is a portion of a multiple structural alignment of the amino acid sequences of 10 members of the chymotrypsin family of serine endopeptidases from Asparagine 148 to Glycine 196 of chymotrypsin, the amino acid sequence of which is at the top.¹⁴⁵

```

1  NANTPDRLQQASLPLLSNTNCKK--YWGTKIKD-AMICAG-AS----GVSSCMGDSG
2  gtSYPDVLKCLKAPILSDSSCKS--AYPGQITS-NMFCAG-Yleg--gKDSCQGDSG
3  -gqLAQTLQQAYLPTVDYAICSsssYWGSTVKN-SMVCAG-Gdg---vRSGCQGDSG
4  gKGQPSVLQVVNLPIVERPVCKD--STriRITD-NMFCAGykpdegkRGDACEGDSG
5  dfeFPDEIQCVQLTLLQNTfcAd--AHpdKVTE-SMLCAG-Ylpg--gKDTCMGDSG
6  -dptsytLREVELRIMDEkacVd--YR--yYEykFQVCVGSPT---tLRAAFMGDSG
7  -----GLRSGSVTGlnatvn--ygssgivy-gMIQTN-----vCAQPGDSG
8  -----GTHSGSVTAlnatvn--ygggdvvy-gMIRTN-----vCAEPGDSG
9  -----GYQCGTITaknvtan--ya--egavrgLTQGN-----aCMGRGDSG
10 -----hGAVQYsgg-----rFT-ip----rgvgGRGDSG

```

The alignment is based on the separate superpositions of crystallographic molecular models of each of the other nine proteins upon the crystallographic molecular model of chymotrypsin, which was chosen as the reference structure for the family. The similarity of the structures to that of chymotrypsin is given by the case and the face of the one-letter code of the amino acids. An uppercase boldface character represents an α carbon that is within a distance of 0.15 nm of the equivalent α carbon in chymotrypsin. An uppercase normal character represents a distance within 0.25 nm, a lowercase boldface character represents a distance within 0.35 nm, and a

lowercase normal character represents a distance of greater than 0.35 nm. A dash represents a gap.

- On a sheet of graph paper, construct a dot matrix for a comparison of the sequence of protein 3 and the sequence of protein 6 between the positions of the first and the third cysteines in the amino acid sequence of chymotrypsin, which is protein 1.
- Trace through the dot matrix the structural alignment of protein 3 and protein 6.
- What is the percentage of identity for the alignment in this segment?

Problem 7-7: The following is the amino acid sequence of the pyruvate kinase from rabbit muscle from Proline 116 to Proline 218.

```

PEIRTGLIKSGTAEVELKKGATLKITLDNAYMEKCDE
NILWLDYKNICKVVDVGSKVYVDDGLISLQVKQKGPDF
LVTEVENGGFLGSKKGVNLPGAAVDLP

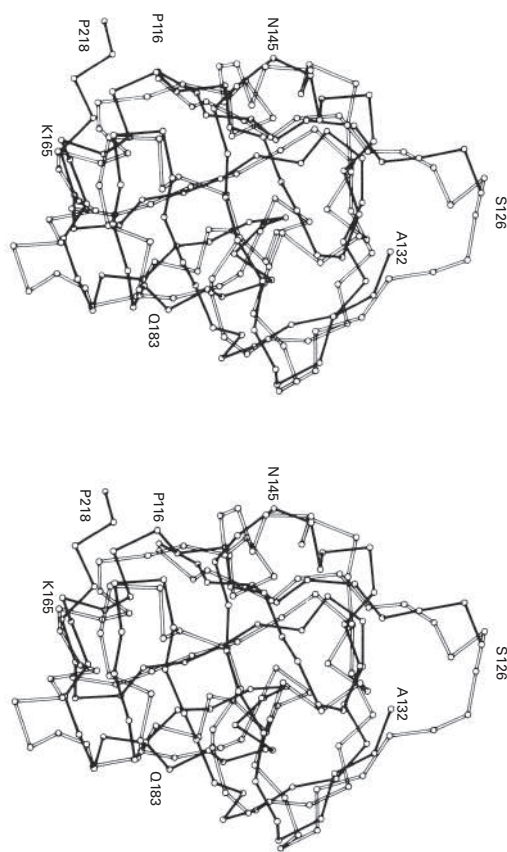
```

The following is an alignment of the amino acid sequences for pyruvate kinase from cat muscle, chicken muscle, rat liver, and yeast over the corresponding segments.

cat muscle	PEIR T GLIKSGTAEVELKKGATLKITLDNAYMEK C DENVLWLD
chicken muscle	PEIR T GLIKSGTAEVELKKGAALKVTLDNAFMENCDENVLWVD
rat liver	PEIR T GV L QGGPES E VEIVKGSQVLVTVDPKFQTRGDAKT V WVK
yeast	PEIR T G--TTTNDVDYPIPPNH E MIFTTDDKYAKACDDKIMYVD
cat muscle	YKNICKVVEVGSKVYVDDGLISLLVKEKG-ADFLVTEVENGG S L
chicken muscle	YKN L IKVIDVGSKIYVDDGLISLLVKEKG-KDFVMTEVENGG M L
rat liver	YHNITRVVAVGGRIYID D GLISLVQKIG-PEGLVTEVEHGG I L
yeast	YKNITKVISAGRIIYVDDGVLSFQVLEVVD D DKTLKV K ALNAGKI
cat muscle	GSKKGVNLP G AAVDLP
chicken muscle	GSKKGVNLP G AAVDLP
rat liver	GS R KGVNLPNTEVDLP
yeast	CS H KGVNLP G TDVDLP

- (A) Why are the amino acid sequences of the proteins from cat and chicken so much more similar to each other than are the sequences from the cat and the rat to each other?
- (B) Align the amino acid sequences of the proteins from rabbit muscle and cat muscle from Proline 116 to Proline 218. What is the percent identity and how many gaps are there?

The following figure¹⁴⁶ is a superposition of the α carbons between Proline 116 and Proline 218 in the crystallographic molecular models of the pyruvate kinases from rabbit muscle and cat muscle.



These portions of the models were superimposed according to the algorithm of Rossmann and Argos.¹⁴⁷ The models of the rabbit and cat enzymes are displayed with filled and unfilled lines, respectively. Those amino acids that are labeled correspond to the protein from rabbit.

- (C) What is wrong with this figure?
- (D) What is the reason that something is wrong with this figure?

Domains

Ferredoxin-NADP⁺ reductase from spinach is a protein composed of a folded polypeptide 314 aa long (Figure 7-12).^{148,149} In the native enzyme, the polypeptide between Asparagine 162 and Tyrosine 314 assumes a doubly wound, parallel β sheet of five strands (upper right of Figure 7-12) and the polypeptide between Glycine 26 and Glutamate 154 assumes an antiparallel β barrel (lower left of Figure 7-12). Doubly wound, parallel β sheets recur in many different proteins and antiparallel β barrels recur in many different proteins, but the two structures are usually not found in the same protein. In the crystallographic molecular model of ferredoxin-NADP⁺ reductase, the doubly wound, parallel β sheet seems to be folded independently from the antiparallel β barrel. Only one strand of polypeptide runs between them.

A large number of observations, among them the ones just described, have led to the conclusion that the native structures of most folded polypeptides can be divided into independent domains. A **domain** is any region within the native tertiary structure of a folded polypeptide for which evidence can be provided of an existence independent of the rest of the polypeptide. There are several types of independent existence that qualify a region within the native structure as a domain.

The most obvious evidence that two regions of the same protein are domains is that either a limited cleavage of the polypeptide or the expression of separate portions of the polypeptide produces independent fragments of the protein that retain their respective native structures. In such instances, the two or more separated fragments would be the **detachable domains** that composed the intact protein. The paradigm of a protein with detachable domains is immunoglobulin G, a circulating antibody responsible for binding to foreign proteins or other antigens. Porter¹⁵⁰ demonstrated that intact immunoglobulin G could be cleaved by the thiol endopeptidase papain into three pieces of almost equal sizes. Two of these detachable domains, the Fab fragments, retained the ability to bind antigens, and the third, the Fc fragment, was stable enough that it crystallized spontaneously during its isolation. The Fab fragments could be readily separated by cation-exchange chromatography from the Fc fragments without any loss of their biological activity. Comparisons of crystallographic molecular models of Fab fragments and Fc fragments with those of intact immunoglobulins G (Figure 7-13)^{151,152} have demonstrated that the detached and separated domains retain the respective structures that they had in the intact molecule before it was cleaved.¹⁵³

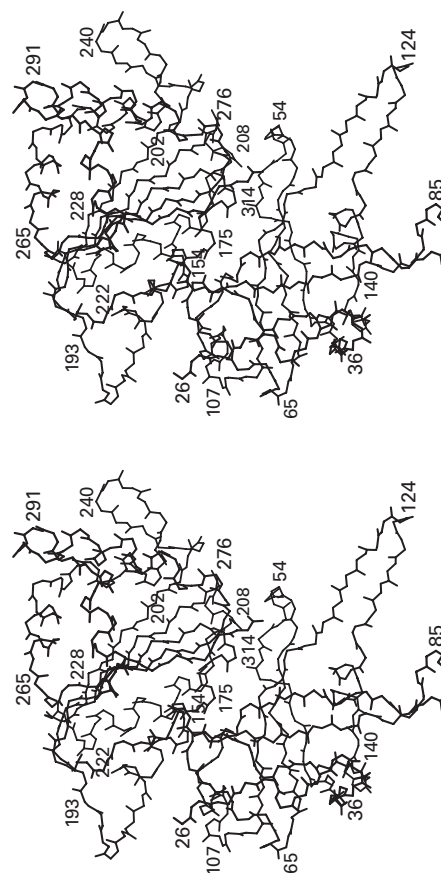
It was unfortunate that an unintended association between limited cleavage with endopeptidases and domains was established with these elegant experiments. The important fact was that Porter separated the detached

domains from each other and demonstrated that each was still structurally intact. It has already been noted that a protein must be prepared for complete digestion with endopeptidases by unfolding it. The reason for this is that most of the **peptide bonds susceptible to digestion** with a particular endopeptidase in the unfolded polypeptide are not susceptible to digestion in the native folded polypeptide. In fact, many native proteins are resistant to digestion over their entire length until they are unfolded.¹⁵⁴ When the digestion of a native protein does occur, it usually occurs at only one or two locations. The sites at which cleavage of a native protein by endopeptidases can occur are exposed, flexible loops of polypeptide on its surface that are rarely situated between domains, and domains are often not connected by such loops.¹⁵⁵⁻¹⁵⁷ If a polypeptide is cut at only one or two positions by an endopeptidase when it is folded in its native conformation, this is not evidence that the fragments observed compose separate domains in that native conformation.¹⁵⁸

If, however, the protein can be digested and the pieces that result can be separated as biologically active or structurally intact moieties, they are detachable domains. Few examples of such **endopeptidolytic detachments** have been reported, and among those are the following. A protein anchored to mammalian cellular membranes contains within its single folded polypeptide the two enzymes peptidylglycine monooxygenase and peptidylamidoglycolate lyase. This protein can be digested either during normal cellular processes or by experimental treatment with an endopeptidase to produce two soluble, detached domains, which can be separated chromatographically, one of which catalyzes the former activity and the other of which catalyzes the latter.¹⁵⁹ The enzyme sulfite oxidase can be cleaved with either trypsin, chymotrypsin, or papain to produce two detached domains that can be separated from each other by molecular exclusion chromatography.¹⁶⁰ One retains the ability to transfer electrons from sulfite to $\text{Fe}(\text{CN})_6^{3-}$; the other retains the spectrum characteristic of the heme in its native environment. The transfer of electrons from sulfite all the way to the ultimate oxidant, cytochrome *c*, can no longer occur because the domain containing the heme is no longer attached to the domain at which the sulfite is oxidized. Anion carrier is a protein in the plasma membrane of erythrocytes and is responsible for anion transport. It can be cleaved with chymotrypsin to produce a water-soluble domain that can be readily separated from the other domain, which remains in the membrane.¹⁶¹ Between them the two detached domains retain the biological functions that are displayed by the intact protein, and each retains the structure it had in the intact protein.

Once the cDNA or genomic DNA encoding a protein has become available and it has become clear exactly where the boundaries between its domains are located, they can often be detached genetically and each

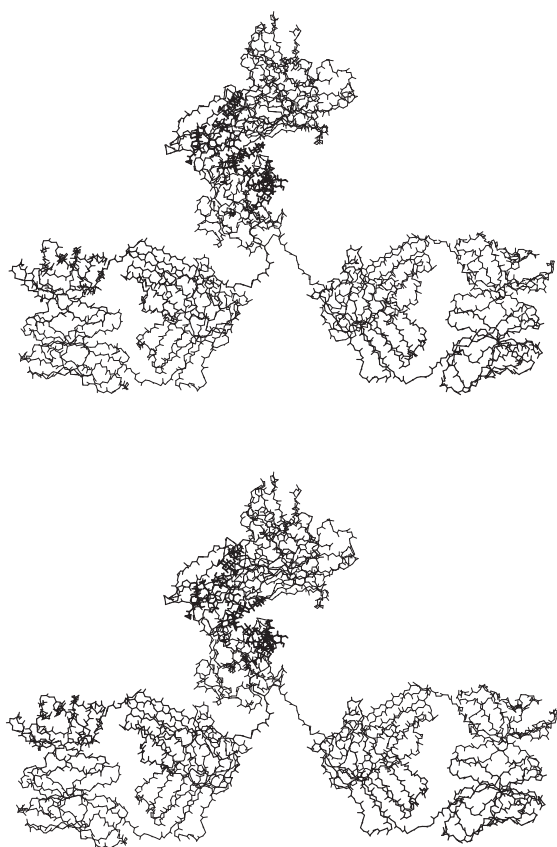
Figure 7-12: Skeletal representation of the backbone of the crystallographic molecular model (Bragg spacing ≥ 0.17 nm) of ferredoxin-NADP⁺ reductase (314 aa) from *S. oleracea*.^{148,149} The published molecular model included Histidine 19 through Tyrosine 314, but the drawing begins at Glycine 26. The doubly wound, parallel β sheet is in the upper right portion of the drawing; the Greek key, antiparallel β barrel, in the lower left. Glutamate 154 identifies the covalent junction between the two domains. Note the irregular cleft between the two domains; only one strand of polypeptide, that containing Glutamate 154, runs across the cleft. This drawing was produced with MolScript.⁴⁰⁹



of them expressed separately. There are many examples of such **genetic detachments**. The two domains that catalyze indole-3-glycerol-phosphate synthase and phosphoribosylanthranilate isomerase within the bifunctional enzyme of *E. coli* can be genetically detached and expressed as separate monofunctional proteins.¹⁶² The two domains of protein S from *Myxococcus xanthus* can be genetically detached and expressed as stable monomeric proteins¹⁶³ that show no tendency to associate with each other and both of which retain their ability to bind Ca^{2+} . The domains of the AraC

protein from *E. coli* can be genetically detached and expressed as separate proteins that exhibit respectively the abilities of the intact native protein to recognize DNA containing its operator and to dimerize in an arabinose-dependent manner.¹⁶⁴ Genetic deletion of the carboxy-terminal 154 aa from cystathionine β -synthase (507 aa) actually increases its enzymatic activity.¹⁶⁵

The advantage of performing a detachment genetically is that it can be **accomplished at the precise boundary** and does not require that the polypeptide between the two domains be as extraordinarily available for digestion by an endopeptidase as are the connecting strands



between the domains in an immunoglobulin G (Figure 7-13). Consequently, there are far more examples of domains that have been successfully detached genetically than have been detached endopeptidolytically. When a domain that is buried in the structure of the native protein and consequently has an excess of hydrophobic side chains on its surface is detached genetically, it is also possible to mutate some of these hydrophobic amino acids to hydrophilic amino acids to prevent the detached domain from aggregating.¹⁶⁶

There is a protein in *E. coli* that is responsible for the two enzymatic activities of aspartate kinase and homoserine dehydrogenase. It is composed of one folded polypeptide 820 aa in length. When this protein is digested with glutamyl endopeptidase from *Streptomyces*

Figure 7-13: Crystallographic molecular model (Bragg spacing ≥ 0.28 nm) of a murine monoclonal immunoglobulin G.¹⁵¹ The two Fab domains are aligned vertically, one above the other on the right-hand side of the figure, and the Fc domain is to the left. There are four folded polypeptides that together compose an intact molecule of immunoglobulin G. In this particular case, they are two identical polypeptides 214 aa long, the light chains, and two identical polypeptides 444 aa long, the heavy chains. In the figure, one heavy chain and one light chain are drawn with thicker lines; and the other heavy chain and the other light chain are drawn with thinner lines. The light chains are each confined to their respective Fab domains, but the heavy chains each form respectively the other half of one of the Fab domains, and they also associate with each other to form the Fc domain. Each of the three detachable domains, the Fc domain and the two Fab domains, is itself formed by four internally repeating domains each about 110 aa long composed of seven antiparallel β -strands. All 12 of these smaller internally repeating domains are superposable, one on the other, and arose from multiple gene duplications. The two most peripheral internally repeating domains in the Fc detachable domain are formed respectively by the carboxy-terminal portions of the two heavy chains and associate directly with each other, as do the four respective pairs of internally repeating domains in the two Fab detachable domains. The two centrally located internally repeating domains in the Fc detachable domain, however, associate with each other through their oligosaccharides, which are drawn with the thickest lines. The two strands of polypeptide connecting each Fab fragment with the Fc fragment are open, flexible, and unstructured; consequently these three detachable domains, each a rigid body, are in constant rearrangement relative to each other when immunoglobulin is in solution. This drawing was produced with MolScript.⁴⁶⁸

griseus, it is cut into two large fragments and a short peptide as a result of cleavages following Glycine 296 and Glutamate 301.¹⁶⁷ Because chymotrypsin digests the protein only at Leucine 294; subtilisin, only at Alanine 297; and trypsin, only at Arginine 299,¹⁶⁷ the region containing these five sites of cleavage must be in a loop of about 10 amino acids on the surface of the protein. The two large fragments from digestion with glutamyl endopeptidase can be separated from each other and from their parent on chromatography by molecular exclusion. The carboxy-terminal fragment retains the homoserine dehydrogenase activity. The amino-terminal fragment, however, shows only 1% or less of the aspartate kinase activity originally present, and this small amount of activity probably results from cross-contamination with uncleaved protein. A fragment retaining normal levels of aspartate kinase activity, however, can be prepared genetically by deleting the carboxy-terminal 45% of the polypeptide.¹⁶⁸ In *B. subtilis*, aspartate kinase and homoserine dehydrogenase are separate proteins. By alignment of the amino acid sequences of these two proteins with that of the bifunctional protein from *E. coli*, it could be shown that the boundary between the two enzymatic domains in the bifunctional protein is a short segment between Phenylalanine 460 and Isoleucine 466.¹⁶⁹

From all of these results it can be concluded that, if

properly detached, each domain of this bifunctional protein from *E. coli* remains folded and enzymatically active. The cleavage by an endopeptidase within the exposed loop in domain 1, which is responsible for aspartate kinase, causes it to unfold and the unfolded amino-terminal fragment to fall away. Domain 2, which is responsible for homoserine dehydrogenase, remains folded and active after the cleavage. If the point of cleavage by glutamyl endopeptidase were a true boundary between two detachable domains, rather than an adventitious loop of polypeptide on the surface of the aspartate kinase, the aspartate kinase activity, which can readily be expressed by the genetically dissected protein, would have been unaffected. Aspartate kinase-homoserine dehydrogenase, then, is an example of a protein that has domains that cannot be detached from each other at their boundary by cleavage with an endopeptidase.

Proteins such as aspartate kinase-homoserine dehydrogenase belong to a class of proteins known as multienzyme complexes. A **multienzyme complex** is a protein that, although it is a single, discrete macromolecule, is able to catalyze two or more enzymatic activities. Usually each of the enzymatic activities in one of these multienzyme complexes is expressed by its own unique domain within the folded polypeptide or one of the folded polypeptides that form the protein. Such an **enzymatic domain** within a larger protein is a domain that is by itself independently responsible for a particular enzymatic activity. The fact that the several enzymatic activities are expressed respectively by several individual proteins in some species yet all of them are expressed by only one protein formed from one folded polypeptide in other species is sufficient evidence of an independent existence to conclude that the multienzyme complex is constructed from enzymatic domains. Proteins constructed from enzymatic domains presumably arose as the result of the fusion of the individual genes that encoded the unfused ancestors of those domains. Many artificial fusions of two genes to produce chimeric proteins have been performed, and the products that result from these artificial fusions seem to be little affected functionally.^{170,171}

A paradigm for a protein containing enzymatic domains is the CAD multienzyme complex in animal tissues that comprises a single folded polypeptide about 2220 aa in length¹⁷² responsible for the enzymatic activities of carbamoyl-phosphate synthase (glutamine hydrolysing), aspartate carbamoyltransferase, and dihydroorotase.¹⁷³ The first enzymatic reaction has two steps, the production of ammonia from the hydrolysis of glutamine at the active site of a glutaminase and the synthesis of carbamoyl phosphate from the resulting ammonia at the active site of a carbamoyl-phosphate synthase (ammonia). Each of the four component enzymatic reactions carried out by the intact complex from animals is carried out by a different discrete protein in *E. coli*. The amino acid sequences of the four **separate bacterial pro-**

teins can be aligned with four consecutive regions in the amino acid sequence of the multifunctional protein from *Mesocricetus auratus*; glutaminase with amino acids 2–355 (40% identity, 1.3 gap percent),^{172,174} carbamoyl-phosphate synthase (ammonia) with amino acids 397–1440 (40% identity, 1.1 gap percent),^{172,174} dihydroorotase¹⁷⁵ with amino acids 1457–1785 (20% identity, 3.5 gap percent),¹⁷⁶ and aspartate carbamoyltransferase with amino acids 1921–2225 (44% identity, 1.6 gap percent).¹⁷⁷ These regions are enzymatic domains 1 through 4 in the CAD multienzyme complex, respectively.

The fact that dihydroorotase has sustained so much more replacement suggests that even within the same polypeptide different domains can suffer **replacement at different rates**. In fact, rates of change can differ so much that one domain in a multienzyme complex can become defunct even as others retain their full function. In the CAD multienzyme complex from yeast, the dihydroorotase domain, although its amino acid sequence is still able to be aligned, has lost the ability to catalyze its enzymatic reaction.¹⁷⁵ A similar loss of function has occurred during the evolution of the fructose-2,6-bisphosphate 2-phosphatase domain of yeast 6-phosphofructo-2-kinase, but its enzymatic activity can be restored by mutating Serine 404 to histidine.¹⁷⁸

Because the amino acid sequences of the four discrete bacterial proteins responsible for the four enzymatic reactions catalyzed by the CAD multienzyme complex from animals can be aligned with the amino acid sequences of its four enzymatic domains, it follows that the tertiary structure of each domain in the animal protein must be superposable on the tertiary structure of the corresponding bacterial protein. Consequently, each domain in the animal protein must be a compact, independently folded unit, and these units must be **strung together** consecutively by the continuity of the polypeptide. This conclusion is supported by the fact that the enzymatically active domains responsible respectively for dihydroorotase^{175,179} and aspartate carbamoyltransferase¹⁷⁷ can be detached either genetically or by cleavage of the protein with endopeptidases at the boundaries of the domains. During the digestion with endopeptidases, however, the activity of carbamoyl-phosphate synthase (glutamine-hydrolysing) is lost and can be associated with none of the fragments smaller than 1700 aa in length. In situations such as this, digestion of one domain, for example, the carbamoyl-phosphate synthase domain, at some point on its surface could cause it to unfold and make the polypeptide much more susceptible to cleavage by endopeptidases in a region forming the boundary between that unfolded domain and a neighboring properly folded domain. An example of such a pruning of an unfolded segment of polypeptide from a properly and compactly folded protein by digestion with an endopeptidase occurred during the production of hybrids of different portions of micrococcal nuclease from *Staphylococcus*.¹⁸⁰

Anthranilate synthase, CTP synthase, phosphoribosylformylglycinamidine synthase, GMP synthase, imidazole glycerol phosphate synthase, glutamine-fructose-6-phosphate transaminase (isomerizing), and aminodeoxychorismate synthase, like the carbamoyl-phosphate synthase (glutamine hydrolysing) incorporated into the CAD multienzyme complex, all contain an enzymatic domain responsible for producing ammonia from glutamine by hydrolysis. The domain can be either

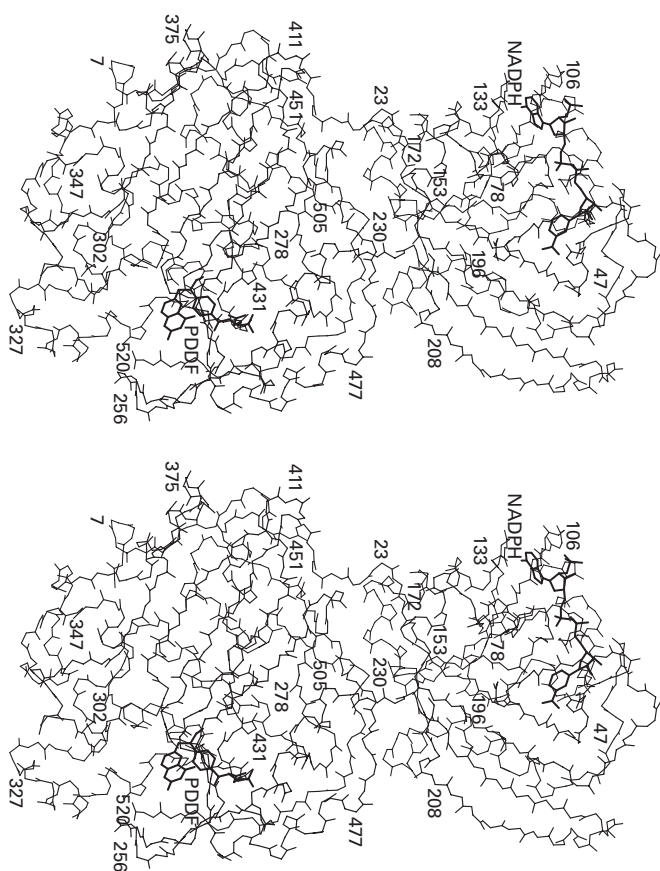


Figure 7-14: Crystallographic molecular model (Bragg spacing ≥ 0.29 nm) of the bifunctional enzyme from *Leishmania major* responsible for dihydrofolate reductase and thymidylate synthase.¹⁸³ The upper enzymatic domain (Serine 23 to Arginine 230) is responsible for dihydrofolate reductase, and its active site is identified by the bound NADPH (drawn with thicker lines). The lower enzymatic domain (Asparagine 231 to Valine 520) is responsible for thymidylate synthase, and its active site is identified by the bound 10-propargyl-5,8-dideazafolate (PDDF), also drawn with thicker lines. Aside from the amino-terminal 22 aa that adventitiously cleave to the outer surface of the opposite domain, that responsible for thymidylate synthase, the two enzymatic domains are independently folded structures separated by an irregular horizontal cleft through which only one strand of polypeptide passes, that containing Arginine 230. This drawing was produced with MolScript.⁴⁰⁸ The atomic coordinates on which this drawing is based were provided by Dave Matthews.

a separate folded polypeptide¹⁸¹ or an enzymatic domain in a longer polypeptide.¹⁸² Because the sequences of these various enzymatic domains catalyzing the hydrolysis of glutamine are homologous, it follows that they share a common ancestor that in its day was a separate, independent protein, presumably a glutaminase. The offspring of this common ancestor were **separately incorporated** into the various multienzyme complexes in which they are now found. Although each of these complexes catalyzes a quite different reaction, each uses the ammonia supplied by the respective glutaminase domain as a substrate.

The crystallographic molecular model of the bifunctional enzyme from *Leishmania major* responsible for dihydrofolate reductase and thymidylate synthase (Figure 7-14)¹⁸³ illustrates the independent existence of its two enzymatic domains. Domain 1, which is responsible for dihydrofolate reductase, comprises the folded polypeptide from Serine 23 to Arginine 230; and domain 2, which is responsible for thymidylate synthase, that from Histidine 234 to Valine 520. The respective active sites are identified by the NADPH and the 10-propargyl-5,8-dideazafolate. Each enzymatic domain is readily superposed on the crystallographic molecular model of the corresponding monofunctional enzyme from *E. coli*, and their respective amino acid sequences can be aligned with those of dihydrofolate reductase (26% identity, 3.8 gap percent) and thymidylate synthase (53% identity, 1.4 gap percent) from *E. coli*. In the bifunctional protein, the two enzymatic domains, although folded separately and once separate unassociated proteins, have nevertheless become **intimately associated** with each other at the interface between themselves.

In *Aspergillus nidulans* there is a multienzyme complex catalyzing 3-dehydroquinate synthase, 3-phosphoshikimate 1-carboxyvinyltransferase, shikimate kinase, 3-dehydroquinate dehydratase, and shikimate dehydrogenase. Although the individual enzymatic domains responsible for dehydroquinate dehydratase and 3-dehydroquinate synthase could be genetically detached as enzymatically active proteins, the enzymatic domain responsible for 3-phosphoshikimate 1-carboxyvinyltransferase was active only when attached to the neighboring domain responsible for 3-dehydroquinate synthase.¹⁸⁴ This result suggests that as time passes following their fusion, two domains may associate with increasing intimacy as a specific interface between them evolves, like that in dihydrofolate reductase-thymidylate synthase (Figure 7-14), and in the end, they may require each other's presence to fold properly as that interface becomes more and more extensive.

In most cases, the enzymatic domains gathered into a multienzyme complex all carry out reactions in the **same metabolic pathway**. The five enzymatic domains in the multienzyme complex from *A. nidulans*, the four enzymatic domains in the CAD multienzyme complex, and the two domains in dihydrofolate reductase-

thymidylate synthase catalyze successive reactions in the biosynthesis of chorismic acid, orotidine 5'-phosphate, and thymidine 5'-phosphate, respectively. Aspartate kinase-homoserine dehydrogenase catalyzes the first and third steps in the biosynthesis of homoserine. It is common in prokaryotes to find the enzymes catalyzing the reactions of a metabolic pathway gathered together in an operon. Such a gathering may have preceded the gene fusions producing multienzyme complexes and facilitated those fusions by placing the genes for the ancestors of the enzymatic domains adjacent to each other in the genome.¹⁸⁵ There are, however, examples of single enzymatic domains responsible for only one enzymatic reaction but inserted into larger proteins. The domain responsible for protein-tyrosine-phosphatase in eukaryotes is a compact enzymatic domain¹⁸⁶ that is fused into a large array of different proteins.

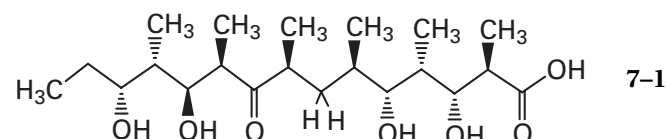
In prokaryotes and plants,^{187,188} the seven enzymes and the acyl carrier protein responsible for the synthesis of fatty acids from acetyl-S-CoA and malonyl-S-CoA are discrete proteins that can be separated and individually purified, but in fungi and animals all of these activities are expressed by a single multienzyme complex. In fungi, the complex is constructed from two folded polypeptides that are encoded by different genes and are completely different in sequence from each other.^{189,190} Their lengths are 1890 and 1980 aa, respectively.^{189,190} The fatty-acid synthase from animals, however, is constructed from only one polypeptide, 2440 aa in length.¹⁹¹ All seven of the enzymatic activities and the acyl carrier protein are located on the single polypeptide comprising the animal enzyme, and the domains responsible for each have been identified in the amino acid sequence.¹⁹¹⁻¹⁹³ Animal fatty-acid synthase has also been dissected both genetically and with the endopeptidase kallikrein¹⁹³ to produce several detached domains that are able to catalyze the enzymatic reactions assigned to them.

The order in which these enzymatic domains occur on the single polypeptide of the animal fatty-acid synthase is unrelated to the orders in which they appear on the two unique polypeptides from fungi.¹⁹⁴ On the basis of this fact, it has been concluded that all or most of the gene fusions that produced the animal protein and the fungal protein, respectively, must have occurred as independent events after the lineages of these two kingdoms diverged from their common ancestor. These separate processes would be ones in which each enzymatic domain has been shuffled into a larger protein. Nevertheless, the individual domains, even though **fused in different orders**, are still homologous to each other because the amino acid sequence of the one responsible for a given activity in the fungal enzyme can be aligned with that of the one responsible for the same activity in the animal enzyme.¹⁹⁴

There are a number of multienzyme complexes that are responsible for the biosynthesis of antibiotics in various fungi and bacteria. For example, the single folded

polypeptide 3131 aa in length¹⁹⁵ responsible for the biosynthesis of enniatin in *Fusarium oxysporin* catalyzes all of the enzymatic reactions required to produce this cyclic hexadepsipeptide,¹⁹⁶ and the multienzyme complex containing four folded polypeptides (3587, 3587, 1274, and 240 aa) responsible for the biosynthesis of surfactin in *B. subtilis* catalyzes all of the enzymatic reactions required to produce this cyclic octadepsipeptide.¹⁹⁷

The multienzyme complex¹⁹⁸ responsible for the biosynthesis of the polyketide 6-deoxyerythronolide B, which is the lactone at the hydroxyl on carbon 13 of the fatty acid



is composed of three distinct but related polypeptides, 3200-3600 aa long. The synthesis of the complete molecule of 6-deoxyerythronolide B proceeds by the successive condensation of six molecules of (2S)-methylmalonyl-S-CoA onto a molecule of propionyl-S-CoA. After each of the six Claisen condensations, the resulting ketone either remains untransformed (carbon 9) or is reduced to the alcohol (carbons 3, 5, 11, 13), which either remains untransformed or is dehydrated and reduced to the alkane (carbon 7). The entire sequence of reactions at each round, from condensation to alkane, requires the successive participation of an acyl carrier protein present as a domain as well as five separate enzymatic domains: an [acyl-carrier-protein] S-acyltransferase, a 3-oxoacyl-[acyl-carrier-protein] synthase, a 3-oxoacyl-[acyl-carrier-protein] reductase, a 3-hydroxyacyl-[acyl-carrier-protein] dehydratase, and an enoyl-[acyl-carrier-protein] reductase. On its three polypeptides erythronolide synthase has 22 enzymatic domains and 6 acyl carrier proteins.

The elongating substrate is passed along the **assembly line** from the first acyl carrier protein to the sixth acyl carrier protein through six successive stations. At each station, it is operated on by the enzymatic domains assembled at that station.¹⁹⁹ After it has been processed at the last station, the product 6-deoxyerythronolide is released from the last acyl carrier protein by an acyl-[acyl-carrier-protein] hydrolase, the last of the 22 enzymatic domains. When one of the domains at a particular station is inactivated or when one or more domains are experimentally added to a particular station, the product of the multienzyme complex changes at the position produced by that station to reflect its altered capacity. When the last station is deleted, a fatty acid lactone shorter by one C₃ unit is produced.²⁰⁰ The enzymatic reactions catalyzed by this large multienzyme complex are homologous to those catalyzed by fatty-acid synthase. Fatty-acid synthase, however, because each of its

successive steps passes through the complete sequence of enzymatic reactions to produce the alkane at each stage, uses only one station for all of the reactions rather than the six stations, one for each of its steps, used by erythronolide synthase.

Coenzymatic domains, such as the acyl carrier proteins carrying 4'-phosphopantetheine that are incorporated into fatty-acid synthase and erythronolide synthase, are domains to which coenzymes are covalently attached. The domains carrying acyl carrier proteins are domains because they are found as independent proteins in prokaryotes and plants; the domains carrying lipoic acid that are incorporated into 2-oxo-acid dehydrogenase complexes are domains because they can be detached.²⁰¹ Domains carrying biotin appear within the longer polypeptides of a number of different biotin-dependent carboxylases.²⁰²

A **functional domain** within a larger protein is a domain that is by itself responsible for a specific function. Enzymatic domains are functional domains, but there are also functional domains that are not enzymatic. Examples of functional domains would be the domains responsible for binding to specific segments of DNA (Figures 6-46, 6-50, and 6-53), each of which is a component of a larger protein responsible for controlling the expression of the gene adjacent to the segment of DNA recognized by the binding domain. Each of the other domains in the larger protein is responsible for a function essential to this control. For example, in each of the proteins that controls genes in response to steroid hormones such as estrogen, testosterone, progesterone, cortisone, and aldosterone, one of these other domains is responsible for binding the respective hormone.²⁰³ There are many examples of domains such as these that are responsible for binding a ligand and that are part of a larger protein. Examples are the two domains responsible for binding cyclic AMP in cyclic AMP-dependent protein kinase²⁰⁴ and the domain responsible for binding flavin mononucleotide and stabilizing its semiquinone radical in sulfite reductase.²⁰⁵ When the amino-terminal 240 aa of 3-phosphoshikimate 1-carboxyvinyltransferase from *E. coli*, which form a compact globular domain in the crystallographic molecular model of the intact protein (427 aa),²⁰⁶ was expressed separately, it folded to form a structure capable of binding shikimate 3-phosphate.²⁰⁷

The domains discussed so far are clearly capable of independent existence or are descended from ancestors that were. There are, however, domains in proteins that either cannot be detached or that are not identified with an independent function. These domains often were joined together so long ago that they have become completely dependent on each other both structurally and functionally. Nevertheless, it is possible to conclude that they once did have an independent existence because they recur in a number of extant proteins. Examples of such recurring domains would be domain 1, the antipar-

allel β barrel,²⁰⁸ and domain 2, the doubly wound parallel β sheet,²⁰⁹ in ferredoxin-NADP⁺ reductase (Figure 7-12), which are also found as domains in a number of different proteins. A **recurring domain** is a domain that is folded with a tertiary structure that can be superposed on the tertiary structures of other domains in other proteins of otherwise entirely different structure. A recurring domain is a compact structure used in more than one distinct situations. Because of its recurrence in different surroundings, there is little doubt that a domain of this type had at one time an independent existence.

Pyruvate kinase is one of the more informative examples of a protein built from recurring domains.²¹⁰ Domain 1 of the protein is an α -helically wound, parallel β barrel that is superposable on the entire folded polypeptide of triose-phosphate isomerase. Inserted into the loop between the third β strand and the third α helix of domain 1 is domain 2, which is a Greek key, antiparallel β barrel.²¹¹ Domain 3, which follows in the sequence of the polypeptide the complete elaboration of domain 1, can be superposed on half of the doubly wound, parallel β sheet found in L-lactate dehydrogenase. (Figure 7-11).²¹⁰ Each of these three structures is a recurring domain. In galactose oxidase from *Dactylium dendroides*, domain 2 is a β propeller (Figure 6-13) that is flanked by domain 1, an eight-stranded jelly roll, antiparallel β barrel,²¹¹ and domain 3, a bundle of seven antiparallel β strands that has the topological arrangement of an immunoglobulin domain.²¹² All three of these structures are also recurring domains.

Recurring domains occasionally appear to be associated with a particular role. For example, benzoate 4-monooxygenase, glucose oxidase, cholesterol oxidase, and glutathione-disulfide reductase contain a recurring domain about 160 aa in length.^{213,214} In all of these enzymes, the domain serves to bind tightly an integral flavin coenzyme, and it has been referred to as the "FAD-binding domain".²¹³

Some secondary structures, such as a β propeller (Figure 6-13)²¹² or a β helix, are self-contained and usually occur as **independent structural entities** in a protein. Because such secondary structures recur in many proteins, they could be considered recurring domains, and they usually do seem to be independent isolated units in a larger protein in which they occur.²¹⁵

An **internally repeating domain** is a member of a set of consecutive segments within the same polypeptide, each homologous in amino acid sequence to the other members of the set or each folded in a tertiary structure superposable on the tertiary structures of the other members of the set. An example of a set of such internally repeating domains are the 12 domains, four in each of the three detachable domains that compose immunoglobulin G (Figure 7-13). Each of these 12 internally repeating domains is a seven-stranded barrel of antiparallel β strands. Each is superposable on all the others and shares statistically significant similarities in

amino acid sequence with some of the others. The two identical long polypeptides in an immunoglobulin G, the heavy chains, each contain four of these domains; and the two identical short polypeptides, the light chains, each contain two. Polypeptides containing such internally repeating domains are quite common. About 10% of polypeptides 200 amino acids in length contain internally repeating domains but the frequency rises steadily to 80% for polypeptides 2000 amino acids in length.²¹⁶

Internally duplicated domains are internally repeating domains that occur only twice in the same polypeptide. Internally duplicated domains arise from the internal duplication of a gene encoding a smaller protein. An **internal duplication** is a gene duplication in which the duplicated genes end up immediately adjacent to each other so that when they are transcribed and translated, the duplicated amino acid sequences remain attached consecutively to each other in the same polypeptide. Because they are the products of internal duplication, the single, unrepeated ancestral amino acid sequence and tertiary structure of each of these duplicated domains must have existed on its own at some time in the past, before the duplication occurred. As with a gene duplication producing two separate proteins, an internal duplication arises in the genome of one individual and then spreads by genetic drift over the whole population. As with gene duplication in general, internal duplications arise often but only rarely spread over the whole population. Following the gene duplication and its spread and fixation by genetic drift, the two internally repeating domains begin to evolve separately.

The two halves of the doubly wound, parallel β sheet as it presently occurs in L-lactate dehydrogenase (Figure 7-11) can be superposed upon each other (Figure 7-15).⁸² It has been proposed that the complete doubly wound, parallel β sheet arose itself from a gene duplication in which the two segments of polypeptides encoded by the duplicated gene remained consecutively attached to each other and then began to evolve independently but within the same protein. That this did happen is supported by the fact that recurring domains are found in the crystallographic molecular models of pyruvate kinase²¹⁰ and phosphoglycerate kinase that superpose on only half of the doubly wound, parallel β sheet from L-lactate dehydrogenase.²¹⁰ The lineages leading to these two shorter domains presumably diverged before the gene duplication that produced the common ancestor of the larger.

The eight-stranded, α -helically wound parallel β barrel that forms the entire molecule of 1-(5-phosphoribosyl)-5-[(5-phosphoribosylamino)methylideneamino]imidazole-4-carboxamide isomerase from *Thermotoga maritima* has a fold typical of this structure, which is usually treated as a single domain (Problem 7-10B). Nevertheless, the amino-terminal half of its crystallographic molecular model can be superposed on the carboxy-terminal half with a root mean square deviation of

0.21 nm, and the amino acid sequences of the two halves can be aligned structurally with a percentage of identity of 23%.²¹⁷ A similar superposition and alignment can be performed with the two halves of the α -helically wound, parallel β barrel of imidazole glycerol phosphate synthase from the same bacterium.^{217,218} Unlike the two halves of the doubly wound, parallel β sheet (Figure 7-15), there are no examples of proteins in which half of an α -helically wound, parallel β barrel is found. Nevertheless, these superpositions and alignments suggest that all α -helically wound, parallel β barrels are also the product of an internal duplication.

The serine endopeptidases,²¹⁹ thiosulfate sulfurtransferase,²²⁰ carbamoyl-phosphate synthase from *E. coli*,²²¹ methionyl aminopeptidase from *E. coli*,²²² and diaminopimelate epimerase from *Haemophilus influenzae*²²³ are each an example of a protein **formed entirely**

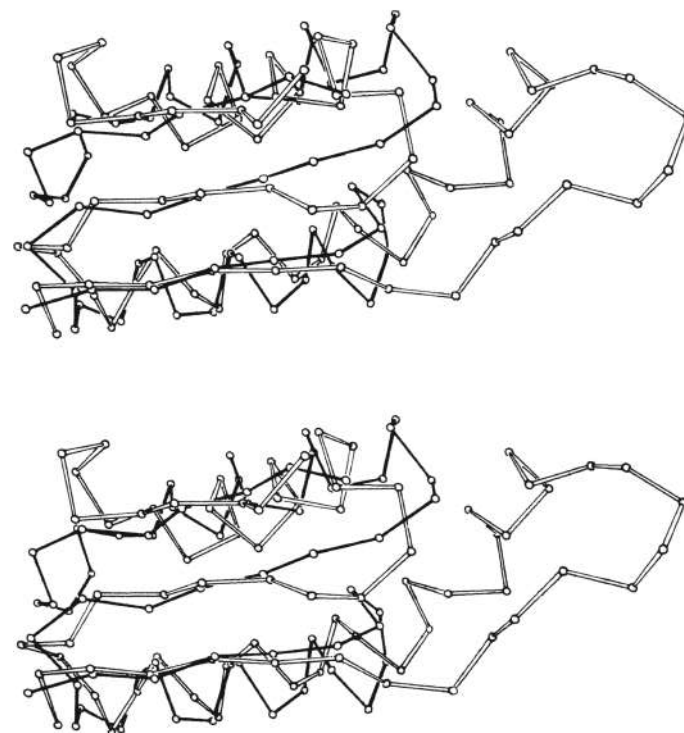


Figure 7-15: One half of the doubly wound, parallel β sheet (Figure 7-11) found in the crystallographic molecular model of L-lactate dehydrogenase from *S. acanthus* (open lines), drawn as an α -carbon diagram, superposed on the other half of the same doubly wound, parallel β sheet (solid lines).⁸² Reprinted with permission from ref 82. Copyright 1974 *Nature*.

by one internal duplication. In UDP-glucose 6-dehydrogenase from *Streptococcus pyogenes*, however, the two domains of an internal duplication (123 and 93 aa) are separated from each other by another domain of 177 aa.²²⁴ Although the bifunctional enzyme phosphoribosylanthranilate isomerase/indole-3-glycerol-phosphate synthase has two consecutive enzymatic domains that are superposable,²²⁵ the two enzymes probably evolved separately from a common ancestor before being fused.

Aside from the alignments of the duplicated amino acid sequences or superpositions of the duplicated tertiary structures, there are other observations supporting the conclusion that duplicated domains at one time had independent existence. The internal duplication in mammalian hexokinase produced two independent enzymatic domains, each containing an active site and each superposable on the other and on the unduplicated enzyme from fungi.^{87,226} The two internally duplicated domains of ovotransferrin²²⁷ can be detached by digestion with endopeptidase, and the resulting amino-terminal domain can be crystallized and shown to have the same structure it had in the intact protein.²²⁸ The aspartic endopeptidase from retroviruses is formed from two identical polypeptides,²²⁹ but those from eukaryotes are formed from a single polypeptide containing an internal duplication of the structure assumed by each polypeptide in the viral protein.^{125,230} Each domain of the enzyme from eukaryotes is superposable on one of the folded viral polypeptides. The two halves of porcine aspartic endopeptidase were expressed separately and, when mixed together, folded to produce an enzymatically active protein formed, as is the retroviral enzyme, from two folded polypeptides²³¹ rather than two internal duplications.

Many proteins contain **more than two** internally repeating domains. Serum albumin^{232,233} is composed of three internally repeating domains; human interstitial retinol-binding protein, of four;²³⁴ gelsolin, of six;²³⁵ human placental ribonuclease inhibitor,²³⁶ hemocyanin from *Octopus dofleini*,²³⁷ and granulin,²³⁸ of seven; and the polymeric globin from *Artemia*, of nine.²³⁹ Such **gene multiplications** are usually not produced by several historically distinct duplications but arise when an initial gene duplication then catalyzes the further multiplication of the gene during successive rounds of recombination. Sometimes the same protein from different species has different numbers of internally repeating domains. Dihydrolipoyllysine-residue acetyltransferase from *E. coli* has three consecutive lipoamide domains; the enzyme from rat, two; and the enzyme from *S. cerevisiae*, one.²⁴⁰ An example of a very recently multiplied gene is the one encoding preproincretin from *Xenopus laevis* in which the identical sequence of 46 aa repeats five times²⁴¹ in the same polypeptide with the replacement of only one amino acid in only one of the repeats. In contrast to such proteins containing amino acid sequences

multiplied so recently that they can be readily aligned²⁴² are proteins in which the internally repeating domains diverged so long ago that their secondary structures, although obviously related, have rearranged significantly.²⁴³

Nebulin and spectrin are examples of proteins with even larger numbers of internally repeating domains. Nebulin is a long protein (6669 aa) composed almost entirely (the first 6480 aa) of 178 internally repeating domains each 30–32 aa long.²⁴⁴ Spectrin is a protein composed of 38 internally repeating domains consecutively occurring in its two unique folded polypeptides.^{245–247} The folded domains sit like beads on a wire to create a long, somewhat flexible protein (Figure 7–16A).^{245,248} Each domain of 106 aa²⁴⁶ is an antiparallel coiled coil of three α helices (Figure 7–16B).^{249–251} Spectrin offers an excellent example of the absence of a correlation between locations at which cleavages with endopeptidases occur upon the surface of a native protein and the boundaries between its domains. Of the 15 cleavages of native spectrin produced by trypsin,²⁵² only four occur that are even near the boundaries of the domains.²⁴⁵ This fact is not surprising, given that the junctions between the domains are continuous α helices (Figure 7–16B).

The internal repeats in a protein, if the multiplications of the gene that produced them have occurred recently enough, can be recognized on a **dot matrix** (Figure 7–2) in which the amino acid sequence of the entire protein is compared to itself. Such a dot matrix of a protein with internally repeating domains contains a set of lines parallel to the central diagonal of identity. The distance between the lines is equal to the length of the internally repeating domains, and the number of lines is one less than the number of domains. In the dot matrix for the self-comparison of the amino acid sequence of human intestinal retinol-binding protein,²³⁴ there are three lines parallel to the central diagonal and the distances between them are 302–310 aa, and in that for human placental ribonuclease inhibitor,²³⁶ there are six lines and the distances between them are 57 aa.

The β propeller (Figure 6–13) is a structure in which four antiparallel β strands form each blade and 6–8 blades form the intact unit. Each blade is superposable on each of the others²⁵³ so an argument might be made that each blade represents an internally repeating domain,^{254,255} but there are no examples of such a small structure having independent existence. There are a number of other repeating structures that are **too small to fold** on their own^{256–259} and consequently should not be considered to be domains. Often short repeating sequences such as those in dragline silk from spiders²⁶⁰ or antifreeze proteins from *Tenebrio molitor*²⁶¹ have been multiplied to produce a protein that is fibrous or that must conform to a repeating molecular structure such as a crystal of ice but that is not formed from a string of independent globular tertiary structures as is spectrin.

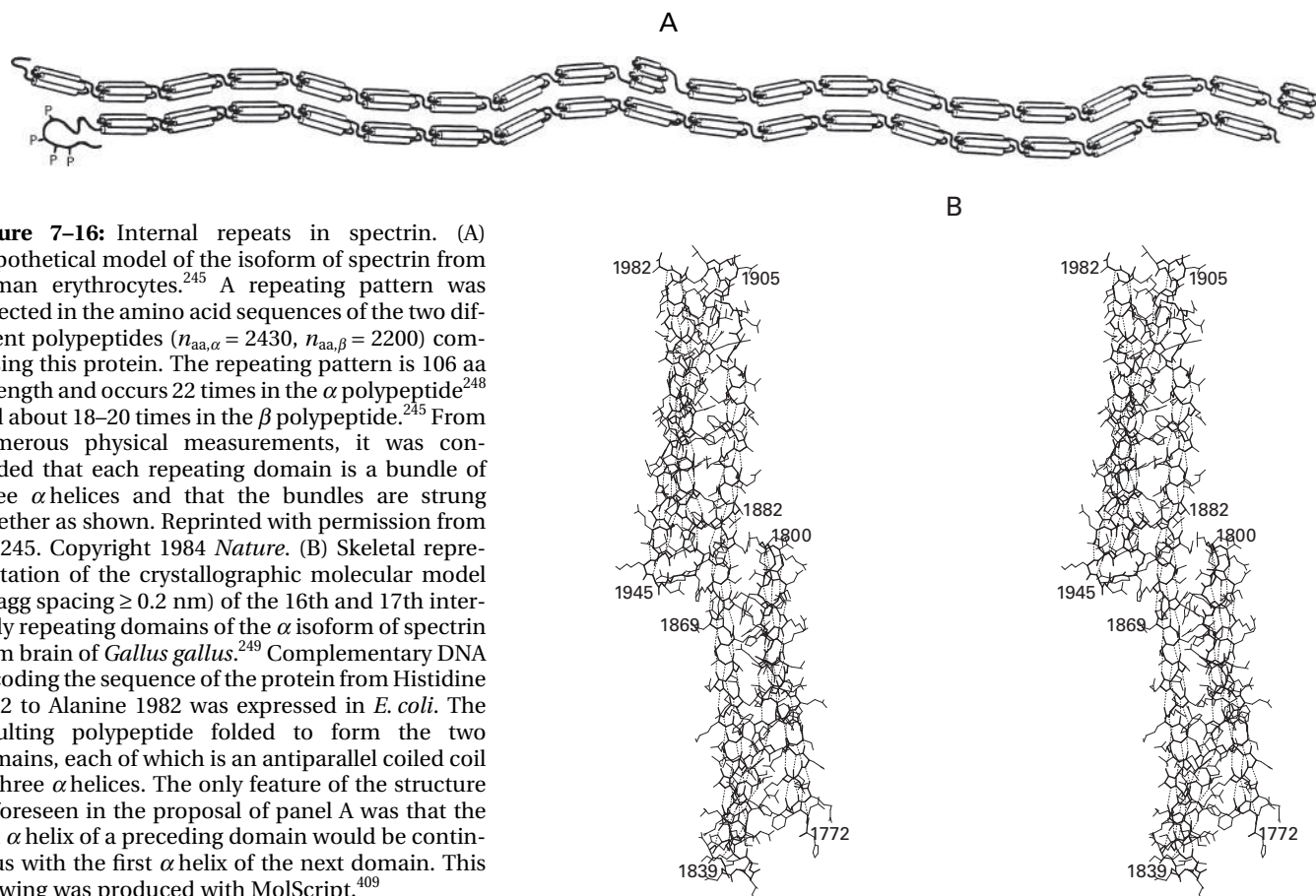


Figure 7-16: Internal repeats in spectrin. (A) Hypothetical model of the isoform of spectrin from human erythrocytes.²⁴⁵ A repeating pattern was detected in the amino acid sequences of the two different polypeptides ($n_{aa,\alpha} = 2430$, $n_{aa,\beta} = 2200$) composing this protein. The repeating pattern is 106 aa in length and occurs 22 times in the α polypeptide²⁴⁸ and about 18–20 times in the β polypeptide.²⁴⁵ From numerous physical measurements, it was concluded that each repeating domain is a bundle of three α helices and that the bundles are strung together as shown. Reprinted with permission from ref 245. Copyright 1984 *Nature*. (B) Skeletal representation of the crystallographic molecular model (Bragg spacing ≥ 0.2 nm) of the 16th and 17th internally repeating domains of the α isoform of spectrin from brain of *Gallus gallus*.²⁴⁹ Complementary DNA encoding the sequence of the protein from Histidine 1772 to Alanine 1982 was expressed in *E. coli*. The resulting polypeptide folded to form the two domains, each of which is an antiparallel coiled coil of three α helices. The only feature of the structure unforeseen in the proposal of panel A was that the last α helix of a preceding domain would be continuous with the first α helix of the next domain. This drawing was produced with MolScript.⁴⁰⁹

The proteins discussed so far are simple multiples of internally repeating domains, but other proteins have internally repeating domains attached to themselves and then to other domains. Sometimes one or the other of the odd domains appears to have evolved from the same common ancestor as the internally repeating domains but has undergone a few more additions, deletions, or rearrangements of its secondary structure,^{262,263} an observation suggesting that two gene duplications occurred at remarkably different times, but usually the odd domain or odd domains are entirely different. For example, pyruvate oxidase from *Lactobacillus plantarum* has two internally repeating domains, each 200 aa long, coupled to a recurring FAD-binding domain in the same polypeptide.²⁶⁴ The extracellular portion of mannose-6-phosphate receptor has 15 contiguous internally repeating domains coupled to a membrane-spanning segment and an intracellular domain.²⁶⁵

Titin is possibly the longest protein built from a single polypeptide. It is a protein greater than $1 \mu\text{m}$ in length that is found in vertebrate muscle.^{266,267} The amino terminus of its polypeptide is embedded in the Z disc, the carboxy terminus of its polypeptide is embedded in the M line, and its elasticity allows it to shorten and lengthen as the muscle contracts and relaxes.^{268,269} As does spectrin (Figure 7-16), titin achieves its remarkable length by using internal repeats. It contains 244

internally repeating domains, each about 100 aa long, that account for 90% of its amino acid sequence of 26,900 aa. Unlike spectrin, however, these internal repeats come in two types, immunoglobulin repeats²⁷⁰ and fibronectin type III repeats. Also, unlike the internally repeating domains of spectrin, these two types of internally repeating domains are widely recurring within a broad class of mosaic eukaryotic proteins that are cobbled together from a particular collection of promiscuous, modular domains. In the amino acid sequences of these proteins, segments have been observed that can be aligned with other segments within the same protein as well as segments in other proteins.²⁷¹ These recurring domains can appear many times in the same protein as internally repeating domains²⁷² or they can appear in combination with other different recurring domains. Because the proteins that contain them seem to have resulted from recent, remarkably active domain shuffling and because their amino acid sequences can usually be aligned readily with those of the other members of their type, these domains are usually considered to be members of a unique group and are referred to as modular domains.

A **modular domain** is a domain that recurs frequently within a group of **mosaic eukaryotic proteins** containing internal repeats of that domain, mixtures of other modular domains, or both internal repeats and

mixtures. The modular domains in such mosaic proteins are usually recognized by sequence alignment, a fact indicating that the proteins containing them have arisen from **recent genetic events**. The modular domains in these proteins can be readily assigned to one of the many types that have been observed. A few of these types are listed in Table 7–7.

A drawing of the crystallographic molecular model of a cohesin domain is presented in Figure 6–21. Although the mosaic proteins constructed from modular domains often contain only one or two types of domain with only a few internal repeats,^{306,317,318} some have four or more types, one or more of which can repeat a number of times (Table 7–8).

Several of the types of modular domains, for example, calcium-binding EF hand, leucine-rich repeats, EGF domains, and ankyrin repeats, are too short (Table 7–7) to fold on their own and are almost always repeated two or more times to produce a large enough structure to permit the polypeptide to fold. Together, in consecutive order, they form structures in which small compact units, each formed by one of the repeats and each of the same structure, are stuck together one against the next.^{319,320} It is the interfaces between the units that bury sufficient numbers of hydrogen-carbon bonds to make the structures stable. An example of this strategy is the β helix in the antifreeze protein from *T. molitor*, in which each turn is an internal repeat of only 12 aa.³²¹ Other short modular domains, clearly related by alignment of their sequences, assume different structures under different circumstances.^{322,323}

Examples of the group of proteins that are mosaics

Table 7–7: Examples of Types of Modular Domains Distributed among Mosaic Eukaryotic Proteins

type of domain	length ^a (aa)	structure ^b
EF hand ^{273,274}	40	$\alpha L\alpha$
immunoglobulin ^{90,275–278}	100	β_7
leucine-rich repeat ^{279,280}	30	$\beta\alpha$
RNA recognition motif ²⁸¹	80	$\beta\alpha\beta_2\alpha\beta$
EGF ^{282–286}	50	RM(Ccy _{3–4})
cohesin ^{287,288}	140	β_9
ankyrin ^{289,290}	40	$\beta_2\alpha_2$
C2 ^{274,291}	120	β_8
SH2 ^{292–295}	100	$\beta\alpha\beta_5\alpha\beta$
SH3 ^{296,297}	60	β_5
Kringle ^{298–300}	80	RM(Ccy ₃)
SAND ³⁰¹	80	$\beta_2\alpha\beta_2\alpha_2\beta\alpha$
pleckstrin ^{302,303}	100	$\beta_7\alpha$
fibronectin type I ^{304–307}	50	β_5
fibronectin type II ^{304,307}	60	$\beta_3\alpha\beta$
armadillo ^{308,309}	50	α_3
fibronectin type III ^{268,304,310–312}	90	β_7
START ³¹³	200	$\alpha\beta_3\alpha_2\beta_6\alpha$
hemopexin ^{314–316}	200	four-bladed β propeller

^aApproximate mean to nearest 10. ^bSecondary structures in the order in which they appear: α , α helix; β , β strand; L, loop; RM, random meander; Ccy, cystine.

Table 7–8: Extreme Examples of Proteins Assembled from Modular Domains^{306,317,318}

protein	order of modular domains ^a
factor XII	F2-EGF-F1-EGF-Kr-Endo ^b
thrombospondin I	Ths-CC-PcC-(Prop) ₃ -(EGF) ₃ -(EF) ₇ -Ths
collagen VI	(vWA) ₉ -Co-(vWA) ₂ -ST-F3-BPTI
aggrecan	(Li) ₂ -KS-Ch1-Ch2-EGF-Lec-Comp
perlecan	Per-(EGF) ₄ -Ig-(EGF) ₁₂ -Ig ₁₈ -[LaC-(EGF) ₂] ₂ -LaC
btK kinase	Plk-SH3-SH2-Kin
phospholipase C γ	Plk-PLC-PY-(SH2) ₂ -SH3-PLC
p120 GTPase activator	SH2-SH3-SH2-Plk-GAP

The order from amino terminus to carboxy terminus in which the modular domains are attached to each other. ^bModular domains (see Table 7–7): F2, fibronectin type II; F1, fibronectin type I; Kr, kringle; CC, coiled coil; PcC, carboxy-terminal procollagen; Prop, properdin; EF, calcium-binding EF hand; vWA, domain A of von Willebrand factor; ST, serine-threonine-enriched; F3, fibronectin type III; BPTI, bovine pancreatic trypsin inhibitor; Li, link protein; KS, keratan sulfate binding; Ch, chondroitin sulfate binding (types 1 and 2); Lec, lectin; Comp, complement control; Ig, immunoglobulin; LaC, carboxy-terminal laminin; Plk, pleckstrin; PY, phosphotyrosine-containing SH2-binding. Domains specific to the particular protein: Endo, endopeptidase; Ths, thrombospondin-specific; Co, (Gly-X-Y)_n repeat of collagen; Per, perlecan; Kin, kinase; PLC, phospholipase C; GAP, GTPase activator.

of modular domains are regulatory kinases and phosphatases, proteins of the extracellular matrix, and endopeptidases of the coagulation system. Within one of these proteins, the modular domains, which are widely recurring, are usually attached to one or more domains that are specific to the function of that protein (Table 7–8). Phosphoinositide phospholipase C δ 1 from *Rattus norvegicus* is a paradigm of such a mosaic protein. It contains, in order from the amino terminus, a pleckstrin domain, four consecutive EF hands, a catalytic domain responsible for the phospholipase activity, and a C2 modular domain (Figure 7–17).^{274,324,325}

Many of the types of modular domains, for example, the immunoglobulin, the EGF, the kringle, the fibronectin, and the hemopexin,^{271,306,316} are usually found on individual exons and are thought to have been distributed among their mosaics by exon shuffling. **Exon shuffling** is a genetic rearrangement in which an intact exon is coupled at one of its ends to the other end of another intact exon from elsewhere in the genome by a mistake in recombination that occurs at sites within the intron following the one exon and the intron preceding the other. The result of the shuffle is that the hybrid intron, containing the front end of one old intron and the back end of the other, now joins the two exons. This new intron is spliced away during the maturation of the messenger RNA so that in the protein the two amino acid sequences encoded by the exons are spliced together. Other types of modular domains, however, do not seem to be distributed by exon shuffling but by other types of genetic rearrangements.^{326–328}

In examining proteins formed from domains, **two periods of construction** can be discerned.²¹⁶ Modular domains and internally repeating domains, the amino

acid sequences of which can be readily aligned, are the products of recent genetic rearrangements. Proteins containing recurring domains and internally duplicated domains the amino acid sequences of which cannot be aligned and the tertiary structures of which have drifted apart significantly are the products of ancient genetic rearrangements by processes that assembled the proteins common to all existing organisms. Many of the recent rearrangements have been produced by exon shuffling, but whether or not any of the early rearrangements also were produced by exon shuffling is unknown.^{329–332}

When the tertiary structure of ferredoxin-NADP⁺ reductase is examined (Figure 7–12), it seems possible to divide it reasonably into two domains. It is now known that both of these are recurring domains, but it has been argued that, even if this were not known, a judicious decision that these were distinct domains could still have been made by inspection of the crystallographic molecular model alone. In this sense, these are two structural domains. A **structural domain** has been defined as a “section of peptide chain that can be enclosed in a compact volume ... by a closed surface ..., and is characterized by possession of two terminal points.”³³³ These two terminal points are the point at which the polypeptide enters the compact volume enclosed by the surface and the point at which it exits. For example, phosphoglycerate kinase is constructed from two structural domains, each formed from one continuous length of polypeptide possessing two terminal points and clearly capable of being enclosed by continuous surfaces surrounding compact volumes.¹⁴² This definition does not possess a requirement for evidence of the independent existence of the domain. It could be argued that evidence will eventually be gathered for the independent existence of each of the structural domains now designated.

The difficulty with the definition of structural domains is that it is **subjective**. Even though the closed surfaces chosen are the ones that seem to be reasonable, in most cases, other choices, which usually produce a greater number of smaller domains, could be made that would satisfy the same definition. In fact, when this basic definition was used to derive a set of objective rules to divide any given crystallographic molecular model into domains, the tertiary structures of the 22 proteins examined could be divided unambiguously into as few as two or as many as 10 structural domains. The number of domains so defined increased monotonically with the lengths of the respective polypeptides, which varied from 58 to 450 aa.³³⁴ The mean length of the polypeptide in these structural domains was about 50 aa, which seems too small to be an evolutionarily significant unit. Most of these irreducible units could be combined with one or more of their neighbors to produce larger structural domains. The relevance of these small segments either to the evolution of one of these proteins or to its structure is not apparent.

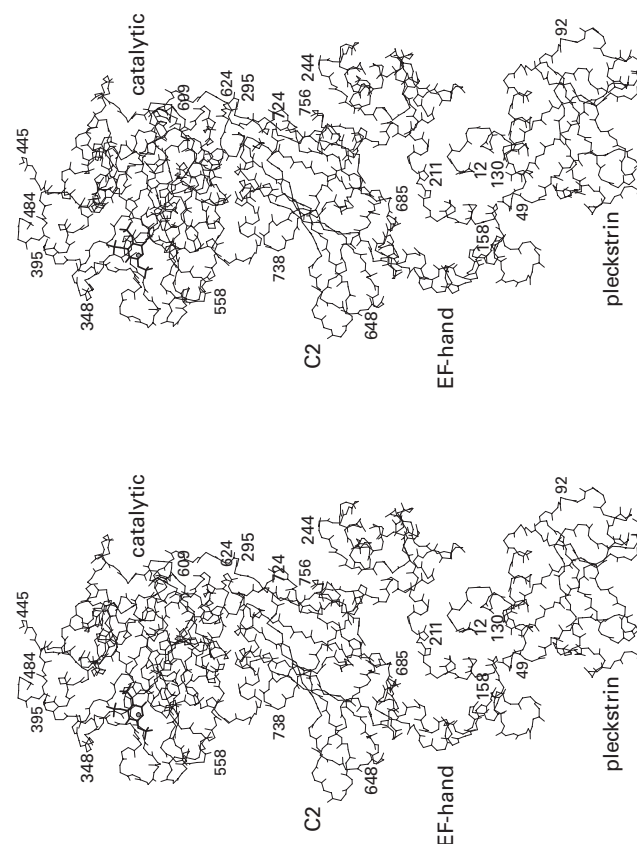


Figure 7–17: Crystallographic molecular model of phosphoinositide phospholipase C α 1 from *R. norvegicus*.^{274,324,325} The amino-terminal pleckstrin domain (Table 7–7) of the protein (Methionine 1 to Histidine 130) was expressed separately and crystallized. A skeletal representation of backbone of that crystallographic molecular model (Bragg spacing ≥ 0.19 nm) is at the bottom of the figure.³²⁵ The remainder of the protein (Glycine 133 to Aspartate 756) was also expressed separately and crystallized. A skeletal representation of that crystallographic molecular model (Bragg spacing ≥ 0.23 nm)³²⁴ is presented above that of the pleckstrin domain. There was no electron density for Glycine 133 through Aspartate 157 in the latter model. The representations of the two crystallographic molecular models are drawn at the same scale and arranged arbitrarily so that the carboxy-terminus of the pleckstrin domain is near the amino terminus of the rest of the molecule. The four EF hands (Table 7–7) comprise positions 133–175, 176–211, 212–245, and 246–282. None has a bound Ca²⁺ even though the cation was present during crystallization, so they do not assume the paradigmatic structure, but except for the first one, which is missing its first α helix, each has an α helix, a loop, and an α helix. The domain responsible for the phospholipase activity (catalytic) comprises Aspartate 299 through Alanine 606. It has a disordered segment (445–484) that produced no electron density, and the active site is occupied by a molecule of 1-D-*myo*-inositol-1,4,5-trisphosphate (thicker lines, upper left), a product of the enzymatic reaction, and a Ca²⁺ cation (gray circle). The C2 modular domain (Table 7–7) comprises Tryptophan 625 through Aspartate 756. This drawing was produced with MolScript.⁴⁰⁹

The intuitive impression persists, nevertheless, that the tertiary structures of most proteins, as revealed in their crystallographic molecular models, can be divided into two or more autonomous structural domains. It seems to be the case that anyone examining these structures would make the same decision, but there is no way to verify this surmise. Some examples of proteins that are thought to contain structural domains are DNA-directed DNA polymerase I,³³⁵ dihydrolipoyl dehydrogenase,³³⁶ and the hemagglutinin glycoprotein of influenza virus.³³⁷ An interesting example of a structural domain for which there is independent evidence of its existence occurs in catalase from *Penicillium vitale*. This protein has a structural domain formed from the carboxy-terminal 160 aa in its sequence (amino acids 510–670)³³⁸ that is missing entirely from bovine liver catalase,³³⁹ even though the two proteins are superposable throughout the other structural domain. Likewise, the carboxy-terminal structural domain found in the two structurally superposable proteins phosphoribosylamine–glycine ligase and biotin carboxylase is missing from the otherwise superposable proteins glutathione synthase and D-alanine–D-alanine ligase.⁹²

As the number of crystallographic molecular models has increased, more and more of the domains that were originally designated structural domains have been found to be recurring domains. It is possible that if the crystallographic molecular models of all of the proteins were known, all structural domains would turn out to be recurring domains, and this latter fact would provide the necessary evidence for their independent existence.

One indication that a structural domain does have independent existence is that its position relative to the rest of the protein shifts when crystallographic molecular models from different crystals of the same protein are compared. Lysozyme from bacteriophage T4 assumes five different structures in five different crystalline environments that differ from each other in the relative positions of its two structural domains.³⁴⁰ Such **independently shifting domains** that reorient within the same protein over milliseconds or seconds should be distinguished from domains that change their orientations over millennia as related proteins diverge from each other during evolution. As two proteins that are derived from a common ancestor diverge with time, structural domains often shift positions relative to each other even though the internal structures of the domains themselves remain superposable. Such **evolutionarily shifting domains** can be documented by superposing the crystallographic molecular models of related proteins. When the crystallographic molecular models of NADH peroxidase and glutathione reductase are compared, each of the four structural domains in these two related proteins superposes on its partner, but their relative positions in the two proteins are significantly shifted.³⁴¹ Significant shifts in the relative positions of the two inter-

nally repeating domains of aspartic endopeptidases have been documented by comparisons of crystallographic molecular models of six different members of the group,²³⁰ and the two structural domains in the related proteins ferredoxin–NADP⁺ reductase from *B. taurus* and thioredoxin–disulfide reductase from *E. coli*,³⁴² although separately superposable, differ in their relative positions by 66°.

One criterion that is often used as evidence for the independence of structural domains in a protein is that they unfold or fold independently. **Separately unfolding domains** are two or more regions of a protein that unfold independently of each other. Fibrinogen is a protein constructed from two copies of each of three polypeptides that are combined in such a way that the intact protein contains a central detachable domain, domain E, and two identical peripheral, detachable domains, domains D. The two domains D are attached to domain E by ropes constructed from three-stranded coiled coils of α helices,^{343–345} and the two domains D can be detached from domain E by cleaving disordered regions in the coiled coils with endopeptidases. When fibrinogen is submitted to differential scanning calorimetry,* two clearly separated transitions can be observed (Figure 7–18).³⁴⁶ These have been assigned to the melting of domains D and E, respectively.³⁴⁶

The melting, or unfolding, of the separately unfolding domains of fibrinogen is an irreversible process under the conditions chosen,³⁴⁶ but the unfolding and the refolding of a protein back to its native structure are often reversible processes, even in a calorimeter. Plasminogen is a protein composed of at least seven domains.³⁰⁰ These are five kringles that repeat consecutively within the entire sequence and two additional segments of polypeptide on each side of this pentuplication. Several of these domains or combinations of these domains can be detached and isolated separately. The reversible unfolding and refolding of five of these detached pieces could be followed by differential scanning calorimetry. These individual measurements could be combined to show that the rather complex, fully reversible calorimetric curve obtained with the intact protein was actually the sum of seven independent transitions.³⁴⁷ It is also possible to observe the independent

* A differential scanning calorimeter is used to measure the difference in the absorption of heat, as the temperature is raised, between a solution containing a protein and an identical solution lacking the protein. Two cells, sample and reference, contain precisely matched coils that introduce identical quantities of heat into each of them and establish a constant rate of temperature increase. The sample cell has an auxiliary coil that provides the additional heat necessary to keep its temperature exactly the same as that of the reference cell. The power supplied to the auxiliary heater is a measure of the excess heat absorbed by the sample, the endothermic heat flow. A protein unfolds, or melts, as the temperature rises, and this transition proceeds with the absorption of heat. This absorption of heat is a convenient way to follow the progress of the unfolding.

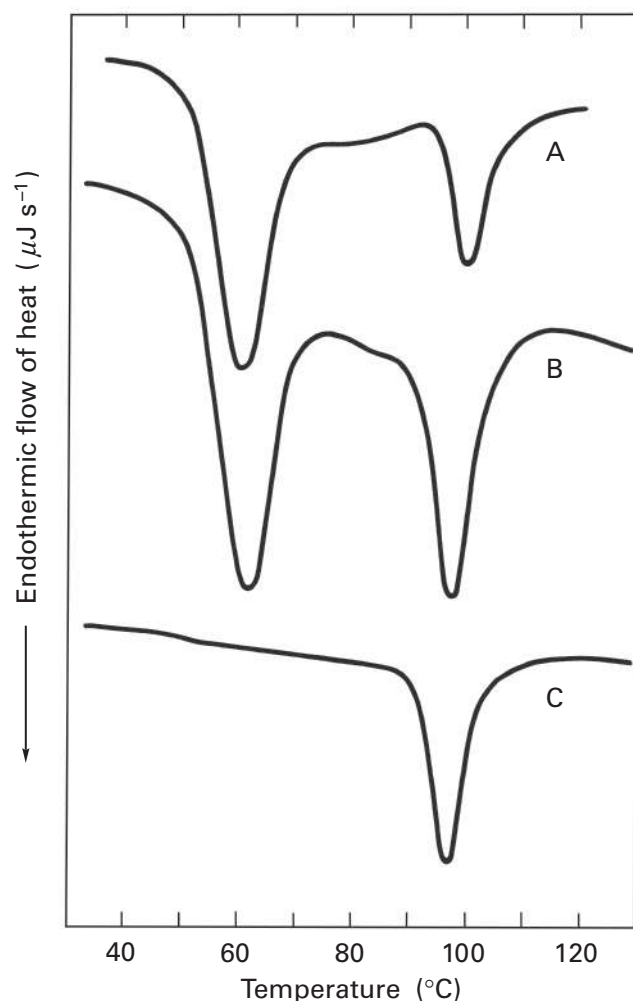


Figure 7-18: Thermal melting of domains D and E of bovine fibrinogen.³⁴⁶ (A) Native intact fibrinogen. A solution (26 μL) of native intact fibrinogen (88 mg mL^{-1}) was introduced into the sample chamber of the differential scanning calorimeter, and endothermic heat flow (microjoules second⁻¹) into the sample in excess of the flow into an identical solution lacking the protein was recorded as a function of temperature (degrees Celsius). (B) A solution (25 μL) of a 2:1 molar mixture of the chromatographically purified domains D and E that had been detached with an endopeptidase was used as the sample at a final concentration of 101 mg mL^{-1} . (C) A solution (22 μL) containing only detached and chromatographically purified domain E was used as the sample at a final concentration of 47 mg mL^{-1} . The scale of the two upper traces (microjoules second⁻¹) is 2.5 times that of the lower. Heating rate for all traces was 10 $^{\circ}\text{C min}^{-1}$. Adapted with permission from ref 346. Copyright 1974 National Academy of Sciences.

unfolding and refolding of separate domains in the same protein by perturbing the equilibrium between folded and unfolded forms of each domain through the addition of a denaturant such as urea³⁴⁸ or guanidinium chloride³⁴⁹ rather than with heat.

When there is no other independent evidence for the possibility that the structural domains in a particular protein at one time had an independent existence, attempts are often made to express one or more of these domains and demonstrate that they are able to fold inde-

pendently. An **independently folding domain** is a portion of a larger protein that is capable of folding by itself into a structure that is the same as the structure that portion assumes when it is within the larger protein. Experimentally, the portion of the protein thought to be an independently folding domain is detached either genetically or with an endopeptidase and shown to be able to fold properly on its own. If the fragment of the larger protein is an enzymatic domain, such as the bisphosphatase domain of 6-phosphofructo-2-kinase/fructose-2,6-bisphosphate 2-phosphatase,³⁵⁰ then its expression as an enzymatically active protein permits one to conclude that it has folded while being expressed to assume a structure that is the same as the structure it assumes in the intact protein.

The same conclusion can often be reached if the detached structural domain exhibits some **function** displayed by the intact protein. A genetically detached structural domain of aspartate transaminase folded independently to produce a globular structure capable of binding pyridoxal phosphate,³⁵¹ and a genetically detached internally repeating domain of human transferrin binds iron with high affinity.³⁵² When a structural domain located at the interface between the two identical subunits in glutathione reductase from *E. coli* was genetically detached, it folded and dimerized, presumably because the domain folded properly to create the face that participates in the dimerization of the native protein.¹⁶⁶ Unfortunately, the situation is often more ambiguous. For example, only when the two genetically detached structural domains accounting for the entire structure of CLC-0 chloride channel from *Torpedo californica* are coexpressed does the protein display function.³⁵³

In the absence of functional activity, a determination of its structure by nuclear magnetic resonance spectroscopy,^{354,355} or a crystallographic molecular model,³⁵⁶ how does one decide, even though it may display transitions characteristic of folding,^{348,357} whether or not the detached fragment is folding to assume the structure it had in the intact protein? A fragment of the polypeptide comprising thermolysin from *Bacillus thermoproteolyticus* can be produced by cleavage with cyanogen bromide and purified by molecular-exclusion chromatography in its unfolded state. This fragment contains the last 111 aa of the intact protein (sequence positions 206–316) and can be induced to refold. The solution that results contains a monomeric, compact, globular protein³⁵⁸ that has an α -helical content, determined spectroscopically, close to that predicted from the crystallographic molecular model of thermolysin.³⁵⁹ The protein in this solution melts, but at a temperature 20 $^{\circ}\text{C}$ below that at which native thermolysin, cut between amino acids 225 and 226, melts. The refolded protein unfolds in solutions of the denaturant guanidinium chloride, but at concentrations of guanidinium half those at which the nicked thermolysin unfolds. If the shorter polypeptide were folding

to assume the same structure it had in native thermolysin, that structure is now much less stable. But the results are also consistent with it assuming a completely different conformation and the protein displays no function that would indicate that it is properly folded.

A claim that an independently folding domain has been produced requires an unambiguous demonstration that the domain can refold into the **native conformation** it assumes in the parent protein. For example, both kringle 4 of plasminogen, detached by digestion with an endopeptidase,³⁶⁰ and kringle 2 of tissue plasminogen activator, detached genetically,³⁶¹ can be unfolded and refolded. The refolded detached domains bind lysine as they do when they are in their native conformation. The refolded kringle 2 also retains its affinity for plasminogen activator inhibitor 1. When kringle 4 is unfolded, its cystines reduced to cysteines, and then refolded, the fact that it has regained its native conformation is demonstrated by the ability of this refolded structure to enforce the formation of only the properly paired cystines upon its exposure to oxygen.³⁶⁰ In this case, the proper pairing of the cysteines, located at distant positions in the amino acid sequence, is the result of their proper juxtaposition in a properly folded polypeptide.

The **segments of polypeptide linking domains** together are of various types. Domains can be joined by flexible links such as those connecting the Fc and Fab domains of immunoglobulin G (Figure 7–13). The segments of polypeptide 35, 15, and 75 aa in length connecting the four enzymatic domains of the CAD multienzyme complex are rich in proline and glycine (30%) and the segments 27 and 24 aa in length connecting the three internally repeating lipoyl domains of dihydrolipoyllysine-residue acetyltransferase from *E. coli* are rich in proline and alanine (73%). All of these links should be **unstructured** and flexible. The amino acid segment –SKSSKEQKKQK– connecting the two functional domains of initiation factor IF3 from *E. coli* has been shown to be randomly disordered in the intact protein in solution,³⁶² and in the map of electron density for RNA recognition motifs from the Sex-lethal protein from *Drosophila melanogaster*, the segment connecting these modular domains is missing owing to its disorder.²⁸¹ It is such extended, disordered segments that are susceptible to endopeptidases when domains are detached by digestions. The long segment of 60 aa connecting the two modular SH2 domains in human protein-tyrosine kinase ZAP-70, however, forms a rigid, antiparallel, two-stranded coiled coil of α helices.²⁹⁵

Domains can also be joined by short **inflexible links**, such as the one in dihydrofolate reductase-thymidylate synthase (Figure 7–14), the interdomain α helix between two spectrin domains (Figure 7–16B), or the two connecting, internally repeating, modular EGF domains 3, 4, and 5 from murine laminin γ 1.²⁸⁶ In other instances, the segment connecting two domains may be structureless, but extensive **contacts between the**

domains cause them to be held tightly together (Figure 7–12). It is in proteins in which the domains were joined long ago that such contacts between domains are the most extensive. In the chaperone protein PapD from *E. coli*, however, it is the random meander of the link between the two domains that forms a hydrophobic core gluing together the two domains.⁹⁰

As a domain, by definition, is a structure that may be now or has been in the past an independent entity, the various categories (detachable, enzymatic, coenzymatic, functional, recurring, internally repeating, modular, structural, independently unfolding, and independently folding) are simply different ways of identifying members of a large group of **fundamental units of protein structure**. This group represents all of the smaller units from which the larger proteins that now exist were constructed, and the various domains that now exist in any one protein were at one time unique, unattached, stable, folded polypeptides that were the ancestors of those portions of the entire polypeptide now containing them. These primordial proteins were then internally multiplied or individually fused together during evolution by natural selection.

It is this role as a **fundamental unit of evolution** that lends luster to the title of domain and elicits the desire to grant it. But the term domain should remain an operational designation, closely tied to the particular evidence presented in each case. Problems can arise when it is applied indiscriminately. In particular, it often happens that when the term is used to describe a region of a protein for a very specific reason, all of the connotations associated with it have a way of attaching themselves to that region. For example, a structural domain subliminally gains the status of an independently folding domain, or an enzymatic domain is assumed to be also a detachable domain. Such confusion should be avoided.

Suggested Reading

- Ploegman, J.H., Drenth, G., Kalk, K.H., & Hol, W.G.J. (1978) Structure of bovine liver rhodanese I. Structure determination at 2.5-Å resolution and a comparison of the conformation and sequence of its two domains, *J. Mol. Biol.* 123, 557–594.
- Porter, R.R. (1959) The hydrolysis of rabbit γ globulin and antibodies with crystalline papain, *Biochem. J.* 73, 119–126.
- Miller, K.I., Cuff, M.E., Lang, W.F., Varga-Weisz, P., Field, K.G., & van Holde, K.E. (1998) Sequence of the *Octopus dofleini* hemocyanin subunit: structural and evolutionary implications, *J. Mol. Biol.* 278, 827–842.

Problem 7–8: There is a protein in vertebrate liver responsible for three enzymatic activities: phosphoribosylamine–glycine ligase, phosphoribosylglycinamide formyltransferase, and phosphoribosylformylglycinamide cyclo-ligase.³⁶³ It is composed of a single polypeptide 1010 aa in length. When the protein was digested with chymotrypsin, two products were produced that could be separated from each other. They were com-

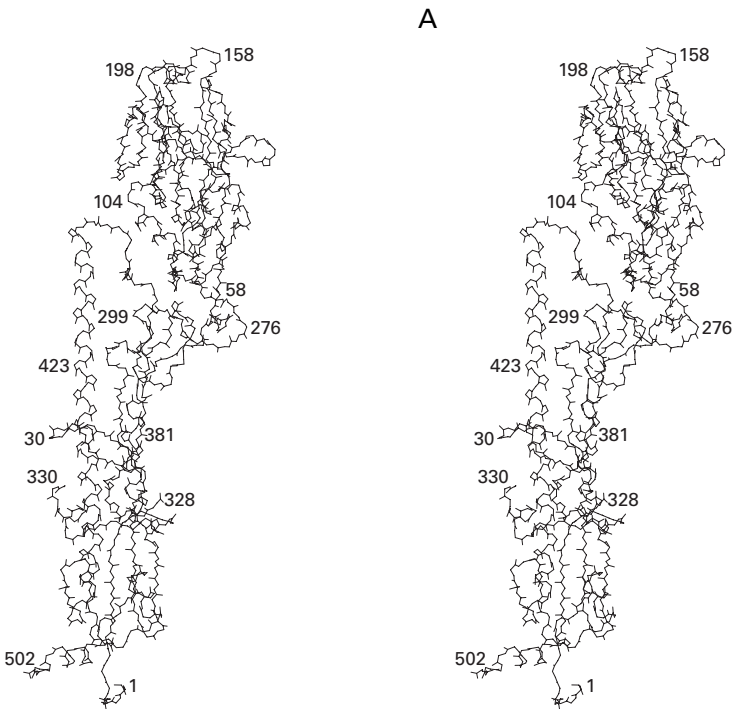
posed of polypeptides 450 and 550 aa in length. The larger retained the phosphoribosylamine–glycine ligase activity; the smaller, the phosphoribosylglycinamide formyltransferase activity. The phosphoribosylformylglycinamidine cyclo-ligase activity was lost. In *E. coli*, the phosphoribosylformylglycinamidine cyclo-ligase reaction is catalyzed by a monofunctional protein composed of a polypeptide 330 aa in length. Discuss and explain these observations in terms of detachable domains and enzymatic domains.

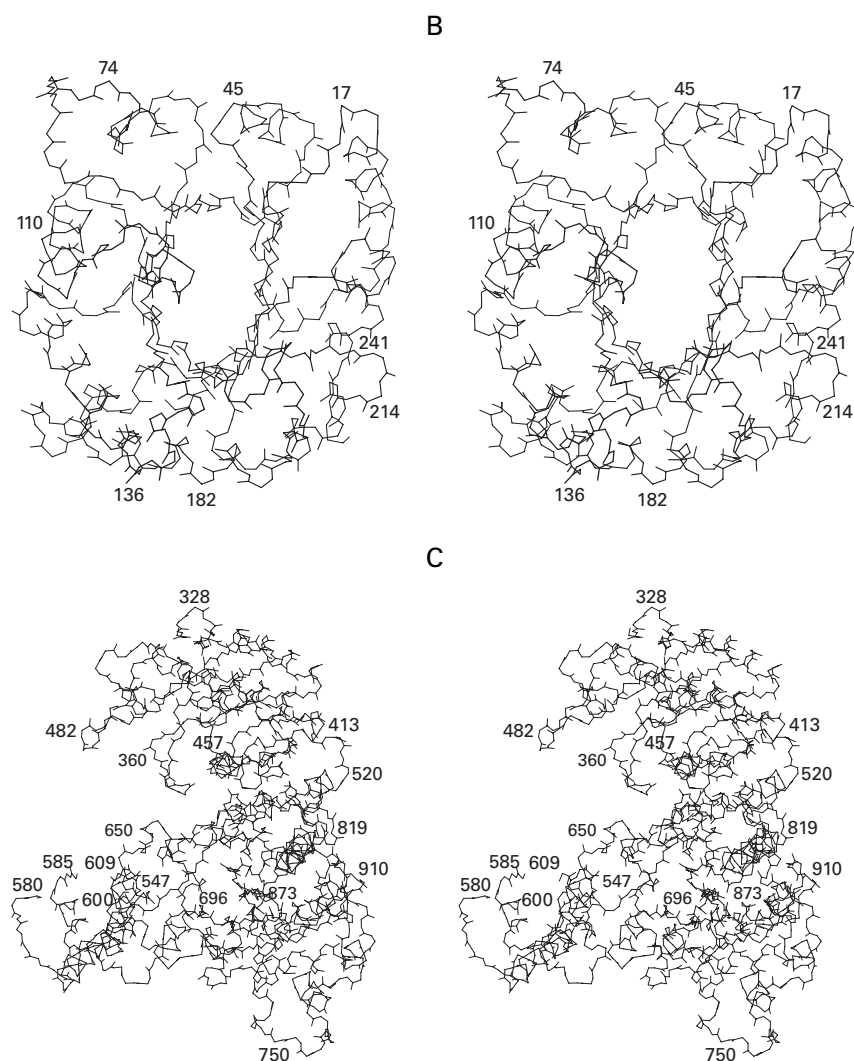
Problem 7–9: What conclusion concerning pantetheine-phosphate adenylyltransferase and dephospho-CoA kinase can be drawn from this table?³⁶⁴

Purification of Pantetheine-Phosphate Adenylyltransferase and Dephospho-CoA Kinase from Porcine Liver (600 g)

purification step	total protein		
	(mg)	transferase specific activity (μmol min ⁻¹ mg ⁻¹)	kinase specific activity (μmol min ⁻¹ mg ⁻¹)
1700 g supernatant	137,000		0.00020
protamine sulfate supernatant	54,000		0.00049
(NH ₄) ₂ SO ₄ fraction + Sephadex G-25	18,000		0.0013
DEAE-cellulose	3,200	0.014	0.0067
procion Red-Sepharose	79	0.54	0.26
blue Sepharose elution with CoA	4.3	7.4	3.6
Sephadex G-150	2.1	7.6	3.7

Problem 7–10: Which of these three tertiary structures have structural domains? How many are there in each?^{365–367} These drawings were produced with MolScript.⁴⁰⁹





Molecular Taxonomy

The proteins observed today have evolved from a much smaller group of less elaborate, primordial proteins, just as the species of organisms observed today have evolved from a much smaller group of less elaborate primordial species. The **primordial proteins** are now represented by the domains of presently existing proteins. Establishing the evolutionary relationships among these species of domains, however, may be far more difficult than establishing the evolutionary relationships among the species of organisms. Unfortunately, there is no fossil record of proteins. It is also quite clear that the evolutionary divergence that produced most of the proteins that are universally distributed among present living organisms, for example, the metabolic enzymes, occurred before the divergence of the organisms themselves. This follows from the observation that the amino acid sequences of the proteins from all living organisms responsible for one particular biological function are usually able to be aligned or their crystallographic molecular models to be superposed, but proteins from

the same organism responsible for two different functions are usually difficult if not impossible to relate to each other. Thus the lineages of these universally distributed proteins have remained almost unbranched since the evolution of the earliest organisms, and the radiation producing these lineages must have occurred before that time. It is also clear, however, from examining amino acid sequences and crystallographic molecular models that more specialized proteins have been arising continuously throughout evolution and are still arising today. These **newer proteins** are usually members of classes peculiar to a particular kingdom or phylum of organisms, and one of the challenges is to identify their ancestral relationships to the more universally distributed proteins.

It is hoped that, as the number of tertiary structures elucidated by crystallography grows, an anatomical collection of the proteins large enough to form the basis for a comprehensive taxonomy can be assembled.²¹¹ When Linnaeus developed his taxonomic system of the organisms, it is possible that he was unaware of the reason for its existence. It was only when taxonomy was connected

to the theory of evolution through natural selection that an exercise in cataloguing became something more profound. At the present time, taxonomy in biology is one of the methods by which evolutionary relationships are established. The desire to establish the evolutionary history of the **speciation of proteins** has led, in an interesting inversion of history, to the formulation of a **taxonomic system of the proteins**.

The fundamental unit in a taxonomic system for proteins is the domain. The history of most of the proteins that now exist is that of the random association of domains, much as wildly different species are assembled into parasitic or symbiotic relationships or into ecosystems. Consequently, attempting to formulate a taxonomic system for proteins, just as assembling a taxonomic system for ecosystems, would be inappropriate. It is the domains that are the equivalent of species of organisms. An **individual domain** is a domain of a particular amino acid sequence in a particular isoform of a particular protein found in a particular species of organism, for example, the doubly wound, parallel β sheet in isoform A of L-lactate dehydrogenase from *S. acanthius* (Figure 7-11). A **species of domains** is a population containing all of the individual domains found in the same relative location in the same protein in all of its isoforms in all of the species of organisms in which it is found. A protein is regarded as the same protein as another protein if both of them perform the same function in their respective organisms and their two respective amino acid sequences can be significantly aligned over their entire length or their complete tertiary structures can be superposed. The doubly wound, parallel β sheets in all of the isoforms of L-lactate dehydrogenases from all of the species of organisms constitute a species of domains, as do all of the globins from all of the species of organisms in which they are found.

As in populations of organisms, individual domains of the same species can differ significantly. To add to the confusion, the names of the individual proteins composed from domains of the same species can often be different, for example, cathepsin K from mammals and papain from plants³⁶⁸ or ferredoxin-NADP⁺ reductase from mammals and thioredoxin-disulfide reductase from bacteria,³⁴² but an examination of their respective functions and an alignment of their respective amino acid sequences or a superposition of their respective crystallographic molecular models establishes they are individuals of the same species. Often the structures of individuals of the same species of domains, such as the single domains constituting the lysozymes from animals and bacteriophage⁸⁸ or the carbonate dehydratases II from animals and bacteria,³⁶⁹ have drifted apart significantly; but their functions identify them. It is the hundreds of thousands of different species of domains that are **hierarchically classified** in the taxonomic system. The sequence of the hierarchy for the taxonomic system of domains is species, family, superfamily, common fold, architecture. This can be compared

to the sequence of the hierarchy of the taxonomic system of the organisms, which is species, genus, family, order, class, phylum, kingdom.

The central concept upon which the present taxonomic systems³⁷⁰⁻³⁷³ for classifying domains are based is the common fold. Although the exact definitions differ among the systems, two or more species of domains share a **common fold** if they have the "same major secondary structures in the same arrangement with the same topological connections." Different species of domains with the same common fold can have "peripheral elements of secondary structure and turn regions that differ in size and conformation."³⁷¹ It is within the cores of their structures that the common fold exists, and loops connecting the elements of the common core can differ significantly in their length and structure. For example, the motor domains of kinesin and myosin share a common fold of an eight-stranded β sheet sandwiched between two sets of three α helices, but the loops connecting these elements of secondary structure differ dramatically in length. The short loop of five amino acids between β strand 6 and β strand 7 and the short loop of 11 amino acids between α helix 4 and α helix 5 in kinesin are 221 and 142 aa long, respectively, in myosin.³⁷⁴ Although such insertions have little influence on the selection of the common fold of a domain, they can have a significant effect on the function of the protein, for example, turning a sulfotransferase into a dehydratase.³⁷⁵

It has been estimated that there are fewer than 1000 common folds in existence,³⁷⁰ the estimate varies depending on the stringency with which a particular taxonomic system divides the species of domains into common folds. The number of common folds and the position of the concept of the common fold in the various hierarchies are both quite close to the number and position of the class in the taxonomic hierarchy of living organisms. *Homo sapiens* belongs to the class Mammalia. The level of the common fold³⁷¹ is also referred to as the topology level,³⁷² the level of the structure type,²¹¹ or the level of structurally unique domains³⁷³ in the different taxonomic systems.

Particular species of domains can be selected as representatives of their common fold (Figure 7-19).²¹¹ L-Lactate dehydrogenase domain 1 (Figure 7-19F) represents the common fold of doubly wound, parallel β sheets containing six β strands in the order 321456 (Figure 7-20).²¹¹ Five representatives from the large set of domains of this common fold³⁷⁶ are listed within the box in Figure 7-20. Domains of the same common fold are often found in proteins with significantly different functions. The catalytic domain of aspartate-tRNA ligase and the domain constituting an entire molecule of asparagine synthase are of the same common fold,³⁷⁷ as well as the domains constituting the entire molecules of tumor necrosis factor and the coat protein of satellite tobacco necrosis virus.³⁷⁸

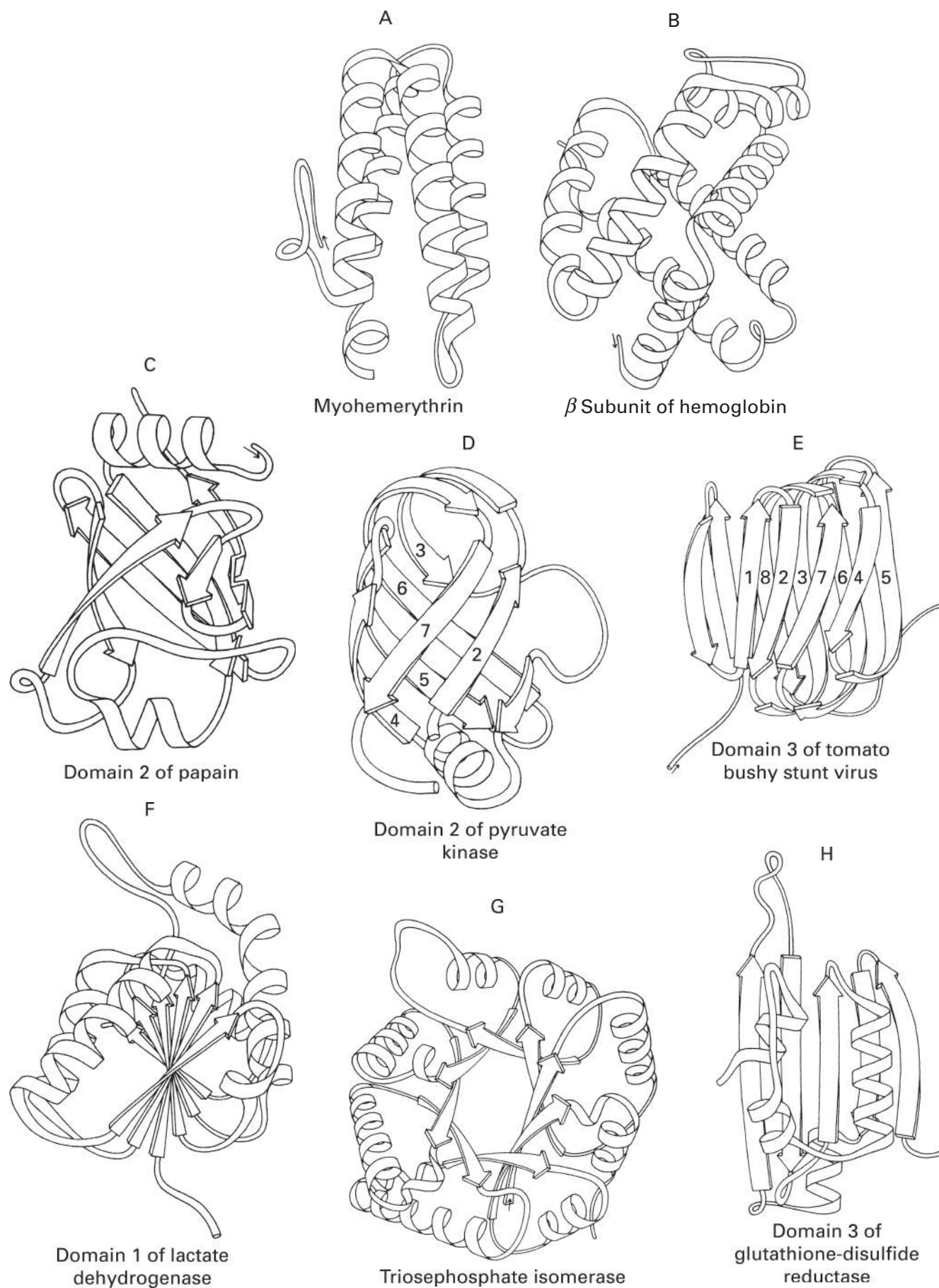
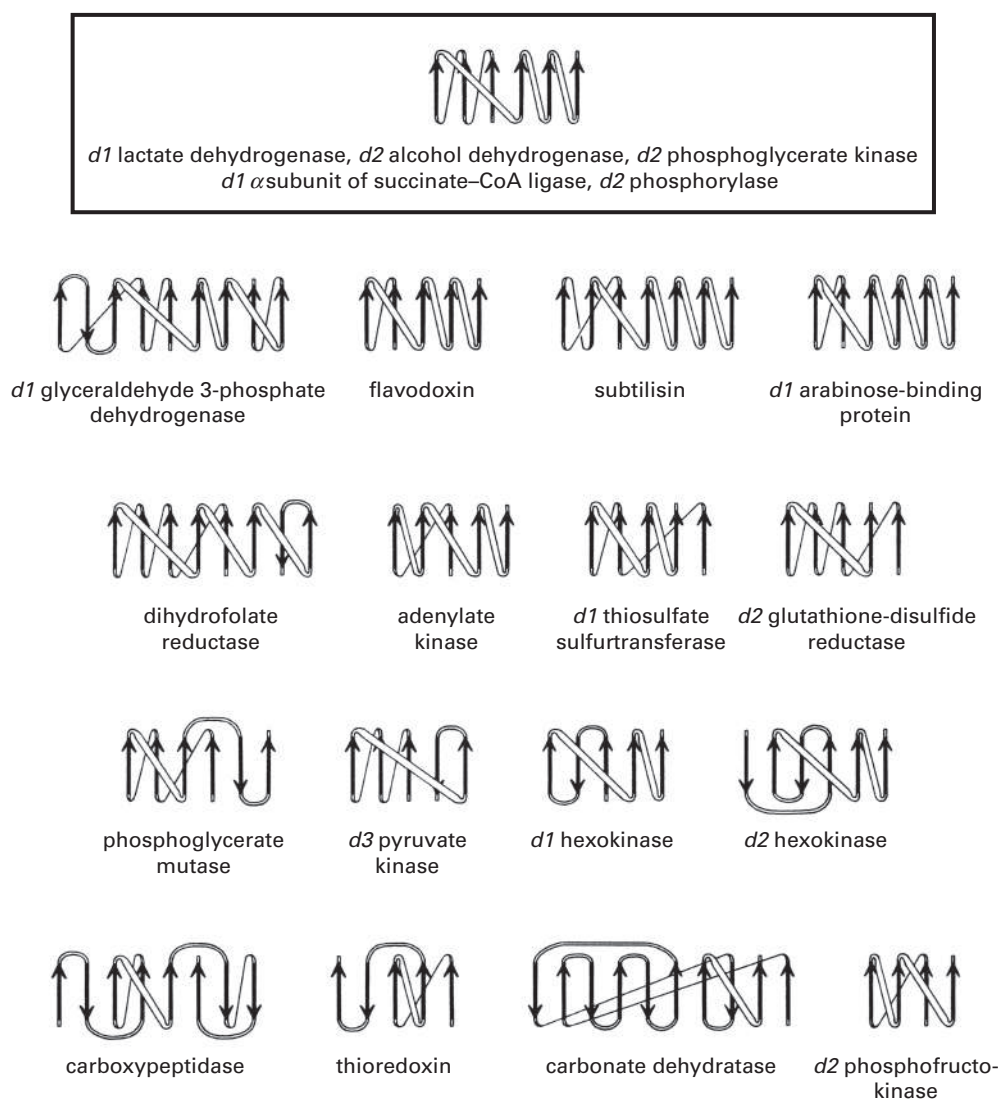


Figure 7-19: A menagerie of representative tertiary structures from eight common folds of domains.²¹¹ Each of these cartoons has been drawn from the respective crystallographic molecular model. The flat arrows represent strands of β structure; and the helical ribbons, α helices. (A) Myohemerythrin is an example of an up-down-up-down, antiparallel α -helical bundle. (B) The β subunit of hemoglobin is an example of a Greek key, antiparallel α -helical bundle. This can be seen if, from the amino terminus (indicated by the arrow), the first long, bent α helix, the second short α helix, and the next four long α helices are numbered 1 through 6, respectively, and it is assumed that α helix 1 has drifted 90° away from being parallel to α helix 6. (C) Domain 2 of papain is an example of an up-down-up-down β barrel if the last two short and bent β strands are ignored. (D) Domain 2 of pyruvate kinase is an example of a Greek key, antiparallel β barrel if the short strand of β structure, β strand 3, between β strands 2 and 4 is ignored. The six strands of the Greek key are numbered as in Figure 7-21. (E) Domain 3 from tomato bushy stunt virus is an example of a jelly roll, antiparallel β barrel if the first two, amino-terminal β strands are ignored. The eight strands of the jelly roll are numbered as in Figure 7-21. (F) Domain 1 of L-lactate dehydrogenase is an example of a doubly wound, parallel β sheet. (G) Triose-phosphate isomerase is an example of an α -helically wound, parallel β barrel. (H) Domain 3 of glutathione-disulfide reductase is an example of an open-faced β sandwich. Adapted with permission from ref 211. Copyright 1981 Academic Press.

Figure 7-20: Topological representations of several members of the architecture of doubly α -helically wound, parallel β sheets.²¹¹ The β sheets seen in each crystallographic molecular model were flattened upon a plane. The order in which the strands of the β sheet occurred in the amino acid sequence of the polypeptide was noted as well as whether the connecting loops, usually α helices, were above or below the plane. The dark arrows represent each β strand in the order in which they appear across the sheet. The open lines represent connections above the plane, and the thin lines represent connections below the plane. Within the box, the pattern of secondary structures displayed by domain 1 of L-lactate dehydrogenase (Figure 7-11), domain 2 of alcohol dehydrogenase, domain 2 of phosphoglycerate kinase, domain 1 of the α subunit of succinate-CoA ligase (ADP-forming), and domain 2 of phosphorylase (Figure 7-11A) represents the common fold to which these five domains belong. Reprinted with permission from ref 211. Copyright 1981 Academic Press.



The next higher level in the taxonomic hierarchy of domains is that of architecture.³⁷² A set of different common folds with the same **architecture** have the same clearly related spatial arrangement of secondary structures even though they differ in the number of individual elements of secondary structure or differ by one or more adjacent interchanges in the order in which those elements of secondary structure are juxtaposed or in both their number and their order. A collection of particular species of domains, each with a different common fold, can represent the architecture of doubly wound, parallel β sheets (Figure 7–20). This particular architecture,³⁷² however, has become so diverse^{379,380} that its systematic reorganization into several different architectures is probably required. A systematic census of the members of the architecture of open-faced β sheets (Figure 7–19H) has been taken.³⁸¹

Whether or not there is a higher level in the hierarchy than architecture is unresolved. For convenience, groupings have been used in which domains are sorted on the basis of whether they are formed entirely from α helices, formed from α helices alternating with β strands, formed from segregated α helices and β sheets, or formed entirely of β structure.³⁷² These groups may have no evolutionary significance. For example, α -helically wound, parallel β barrels (Figure 7–19G) would be in the alternating $\alpha\beta$ group, but glucan 1,4- α -glucosidase from *Aspergillus awamori* is an α -helically wound, parallel α barrel³⁸² that could be more closely related to α -helically wound, parallel β barrels than to any entirely α -helical domain. If so, this would demonstrate that β strands can become α helices, a possibility for which there is evidence on the much smaller scale of single elements of secondary structure.^{91,383} If such a transformation turns out to be common, higher groups based on topological arrangements rather than type of secondary structure might turn out to be more appropriate.

The level in the taxonomic hierarchy of domains above that of species is that of family. The central criterion on which the level of family is usually based in the classification of domains is that of the function and the structure of the protein containing the domain. A **family of domains** is a set containing all of the domains of the same common fold that are found at the same position in the complete set of those proteins that have coincident structures and that perform related functions. Two proteins have **coincident structures** when they are both composed of the same number of domains, and the domains found at the same respective positions in the two proteins have the same common fold. The crystallographic molecular models of proteins with coincident structures are superposable, domain by domain, over their entire length. Each of the consecutive folds can be, but is not necessarily, different. For example, the first and second domains in benzoylformate decarboxylase and the first and second domains in pyruvate decarboxylase all share the same common fold, but the third domains

from these two proteins with coincident structures share a different common fold.^{262,263}

Examples of domains in the same family illustrate the classification. The domains constituting the entire molecules of the hydrolases carboxymethylenebutenolidase, alkylhalidase, and carboxypeptidase D³⁸⁴ all belong to the same family. The single domains constituting the entire molecules of the mammalian endopeptidases factor D and trypsin are in the same family,³⁸⁵ as are those constituting the entire molecules of adenosine kinase and ribokinase.⁸⁹ Phosphoglycerate dehydrogenase, L-2-hydroxyisocaproate dehydrogenase, D-lactate dehydrogenase, erythronate-4-phosphate dehydrogenase, and glycerate dehydrogenase all have coincident structures and catalyze similar reactions.³⁸⁶ All of their domains 1 have one common fold and belong to one family, and all of their domains 2, which have a different common fold from that of the domains 1, belong to another family. The corresponding domains from cyclin-dependent protein kinase 2, MAP protein kinase ERK2, and cyclic-AMP dependent protein kinase are in the same respective families,^{387,388} as are the corresponding domains from aspartate-semialdehyde dehydrogenase and glyceraldehyde-3-phosphate dehydrogenase.¹²⁰

The **elaborations of the common fold** of the domains within the same family can be dramatic. For example, domains 1 of tyrosine phenol-lyase, cystathionine β -lyase, ornithine decarboxylase, aspartate transaminase, phosphoserine transaminase, and adenosylmethionine-8-amino-7-oxononate transaminase are members of the same family because they share the same common fold in which five β strands form the core, their pyridoxal phosphates are located in the same positions relative to the core, and the reactions they catalyze are of the same type. The elements of the common fold beyond the core, however, have drifted so far apart that they cannot be superposed, and there are a number of additional peripheral elements of secondary structure found in some species in the family but not in other species.³⁸⁹ This example illustrates the ambiguity of the upper limit for the level of family in the hierarchy.

Between the level of a family of domains, which is anchored in the coincident structures of the proteins containing the domains and their related functions, and the level of the common fold, which is anchored in the topological identity of their structures, is a region in the taxonomic hierarchy in which there are, at the moment, no consistent rules. This region is vaguely referred to as the level of the **superfamily**. Domains are grouped in a superfamily if the proteins that contain them have coincident structures but significantly different functions. For example, thiamine pyridinylase and maltose-binding protein have coincident structures but significantly different functions.⁹¹ Although all of the members of the enolase superfamily do share the function of abstracting a proton from a carbon α to a carboxylate, the superfamily has been divided into three families.^{118,390} Domains

are also grouped in a superfamily if the proteins that contain them have only partially coincident structures. The domains 2 and 3, respectively, from biotin carboxylase, phosphoribosylamine–glycine ligase, synapsin Ia, D-alanine–D-alanine ligase, and glutathione synthase have the same common folds, and all of the domains 1 have the same architecture,^{92,391} but the five proteins do not have coincident structures. Whether their domains 2 and 3 are of the same respective superfamilies or only of the same respective common folds is a question that illustrates the ambiguity of the upper limit of the level of superfamily in the hierarchy.

There are hundreds of common folds of domains, so a comprehensive discussion of even the architectures into which these common folds are arranged is not possible. There are some common folds and architectures, however, that have regular structural patterns that stand out. Both the **β helix** (Figure 6–12) and the **β propeller** (Figure 6–13) define architectures of domains. One architecture contains right-handed β helices; another, left-handed β helices.³⁹² Within the architecture of β propellers, the number of blades determines the common fold to which a particular member belongs.^{393,394}

A **parallel β barrel** (Figure 6–11) is usually wound completely by α helices (Problem 7–10B), as is the parallel β barrel constituting the entire folded polypeptide of triose-phosphate isomerase (Figure 7–19G). An α helix connects each stave of the barrel to the next. The number of staves in such a regularly α -helically wound, parallel β barrel determines the common fold to which it belongs. The common fold with the largest population is that in which the barrel has eight β strands. There are a large number of enzymes each of whose entire structure³⁹⁵ or the majority of each of whose structure³⁹⁰ is an α -helically wound, parallel β barrel of eight strands. There are often additional elements of secondary structure found in the loops connecting an α helix of the winding to a β strand in an α -helically wound, parallel β barrel,^{396,397} but these are found at the periphery of the structure and do not disrupt the common fold. α -Helically wound, parallel β barrels can stand alone or have other domains attached to them as in pyruvate kinase, phosphopyruvate hydratase,³⁹⁸ and the R1 protein of ribonucleoside-diphosphate reductase.³⁹⁷

In a number of the common folds of domains (Figure 7–19A–E), the structures observed seem to arise from a reasonable **topological operation** (Figure 7–21)²¹¹ that could explain their creation. Consider a polar curve that doubles back upon itself to form a hairpin. Twist the hairpin thus formed so that it folds into two turns of a right-handed superhelix (Figure 7–21A). Compress this superhelix until its neighboring segments both in front and behind come in contact and then incorporate the segments into the surface of a flattened cylinder (Figure 7–21B). This produces a flattened barrel with eight staves the polarities of which alternate as one proceeds around

the structure. This flattened cylinder can be rolled as the tread on a caterpillar tractor to produce eight different barrels that resemble each other but that vary in the juxtapositions of the staves across the center. The connections between the segments, which define the topological relations of the curve, remain unaltered during such rolling.

If the flattened cylinder in any of its guises is cut between the first and second segments in the hairpin (segments 1 and 2 in Figure 7–21) and spread upon a plane, a **jelly roll**²¹¹ (Figure 7–21C) is produced. Shorten the hairpin by removing the two most peripheral segments (cuts ① in Figure 7–21 to remove segments 1 and 8). A new flattened barrel is created (dotted lines in Figure 7–21B) with six staves that alternate in polarity. If this smaller flattened cylinder is cut between the first and last segments in the hairpin (segments 2 and 7 in Figure 7–21) and spread upon a plane, a **Greek key**²¹¹ (Figure 7–21D) is produced. Shorten the hairpin by removing the two most peripheral segments (cuts ② in Figure 7–21 to remove segments 2 and 7). A new flattened barrel is created with four staves that alternate in polarity. If this cylinder is cut between the first and last segments in the hairpin (segments 3 and 6 in Figure 7–21) and flattened upon a plane, an **up-down-up-down**²¹¹ pattern is produced.

The polar curve in the topological exercise can be substituted with a polypeptide either as a strand of β structure or in an α helix, and the staves of the flattened barrel will be either strands of β structure or α helices, respectively. If they are strands of β structure, they are gathered as systematically antiparallel (Figure 7–21B) pleated sheets (Figure 4–16C) into an **antiparallel β barrel** (Figures 7–12 and 7–13). The β jelly roll is represented by the cohesin domain (Figure 6–21)²⁸⁸ and domain 3 of the coat protein of tomato bushy stunt virus (Figure 7–19E). There are a large number of viral coat proteins each containing a domain of this class.³⁹⁹ There are elaborations on this topological arrangement; for example, each of the spermadhesins from seminal fluid is composed of a single domain that represents a class of domains in which two consecutive antiparallel β strands are added to a jelly roll (Figure 7–21C) at the amino-terminal end of the polypeptide (strands 0 and –1). These additional β strands are inserted into the barrel (Figure 7–21B) between staves 1 and 2.^{400,401}

The β Greek key is represented by domain 2 of pyruvate kinase (Figure 7–19D). There are also elaborations on this theme. The members of the class of immunoglobulin modular domains (Figure 7–13 and Table 7–7) are β Greek keys (Figure 7–21D) in which an antiparallel β strand is added at the carboxy-terminal end of the polypeptide (strand 8), but unlike the strand 8 in a jelly roll, which would be located between strands 2 and 3 of the barrel if strand 1 were deleted (Figure 7–21B), the strand 8 in the immunoglobulin class is found between strands 2 and 7 of the barrel.^{90,402–405} The class of

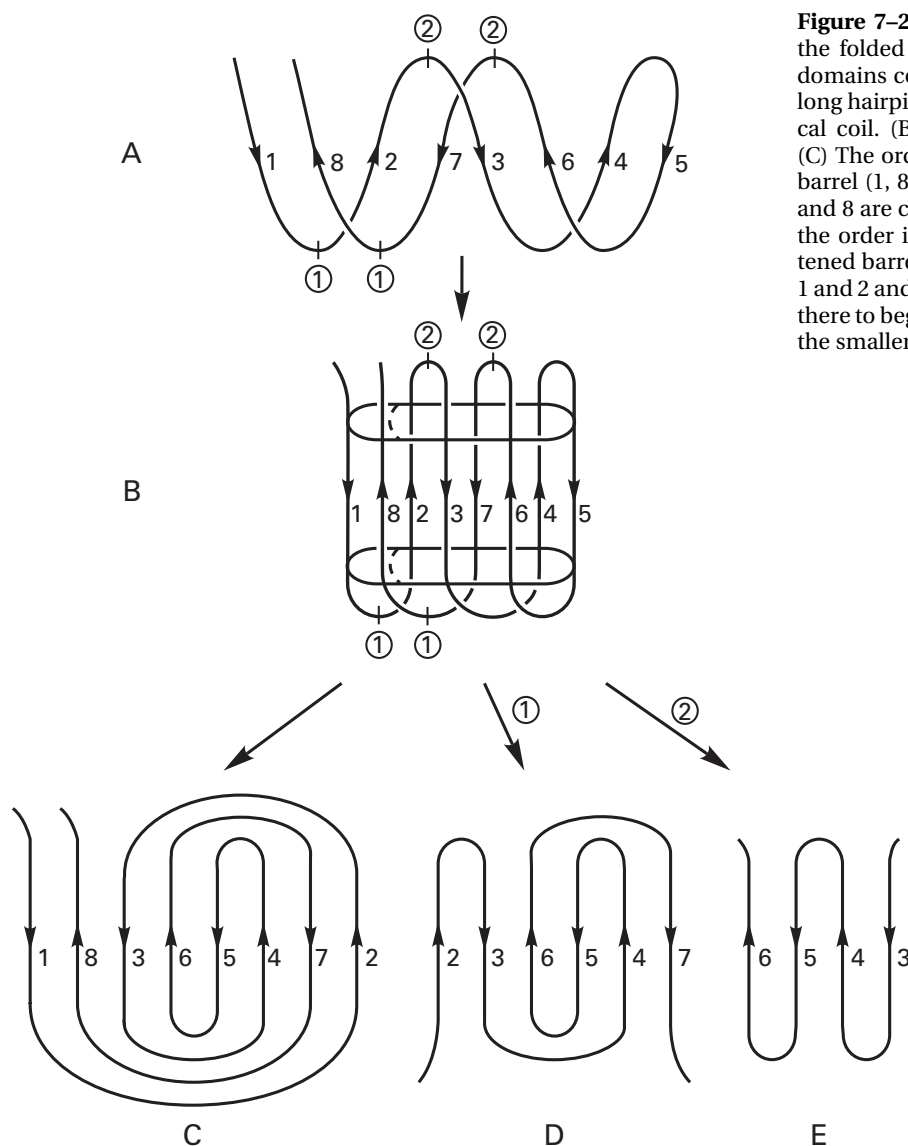


Figure 7-21: Topological explanation of how the patterns of the folded polypeptides defining six of the common folds of domains could have arisen by a common mechanism.²¹¹ (A) A long hairpin of β structure or α helix is twisted into a superhelical coil. (B) That superhelical coil is flattened into a barrel. (C) The order in which the strands occur around the flattened barrel (1, 8, 3, 6, 5, 4, 7, 2) is that of a jelly roll. (D) If strands 1 and 8 are cut away (cuts ①), or were never there to begin with, the order in which the strands occur around the smaller flattened barrel (2, 3, 6, 5, 4, 7) is that of a Greek key. (E) If strands 1 and 2 and strands 7 and 8 are cut away (cuts ②), or were never there to begin with, the order in which the strands occur around the smaller flattened barrel (6, 5, 4, 3) is up-down-up-down.

β up-down-up-down domains is represented by domain 2 of papain (Figure 7-19C).

Not all of the classes of domains in which the polypeptide forms an antiparallel β barrel have topological arrangements derived from a superhelical hairpin. For example, in the β barrel of licheninase, the structure is so far removed from a compressed superhelical hairpin that over all of the possible combinations of its 14 strands, there are only two turns of superhelical hairpin, one turn involving strands 9 through 12 and one turn involving strands 1 through 4.⁴⁰⁶

The antiparallel α Greek key is represented by the β subunit of hemoglobin in Figure 7-19B, but a much more regular representative of this class in which all six α -helical staves of the barrel are aligned more regularly is found in the caspase recruitment domain of the caspase activator Apaf1.⁴⁰⁷ The class of α up-down-up-down domains is cleanly represented by myohemerythrin (Figure 7-19A).

When these various representatives (Figure 7-19 A-E) are examined closely, it is obvious that if the topological scheme displayed in Figure 7-21 was the initial mechanism of folding, individual staves of the barrels have drifted significantly from their original positions (as do the helices in Figure 7-10), and more recent secondary structures have arisen at the ends of the barrels and in the loops connecting the staves.

There are two stereochemical properties producing alternative topological arrangements of the helical hairpin (Figure 7-21A) that generates the barrel (Figure 7-21B). Because the polypeptide is polar, amino terminus to carboxy terminus, there are two distinct **polarities to the hairpin**, the one shown by the arrowheads in Figure 7-21A and the one in which the polypeptide runs in the opposite direction. For example, in the α Greek keys of both the β subunit of hemoglobin and the caspase recruitment domain of the caspase activator Apaf-1, the polypeptide runs through the barrels with a

polarity opposite to that shown in Figure 7–21B. Because the barrel is generated by a helical conformation of the hairpin, there are two possible **twists to the helix**, the right-handed one shown in Figure 7–21A and the left-handed one. For example, the six-stranded β barrel forming the common fold in the family of domains containing growth hormones, interleukins, and granulocyte-colony-stimulating factor is a β Greek key in which the superhelix is left-handed.⁴⁰⁸ Because two polarities are possible and two twists are possible, there are four distinct geometries for the jelly roll, four for the Greek key, and two for the up-down-up-down conformation.

It is possible that the regular structures such as the β helix, the β propeller, the jelly roll, and the Greek key represent **evolutionarily efficient topological solutions** to the problem of folding a polypeptide. Within an architecture of domains or even among the members of the same common fold, a good deal of variation is observed; either individual elements of secondary structure in the common topological pattern do not superpose very well or extensive peripheral elements of secondary structure are found in one member of a class that are not found in others. If these variations represent the drift in the locations of the elements of secondary structure from those they had in the common ancestor or the insertion of elements following the divergence from a common ancestor, they reflect the degree to which these domains are evolutionarily related to each other, and the taxonomic system for domains parallels the taxonomic system for species. If many of the variations observed, however, state that the domains being compared differ so dramatically because they do not share a common ancestor, they reflect the fact that the structure is a particularly favorable solution to folding a polypeptide that has been exploited many times by convergent evolution and the taxonomic systems for domains overstate their phylogenetic information.

Suggested Reading

Richardson, J.S. (1981) Protein anatomy, *Adv. Protein Chem.* 34, 167–339.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., & Thornton, J.M. (1997) CATH—A hierarchic classification of protein domain structures, *Structure* 5, 1093–1108.

Problem 7–11: Construct a phylogenetic tree from Figure 7–20.

References

- Hardy, D.O., Bender, P.K., & Kretsinger, R.H. (1988) *J. Mol. Biol.* 199, 223–227.
- Dayhoff, M.O. (1972) *Atlas of Protein Sequence and Structure*, Vol. 5, National Biomedical Research Foundation, Silver Spring, MD.
- Dayhoff, M.O. (1978) *Atlas of Protein Sequence and Structure*, Vol. 5, Suppl. 3, National Biomedical Research Foundation, Silver Spring, MD.

- Gray, J.V., Golinelli-Pimpaneau, B., & Knowles, J.R. (1990) *Biochemistry* 29, 376–383.
- Doolittle, R.F. (1979) in *The Proteins* (Neurath, H., & Hill, R.L., Eds.) Vol. IV, pp 1–118, Academic Press, New York.
- King, J.L., & Jukes, T.H. (1969) *Science* 164, 788–798.
- Perutz, M.F., & Lehmann, H. (1968) *Nature* 219, 902–909.
- Reidhaar-Olson, J.F., & Sauer, R.T. (1988) *Science* 241, 53–57.
- Xu, Z., Bernlohr, D.A., & Banaszak, L.J. (1992) *Biochemistry* 31, 3484–3492.
- Drinkwater, C.C., Evans, B.A., & Richards, R.I. (1988) *J. Biol. Chem.* 263, 8565–8568.
- Pohjanjoki, P., Lahti, R., Goldman, A., & Cooperman, B.S. (1998) *Biochemistry* 37, 1754–1761.
- Zhang, M., Van Etten, R.L., & Stauffacher, C.V. (1994) *Biochemistry* 33, 11097–11105.
- Kanaya, S., Kohara, A., Miura, Y., Sekiguchi, A., Iwai, S., Inoue, H., Ohtsuka, E., & Ikehara, M. (1990) *J. Biol. Chem.* 265, 4615–4621.
- Hege, T., & Baumann, U. (2002) *J. Mol. Biol.* 314, 181–186.
- Lavrukhin, O.V., & Lloyd, R.S. (2000) *Biochemistry* 39, 15266–15271.
- Taton, M., Hüsselstein, T., Benveniste, P., & Rahier, A. (2000) *Biochemistry* 39, 701–711.
- Laskowski, M., Jr., Kato, I., Ardelt, W., Cook, J., Denton, A., Empie, M.W., Kohr, W.J., Park, S.J., Parks, K., Schatzley, B.L., et al. (1987) *Biochemistry* 26, 202–221.
- Margoliash, E., & Schejter, A. (1966) *Adv. Protein Chem.* 21, 113–286.
- Forry-Schaudies, S., Maihle, N.J., & Hughes, S.H. (1990) *J. Mol. Biol.* 211, 321–330.
- Hendriks, W., Sanders, J., de Leij, L., Ramaekers, F., Bloemendal, H., & de Jong, W.W. (1988) *Eur. J. Biochem.* 174, 133–137.
- Noguchi, T., Yamada, K., Inoue, H., Matsuda, T., & Tanaka, T. (1987) *J. Biol. Chem.* 262, 14366–14371.
- Weiss, C., Zeng, Y., Huang, J., Sobocka, M.B., & Rushbrook, J.I. (2000) *Biochemistry* 39, 1807–1816.
- Haase, G.H., Brune, M., Reinstein, J., Pai, E.F., Pingoud, A., & Wittinghofer, A. (1989) *J. Mol. Biol.* 207, 151–162.
- MacKintosh, R.W., Haycox, G., Hardie, D.G., & Cohen, P.T. (1990) *FEBS Lett.* 276, 156–160.
- Doolittle, R.F. (1981) *Science* 214, 149–159.
- Needleman, S.B., & Wunsch, C.D. (1970) *J. Mol. Biol.* 48, 443–453.
- Gibbs, A.J., & McIntyre, G.A. (1970) *Eur. J. Biochem.* 16, 1–11.
- Vogt, G., Etzold, T., & Argos, P. (1995) *J. Mol. Biol.* 249, 816–831.
- Feng, D.F., Johnson, M.S., & Doolittle, R.F. (1984) *J. Mol. Evol.* 21, 112–125.
- McLachlan, A.D. (1972) *J. Mol. Biol.* 64, 417–437.
- Staden, R. (1982) *Nucleic Acids Res.* 10, 2951–2961.
- Brenner, S.E., Chothia, C., & Hubbard, T.J. (1998) *Proc. Natl. Acad. Sci. U.S.A.* 95, 6073–6078.
- Jue, R.A., Woodbury, N.W., & Doolittle, R.F. (1980) *J. Mol. Evol.* 15, 129–148.
- Johnson, M.S., & Doolittle, R.F. (1986) *J. Mol. Evol.* 23, 267–278.

35. Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., & Chothia, C. (1998) *J. Mol. Biol.* 284, 1201–1210.
36. Pearson, W.R. (1998) *J. Mol. Biol.* 276, 71–84.
37. Altschul, S.F., & Gish, W. (1996) *Methods Enzymol.* 266, 460–480.
38. Thayer, M.M., Flaherty, K.M., & McKay, D.B. (1991) *J. Biol. Chem.* 266, 2864–2871.
39. Hyde, S.C., Emsley, P., Hartshorn, M.J., Mimmack, M.M., Gileadi, U., Pearce, S.R., Gallagher, M.P., Gill, D.R., Hubbard, R.E., & Higgins, C.F. (1990) *Nature* 346, 362–365.
40. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., & Lipman, D.J. (1990) *J. Mol. Biol.* 215, 403–410.
41. Pearson, W.R. (1990) *Methods Enzymol.* 183, 63–98.
42. Smith, T.F., & Waterman, M.S. (1981) *J. Mol. Biol.* 147, 195–197.
43. Philpott, C.C., Klausner, R.D., & Rouault, T.A. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 7321–7325.
44. Read, J., Pearce, J., Li, X., Muirhead, H., Chirgwin, J., & Davies, C. (2001) *J. Mol. Biol.* 309, 447–463.
45. Takahashi, S., Kuzuyama, T., Watanabe, H., & Seto, H. (1998) *Proc. Natl. Acad. Sci. U.S.A.* 95, 9879–9884.
46. Ziegler, G.A., & Schulz, G.E. (2000) *Biochemistry* 39, 10986–10995.
47. Klenk, H.P., Clayton, R.A., Tomb, J.F., White, O., Nelson, K.E., Ketchum, K.A., Dodson, R.J., Gwinn, M., Hickey, E.K., Peterson, J.D., Richardson, D.L., Kerlavage, A.R., Graham, D.E., Kyrpides, N.C., Fleischmann, R.D., Quackenbush, J., Lee, N.H., Sutton, G.G., Gill, S., Kirkness, E.F., Dougherty, B.A., McKenney, K., Adams, M.D., Loftus, B., Venter, J.C., et al. (1997) *Nature* 390, 364–370.
48. Moore, G.W., Goodman, M., Callahan, C., Holmquist, R., & Moise, H. (1976) *J. Mol. Biol.* 105, 15–37.
49. Feng, D.F., Cho, G., & Doolittle, R.F. (1997) *Proc. Natl. Acad. Sci. U.S.A.* 94, 13028–13033.
50. Lenstra, J.A., & Beintema, J.J. (1979) *Eur. J. Biochem.* 98, 399–408.
51. Bishop, J.G., Dean, A.M., & Mitchell-Olds, T. (2000) *Proc. Natl. Acad. Sci. U.S.A.* 97, 5322–5327.
52. Ambler, R.P., & Wynn, M. (1973) *Biochem. J.* 131, 485–498.
53. Grishin, N.V. (1995) *J. Mol. Evol.* 41, 675–679.
54. Feng, D.F., & Doolittle, R.F. (1997) *J. Mol. Evol.* 44, 361–370.
55. Fitch, W.M., & Margoliash, E. (1967) *Science* 155, 279–284.
56. Ayala, F.J., Rzhetsky, A., & Ayala, F.J. (1998) *Proc. Natl. Acad. Sci. U.S.A.* 95, 606–611.
57. Feng, D.F., & Doolittle, R.F. (1987) *J. Mol. Evol.* 25, 351–360.
58. Venkatesh, B., Erdmann, M.V., & Brenner, S. (2001) *Proc. Natl. Acad. Sci. U.S.A.* 98, 11382–11387.
59. Baldauf, S.L., Roger, A.J., Wenk-Siefert, I., & Doolittle, W.F. (2000) *Science* 290, 972–977.
60. Loytynoja, A., & Milinkovitch, M.C. (2001) *Proc. Natl. Acad. Sci. U.S.A.* 98, 10202–10207.
61. Mross, G.A., & Doolittle, R.F. (1967) *Arch. Biochem. Biophys.* 122, 674–684.
62. Jermann, T.M., Opitz, J.G., Stackhouse, J., & Benner, S.A. (1995) *Nature* 374, 57–59.
63. Bradac, J.A., Gruber, C.E., Forry-Schaudies, S., & Hughes, S.H. (1989) *Mol. Cell Biol.* 9, 185–192.
64. Monteiro, M.J., & Cleveland, D.W. (1988) *J. Mol. Biol.* 199, 439–446.
65. Meyer, U., Benghezal, M., Imhof, I., & Conzelmann, A. (2000) *Biochemistry* 39, 3461–3471.
66. Crawford, D.L., Constantino, H.R., & Powers, D.A. (1989) *Mol. Biol. Evol.* 6, 369–383.
67. Li, S.S., Fitch, W.M., Pan, Y.C., & Sharief, F.S. (1983) *J. Biol. Chem.* 258, 7029–7032.
68. Xu, X., & Doolittle, R.F. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87, 2097–2101.
69. Wu, G., Fiser, A., ter Kuile, B., Sali, A., & Muller, M. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 6285–6290.
70. Weber, K., Plessmann, U., & Ulrich, W. (1989) *EMBO J.* 8, 3221–3227.
71. Zimniak, L., Dittrich, P., Gogarten, J.P., Kibak, H., & Taiz, L. (1988) *J. Biol. Chem.* 263, 9102–9112.
72. Griffin, T.A., Lau, K.S., & Chuang, D.T. (1988) *J. Biol. Chem.* 263, 14008–14014.
73. Tomkinson, B., & Jonsson, A.K. (1991) *Biochemistry* 30, 168–174.
74. Chang, Y.Y., Wang, A.Y., & Cronan, J.E., Jr. (1993) *J. Biol. Chem.* 268, 3911–3919.
75. Vlahos, C.J., & Dekker, E.E. (1988) *J. Biol. Chem.* 263, 11683–11691.
76. Banfield, M.J., & Brady, R.L. (2000) *J. Mol. Biol.* 297, 1159–1170.
77. Haller, T., Buckel, T., Retey, J., & Gerlt, J.A. (2000) *Biochemistry* 39, 4622–4629.
78. Bond, M.D., & Strydom, D.J. (1989) *Biochemistry* 28, 6110–6113.
79. Wistow, G., & Piatigorsky, J. (1987) *Science* 236, 1554–1556.
80. Gulick, A.M., Palmer, D.R., Babbitt, P.C., Gerlt, J.A., & Rayment, I. (1998) *Biochemistry* 37, 14358–14368.
81. Mross, G.A., Doolittle, R.F., & Roberts, B.F. (1970) *Science* 170, 468–470.
82. Rossmann, M.G., Moras, D., & Olsen, K.W. (1974) *Nature* 250, 194–199.
83. Rossmann, M.G., & Argos, P. (1976) *J. Mol. Biol.* 105, 75–95.
84. Remington, S.J., Woodbury, R.G., Reynolds, R.A., Matthews, B.W., & Neurath, H. (1988) *Biochemistry* 27, 8097–8105.
85. Bazan, J.F., Weaver, L.H., Roderick, S.L., Huber, R., & Matthews, B.W. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 2473–2477.
86. Lah, M.S., Dixon, M.M., Patridge, K.A., Stallings, W.C., Fee, J.A., & Ludwig, M.L. (1995) *Biochemistry* 34, 1646–1660.
87. Aleshin, A.E., Zeng, C., Bourenkov, G.P., Bartunik, H.D., Fromm, H.J., & Honzatko, R.B. (1998) *Structure* 6, 39–50.
88. Evrard, C., Fastrez, J., & Declercq, J.P. (1998) *J. Mol. Biol.* 276, 151–164.
89. Mathews, I.L., Erion, M.D., & Ealick, S.E. (1998) *Biochemistry* 37, 15607–15620.
90. Holmgren, A., & Branden, C.I. (1989) *Nature* 342, 248–251.
91. Campobasso, N., Costello, C.A., Kinsland, C., Begley, T.P., & Ealick, S.E. (1998) *Biochemistry* 37, 15981–15989.
92. Wang, W., Kappock, T.J., Stubbe, J., & Ealick, S.E. (1998) *Biochemistry* 37, 15647–15662.

93. Przylas, I., Tomoo, K., Terada, Y., Takaha, T., Fujii, K., Saenger, W., & Strater, N. (2000) *J. Mol. Biol.* 296, 873–886.
94. Takano, T. (1977) *J. Mol. Biol.* 110, 537–568.
95. Poljak, R.J. (1978) *CRC Crit. Rev. Biochem.* 5, 45–84.
96. Biesecker, G., Harris, J.I., Thierry, J.C., Walker, J.E., & Wonacott, A.J. (1977) *Nature* 266, 328–333.
97. Kossiakoff, A.A., Chambers, J.L., Kay, L.M., & Stroud, R.M. (1977) *Biochemistry* 16, 654–664.
98. Tang, J., James, M.N., Hsu, I.N., Jenkins, J.A., & Blundell, T.L. (1978) *Nature* 271, 618–621.
99. Bolin, J.T., Filman, D.J., Matthews, D.A., Hamlin, R.C., & Kraut, J. (1982) *J. Biol. Chem.* 257, 13650–13662.
100. Hynes, T.R., Randal, M., Kennedy, L.A., Eigenbrot, C., & Kossiakoff, A.A. (1990) *Biochemistry* 29, 10018–10022.
101. Chothia, C., & Lesk, A.M. (1986) *EMBO J.* 5, 823–826.
102. Sielecki, A.R., Hayakawa, K., Fujinaga, M., Murphy, M.E., Fraser, M., Muir, A.K., Carilli, C.T., Lewicki, J.A., Baxter, J.D., & James, M.N. (1989) *Science* 243, 1346–1351.
103. MacRae, I.J., Segel, I.H., & Fisher, A.J. (2000) *Biochemistry* 39, 1613–1621.
104. Serrano, R., Kielland-Brandt, M.C., & Fink, G.R. (1986) *Nature* 319, 689–693.
105. Kyte, J. (1981) *Nature* 292, 201–204.
106. Ochi, H., Hata, Y., Tanaka, N., Kakudo, M., Sakurai, T., Aihara, S., & Morita, Y. (1983) *J. Mol. Biol.* 166, 407–418.
107. Meyer, T.E., & Kamen, M.D. (1982) *Adv. Protein Chem.* 35, 105–212.
108. Almasy, R.J., & Dickerson, R.E. (1978) *Proc. Natl. Acad. Sci. U.S.A.* 75, 2674–2678.
109. Benini, S., Gonzalez, A., Rypniewski, W.R., Wilson, K.S., Van Beeumen, J.J., & Ciurli, S. (2000) *Biochemistry* 39, 13115–13126.
110. Goddette, D.W., Paech, C., Yang, S.S., Mielenz, J.R., Bystroff, C., Wilke, M.E., & Fletterick, R.J. (1992) *J. Mol. Biol.* 228, 580–595.
111. Tahirov, T.H., Oki, H., Tsukihara, T., Ogasahara, K., Yutani, K., Ogata, K., Izu, Y., Tsunasawa, S., & Kato, I. (1998) *J. Mol. Biol.* 284, 101–124.
112. Endrizzi, J.A., Breddam, K., & Remington, S.J. (1994) *Biochemistry* 33, 11106–11120.
113. Sagermann, M., Baase, W.A., & Matthews, B.W. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 6078–6083.
114. Timkovich, R., & Dickerson, R.E. (1976) *J. Biol. Chem.* 251, 4033–4046.
115. Chothia, C., & Lesk, A.M. (1985) *J. Mol. Biol.* 182, 151–158.
116. Eriksson, A.E., Cousens, L.S., Weaver, L.H., & Matthews, B.W. (1991) *Proc. Natl. Acad. Sci. U.S.A.* 88, 3441–3445.
117. Lesk, A.M., & Chothia, C. (1980) *J. Mol. Biol.* 136, 225–270.
118. Babbitt, P.C., Hasson, M.S., Wedekind, J.E., Palmer, D.R., Barrett, W.C., Reed, G.H., Rayment, I., Ringe, D., Kenyon, G.L., & Gerlt, J.A. (1996) *Biochemistry* 35, 16489–16501.
119. Volbeda, A., Lahm, A., Sakiyama, F., & Suck, D. (1991) *EMBO J.* 10, 1607–1618.
120. Hadfield, A., Kryger, G., Ouyang, J., Petsko, G.A., Ringe, D., & Viola, R. (1999) *J. Mol. Biol.* 289, 991–1002.
121. Krishnaswamy, S., & Rossmann, M.G. (1990) *J. Mol. Biol.* 211, 803–844.
122. Al-Lazikani, B., Sheinerman, F.B., & Honig, B. (2001) *Proc. Natl. Acad. Sci. U.S.A.* 98, 14796–14801.
123. Risler, J.L., Delorme, M.O., Delacroix, H., & Henaut, A. (1988) *J. Mol. Biol.* 204, 1019–1029.
124. Johnson, M.S., & Overington, J.P. (1993) *J. Mol. Biol.* 233, 716–738.
125. Sielecki, A.R., Fedorov, A.A., Boodhoo, A., Andreeva, N.S., & James, M.N. (1990) *J. Mol. Biol.* 214, 143–170.
126. Louie, G.V., & Brayer, G.D. (1990) *J. Mol. Biol.* 214, 527–555.
127. Bilwes, A., Rees, B., Moras, D., Menez, R., & Menez, A. (1994) *J. Mol. Biol.* 239, 122–136.
128. Louie, G.V., Hutcheon, W.L., & Brayer, G.D. (1988) *J. Mol. Biol.* 199, 295–314.
129. Phillips, S.E. (1980) *J. Mol. Biol.* 142, 531–554.
130. Chothia, C., & Lesk, A.M. (1987) *Cold Spring Harbor Symp. Quant. Biol.* 52, 399–405.
131. Scott, D.L., White, S.P., Otwinowski, Z., Yuan, W., Gelb, M.H., & Sigler, P.B. (1990) *Science* 250, 1541–1546.
132. Xue, Y., Lipscomb, W.N., Graf, R., Schnappauf, G., & Braus, G. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 10814–10818.
133. Gourley, D.G., Shrive, A.K., Polikarpov, I., Krell, T., Coggins, J.R., Hawkins, A.R., Isaacs, N.W., & Sawyer, L. (1999) *Nat. Struct. Biol.* 6, 521–525.
134. Tainer, J.A., Getzoff, E.D., Beem, K.M., Richardson, J.S., & Richardson, D.C. (1982) *J. Mol. Biol.* 160, 181–217.
135. Cooper, J.B., McIntyre, K., Badasso, M.O., Wood, S.P., Zhang, Y., Garbe, T.R., & Young, D. (1995) *J. Mol. Biol.* 246, 531–544.
136. Yeh, A.P., Chatelet, C., Soltis, S.M., Kuhn, P., Meyer, J., & Rees, D.C. (2000) *J. Mol. Biol.* 300, 587–595.
137. Crane, B.R., Arvai, A.S., Gachhui, R., Wu, C., Ghosh, D.K., Getzoff, E.D., Stuehr, D.J., & Tainer, J.A. (1997) *Science* 278, 425–431.
138. McPhalen, C.A., & James, M.N. (1988) *Biochemistry* 27, 6582–6598.
139. Baker, P.J., Sawa, Y., Shibata, H., Sedelnikova, S.E., & Rice, D.W. (1998) *Nat. Struct. Biol.* 5, 561–567.
140. Mattevi, A., Vanoni, M.A., Todone, F., Rizzi, M., Teplyakov, A., Coda, A., Bolognesi, M., & Curti, B. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 7496–7501.
141. Ohlsson, I., Nordstrom, B., & Breandaen, C.I. (1974) *J. Mol. Biol.* 89, 339–354.
142. Banks, R.D., Blake, C.C., Evans, P.R., Haser, R., Rice, D.W., Hardy, G.W., Merrett, M., & Phillips, A.W. (1979) *Nature* 279, 773–777.
143. Fletterick, R.J., Sygusch, J., Semple, H., & Madsen, N.B. (1976) *J. Biol. Chem.* 251, 6142–6146.
144. Kolker, E., & Trifonov, E.N. (1995) *Proc. Natl. Acad. Sci. U.S.A.* 92, 557–560.
145. Tong, L., Wengler, G., & Rossmann, M.G. (1993) *J. Mol. Biol.* 230, 228–247.
146. Larsen, T.M., Laughlin, L.T., Holden, H.M., Rayment, I., & Reed, G.H. (1994) *Biochemistry* 33, 6301–6309.
147. Rossmann, M.G., & Argos, P. (1975) *J. Biol. Chem.* 250, 7525–7532.
148. Karplus, P.A., Daniels, M.J., & Herriott, J.R. (1991) *Science* 251, 60–66.

149. Bruns, C.M., & Karplus, P.A. (1995) *J. Mol. Biol.* 247, 125–145.
150. Porter, R.R. (1959) *Biochem. J.* 73, 119–126.
151. Harris, L.J., Larson, S.B., Hasel, K.W., & McPherson, A. (1997) *Biochemistry* 36, 1581–1597.
152. Saphire, E.O., Parren, P.W., Pantophlet, R., Zwick, M.B., Morris, G.M., Rudd, P.M., Dwek, R.A., Stanfield, R.L., Burton, D.R., & Wilson, I.A. (2001) *Science* 293, 1155–1159.
153. Silverton, E.W., Navia, M.A., & Davies, D.R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* 74, 5140–5144.
154. Betton, J.M., Desmadril, M., & Yon, J.M. (1989) *Biochemistry* 28, 5421–5428.
155. Yamaguchi, H., Kato, H., Hata, Y., Nishioka, T., Kimura, A., Oda, J., & Katsube, Y. (1993) *J. Mol. Biol.* 229, 1083–1100.
156. Tomaszek, T.A., Jr., Moore, M.L., Strickler, J.E., Sanchez, R.L., Dixon, J.S., Metcalf, B.W., Hassell, A., Dreyer, G.B., Brooks, I., Debouck, C., Meek, T.D., & Lewis, M. (1992) *Biochemistry* 31, 10153–10168.
157. Baldwin, E.T., Bhat, T.N., Gulnik, S., Hosur, M.V., Sowder, R.C.N., Cachau, R.E., Collins, J., Silva, A.M., & Erickson, J.W. (1993) *Proc. Natl. Acad. Sci. U.S.A.* 90, 6796–6800.
158. Rubenstein, D.S., Enghild, J.J., & Pizzo, S.V. (1991) *J. Biol. Chem.* 266, 11252–11261.
159. Husten, E.J., & Eipper, B.A. (1991) *J. Biol. Chem.* 266, 17004–17010.
160. Southerland, W.M., Winge, D.R., & Rajagopalan, K.V. (1978) *J. Biol. Chem.* 253, 8747–8752.
161. Appell, K.C., & Low, P.S. (1981) *J. Biol. Chem.* 256, 11104–11111.
162. Eberhard, M., Tsai-Pflugfelder, M., Bolewska, K., Hommel, U., & Kirschner, K. (1995) *Biochemistry* 34, 5419–5428.
163. Wenk, M., Baumgartner, R., Holak, T.A., Huber, R., Jaenicke, R., & Mayr, E.M. (1999) *J. Mol. Biol.* 286, 1533–1545.
164. Bustos, S.A., & Schleif, R.F. (1993) *Proc. Natl. Acad. Sci. U.S.A.* 90, 5638–5642.
165. Jhee, K.H., McPhie, P., & Miles, E.W. (2000) *Biochemistry* 39, 10548–10556.
166. Leistler, B., & Perham, R.N. (1994) *Biochemistry* 33, 2773–2781.
167. Sibilli, L., Le Bras, G., Le Bras, G., & Cohen, G.N. (1981) *J. Biol. Chem.* 256, 10228–10230.
168. Vaeron, M., Falcoz-Kelly, F., & Cohen, G.N. (1972) *Eur. J. Biochem.* 28, 520–527.
169. Parsot, C., & Cohen, G.N. (1988) *J. Biol. Chem.* 263, 14654–14660.
170. Rechler, M.M., & Bruni, C.B. (1971) *J. Biol. Chem.* 246, 1806–1813.
171. Chen, H.P., & Marsh, E.N. (1997) *Biochemistry* 36, 14939–14945.
172. Bein, K., Simmer, J.P., & Evans, D.R. (1991) *J. Biol. Chem.* 266, 3791–3799.
173. Coleman, P.F., Suttle, D.P., & Stark, G.R. (1977) *J. Biol. Chem.* 252, 6379–6385.
174. Simmer, J.P., Kelly, R.E., Rinker, A.G., Jr., Scully, J.L., & Evans, D.R. (1990) *J. Biol. Chem.* 265, 10395–10402.
175. Zimmermann, B.H., & Evans, D.R. (1993) *Biochemistry* 32, 1519–1527.
176. Simmer, J.P., Kelly, R.E., Rinker, A.G., Jr., Zimmermann, B.H., Scully, J.L., Kim, H., & Evans, D.R. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87, 174–178.
177. Simmer, J.P., Kelly, R.E., Scully, J.L., Grayson, D.R., Rinker, A.G., Jr., Bergh, S.T., & Evans, D.R. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 4382–4386.
178. Kretschmer, M., Langer, C., & Prinz, W. (1993) *Biochemistry* 32, 11143–11148.
179. Mally, M.I., Grayson, D.R., & Evans, D.R. (1981) *Proc. Natl. Acad. Sci. U.S.A.* 78, 6647–6651.
180. Taniuchi, H., & Anfinsen, C.B. (1971) *J. Biol. Chem.* 246, 2291–2301.
181. Guillou, F., Rubino, S.D., Markovitz, R.S., Kinney, D.M., & Lusty, C.J. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 8304–8308.
182. Teplyakov, A., Obmolova, G., Badet, B., & Badet-Denisot, M.A. (2001) *J. Mol. Biol.* 313, 1093–1102.
183. Knighton, D.R., Kan, C.C., Howland, E., Janson, C.A., Hostomska, Z., Welsh, K.M., & Matthews, D.A. (1994) *Nat. Struct. Biol.* 1, 186–194.
184. Hawkins, A.R., & Smith, M. (1991) *Eur. J. Biochem.* 196, 717–724.
185. Keesey, J.K., Jr., Bigelis, R., & Fink, G.R. (1979) *J. Biol. Chem.* 254, 7427–7433.
186. Barford, D., Flint, A.J., & Tonks, N.K. (1994) *Science* 263, 1397–1404.
187. Caughey, I., & Kekwick, R.G. (1982) *Eur. J. Biochem.* 123, 553–561.
188. Shimakata, T., & Stumpf, P.K. (1982) *Arch. Biochem. Biophys.* 218, 77–91.
189. Mohamed, A.H., Chirala, S.S., Mody, N.H., Huang, W.Y., & Wakil, S.J. (1988) *J. Biol. Chem.* 263, 12315–12325.
190. Chirala, S.S., Kuziora, M.A., Spector, D.M., & Wakil, S.J. (1987) *J. Biol. Chem.* 262, 4231–4240.
191. Holzer, K.P., Liu, W., & Hammes, G.G. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 4387–4391.
192. Joshi, A.K., & Smith, S. (1993) *J. Biol. Chem.* 268, 22508–22513.
193. Chirala, S.S., Huang, W.Y., Jayakumar, A., Sakai, K., & Wakil, S.J. (1997) *Proc. Natl. Acad. Sci. U.S.A.* 94, 5588–5593.
194. Chang, S.I., & Hammes, G.G. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 8373–8376.
195. Haese, A., Schubert, M., Herrmann, M., & Zocher, R. (1993) *Mol. Microbiol.* 7, 905–914.
196. Billich, A., & Zocher, R. (1987) *Biochemistry* 26, 8417–8423.
197. Weinreb, P.H., Quadri, L.E., Walsh, C.T., & Zuber, P. (1998) *Biochemistry* 37, 1575–1584.
198. Pieper, R., Ebert-Khosla, S., Cane, D., & Khosla, C. (1996) *Biochemistry* 35, 2054–2060.
199. Wu, N., Kudo, F., Cane, D.E., & Khosla, C. (2000) *J. Am. Chem. Soc.* 122, 4847–4852.
200. Xue, Y., & Sherman, D.H. (2000) *Nature* 403, 571–575.
201. Berg, A., Vervoort, J., & de Kok, A. (1996) *J. Mol. Biol.* 261, 432–442.
202. Lim, F., Morris, C.P., Occhiodoro, F., & Wallace, J.C. (1988) *J. Biol. Chem.* 263, 11493–11497.
203. Chang, C., Kokontis, J., & Liao, S. (1988) *Proc. Natl. Acad. Sci. U.S.A.* 85, 7211–7215.

204. Ringheim, G.E., & Taylor, S.S. (1990) *J. Biol. Chem.* 265, 4800–4808.
205. Coves, J., Zeghouf, M., Macherel, D., Guigliarelli, B., Asso, M., & Fontecave, M. (1997) *Biochemistry* 36, 5921–5928.
206. Stallings, W.C., Abdel-Meguid, S.S., Lim, L.W., Shieh, H.S., Dayringer, H.E., Leimgruber, N.K., Stegeman, R.A., Anderson, K.S., Sikorski, J.A., Padgett, S.R., & Kishore, G.M. (1991) *Proc. Natl. Acad. Sci. U.S.A.* 88, 5046–5050.
207. Stauffer, M.E., Young, J.K., & Evans, J.N. (2001) *Biochemistry* 40, 3951–3957.
208. Yee, V.C., Pedersen, L.C., Le Trong, I., Bishop, P.D., Stenkamp, R.E., & Teller, D.C. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 7296–7300.
209. Waldrop, G.L., Rayment, I., & Holden, H.M. (1994) *Biochemistry* 33, 10249–10256.
210. Levine, M., Muirhead, H., Stammers, D.K., & Stuart, D.I. (1978) *Nature* 271, 626–630.
211. Richardson, J.S. (1981) *Adv. Protein Chem.* 34, 167–339.
212. Ito, N., Phillips, S.E.V., Yadav, K.D.S., & Knowles, P.F. (1994) *J. Mol. Biol.* 238, 794–814.
213. Wierenga, R.K., Drenth, J., & Schulz, G.E. (1983) *J. Mol. Biol.* 167, 725–739.
214. Hecht, H.J., Kalisz, H.M., Hendle, J., Schmid, R.D., & Schomburg, D. (1993) *J. Mol. Biol.* 229, 153–172.
215. Olsen, L.R., & Roderick, S.L. (2001) *Biochemistry* 40, 1913–1921.
216. Marcotte, E.M., Pellegrini, M., Yeates, T.O., & Eisenberg, D. (1999) *J. Mol. Biol.* 293, 151–160.
217. Lang, D., Thoma, R., Henn-Sax, M., Sterner, R., & Wilmanns, M. (2000) *Science* 289, 1546–1550.
218. Hocker, B., Beismann-Driemeyer, S., Hettwer, S., Lustig, A., & Sterner, R. (2001) *Nat. Struct. Biol.* 8, 32–36.
219. McLachlan, A.D. (1979) *J. Mol. Biol.* 128, 49–79.
220. Ploegman, J.H., Drent, G., Kalk, K.H., & Hol, W.G. (1978) *J. Mol. Biol.* 123, 557–594.
221. Nyunoya, H., & Lusty, C.J. (1983) *Proc. Natl. Acad. Sci. U.S.A.* 80, 4629–4633.
222. Roderick, S.L., & Matthews, B.W. (1993) *Biochemistry* 32, 3907–3912.
223. Cirilli, M., Zheng, R., Scapin, G., & Blanchard, J.S. (1998) *Biochemistry* 37, 16452–16458.
224. Campbell, R.E., Mosimann, S.C., van De Rijn, I., Tanner, M.E., & Strynadka, N.C. (2000) *Biochemistry* 39, 7012–7023.
225. Wilmanns, M., Priestle, J.P., Niermann, T., & Jansonius, J.N. (1992) *J. Mol. Biol.* 223, 477–507.
226. Schwab, D.A., & Wilson, J.E. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 2563–2567.
227. Kurokawa, H., Mikami, B., & Hirose, M. (1995) *J. Mol. Biol.* 254, 196–207.
228. Lindley, P.F., Bajaj, M., Evans, R.W., Garratt, R.C., Hasnain, S.S., Jhoti, H., Kuser, P., Neu, M., Patel, K., et al. (1993) *Acta Crystallogr., Sect. D: Biol. Crystallogr.* D49, 292–304.
229. Jaskolski, M., Miller, M., Rao, J.K., Leis, J., & Wlodawer, A. (1990) *Biochemistry* 29, 5889–5898.
230. Newman, M., Watson, F., Roychowdhury, P., Jones, H., Badasso, M., Cleasby, A., Wood, S.P., Tickle, I.J., & Blundell, T.L. (1993) *J. Mol. Biol.* 230, 260–283.
231. Lin, X.L., Lin, Y.Z., Koelsch, G., Gustchina, A., Wlodawer, A., & Tang, J. (1992) *J. Biol. Chem.* 267, 17257–17263.
232. McLachlan, A.D., & Walker, J.E. (1977) *J. Mol. Biol.* 112, 543–558.
233. He, X.M., & Carter, D.C. (1992) *Nature* 358, 209–215.
234. Fong, S.L., & Bridges, C.D. (1988) *J. Biol. Chem.* 263, 15330–15334.
235. Pope, B., Maciver, S., & Weeds, A. (1995) *Biochemistry* 34, 1583–1588.
236. Lee, F.S., Fox, E.A., Zhou, H.M., Strydom, D.J., & Vallee, B.L. (1988) *Biochemistry* 27, 8545–8553.
237. Miller, K.I., Cuff, M.E., Lang, W.F., Varga-Weisz, P., Field, K.G., & van Holde, K.E. (1998) *J. Mol. Biol.* 278, 827–842.
238. Bhandari, V., Palfree, R.G., & Bateman, A. (1992) *Proc. Natl. Acad. Sci. U.S.A.* 89, 1715–1719.
239. Manning, A.M., Trotman, C.N., & Tate, W.P. (1990) *Nature* 348, 653–656.
240. Niu, X.D., Browning, K.S., Behal, R.H., & Reed, L.J. (1988) *Proc. Natl. Acad. Sci. U.S.A.* 85, 7546–7550.
241. Terry, A.S., Poulter, L., Williams, D.H., Nutkins, J.C., Giovannini, M.G., Moore, C.H., & Gibson, B.W. (1988) *J. Biol. Chem.* 263, 5745–5751.
242. Pepinsky, R.B., Tizard, R., Mattaliano, R.J., Sinclair, L.K., Miller, G.T., Browning, J.L., Chow, E.P., Burne, C., Huang, K.S., Pratt, D., et al. (1988) *J. Biol. Chem.* 263, 10799–10811.
243. Han, S., Eltis, L.D., Timmis, K.N., Muchmore, S.W., & Bolin, J.T. (1995) *Science* 270, 976–980.
244. Labeit, S., & Kolmerer, B. (1995) *J. Mol. Biol.* 248, 308–315.
245. Speicher, D.W., & Marchesi, V.T. (1984) *Nature* 311, 177–180.
246. Dubreuil, R.R., Byers, T.J., Sillman, A.L., Bar-Zvi, D., Goldstein, L.S., & Branton, D. (1989) *J. Cell Biol.* 109, 2197–2205.
247. Winograd, E., Hume, D., & Branton, D. (1991) *Proc. Natl. Acad. Sci. U.S.A.* 88, 10788–10791.
248. Sahr, K.E., Laurila, P., Kotula, L., Scarpa, A.L., Coupal, E., Leto, T.L., Linnenbach, A.J., Winkelmann, J.C., Speicher, D.W., Marchesi, V.T., Curtis, P.J., & Forget, B.G. (1990) *J. Biol. Chem.* 265, 4434–4443.
249. Grum, V.L., Li, D., MacDonald, R.I., & Mondragon, A. (1999) *Cell* 98, 523–535.
250. Pascual, J., Pfuhl, M., Walther, D., Saraste, M., & Nilges, M. (1997) *J. Mol. Biol.* 273, 740–751.
251. Yan, Y., Winograd, E., Viel, A., Cronin, T., Harrison, S.C., & Branton, D. (1993) *Science* 262, 2027–2030.
252. Speicher, D.W., Morrow, J.S., Knowles, W.J., & Marchesi, V.T. (1982) *J. Biol. Chem.* 257, 9093–9101.
253. Renault, L., Nassar, N., Vetter, I., Becker, J., Klebe, C., Roth, M., & Wittinghofer, A. (1998) *Nature* 392, 97–101.
254. Oubrie, A., Rozeboom, H.J., Kalk, K.H., Duine, J.A., & Dijkstra, B.W. (1999) *J. Mol. Biol.* 289, 319–333.
255. Sondek, J., Bohm, A., Lambright, D.G., Hamm, H.E., & Sigler, P.B. (1996) *Nature* 379, 369–374.
256. Habazettl, J., Gondol, D., Wiltsccheck, R., Otlewski, J., Schleicher, M., & Holak, T.A. (1992) *Nature* 359, 855–858.
257. Onesti, S., Brick, P., & Blow, D.M. (1991) *J. Mol. Biol.* 217, 153–176.

258. Saper, M.A., Bjorkman, P.J., & Wiley, D.C. (1991) *J. Mol. Biol.* 219, 277–319.
259. Wright, C.S. (1992) *J. Biol. Chem.* 267, 14345–14352.
260. Prince, J.T., McGrath, K.P., DiGirolamo, C.M., & Kaplan, D.L. (1995) *Biochemistry* 34, 10879–10885.
261. Liou, Y.C., Thibault, P., Walker, V.K., Davies, P.L., & Graham, L.A. (1999) *Biochemistry* 38, 11415–11424.
262. Hasson, M.S., Muscate, A., McLeish, M.J., Polovnikova, L.S., Gerlt, J.A., Kenyon, G.L., Petsko, G.A., & Ringe, D. (1998) *Biochemistry* 37, 9918–9930.
263. Arjunan, P., Umland, T., Dyda, F., Swaminathan, S., Furey, W., Sax, M., Farrenkopf, B., Gao, Y., Zhang, D., & Jordan, F. (1996) *J. Mol. Biol.* 256, 590–600.
264. Muller, Y.A., Schumacher, G., Rudolph, R., & Schulz, G.E. (1994) *J. Mol. Biol.* 237, 315–335.
265. Lobel, P., Dahms, N.M., & Kornfeld, S. (1988) *J. Biol. Chem.* 263, 2563–2570.
266. Wang, K., McClure, J., & Tu, A. (1979) *Proc. Natl. Acad. Sci. U.S.A.* 76, 3698–3702.
267. Pan, K.-M., Damodaran, S., & Greaser, M.L. (1994) *Biochemistry* 33, 8255–8261.
268. Labeit, S., & Kolmerer, B. (1995) *Science* 270, 293–296.
269. Linke, W.A., Ivemeyer, M., Olivieri, N., Kolmerer, B., Ruegg, J.C., & Labeit, S. (1996) *J. Mol. Biol.* 261, 62–71.
270. Politou, A.S., Gautel, M., Pfuhl, M., Labeit, S., & Pastore, A. (1994) *Biochemistry* 33, 4730–4737.
271. Doolittle, R.F. (1985) *Trends Biochem. Sci. (Pers. Ed.)* 10, 233–237.
272. Doolittle, R.F. (1989) in *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G. D., Ed.) pp 599–623, Plenum, New York.
273. Kretsinger, R.H., & Nockolds, C.E. (1973) *J. Biol. Chem.* 248, 3313–3326.
274. Essen, L.O., Perisic, O., Cheung, R., Katan, M., & Williams, R.L. (1996) *Nature* 380, 595–602.
275. Wang, J.H., Yan, Y.W., Garrett, T.P., Liu, J.H., Rodgers, D.W., Garlick, R.L., Tarr, G.E., Husain, Y., Reinherz, E.L., & Harrison, S.C. (1990) *Nature* 348, 411–418.
276. Su, X.D., Gastinel, L.N., Vaughn, D.E., Faye, I., Poon, P., & Bjorkman, P.J. (1998) *Science* 281, 991–995.
277. Holden, H.M., Ito, M., Hartshorne, D.J., & Rayment, I. (1992) *J. Mol. Biol.* 227, 840–851.
278. Jones, E.Y., Harlos, K., Bottomley, M.J., Robinson, R.C., Driscoll, P.C., Edwards, R.M., Clements, J.M., Dudgeon, T.J., & Stuart, D.I. (1995) *Nature* 373, 539–544.
279. Schneider, R., Schneider-Scherzer, E., Thurnher, M., Auer, B., & Schweiger, M. (1988) *EMBO J.* 7, 4151–4156.
280. Kobe, B., & Deisenhofer, J. (1993) *Nature* 366, 751–756.
281. Crowder, S.M., Kanaar, R., Rio, D.C., & Alber, T. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 4892–4897.
282. Graves, B.J., Crowther, R.L., Chandran, C., Rumberger, J.M., Li, S., Huang, K.S., Presky, D.H., Familletti, P.C., Wolitzky, B.A., & Burns, D.K. (1994) *Nature* 367, 532–538.
283. Sasaki, T., Kostka, G., Gohring, W., Wiedemann, H., Mann, K., Chu, M.L., & Timpl, R. (1995) *J. Mol. Biol.* 245, 241–250.
284. Dahlback, B., Hildebrand, B., & Linse, S. (1990) *J. Biol. Chem.* 265, 18481–18489.
285. Huang, L.H., Cheng, H., Pardi, A., Tam, J.P., & Sweeney, W.V. (1991) *Biochemistry* 30, 7402–7409.
286. Stetefeld, J., Mayer, U., Timpl, R., & Huber, R. (1996) *J. Mol. Biol.* 257, 644–657.
287. Tavares, G.A., Beguin, P., & Alzari, P.M. (1997) *J. Mol. Biol.* 273, 701–713.
288. Spinelli, S., Fierobe, H.P., Belaich, A., Belaich, J.P., Henrissat, B., & Cambillau, C. (2000) *J. Mol. Biol.* 304, 189–200.
289. Foord, R., Taylor, I.A., Sedgwick, S.G., & Smerdon, S.J. (1999) *Nat. Struct. Biol.* 6, 157–165.
290. Michaely, P., & Bennett, V. (1993) *J. Biol. Chem.* 268, 22703–22709.
291. Sutton, R.B., Davletov, B.A., Berghuis, A.M., Sudhof, T.C., & Sprang, S.R. (1995) *Cell* 80, 929–938.
292. Mikol, V., Baumann, G., Keller, T.H., Manning, U., & Zurini, M.G. (1995) *J. Mol. Biol.* 246, 344–355.
293. Tong, L., Warren, T.C., King, J., Betageri, R., Rose, J., & Jakes, S. (1996) *J. Mol. Biol.* 256, 601–610.
294. Waksman, G., Kominos, D., Robertson, S.C., Pant, N., Baltimore, D., Birge, R.B., Cowburn, D., Hanafusa, H., Mayer, B.J., Overduin, M., et al. (1992) *Nature* 358, 646–653.
295. Hatada, M.H., Lu, X., Laird, E.R., Green, J., Morgenstern, J.P., Lou, M., Marr, C.S., Phillips, T.B., Ram, M.K., Theriault, K., Zoller, M.J., & Karas, J.L. (1995) *Nature* 377, 32–38.
296. Musacchio, A., Saraste, M., & Wilmanns, M. (1994) *Nat. Struct. Biol.* 1, 546–551.
297. Maignan, S., Guilloteau, J.P., Fromage, N., Arnoux, B., Becquart, J., & Ducruix, A. (1995) *Science* 268, 291–293.
298. Stec, B., Yamano, A., Whitlow, M., & Teeter, M.M. (1997) *Acta Crystallogr., Sect. D, Pt. 2* 53, 169–178.
299. Pennica, D., Holmes, W.E., Kohr, W.J., Harkins, R.N., Vehar, G.A., Ward, C.A., Bennett, W.F., Yelverton, E., Seeburg, P.H., Heyneker, H.L., Goeddel, D.V., & Collen, D. (1983) *Nature* 301, 214–221.
300. Claeys, H., Sottrup-Jensen, L., Zajdel, M., Petersen, T.E., & Magnusson, S. (1976) *FEBS Lett.* 61, 20–24.
301. Bottomley, M.J., Collard, M.W., Huggenvik, J.I., Liu, Z., Gibson, T.J., & Sattler, M. (2001) *Nat. Struct. Biol.* 8, 626–633.
302. Timm, D., Salim, K., Gout, I., Guruprasad, L., Waterfield, M., & Blundell, T. (1994) *Nat. Struct. Biol.* 1, 782–788.
303. Gibson, T.J., Hyvonen, M., Musacchio, A., Saraste, M., & Birney, E. (1994) *Trends Biochem. Sci.* 19, 349–353.
304. Petersen, T.E., Thogersen, H.C., Skorstengaard, K., Vibe-Pedersen, K., Sahl, P., Sottrup-Jensen, L., & Magnusson, S. (1983) *Proc. Natl. Acad. Sci. U.S.A.* 80, 137–141.
305. Kornblihtt, A.R., Umezawa, K., Vibe-Pedersen, K., & Baralle, F.E. (1985) *EMBO J.* 4, 1755–1759.
306. Baron, M., Norman, D.G., & Campbell, I.D. (1991) *Trends Biochem. Sci.* 16, 13–17.
307. Pickford, A.R., Smith, S.P., Staunton, D., Boyd, J., & Campbell, I.D. (2001) *EMBO J.* 20, 1519–1529.
308. Gorlich, D., Prehn, S., Laskey, R.A., & Hartmann, E. (1994) *Cell* 79, 767–778.
309. Conti, E., Uy, M., Leighton, L., Blobel, G., & Kuriyan, J. (1998) *Cell* 94, 193–204.
310. Muller, Y.A., Ultsch, M.H., & de Vos, A.M. (1996) *J. Mol. Biol.* 256, 144–159.

311. Streuli, M., Krueger, N.X., Tsai, A.Y., & Saito, H. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 8698–8702.
312. Dickinson, C.D., Veerapandian, B., Dai, X.P., Hamlin, R.C., Xuong, N.H., Ruoslahti, E., & Ely, K.R. (1994) *J. Mol. Biol.* 236, 1079–1092.
313. Tsujishita, Y., & Hurley, J.H. (2000) *Nat. Struct. Biol.* 7, 408–414.
314. Gomis-Ruth, F.X., Gohlke, U., Betz, M., Knauper, V., Murphy, G., Lopez-Otin, C., & Bode, W. (1996) *J. Mol. Biol.* 264, 556–566.
315. Faber, H.R., Groom, C.R., Baker, H.M., Morgan, W.T., Smith, A., & Baker, E.N. (1995) *Structure* 3, 551–559.
316. Jenne, D., & Stanley, K.K. (1987) *Biochemistry* 26, 6735–6742.
317. Pawson, T. (1995) *Nature* 373, 573–580.
318. Engel, J., Efimov, V.P., & Maurer, P. (1994) *Development, Suppl. (Evol. Dev. Mech.)*, 35–42.
319. Evdokimov, A.G., Anderson, D.E., Routzahn, K.M., & Waugh, D.S. (2001) *J. Mol. Biol.* 312, 807–821.
320. Wu, H., Maciejewski, M.W., Marintchev, A., Benashski, S.E., Mullen, G.P., & King, S.M. (2000) *Nat. Struct. Biol.* 7, 575–579.
321. Liou, Y.-C., Tocilj, A., Davies, P.L., & Jia, Z. (2000) *Nature* 406, 322–325.
322. White, C.E., Hunter, M.J., Meininger, D.P., Garrod, S., & Komives, E.A. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 10177–10182.
323. Kumar, A., Roach, C., Hirsh, I.S., Turley, S., deWalque, S., Michels, P.A., & Hol, W.G. (2001) *J. Mol. Biol.* 307, 271–282.
324. Essen, L.O., Perisic, O., Katan, M., Wu, Y., Roberts, M.F., & Williams, R.L. (1997) *Biochemistry* 36, 1704–1718.
325. Ferguson, K.M., Lemmon, M.A., Schlessinger, J., & Sigler, P.B. (1995) *Cell* 83, 1037–1046.
326. Kretsinger, R.H., & Nakayama, S. (1993) *J. Mol. Evol.* 36, 477–488.
327. Kobe, B., & Deisenhofer, J. (1994) *Trends Biochem. Sci.* 19, 415–421.
328. Tornero, P., Mayda, E., Gomez, M.D., Canas, L., Conejero, V., & Vera, P. (1996) *Plant J.* 10, 315–330.
329. Logsdon, J.M., Jr., Tyshenko, M.G., Dixon, C., D-Jafari, J., Walker, V.K., & Palmer, J.D. (1995) *Proc. Natl. Acad. Sci. U.S.A.* 92, 8507–8511.
330. de Souza, S.J., Long, M., Klein, R.J., Roy, S., Lin, S., & Gilbert, W. (1998) *Proc. Natl. Acad. Sci. U.S.A.* 95, 5094–5099.
331. Stoltzfus, A., Spencer, D.F., Zuker, M., Logsdon, J.M., Jr., & Doolittle, W.F. (1994) *Science* 265, 202–207.
332. Rzhetsky, A., Ayala, F.J., Hsu, L.C., Chang, C., & Yoshida, A. (1997) *Proc. Natl. Acad. Sci. U.S.A.* 94, 6820–6825.
333. Wetlaufer, D.B. (1973) *Proc. Natl. Acad. Sci. U.S.A.* 70, 697–701.
334. Rose, G.D. (1979) *J. Mol. Biol.* 134, 447–470.
335. Ollis, D.L., Brick, P., Hamlin, R., Xuong, N.G., & Steitz, T.A. (1985) *Nature* 313, 762–766.
336. Mande, S.S., Sarfaty, S., Allen, M.D., Perham, R.N., & Hol, W.G. (1996) *Structure* 4, 277–286.
337. Wilson, I.A., Skehel, J.J., & Wiley, D.C. (1981) *Nature* 289, 366–373.
338. Vainshtein, B.K., Melik-Adamyan, W.R., Barynin, V.V., Vagin, A.A., Grebenko, A.I., Borisov, V.V., Bartels, K.S., Fita, I., & Rossmann, M.G. (1986) *J. Mol. Biol.* 188, 49–61.
339. Murthy, M.R., Reid, T.J.d., Sicignano, A., Tanaka, N., & Rossmann, M.G. (1981) *J. Mol. Biol.* 152, 465–499.
340. Faber, H.R., & Matthews, B.W. (1990) *Nature* 348, 263–266.
341. Stehle, T., Ahmed, S.A., Claiborne, A., & Schulz, G.E. (1991) *J. Mol. Biol.* 221, 1325–1344.
342. Ziegler, G.A., Vornrhein, C., Hanukoglu, I., & Schulz, G.E. (1999) *J. Mol. Biol.* 289, 981–990.
343. Doolittle, R.F., Goldbaum, D.M., & Doolittle, L.R. (1978) *J. Mol. Biol.* 120, 311–325.
344. Williams, R.C. (1981) *J. Mol. Biol.* 150, 399–408.
345. Yang, Z., Kollman, J.M., Pandi, L., & Doolittle, R.F. (2001) *Biochemistry* 40, 12515–12523.
346. Donovan, J.W., & Mihalyi, E. (1974) *Proc. Natl. Acad. Sci. U.S.A.* 71, 4125–4128.
347. Novokhatny, V.V., Kudinov, S.A., & Privalov, P.L. (1984) *J. Mol. Biol.* 179, 215–232.
348. Rudolph, R., Siebendritt, R., Nessler, G., Sharma, A.K., & Jaenicke, R. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87, 4625–4629.
349. Betton, J.M., Desmadril, M., Mitraki, A., & Yon, J.M. (1984) *Biochemistry* 23, 6654–6661.
350. Tauler, A., Rosenberg, A.H., Colosia, A., Studier, F.W., & Pilakis, S.J. (1988) *Proc. Natl. Acad. Sci. U.S.A.* 85, 6642–6646.
351. Herold, M., Leistler, B., Hage, A., Luger, K., & Kirschner, K. (1991) *Biochemistry* 30, 3612–3620.
352. Funk, W.D., MacGillivray, R.T., Mason, A.B., Brown, S.A., & Woodworth, R.C. (1990) *Biochemistry* 29, 1654–1660.
353. Maduke, M., Williams, C., & Miller, C. (1998) *Biochemistry* 37, 1315–1321.
354. Kay, L.E., Forman-Kay, J.D., McCubbin, W.D., & Kay, C.M. (1991) *Biochemistry* 30, 4323–4333.
355. Robien, M.A., Clore, G.M., Omichinski, J.G., Perham, R.N., Appella, E., Sakaguchi, K., & Gronenborn, A.M. (1992) *Biochemistry* 31, 3463–3471.
356. Spraggon, G., Applegate, D., Everse, S.J., Zhang, J.Z., Veerapandian, L., Redman, C., Doolittle, R.F., & Grieneringer, G. (1998) *Proc. Natl. Acad. Sci. U.S.A.* 95, 9099–9104.
357. Missiakas, D., Betton, J.M., Minard, P., & Yon, J.M. (1990) *Biochemistry* 29, 8683–8689.
358. Vita, C., Fontana, A., & Jaenicke, R. (1989) *Eur. J. Biochem.* 183, 513–518.
359. Vita, C., Fontana, A., & Chaiken, I.M. (1985) *Eur. J. Biochem.* 151, 191–196.
360. Trexler, M., & Patthy, L. (1983) *Proc. Natl. Acad. Sci. U.S.A.* 80, 2457–2461.
361. Wilhelm, O.G., Jaskunas, S.R., Vlahos, C.J., & Bang, N.U. (1990) *J. Biol. Chem.* 265, 14606–14611.
362. Moreau, M., de Cock, E., Fortier, P.L., Garcia, C., Albaret, C., Blanquet, S., Lallemand, J.Y., & Dardel, F. (1997) *J. Mol. Biol.* 266, 15–22.
363. Daubner, S.C., Schrimsher, J.L., Schendel, F.J., Young, M., Henikoff, S., Patterson, D., Stubbe, J., & Benkovic, S.J. (1985) *Biochemistry* 24, 7059–7062.
364. Worrall, D.M., & Tubbs, P.K. (1983) *Biochem. J.* 215, 153–157.

365. Weis, W.I., Brunger, A.T., Skehel, J.J., & Wiley, D.C. (1990) *J. Mol. Biol.* 212, 737–761.
366. Banner, D.W., Bloomer, A., Petsko, G.A., Phillips, D.C., & Wilson, I.A. (1976) *Biochem. Biophys. Res. Commun.* 72, 146–155.
367. Beese, L.S., Derbyshire, V., & Steitz, T.A. (1993) *Science* 260, 352–355.
368. Zhao, B., Janson, C.A., Amegadzie, B.Y., D'Alessio, K., Griffin, C., Hanning, C.R., Jones, C., Kurdyla, J., McQueney, M., Qiu, X., Smith, W.W., & Abdel-Meguid, S.S. (1997) *Nat. Struct. Biol.* 4, 109–111.
369. Huang, S., Xue, Y., Sauer-Eriksson, E., Chirica, L., Lindskog, S., & Jonsson, B.H. (1998) *J. Mol. Biol.* 283, 301–310.
370. Zhang, C., & DeLisi, C. (1998) *J. Mol. Biol.* 284, 1301–1305.
371. Murzin, A.G., Brenner, S.E., Hubbard, T., & Chothia, C. (1995) *J. Mol. Biol.* 247, 536–540.
372. Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., & Thornton, J.M. (1997) *Structure* 5, 1093–1108.
373. Holm, L., & Sander, C. (1996) *Science* 273, 595–603.
374. Kull, F.J., Sablin, E.P., Lau, R., Fletterick, R.J., & Vale, R.D. (1996) *Nature* 380, 550–555.
375. Pakhomova, S., Kobayashi, M., Buck, J., & Newcomer, M.E. (2001) *Nat. Struct. Biol.* 8, 447–451.
376. Morais, M.C., Zhang, W., Baker, A.S., Zhang, G., Dunaway-Mariano, D., & Allen, K.N. (2000) *Biochemistry* 39, 10385–10396.
377. Nakatsu, T., Kato, H., & Oda, J. (1998) *Nat. Struct. Biol.* 5, 15–19.
378. Jones, E.Y., Stuart, D.I., & Walker, N.P. (1989) *Nature* 338, 225–228.
379. Spurlino, J.C., Lu, G.Y., & Quirocho, F.A. (1991) *J. Biol. Chem.* 266, 5202–5219.
380. Lamzin, V.S., Aleshin, A.E., Strokopytov, B.V., Yuhnevich, M.G., Popov, V.O., Harutyunyan, E.H., & Wilson, K.S. (1992) *Eur. J. Biochem.* 206, 441–452.
381. Zhang, C., & Kim, S.H. (2000) *J. Mol. Biol.* 299, 1075–1089.
382. Aleshin, A., Golubev, A., Firsov, L.M., & Honzatko, R.B. (1992) *J. Biol. Chem.* 267, 19291–19298.
383. Cordes, M.H., Burton, R.E., Walsh, N.P., McKnight, C.J., & Sauer, R.T. (2000) *Nat. Struct. Biol.* 7, 1129–1132.
384. Ollis, D.L., Cheah, E., Cygler, M., Dijkstra, B., Frolow, F., Franken, S.M., Harel, M., Remington, S.J., Silman, I., Schrag, J., et al. (1992) *Protein Eng.* 5, 197–211.
385. Narayana, S.V., Carson, M., el-Kabbani, O., Kilpatrick, J.M., Moore, D., Chen, X., Bugg, C.E., Volanakis, J.E., & DeLucas, L.J. (1994) *J. Mol. Biol.* 235, 695–708.
386. Goldberg, J.D., Yoshida, T., & Brick, P. (1994) *J. Mol. Biol.* 236, 1123–1140.
387. De Bondt, H.L., Rosenblatt, J., Jancarik, J., Jones, H.D., Morgan, D.O., & Kim, S.H. (1993) *Nature* 363, 595–602.
388. Zhang, F., Strand, A., Robbins, D., Cobb, M.H., & Goldsmith, E.J. (1994) *Nature* 367, 704–711.
389. Kack, H., Sandmark, J., Gibson, K., Schneider, G., & Lindqvist, Y. (1999) *J. Mol. Biol.* 291, 857–876.
390. Thompson, T.B., Garrett, J.B., Taylor, E.A., Meganathan, R., Gerlt, J.A., & Rayment, I. (2000) *Biochemistry* 39, 10662–10676.
391. Thoden, J.B., Firestine, S., Nixon, A., Benkovic, S.J., & Holden, H.M. (2000) *Biochemistry* 39, 8791–8802.
392. Steinbacher, S., Baxa, U., Miller, S., Weintraub, A., Seckler, R., & Huber, R. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 10584–10588.
393. Xia, Z., Dai, W., Zhang, Y., White, S.A., Boyd, G.D., & Mathews, F.S. (1996) *J. Mol. Biol.* 259, 480–501.
394. Crennell, S.J., Garman, E.F., Philippon, C., Vasella, A., Laver, W.G., Vimr, E.R., & Taylor, G.L. (1996) *J. Mol. Biol.* 259, 264–280.
395. Barbosa, J.A., Smith, B.J., DeGori, R., Ooi, H.C., Marcuccio, S.M., Campi, E.M., Jackson, W.R., Brossmer, R., Sommer, M., & Lawrence, M.C. (2000) *J. Mol. Biol.* 303, 405–421.
396. Khurana, S., Powers, D.B., Anderson, S., & Blaber, M. (1998) *Proc. Natl. Acad. Sci. U.S.A.* 95, 6768–6773.
397. Uhlin, U., & Eklund, H. (1994) *Nature* 370, 533–539.
398. Stec, B., & Lebioda, L. (1990) *J. Mol. Biol.* 211, 235–248.
399. Chelvanayagam, G., Heringa, J., & Argos, P. (1992) *J. Mol. Biol.* 228, 220–242.
400. Romao, M.J., Kolln, I., Dias, J.M., Carvalho, A.L., Romero, A., Varela, P.F., Sanz, L., Topfer-Petersen, E., & Calvete, J.J. (1997) *J. Mol. Biol.* 274, 650–660.
401. Varela, P.F., Romero, A., Sanz, L., Romao, M.J., Topfer-Petersen, E., & Calvete, J.J. (1997) *J. Mol. Biol.* 274, 635–649.
402. Deisenhofer, J. (1981) *Biochemistry* 20, 2361–2370.
403. Juy, M., Amit, A.G., Alzari, P.M., Poljak, R.J., Claeysens, M., Beguin, P., & Aubert, J.P. (1992) *Nature* 357, 89–91.
404. Schreuder, H., Tardif, C., Trump-Kallmeyer, S., Soffientini, A., Sarubbi, E., Akesson, A., Bowlin, T., Yanofsky, S., & Barrett, R.W. (1997) *Nature* 386, 194–200.
405. Vigers, G.P., Anderson, L.J., Caffes, P., & Brandhuber, B.J. (1997) *Nature* 386, 190–194.
406. Keitel, T., Simon, O., Borriss, R., & Heinemann, U. (1993) *Proc. Natl. Acad. Sci. U.S.A.* 90, 5287–5291.
407. Vaughn, D.E., Rodriguez, J., Lazebnik, Y., & Joshua-Tor, L. (1999) *J. Mol. Biol.* 293, 439–447.
408. Hill, C.P., Osslund, T.D., & Eisenberg, D. (1993) *Proc. Natl. Acad. Sci. U.S.A.* 90, 5167–5171.
409. Kraulis, P.J. (1991) *J. Applied Crystallogr.* 24, 946–950.

Chapter 8

Counting Polypeptides

Almost all of the proteins found in a living organism are multimeric proteins. A **multimeric protein** is a protein containing more than one folded polypeptide. Each of these folded polypeptides was originally synthesized by a ribosome from messenger RNA that encoded a sequence of amino acids of a precise, finite length. These polypeptides folded into defined conformations and were posttranslationally modified. Each of the folded, posttranslationally modified polypeptides in a multimeric protein is one of its **subunits**. Usually, only a specific and well-defined number of these subunits are gathered together to form the macromolecular complex that is the finished, existing molecule of the multimeric protein. An **oligomeric protein** is a multimeric protein with a fixed, invariant number of subunits. In a few instances, such as the proteins actin, keratin, and collagen, a large and undefined number of subunits combine to form a polymeric protein, which in theory could continue to polymerize indefinitely. A **polymeric protein** is a protein with many subunits, the number of which varies from molecule to molecule of that protein. Polymeric proteins are the exception; most proteins are oligomeric.

The **stoichiometry of the subunits** of a protein is the number of each type of folded, posttranslationally modified polypeptide that are combined to produce the specific structure. At this level of definition, each of the polypeptides is identified only by its length. The **length of a polypeptide** is the number of amino acids it contains, n_{aa} , an integer that is either dimensionless or has the units of amino acids (molecule of polypeptide)⁻¹ or moles of amino acids (mole of polypeptide)⁻¹. The length of a polypeptide is usually a precisely known quantity because its amino acid sequence is usually available and any posttranslational modifications have usually been defined.

A great deal of effort has been expended in discovering the stoichiometries of the subunits of proteins. The original approach to this information was to determine the molar mass of the intact protein, to separate the individual polypeptides composing the protein, to quantify the mass ratio among the various polypeptides, and to determine the molar mass of each of the separated polypeptides. The measurement of the molar masses of intact proteins was at one time a major area of biophysical research, but this pursuit presently attracts much less attention.

The individual subunits composing an intact native

protein are separated and catalogued analytically by electrophoresis on polyacrylamide gels cast in solutions of the detergent sodium dodecyl sulfate. The separation that is effected by these polyacrylamide gels relies on their ability to sieve the unfolded polypeptides. The constituent polypeptides of a protein are separated preparatively by chromatography that depends either on sieving or on ion exchange of the unfolded, dissociated polymers. The separated polypeptides are shown to be homogeneous and unique by peptide mapping.

The major weaknesses of the original approach to defining the stoichiometry of the subunits of a protein were the extreme care with which the initial measurements of the molar mass of the intact protein had to be performed and the unreliability of the assessments of the mass ratios among the constituent polypeptides and of their molar masses. The present approach to defining the stoichiometry of the subunits of a protein avoids these problems. The individual polypeptides composing a protein are still separated and catalogued by electrophoresis and shown to be unique and homogeneous by peptide mapping. The length of each of the constituent polypeptides is assessed either by the electrophoresis itself or, preferably, by sequencing the appropriate cDNA. Any glycosylation is quantified analytically. The number of each polypeptide composing the intact protein is determined by covalently cross-linking the protein to various degrees of completion and identifying the various intermediate covalent complexes and the limit complex.

The different subunits in an oligomeric protein are defined by their lengths and distinguished by assigning them consecutive letters of the Greek alphabet. For example, deoxyhemoglobin at its normal concentrations is constructed from two subunits, α and β , each present in two copies to produce the complex $(\alpha\beta)_2$. Nicotinic acetylcholine receptor has the composition $\alpha_2\beta\gamma\delta$; L-lactate dehydrogenase, $(\alpha_2)_2$; DNA-directed RNA polymerase from *Escherichia coli*, $\alpha_2\beta\gamma\delta$; and 2-dehydro-3-deoxy-phosphogluconate aldolase, α_3 . The grouping of subunits into subsets, for example, the groups of two subunits in L-lactate dehydrogenase, arises from the symmetries in which the subunits are arranged within the intact molecule of an oligomeric protein.

Some oligomeric proteins, when they are dissolved at certain concentrations, are mixtures of two different combinations of subunits in equilibrium with each other. For example, oxygenated hemoglobin is an equi-

librium mixture of $\alpha\beta$ dimers and $(\alpha\beta)_2$ tetramers. Most oligomeric proteins, however, have a particular composition of subunits that does not vary unless harsh conditions are applied. A solution of a pure oligomeric protein will usually be monodisperse. A monodisperse solution of a particular macromolecule is a solution in which every one of those macromolecules is of the same size and shape, and each is compact and unique and remains dissolved and unassociated with its neighbors.

When the map of electron density from a crystal of an oligomeric protein is examined, the complete molecule can be discerned. It is recognized as a large, independent feature in the map that is formed from several folded tubes of electron density. Although not always essential, it is reassuring to know before the map is examined how many subunits are combined to produce the protein that has been crystallized. This determination can be made by a combination of sequencing, molecular sieving, and cross-linking. It is now so routine to do this that few oligomeric proteins the subunit stoichiometry of which have not already been established are examined crystallographically.

Molar Mass^{1,2}

The only indubitably reliable method for determining the length and amino acid composition of a polypeptide, and consequently, its precise molar mass, is to **sequence** it correctly. At the moment, the sequences of a large array of readily available polypeptides are known, and they form a collection of standards each of the lengths of which, n_{aa} , is an exact quantity. It is often, but not always, the case that the amino acid sequence of a newly purified polypeptide is known from the nucleotide sequence of its cDNA before enough of it becomes available to study its physical properties in detail. This situation has inverted the classical strategy of physical measurements in which the molar mass of a protein was one of the ultimate discoveries rather than something precisely known from the beginning. Unfortunately, however, the extensive effort expended in determinations of molar mass, although expended decades ago, still influences the importance attached to molar mass.

The **molar mass** of a protein, M_{prot} , is the number of grams in a mole of that protein. The **molecular mass** of a protein is the mass of a single molecule of that protein expressed in relative units that are referred to as either atomic mass units (amu) or daltons (Da). Both an atomic mass unit and a dalton (1.6606×10^{-24} g) are $1/12$ the mass of carbon isotope 12. Because Avogadro's number is the number of carbon atoms of isotope 12 in 12 g of carbon atoms of isotope 12, the numerical values of molar mass and molecular mass for the same molecule are the same, but not the units attached to the numbers. Molar mass is the quantity that is determined by the measurements of physical behavior such as osmotic pressure, sedimenta-

tion equilibrium, and light scattering because the units on the final quantity are usually grams mole⁻¹.

Because both sedimentation equilibrium and light scattering are alternative measurements of osmotic pressure, all three techniques determine the same fundamental physical property of a protein in a solution, namely, its chemical potential. Osmotic pressure is a colligative property of the solution. A **colligative property** of a solute is a physical property that is a function only of the moles of independent particles of that solute in a standard volume of the solution. If the solution were monodisperse, if the only osmotically active particles present were individual molecules of the protein, and if the concentration of the protein in grams centimeter⁻³ were known precisely, the measured molar concentration could be used to calculate the molar mass of the protein.

Osmotic pressure is the pressure exerted by impermeant solutes when a solution containing those impermeant solutes is separated by a semipermeable membrane from another solution identical in every way to the first except that it lacks the impermeant solutes. A solute is **impermeant** to a particular barrier if it cannot pass through that barrier. A **semipermeable membrane** is a membrane through which all of the components of the two solutions can pass freely except the impermeant solutes. The chemical and physical properties of the membrane define which solutes are permeant and which are impermeant, which are osmotically silent and which are osmotically active, respectively. In the case of experimental measurements of the osmotic pressure of solutions of proteins, a membrane is chosen that is porous to the small molecules in the solution but the pores of which are too small to pass the molecules of protein. This is usually a sheet formed from a polymeric material that has spaces between the strands of polymer wide enough to pass small molecules and ions but too narrow to pass macromolecules.

Pressure is the force that results from the tendency of molecules to expand the confines in which they are contained and fill a larger volume. The pressure they exert is a direct measurement of the **chemical potential** of a population of molecules. When molecules such as impermeant solutes exert an osmotic pressure, they cannot expand their confines, as can gases, by entering the vacant space above the solution because they are held by intermolecular forces within a condensed phase. The volume of solution in which they are confined, however, can expand if solution from the other side passes across the semipermeable membrane into the solution containing the impermeant solutes. This is the liquid analogy to a balloon expanding to fill more space. Just as the balloon expands until the external pressure is equivalent to the internal pressure of the trapped gas, the solution containing the impermeant molecules expands until an external pressure equal to its osmotic pressure is applied to it. Operationally, the osmotic pressure of a

solution is the external pressure that must be applied to the solution containing the impermeant solute to prevent any expansion in its volume by the net movement of fluid through the semipermeable membrane. In an apparatus for measuring osmotic pressure, the difference in pressure between the chamber containing the protein and the chamber lacking the protein can be maintained and continuously monitored with a pressure transducer.³

It can be shown⁴ that the osmotic pressure exerted by a solution of impermeant solutes is formally equivalent to the pressure exerted on the walls of a container filled with gases that are impermeant to those walls. The molecules of the impermeant solute are formally equivalent to the molecules of the gas, and those of the solvent and permeant solutes are as physically silent as the vacuum in which the gas is suspended. The **ideal gas law** is

$$P = \frac{n_m RT}{V} \quad (8-1)$$

where P is the pressure (newtons centimeter⁻²) exerted by the gas, R is the gas constant (831.5 N cm K⁻¹ mol⁻¹), T is the temperature (kelvins), n_m is the number of moles of the gas, and V is the volume (centimeters³) in which it is confined. The pressure exerted by a real, nonideal gas, however, is

$$P = RT \left[\frac{n_m}{V} + B \left(\frac{n_m}{V} \right)^2 + C \left(\frac{n_m}{V} \right)^3 \dots \right] \quad (8-2)$$

where $n_m V^{-1}$ is the concentration (moles centimeter⁻³) of the gas and the coefficients B , C , and so forth are referred to, respectively, as the second virial coefficient, the third virial coefficient, and so forth. The **virial coefficients** provide the necessary corrections for the behavior of the nonideal gas due to the specific properties that make it nonideal, such as the finite dimensions of the molecules that fill, or exclude, some of the volume of the container, the intermolecular forces between the molecules of the gas, and the tendency of molecules of the gas to dimerize or polymerize. For all of the same reasons,⁴ the osmotic pressure Π (newtons centimeter⁻²), exerted by a nonideal, impermeant solute S is

$$\Pi = RT([S] + B[S]^2 + C[S]^3 \dots) \quad (8-3)$$

where $[S]$ is the molar concentration of the solute. If the impermeant solute S is protein i

$$\lim_{[\text{protein } i] \rightarrow 0} \Pi = RT[\text{protein } i] \quad (8-4)$$

Equation 8-4 states that, at low enough concentrations of protein, the osmotic pressure observed will be directly proportional to the molar concentration of the protein.

The molar concentration of the protein in the solution cannot be known if the molar mass of the protein is unknown, but the **concentration of the protein** in the solution must be known in some type of units. From the ultraviolet spectrum of the solution, the concentration (moles centimeter⁻³) of the tryptophan and tyrosine in the protein might be known.⁵ From a colorimetric assay, the concentration (moles centimeter⁻³) of peptide bonds in the solution might be known.⁶ From total amino acid analysis,⁷ the concentration (moles centimeter⁻³) of amino acids in the solution might be known. From a dry weight measurement of the protein, the grams of dry weight of protein in a milliliter of solution (grams centimeter⁻³) might be known. Regardless of the units, the value for the concentration of protein in the units of the quantity measured can be designated as C_{prot} (units centimeter⁻³). It follows that

$$W[\text{protein } i] = C_{\text{prot}} \quad (8-5)$$

where W is a constant of proportionality.

The units on W are moles of tryptophan and tyrosine (mole of protein)⁻¹, moles of peptide bonds (mole of protein)⁻¹, moles of amino acids (mole of protein)⁻¹, or grams of dry weight (mole of protein)⁻¹, respectively. It is only coincidental that the last is usually chosen. It is this exercise that defines what a colligative property is and illustrates that osmotic pressure does not measure molar mass directly. The final result can be always traced back to an independent measurement of the concentration of the protein. When Equation 8-5 is incorporated into Equation 8-4

$$\lim_{C_{\text{prot}} \rightarrow 0} \frac{\Pi}{C_{\text{prot}}} = \frac{RT}{W} \quad (8-6)$$

and the intercept of Π/C_{prot} as C_{prot} is decreased to 0 provides the value of W (Figure 8-1).^{8*} If the units of the concentration C_{prot} were grams of protein centimeter⁻³ and the only osmotically active species present were molecules of the protein of interest in a monodisperse solution, W would be the molar mass of the protein in grams mole⁻¹.

The complete equation describing the actual behavior of the osmotic pressure at low concentrations of protein is

$$\Pi = RT \left(\frac{C_{\text{prot}}}{W} + B C_{\text{prot}}^2 + C C_{\text{prot}}^3 \dots \right) \quad (8-7)$$

where B , C , ... are the virial coefficients expressed in appropriate units. In Figure 8-1, the chosen concentra-

* The units of concentration used by the authors for the data in Figures 8-1, 8-2, and 8-5 were grams centimeter⁻³. The established symbol for concentration in units of mass volume⁻¹ is γ .

tions of the protein, bovine serum albumin, are in the range where only the first virial coefficient, B , is significant, and the results are presented as if the equation were

$$\frac{\Pi}{C_{\text{prot}}} = RT \left(\frac{1}{W} + BC_{\text{prot}} \right) \quad (8-8)$$

It can be seen that when the charge on the protein was changed by changing the pH, the value of the second virial coefficient, B , and hence the slope of the line, changed in response to changes in the various parameters of the solution, but in agreement with Equation 8-6, the intercept seems to have remained the same.

The fact that the slopes of the two lines in Figure 8-1, and hence the two values of the second virial coefficient, are different arises from the **Donnan effect**. Because a molecule of protein bears a net charge at a given pH and because the two solutions on either side of the semipermeable membrane must be electroneutral, the concentrations of electrolytes cannot be the same in

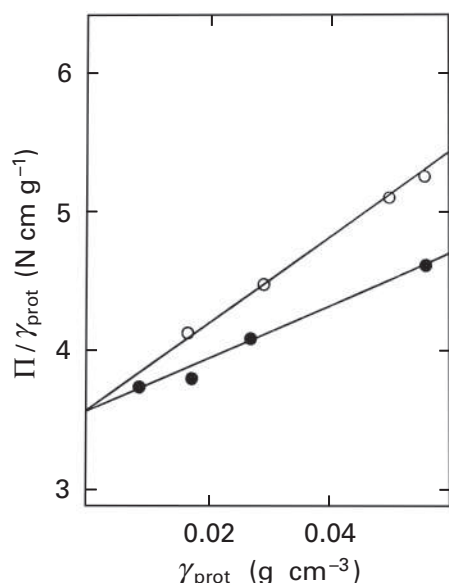


Figure 8-1: Osmotic pressure of solutions of bovine serum albumin.⁸ The incremental osmotic pressure (Π/γ_{prot}), where the osmotic pressure Π is in newtons centimeter⁻² and the concentration of protein γ_{prot} is in grams of protein centimeter⁻³, is plotted as a function of γ_{prot} . The apparatus was at 25 °C, and within the apparatus the solution of protein was separated from the solution lacking the protein by a membrane of nitrocellulose polymer. The osmotic pressure of the solution was the external pressure that had to be applied to the solution of protein, in excess of the atmospheric pressure on the solution lacking the protein, to prevent its expansion. The excess pressure was measured with a toluene manometer, and millimeters of toluene was converted to newtons centimeter⁻². The concentration of protein, following the measurement, in each of the solutions containing protein was determined by a dry weight analysis that was corrected for the weight of other solutes present. The solutions contained 0.15 M NaCl as supporting electrolyte. The behaviors of the osmotic pressure at pH 7.0 (○) and pH 5.37 (●) are shown. Adapted with permission from ref 8. Copyright 1946 American Chemical Society.

the two solutions. Suppose that the only electrolyte in the solution is KCl, the solution in the compartment containing the protein is designated α , and the solution in the other compartment is designated β . To preserve electroneutrality

$$[K^+]_{\beta} = [Cl^-]_{\beta} \quad (8-9)$$

and

$$\bar{Z}_i [\text{protein } i] + [K^+]_{\alpha} = [Cl^-]_{\alpha} \quad (8-10)$$

where \bar{Z}_i is the mean net charge number on protein i . If activity coefficients are ignored for the moment,

$$\mu_{\text{KCl}} = \mu_{\text{KCl}}^{\circ} + RT \ln [K^+][Cl^-] \quad (8-11)$$

where μ_{KCl} is the chemical potential of the KCl. The chemical potential of the KCl must be the same on both sides of the membrane, and μ_{KCl}° , the standard chemical potential of KCl, is always the same, so

$$[K^+]_{\alpha} [Cl^-]_{\alpha} = [K^+]_{\beta} [Cl^-]_{\beta} \quad (8-12)$$

at equilibrium. Combining these equations

$$[K^+]_{\beta} = [K^+]_{\alpha} \left(1 + \frac{\bar{Z}_i [\text{protein } i]}{[K^+]_{\alpha}} \right)^{1/2} \quad (8-13)$$

$$[Cl^-]_{\beta} = [Cl^-]_{\alpha} \left(1 - \frac{\bar{Z}_i [\text{protein } i]}{[Cl^-]_{\alpha}} \right)^{1/2} \quad (8-14)$$

Equations 8-13 and 8-14 state that if the protein is positively charged, there will be more chloride and less potassium in compartment α than in compartment β , or that if the protein is negatively charged, there will be more potassium and less chloride in compartment α than in compartment β . These imbalances of molecular species will affect the osmotic pressure because the electrolytes cannot distribute freely across the semipermeable membrane and consequently they become osmotically active. Although there is no exact solution to these equations, at low concentrations of protein and at small values of \bar{Z}_i ⁹

$$\lim_{[\text{protein } i] \rightarrow 0} \Pi \cong RT [\text{protein } i] \left(1 + \frac{\bar{Z}_i^2 [\text{protein } i]}{4[KCl]} \right) \quad (8-15)$$

For molecules with large values of \bar{Z}_i , such as nucleic acids, this approximation fails badly.

From Equation 8-15 it can be seen that the Donnan effect is expressed in the second virial coefficient, B , as demonstrated in Figure 8-1; but when $[KCl] > 5\bar{Z}_i^2[\text{pro-}]$

tein i], the Donnan effect has less than a 5% effect on the osmotic pressure and as [protein i] approaches 0, the Donnan effect also approaches 0. For the measurements presented in Figures 8-2 and 8-5, but not those in Figure 8-1, the addition of 0.1 M KCl would satisfy this inequality.

Another way to consider the imbalances of counterions created by the Donnan effect is to assume that the charge on the impermeant protein creates an electrical potential and the permeant ions redistribute in response to this potential. This **Donnan potential** complicates measurements performed by sedimentation equilibrium and light scattering as well as osmotic pressure. In the absence of added electrolytes, gradients or discontinuities of electrical potential would form during all three of these procedures as the concentration of the protein varies within the chambers of the apparatuses. These gradients or discontinuities of electrical potential, because they arise from the Donnan effect, can be eliminated in the case of most proteins by adding a simple electrolyte such as KCl to the solution to a concentration of around 0.1 M. One way to understand the effect of adding salt is to imagine that the increase in ionic strength decreases significantly the thickness of the ionic double layer (Equations 1-71 and 1-77) so that it encloses the molecule of protein tightly, effectively neutralizes its charge, and turns it into an apparently neutral macromolecule. The added **electrolyte** also eliminates any local gradients of electrical potential caused either by separating two phases by a semipermeable membrane,¹ as in measurements of osmotic pressure, or by differential gravitational forces exerted on ions of unlike mass, as in sedimentation equilibrium.¹⁰ In addition to an electrolyte, a buffer may also be added to the solution to maintain the pH.

When a solution of protein is submitted to **sedimentation equilibrium**, it is placed in a chamber within the strong centrifugal field created by a spinning rotor, and its distribution through the chamber is allowed to reach equilibrium. Because the centrifugal force in this chamber in the rotor is a function only of the radial distance r from the center of rotation, the distribution of the protein at equilibrium is a function only of r . At equilibrium, the centrifugal force upon the protein at each point in the solution is equal but opposite in sign to the force of diffusion that arises from the gradient of its concentration. If negligible gradients of electrical potential form because sufficient electrolyte has been added to the solution, the protein will redistribute until the differential of the **chemical potential** it experiences at each position in the chamber balances the differential of the **centrifugal potential** that it experiences. For a protein, the equality produced by this balance can be expressed as

$$\left(\frac{d\Pi}{dC_{\text{prot}}} \right) \left(\frac{dC_{\text{prot}}}{dr} \right) = \omega^2 r C_{\text{prot}} \left(\frac{\partial \rho}{\partial C_{\text{prot}}} \right)_{T,P,\mu} \quad (8-16)$$

from which the exact relationship^{2,10,11} follows by rearrangement

$$\frac{d \ln C_{\text{prot}}}{dr^2} = \left(\frac{\omega^2}{2} \right) \left(\frac{\partial \rho}{\partial C_{\text{prot}}} \right)_{T,P,\mu} \left(\frac{d\Pi}{dC_{\text{prot}}} \right)^{-1} \quad (8-17)$$

where C_{prot} is the concentration of protein expressed in any units, r is the distance (centimeters) of any point in the chamber from the center of the rotor, ω is the angular velocity (radians second⁻¹) of the rotor, ρ is the density (grams centimeter⁻³) of the solution of protein, and $(\partial \rho / \partial C_{\text{prot}})_{T,P,\mu}$ is the change in density of the solution as a function only of the change in the concentration of protein. Because at sedimentation equilibrium the chemical potential through the entire solution must be the same, this change in the density of the solution is at constant chemical potential of solvent and all solutes other than the protein, as indicated by the subscript μ .

Because the derivative on the left in Equation 8-17 is that of the natural logarithm of the concentration of protein, it will have the same numerical value regardless of the units chosen to express the concentration of the protein, and the units of concentration cancel on the right. Because

$$\frac{d\Pi}{dC_{\text{prot}}} = RT \left(\frac{1}{W} + 2BC_{\text{prot}} + 3CC_{\text{prot}}^2 \dots \right) \quad (8-18)$$

and

$$\lim_{C_{\text{prot}} \rightarrow 0} \left(\frac{1}{W} + 2BC_{\text{prot}} + 3CC_{\text{prot}}^2 \dots \right) = \frac{1}{W} \quad (8-19)$$

Equation 8-17 can be combined with Equations 8-18 and 8-19 and rearranged to give^{2,10}

$$\lim_{C_{\text{prot}} \rightarrow 0} \frac{d \ln C_{\text{prot}}}{dr^2} = \frac{\omega^2}{2RT} \left(\frac{\partial \rho}{\partial C_{\text{prot}}} \right)_{T,P,\mu} W \quad (8-20)$$

As predicted by Equation 8-20, when a monodisperse solution of a particular protein is submitted to centrifugation and the distribution of that protein is allowed to reach equilibrium, the **gradient of concentration** that forms is such that a plot of $\ln C_{\text{prot}}$ against r^2 is a straight line (Figure 8-2).¹² From the slope of this plot and a value for $(\partial \rho / \partial C_{\text{prot}})_{T,P,\mu}$, the value of W can be calculated.

The independent measurement of the partial derivative $(\partial \rho / \partial C_{\text{prot}})_{T,P,\mu}$ in Equation 8-20 requires the tabula-

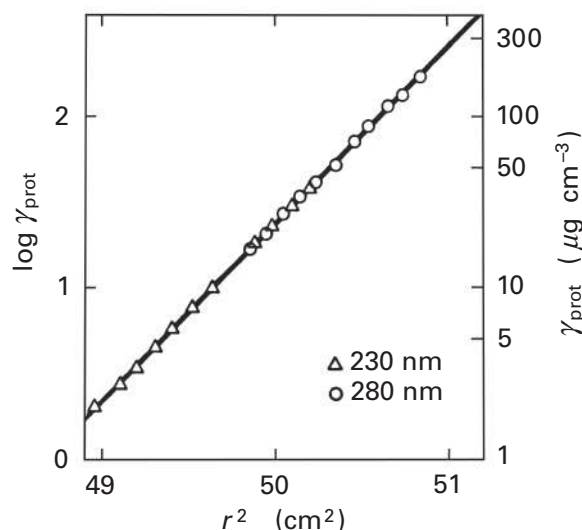


Figure 8-2: Sedimentation equilibrium of bovine serum albumin.¹² A solution of bovine serum albumin was placed in an optical cell in a rotor. The rotor was placed in an ultracentrifuge and spun at 24,630 rpm. The distribution of protein in the optical cell was followed by scanning the absorbance of the solution along the radial axis of the cell. After 18 h, the distribution of protein was no longer changing and equilibrium had been reached. The absorbance as a function of radial distance from the center of rotation of the rotor was measured at wavelengths of 280 nm (○) and 230 nm (△). Absorbance was converted to concentration of protein (γ_{prot} ; micrograms centimeter⁻³) and the logarithm to the base 10 of the concentration ($\log \gamma_{\text{prot}}$) was taken. Radial distance, r , was measured in centimeters and the respective values were squared (r^2). Adapted with permission from ref 12. Copyright 1966 American Chemical Society.

tion of the macroscopic density of a series of solutions containing increasing, precisely known concentrations of protein and brought to equilibrium, at the appropriate osmotic pressure, with a solution identical to the solution used to prepare the samples for centrifugation but lacking the protein. A procedure has been devised for measuring the required densities both at constant chemical potential of solvent and all diffusible solutes and at constant composition of the solution,^{13,14} and if this procedure is followed it permits the most accurate and reliable values for the molar mass of a protein to be calculated. Usually, however, the **approximation**

$$\left(\frac{\partial \rho}{\partial \gamma_{\text{prot}}} \right)_{T,P,\mu} \cong 1 - \bar{v}_{\text{prot}} \rho_{\text{sol}} \quad (8-21)$$

is made, where γ_{prot} is the concentration of protein in units of grams centimeter⁻³, \bar{v}_{prot} is the partial specific volume (centimeters³ gram⁻¹) of the protein in the particular solution chosen, and ρ_{sol} is the density (grams centimeter⁻³) of the solution in the absence of the protein. Unfortunately, because the decision to use this approximation was made for other reasons, many investigators are unaware that their use of the right-hand term in Equation 8-21 is only an approximation. This lack of

awareness can lead to some confusion.¹⁵ Furthermore, the approximation fails significantly under certain circumstances. For example, it is in error by 14% for DNA at 1 M sodium chloride and by 15% for protein in 4 M guanidinium chloride.¹⁶

The units on the term $(\partial \rho / \partial C_{\text{prot}})_{T,P,\mu}$, as is the case with the term Π / C_{prot} found in Equation 8-6, are determined by the units chosen for C_{prot} , the concentration of protein. By chance, it happens that if C_{prot} is expressed in units of grams centimeter⁻³ (γ_{prot}) and ρ is expressed in units of grams centimeter⁻³, $(\partial \rho / \partial \gamma_{\text{prot}})_{T,P,\mu}$ is dimensionless. Approximation 8-21, however, requires that C_{prot} be expressed in grams centimeter⁻³ and implicitly dictates a choice of these units for the entire equation. This assumption remains hidden in the definition of the partial specific volume

$$\bar{v}_{\text{prot}} \equiv \left(\frac{\partial V}{\partial m_{\text{prot}}} \right)_{T,P,m_j} \quad (8-22)$$

where m_{prot} is the mass (grams) of protein added to a solution, V is the resulting volume of the solution (centimeters³), and the subscript m_j states that the masses of all of the other components j in the solution must remain constant. The accuracy of this measurement is no greater than the accuracy with which the mass of the protein in grams can be known. Nor is this requirement avoided by the use of values of \bar{v}_{prot} calculated from the amino acid composition.¹⁷ In effect, this latter approximation simply relies on the care with which protein concentrations in grams centimeter⁻³ were determined in the earlier experiments validating such a calculation and demonstrating that it was reliable.¹⁸

Equation 8-20 requires an extrapolation to C_{prot} equal to 0. The purpose of the extrapolation is to eliminate the effect of the **virial coefficients** (Equation 8-19) on $d\Pi/dC_{\text{prot}}$. The same protein, bovine serum albumin, under similar conditions ($I_c = 0.1$ – 0.15 M), was used in the experiments described in Figures 8-1 and 8-2. From the values of the second virial coefficients, B , of Figure 8-1, it can be calculated that at the actual concentrations examined in Figure 8-2, the nonideality of the solution should only have affected the molar mass determined from the slope of the line by less than 0.1% of its value. Consequently, it is not surprising that the molar mass calculated¹² from Figure 8-2, 64,500 g mol⁻¹, agrees quite closely with the actual value of the molar mass of bovine serum albumin, 66,430 g mol⁻¹, a value that was subsequently established by its amino acid sequence.¹⁹ Uncoiled polypeptides or highly charged macromolecules such as DNA, however, have much larger virial coefficients, and sedimentation equilibrium of such species can be significantly affected by those virial coefficients.

At the present time, instruments that register the concentration of protein by its **absorbance** at 280 nm are

used to monitor its distribution over the cell within the rotor of the ultracentrifuge at sedimentation equilibrium.¹² Consequently, as the example of serum albumin in Figure 8–2 demonstrates, because the concentration of protein is so low, the virial coefficients can usually be ignored, but it is nevertheless prudent to perform measurements at several different concentrations of protein or several different angular velocities of the rotor or both to validate their insignificance.^{20,21}

If the virial coefficients can be disregarded so that the limit can be assumed, Equation 8–20 can be integrated, and because the concentration of protein (C_{prot}) is directly proportional to absorbance at 280 nm (A_{280})

$$A_{280} = A_{0,280} \exp \left[\frac{\omega^2 M_{\text{prot}}}{2RT} \left(\frac{\partial \rho}{\partial C_{\text{prot}}} \right)_{T,P,\mu} (r^2 - r_0^2) \right] \quad (8-23)$$

where $A_{0,280}$ is the absorbance the solution has at a reference position within the cell at which, by definition, $r = r_0$. This equation is then fit by nonlinear least squares to the distribution of absorbance as a function of radius (Figure 8–3).^{15,22,23} From the fit, a numerical value for the quantity $\omega^2 M_{\text{prot}} (\partial \rho / \partial C_{\text{prot}}) / 2RT$ is obtained. While ω^2 , R , and T are known precisely, M_{prot} and $(\partial \rho / \partial C_{\text{prot}})_{T,P,\mu}$ or its surrogate \bar{v}_{prot} (Equation 8–21) may or may not be.

Usually, independent measurements of either $(\partial \rho / \partial C_{\text{prot}})_{T,P,\mu}$ or \bar{v}_{prot} are made because the reason for the experiment is to determine M_{prot} . In the case of the p51 subunit of RNA-directed DNA polymerase from human immunodeficiency virus 1 (Figure 8–3), however, which is a monomer containing a single subunit, its molar mass M_{prot} (49,660 g mol^{–1}) was already known precisely from its amino acid sequence, and the purpose of the measurement of sedimentation equilibrium was to estimate $(\partial \rho / \partial \gamma_{\text{prot}})_{T,P,\mu}$ (0.225), a quantity that was needed for the later experiments reported. It is also possible to use the sedimentation equilibrium of a protein of known molar mass at different concentrations of another solute to determine a value for the preferential hydration of the protein, $(\partial m_{\text{H}_2\text{O}} / \partial m_{\text{prot}})_{T,P,\mu}$ in the presence of that solute.²⁴

From the molar mass of a subunit of a protein, which has been determined precisely by sequencing its constituent polypeptide or polypeptides, and the molar mass of the intact protein, which has been estimated by sedimentation equilibrium, the **number of subunits** in an intact protein can be assessed. For example, by sedimentation equilibrium, the molar mass of the UvsY recombination protein from bacteriophage T4 has been estimated to be 98.6 ± 3.9 kg mol^{–1}²⁵ and the molar mass of each of its identical constituent polypeptides is 15.84 kg mol^{–1}. The molar mass of carbon-monoxide dehydrogenase from *Moorella thermoacetica* was estimated to be 300 ± 30 kg mol^{–1} by sedimentation equilibrium;²⁶ the molar masses of its constituent polypeptides,

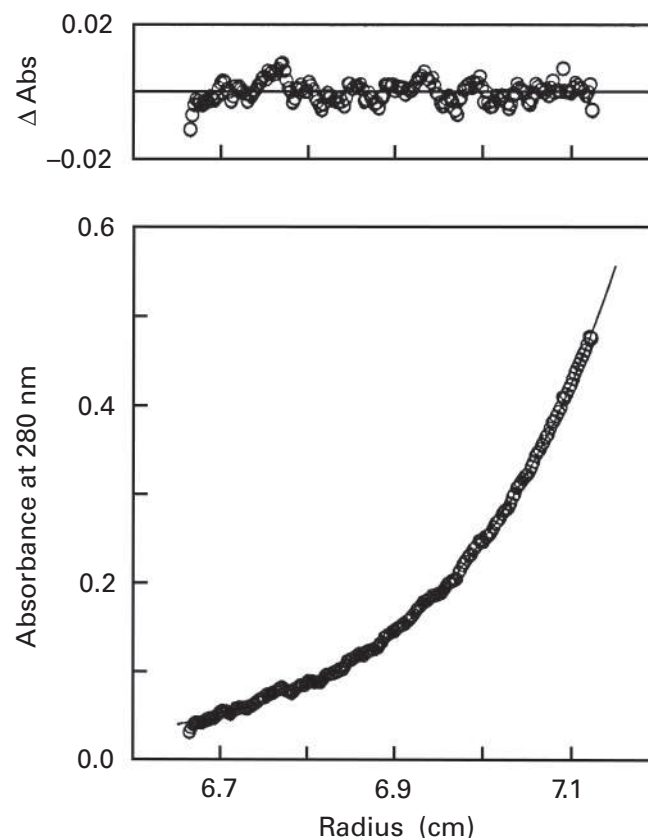


Figure 8–3: Sedimentation equilibrium of a recombinant form of the p51 subunit ($M = 49,660$ g mol^{–1}, $n_{\text{aa}} = 426$) of RNA-directed DNA polymerase from human immunodeficiency virus 1.¹⁵ The optical cell in the rotor was loaded with a solution of the protein with an initial absorbance at 280 nm of 0.29. Centrifugation was performed for 74 h at 12,000 rpm to reach equilibrium. The distribution of absorbance at 280 nm over the cell is plotted in the lower panel as a function of radius (centimeters) from the center of rotation. The line drawn through the points is a nonlinear least-squares fit of Equation 8–23 to the data, using as the reference position r_0 the point of highest measured absorbance, $A_{0,280}$, at the bottom of the optical cell (to the right of the graph). The upper panel is the deviation of the experimental absorbance (ΔAbs) for each point from the value of the fit at that point. The deviations are distributed at random around a value of 0.

α and β , are 81.73 and 72.92 kg mol^{–1} respectively; and the measured mass ratio of the two polypeptides in the intact protein is 1.18 g g^{–1}.

There is some confusion between the use in the present instance of a centrifugal field to create a gradient of the molar concentration of a protein and its use to measure the sedimentation coefficient, a hydrodynamic property of an individual molecule of the protein. Because the centrifugal potential at each point in the chamber can be calculated directly, at equilibrium the chemical potential of the solute, and hence its molar concentration at each point, can also be calculated. In contrast to this measurement at equilibrium, a measurement of a hydrodynamic property of a molecule of protein is, as the name implies, a kinetic measurement. In such a kinetic measurement, the rate of movement of the

molecule of protein under an applied force is measured. Free electrophoretic mobility is an example of such a hydrodynamic property. The confusion arises because, in addition to its use in sedimentation equilibrium, centrifugal force can be used to move a molecule of protein through a solution. This use of centrifugal force is unrelated to the use of centrifugal force to create a gradient of concentration. The confusion also arises because the same instrument, an analytical ultracentrifuge, is used to make each of the measurements, even though they are unrelated to each other.

The distribution of concentration as a function of radius at sedimentation equilibrium is often used to obtain a value for the **dissociation constant between two oligomeric states** of a protein that are in equilibrium with each other. For example, the protein could be an equilibrium mixture of a monomer and a dimer



for which there is a dissociation constant

$$K_{d,\alpha_2} = \frac{[\alpha]^2}{[\alpha_2]} \quad (8-25)$$

To estimate the numerical value of this dissociation constant, the mass-average molar masses can be calculated at selected radii from the respective slopes of the distribution of concentration at those radii at sedimentation equilibrium (Equation 8-20), and a plot of molar mass against concentration of protein can be fit by an equation incorporating the equilibrium between monomer and dimer.²²

Usually, however, a value for the dissociation constant is extracted directly from the distribution of the concentration of protein as a function of radius at sedimentation equilibrium. The purpose of creating the centrifugal field is to form predictably a continuous gradient in the molar concentration of the protein. If the protein is involved in an equilibrium between two forms that have different stoichiometries of subunits, the ratio between the molar concentrations of the two forms at a particular position in the sample will be a function of the total concentration of protein, $[\text{protein}]_{\text{TOT}}$, at that position. For example, in the case of an equilibrium between monomer and dimer

$$\frac{[\alpha_2]}{[\alpha]} = \left(\frac{1}{4} - \frac{2[\text{protein}]_{\text{TOT}}}{K_{d,\alpha_2}} \right)^{1/2} - \frac{1}{2} \quad (8-26)$$

Because the molar concentration of particles of protein at any position in the sample is $[\alpha] + [\alpha_2]$, the equilibrium between monomer and dimer affects the chemical potential of the protein and hence the balance between chemical potential and centrifugal potential. The result

of this effect is that the distribution of the concentration of protein as a function of radius is perturbed by the equilibrium between the two forms of the protein.

The distribution of the concentration of protein in the chamber as a function of the radial distance from the center of rotation can be fit by numerical analysis with an equation incorporating both the centrifugal potential and the perturbation caused by the equilibrium between the two forms of the protein¹⁵ to obtain simultaneously estimates of both the molar masses of the two forms and the numerical value of the equilibrium constant. For example, the distribution of the concentration of the p66 subunit of RNA-directed DNA polymerase from human immunodeficiency virus 1 at sedimentation equilibrium is consistent with the existence in the solution of an uncomplicated equilibrium between a monomer and a dimer of the subunit with a dissociation constant of 2×10^{-5} M,¹⁵ that of the subunit of chaperonin GroES is consistent with an uncomplicated equilibrium between monomer and heptamer of the subunit with a dissociation constant of 1×10^{-38} M^{6,27} and that of the subunit of CTP synthase from *E. coli* is consistent with an equilibrium among monomer, dimer, and tetramer of the subunit.²⁸ The dissociation constant ($3 \mu\text{M}$) for the equilibrium between monomers and dimers of DNA helicase II measured by sedimentation equilibrium²⁹ agreed with that ($1.4 \mu\text{M}$) estimated by counting monomers and dimers directly in an atomic force microscope.³⁰

The perturbations of the distributions of concentration at sedimentation equilibrium caused by such equilibria are slight, so a detailed analysis of the deviations of the data from the curve that has been fit to them must be made to insure that those deviations are at random (Figure 8-3) rather than systematic. Better yet, the measurements should be performed at several different concentrations of protein or several different rotor speeds or both to demonstrate that the same value of the measured dissociation constant is consistent with each of the distributions of protein observed under these different conditions.^{31,32}

Sedimentation equilibrium can also be used to rule out the existence of an equilibrium among oligomers and demonstrate that there is only one form of the protein present in the solution³³ or that there are two forms of the protein present with different stoichiometries of subunits that are not in equilibrium with each other and that are distributing independently of each other.³⁴ Again, in order to bolster the conclusion that no equilibration among oligomers is occurring, it should be demonstrated that the calculated molar mass or masses are affected neither by changing the concentration of the protein added to the cell nor by changing the speed of the rotor.^{20,21}

Light scattering is a property of any fluid. It arises because a fluid is a collection of molecules undergoing random movements rather than a rigid solid or a uniform continuum of electrons. Scattered light emerges from a

fluid at all angles to the incident direction of a beam of light passing through the fluid. The source of this scattered light is the electrons in the fluid that oscillate in response to the alternating electric field of the light and in turn emit light. The magnitude of the susceptibility of electrons in a molecule to this phenomenon arises from their respective polarizabilities, which are reflected in the refractive index of the molecule. The scattered light is emitted in directions other than that of the incident light.

The **emission of scattered light** from a solution arises from regional fluctuations in polarizability on a scale smaller than the wavelength of the light. If a fluid were a uniform, unfluctuating distribution of electrons, the scattered light from its constituent electrons would always be canceled by interference and hence no emission would result. The fluctuations producing the net scattering are related to local fluctuations in the concentrations of the components of the solution and hence to the chemical potential of those components. The majority of the electrons in a solution of protein are on molecules of water. Fluctuations in the local concentrations of water are the major contributors to the light scattered from the solution in the absence of the protein. When protein is present, the scattering arising from the molecules of protein in the solution is in addition to this background scattering.

The scattering of a beam of collimated, unpolarized light is measured by placing a detector at an angle θ to the beam of unscattered light passing through the sample (Figure 8-4) and at a distance r (centimeters) from the sample. The **incremental scattering**, i_θ , is the difference in intensity between light scattered by a unit volume of the solution of protein [photons second⁻¹

(centimeter³ of solution)⁻¹] and that scattered by an identical solution (also in photons second⁻¹ centimeter⁻³) not containing protein, and it is reported relative to the intensity of the incident light, I_0 (photons second⁻¹), of which it is a very small fraction. It can be shown^{1,35,36} that

$$\lim_{\theta \rightarrow 0} \frac{r^2 i_\theta}{I_0 (1 + \cos^2 \theta)} = RTKC_{\text{prot}} \left(\frac{\partial \tilde{n}}{\partial C_{\text{prot}}} \right)_{T,P,\mu}^2 \left(\frac{d\Pi}{dC_{\text{prot}}} \right)^{-1} \quad (8-27)$$

where \tilde{n} is the refractive index (dimensionless) of the solution of protein and the optical constant K (moles centimeter⁻⁴) is defined by

$$K \equiv \frac{2\pi^2 \tilde{n}_0^2}{\lambda_0^4 N_A} \quad (8-28)$$

where \tilde{n}_0 is the refractive index of the solution in the absence of the protein, λ_0 is the wavelength (centimeters) of the light in a vacuum, and N_A is Avogadro's number ($6.022 \times 10^{23} \text{ mol}^{-1}$). It should be noted that the units on C_{prot} , the concentration of protein, cancel in Equation 8-27. This cancellation again illustrates that the concentration of protein can be expressed in any units. Equation 8-27 also illustrates that light scattering is a measurement of the partial derivative of the chemical potential and hence the osmotic pressure of the solution of protein. The limit in Equation 8-27 is taken to eliminate any optical interference that might arise if the dimensions of the protein are close to the magnitude of the wavelength of the light.

The partial derivative $(\partial \tilde{n} / \partial C_{\text{prot}})_{T,P,\mu}$ is the change in the refractive index of the solution as only the concentration of protein is increased, at constant chemical potential of the other solutes such as electrolytes and buffers. Each of the solutions of protein used to make the determination of $(\partial \tilde{n} / \partial C_{\text{prot}})_{T,P,\mu}$ as well as the solution used in the determination of the light scattering itself, should be equilibrated by dialysis at the appropriate osmotic pressure against a solution identical except for the protein to obtain a constant chemical potential of the other solutes throughout. A procedure for measuring the required refractive indices at constant chemical potentials of diffusible components has been devised.^{2,37}

At the present time, the light source usually used for measurements of light scattering is a laser; and, if it is not already polarized, the light is passed through a polarizer. The intensity of the scattered light from a source of **polarized light** has a different angular dependence than that from a source of unpolarized light. The oscillating electric vector of the polarized light defines the z axis of a coordinate system with the sample at the origin (Figure 8-4). As defined, the x,y plane is normal to the oscillating electric vector. If ϕ is the angle between the z axis and the ray of scattered light entering the detector, then

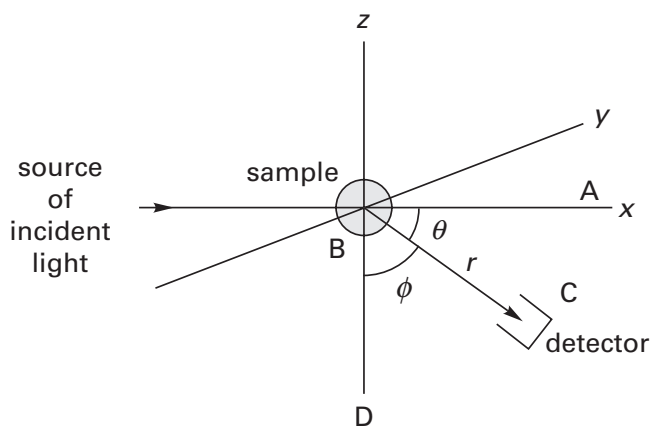


Figure 8-4: Angular dependence of scattered light. The angular dependence of scattered light is related to a coordinate system in which the x axis is along the beam of incident light, the z axis is parallel to the electric vector of the light from a polarized source, and the origin is the center of the sample. The angle θ is the angle ABC where A is a point on the x axis beyond the sample, AB is along the x axis, B is the origin, and C is the position of the detector. The angle ϕ is the angle DBC where D is any point on the z axis, B is the origin, and C is the position of the detector. The distance r is the distance from the origin to the detector.

$$\lim_{\theta \rightarrow 0} \frac{r^2 i_\theta}{I_0 (2 \sin^2 \phi)} = RTKC_{\text{prot}} \left(\frac{\partial \tilde{n}}{\partial C_{\text{prot}}} \right)_{T,P,\mu}^2 \left(\frac{d\Pi}{dC_{\text{prot}}} \right)^{-1} \quad (8-29)$$

where θ is still the angle between the beam of unscattered light emerging from the sample and the ray of scattered light entering the detector. If the detector is confined to the x,y plane (Figure 8-4), the angle ϕ is always 90° and $\sin^2 \phi = 1$. In this configuration, to perform the necessary extrapolation to $\theta = 0$, the angle θ can be varied over all values without changing the angle ϕ . When unpolarized light is used, $\cos^2 \theta$ changes continuously as the limit of $\theta \rightarrow 0$ is taken, and this complicates the extrapolation.

It is convenient to define a quantity, R_θ , known as **Rayleigh's ratio** (centimeters⁻¹), to eliminate the dimensions of the apparatus from the calculation. For unpolarized light

$$R_\theta \equiv \frac{r^2 i_\theta}{I_0 (1 + \cos^2 \theta)} \quad (8-30)$$

and for polarized light when $\phi = 90^\circ$

$$R_\theta \equiv \frac{r^2 i_\theta}{2I_0} \quad (8-31)$$

When Equation 8-27 or 8-29 is combined with Equation 8-30 or 8-31, respectively, as well as Equations 8-18 and 8-19

$$\lim_{\substack{\theta \rightarrow 0 \\ C_{\text{prot}} \rightarrow 0}} R_\theta = KC_{\text{prot}} \left(\frac{\partial \tilde{n}}{\partial C_{\text{prot}}} \right)_{T,P,\mu}^2 W \quad (8-32)$$

The double limit in Equation 8-32 is often taken by a procedure known as the Zimm plot,³⁸ but with proteins of normal dimensions, the variation of R_θ with θ is small and inconsequential, and the extrapolation to $\theta = 0$ is a minor correction and often ignored. The extrapolation to $C_{\text{prot}} = 0$ is always required because the virial coefficients (Equation 8-7) can be appreciable (Figure 8-5).³⁹ To perform this extrapolation, a rearrangement of Equations 8-29 and 8-32, incorporating the virial coefficients explicitly, is performed:

$$\lim_{\theta \rightarrow 0} \frac{C_{\text{prot}}}{R_\theta} = \frac{1}{K} \left(\frac{\partial \tilde{n}}{\partial C_{\text{prot}}} \right)_{T,P,\mu}^{-2} \left(\frac{1}{W} + 2BC_{\text{prot}} + 3CC_{\text{prot}}^2 \dots \right) \quad (8-33)$$

In Figure 8-5, two experiments with bovine serum albumin under different conditions, with and without added electrolyte, are presented. As with the measurements shown in Figure 8-1, it can be seen that the virial coefficient, B , changes appreciably with changes in conditions; in this case it even inverts in sign because the protein is participating in a concentration-dependent oligomerization in the absence of electrolyte. Bovine serum albumin readily self-associates to form adventitious dimers, trimers, and higher oligomers.⁴⁰ The intercepts, however, again remain the same. It has been shown that under the same conditions with the same protein, the same values of the second **virial coefficient** are obtained by measurements of either osmotic pressure or light scattering.⁴¹ This result demonstrates that the virial coefficient is a property of the solution itself rather than the method of measurement, and it provides further evidence that these techniques are both measuring the same property of the solution. From the extrapolations in Figure 8-5, the molar mass of bovine serum albumin was estimated to be $70,200 \text{ g mol}^{-1}$, which is within 6% of the actual value of $66,430 \text{ g mol}^{-1}$.

From an examination of Equation 8-32, it is clear again that the units chosen for C_{prot} , the concentration of protein (units centimeter⁻³), determine the units (units mole⁻¹) of the parameter W . If the concentration of protein is known in grams centimeter⁻³, the units on W will be grams mole⁻¹, or molar mass. The determination of the molar mass, however, will only be as accurate as the measurement of the concentration of protein.

Electrospray mass spectrometry⁴² is widely used to

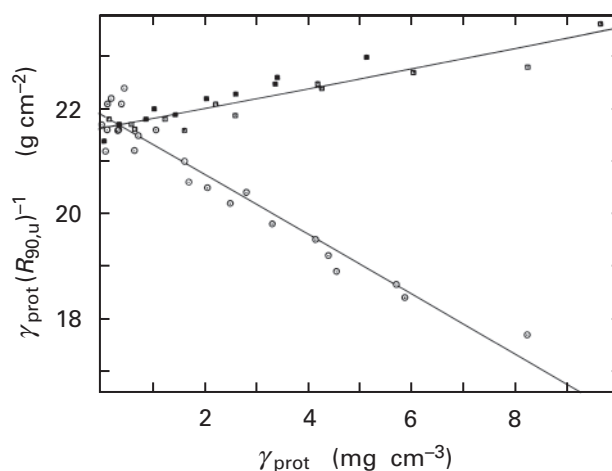


Figure 8-5: Light scattering by solutions of bovine serum albumin.³⁹ The Rayleigh ratio $R_{90,u}$ (centimeters⁻¹) was determined for each of a series of solutions of bovine serum albumin, each at a different concentration (γ_{prot} ; milligrams centimeter⁻³), by measuring the scattering of unpolarized light (u) at an angle θ of 90° . The quotient $\gamma_{\text{prot}}(R_{90,u})^{-1}$ was calculated for each measurement and plotted as a function of the concentration of protein (γ_{prot}). Measurements were made in 0.15 M sodium chloride (upper line) or in water (lower line) of solutions prepared by diluting an isoionic solution of albumin into the appropriate solutions. Adapted with permission from ref 39. Copyright 1954 American Chemical Society.

determine the molecular masses of proteins. Individual vaporized molecules of protein, each molecule with its own particular charge, are submitted to mass spectrometry (Figure 8-6).⁴³ From the envelope of individual peaks, precise estimates of the mass of the molecule of protein can be calculated. For example, it was possible to show that the molecular mass of the blue copper protein rusticyanin from *Thiobacillus ferrooxidans* was 16,552 Da, which is within 1 Da of the mass calculated from its amino acid sequence.⁴⁴

Major **applications** of mass spectrometry are the analysis of posttranslational modification, the verification of the integrity of a preparation of protein, and the assessment of the number of subunits in an oligomer. For example, in the case of rusticyanin the conclusion drawn from the mass spectrometric analysis was not that the molecular mass of the protein was 16,552 Da, which was already known, but that the protein lacked posttranslational modifications. In the case of the L1 metallo- β lactamase of *Stenotrophomonas maltophilia*, mass spectrometric analysis demonstrated that the normal posttranslational removal of the 21 amino-terminal amino acids had occurred,⁴⁵ and in the case of subunit V of ubiquinol-cytochrome-*c* reductase, mass spectrometric analysis demonstrated that the iron-sulfur cluster was

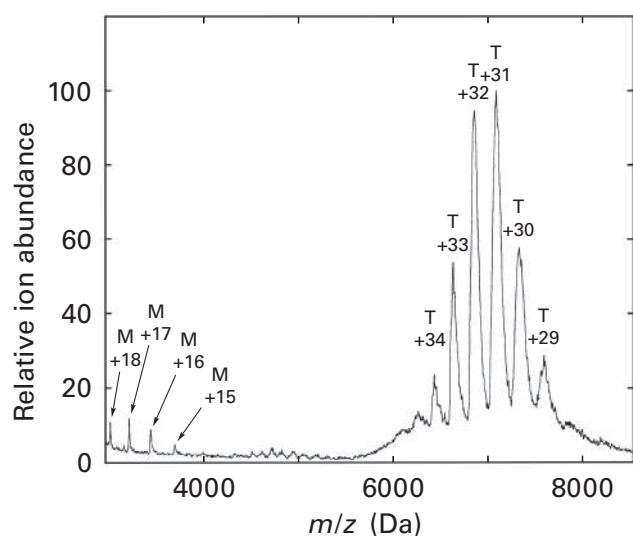


Figure 8-6: Mass spectrum of aldehyde dehydrogenase (NAD) from rat liver.⁴³ The purified protein was vaporized by electrospray, and the resulting gaseous ions were passed into a mass spectrometer. The relative abundance of each ion is plotted as a function of the ratio (m/z) of its molecular mass to its charge. Four ions of the single subunit (54.87 kDa), bearing charges of +15 (m/z = 3658 Da), +16 (m/z = 3430 Da), +17 (m/z = 3228 Da), and +18 (m/z = 3049 Da), are labeled M for monomeric. The six ions labeled T, bearing assigned charges of +29 (m/z = 7568 Da), +30 (m/z = 7316 Da), +31 (m/z = 7080 Da), +32 (m/z = 6859 Da), +33 (m/z = 6651 Da), and +34 (m/z = 6455 Da) were designated as molecular ions of a tetramer of four subunits (219.48 kDa) because the monomer could have produced only ions of +7 (m/z = 7839 Da), +8 (m/z = 6859 Da), and +9 (m/z = 6096 Da) in this range of masses. Because three peaks lie between each of these positions, the species producing all of these peaks must be ions of a tetramer.

present.⁴⁶ Electrospray mass spectrometry was used to verify that each member of a set of site-directed mutants of dethiobiotin synthase from *E. coli* contained the intended amino acid replacement.⁴⁷ Mass spectrometry can also discover errors in a sequence. For example, in the case of subunit I of bovine ubiquinol-cytochrome-*c* reductase, electrospray mass spectrometry indicated correctly that the published amino acid sequence of the protein was missing 27% of its amino acids.⁴⁶ Intact molecules of a protein containing several subunits (Figure 8-6) can be vaporized to obtain the molecular mass of the oligomer,⁴⁸ and membrane-bound proteins can be vaporized after the phospholipid in which they are normally dissolved has been removed from them.⁴⁹

Now that the sequences of so many proteins are known, as well as the stoichiometries of their subunits, precise values of molar mass can be calculated for a large array of proteins from their atomic compositions. These can be compared with values determined by osmotic pressure, sedimentation equilibrium, and light scattering before the actual molar masses were known (Table 8-1). By and large, the agreement between the actual values of molar mass and the measured values is quite close, and this in itself validates the methods.

The problem with molar mass, no matter how accurately it can be determined, is that it means very little to most people. Once it was clear that proteins were polymers of amino acids, sometimes posttranslationally modified, the reason behind all determinations of molar mass has been to estimate the **number of amino acids** that are contained in a given polypeptide and the number of polypeptides that are contained in a given protein. These are quantities that anyone can understand. Unfortunately, the results have seldom been presented in these terms even though they always could have been.

As it happens, the mean molar mass of an amino acid in most proteins is a reasonably constant number, $110 \pm 3 \text{ g mol}^{-1}$ (Table 8-2). When the **mean molar mass of an amino acid** is calculated from the amino acid composition of a set of proteins containing a total of 43,250 amino acids (Table 7-4),⁷² the value obtained is 109.3 g mol^{-1} . It follows that the number of amino acids in most proteins can be estimated by simply dividing an estimate of its molar mass by 110 g mol^{-1} , after glycosylation and other posttranslational modifications are accounted for.

It should not be imagined, however, that the molar mass is a more fundamental number while the number of amino acids in a protein is somehow an approximation. It has already been pointed out that expressing C_{prot} in the units of grams centimeter⁻³ during determinations of molar mass is an arbitrary choice. Were C_{prot} expressed in terms of moles of amino acids centimeter⁻³, each of the procedures would have necessarily provided as accurate a value of W in terms of moles of amino acids (mole of protein)⁻¹ rather than grams (mole of protein)⁻¹. The ability to determine molar masses by physical measure-

Table 8-1: Comparison of the Actual Molar Masses of Selected Proteins and the Molar Masses Determined by Light Scattering, Osmotic Pressure, and Sedimentation Equilibrium

protein	subunit stoichiometry	molar mass (kg mol ⁻¹)			
		actual ^a	light scattering	osmotic pressure	sedimentation equilibrium
bovine serum albumin ^{1,50}	α	66.47	70	69	66
bovine pancreatic ribonuclease ^{1,51,52}	α	13.69			13.7, 13.0
bovine β -lactoglobulin ^{1,51,52}	α_2	36.56	36	35, 39	38
lysozyme from <i>Gallus gallus</i> ^{51,52}	α	14.31	14.8	17.5, 16.6	
porcine L-lactate dehydrogenase ^{50,53}	α_4	145.9	143	146	141
phosphorylase <i>b</i> from <i>Oryctolagus cuniculus</i> ^{54,55}	α_2	194.3			185
mammalian glyceraldehyde-3-phosphate dehydrogenase ^{52,56}	α_4	143	145		
bovine catalase ⁵⁷⁻⁵⁹	α_4	232.5			240
fructose-bisphosphate aldolase from muscle of <i>O. cuniculus</i> ^{60,61}	α_4	156.8			142, 158
mammalian apoferritin ^{62,63}	$\alpha_n\beta_m$ ($m + n = 24$)	510	430		
aspartate carbamoyltransferase from <i>E. coli</i> ^{52,64,65}	$\alpha_6\beta_6$	307.7			310
bovine chymotrypsinogen ^{51,52}	α	25.7		36	
porcine pepsin A ⁵¹	α	34.62		36	39
2-dehydro-3-deoxy-phosphogluconate aldolase from <i>Pseudomonas putida</i> ⁶⁶	α_3	71.81			73
aspartate kinase I-homoserine dehydrogenase I from <i>E. coli</i> ^{67,68}	α_4	356.5	358		360
bovine glutamate dehydrogenase ⁶⁹⁻⁷¹	α_6	333.4	313		320

^aThe actual molar mass of each protein was calculated from the amino acid sequences of its constituent polypeptides and their stoichiometry in the complex.

Table 8-2: Tabulation of the Mean Grams (Mole of Amino Acid)⁻¹ in a Set of Proteins

polypeptide ^a	type of protein	length ^b	total number of amino acids in protein	polypeptide stoichiometry	grams (mole of amino acid) ^{-1 c}
parvalbumin from <i>Gadus callarias</i>	cytoplasmic	113	113	α	107.1
lysozyme from <i>G. gallus</i>	extracytoplasmic, enzymatic	129	129	α	111.0
R17 coat protein	virus coat	129			106.4
human hemoglobin ($\alpha + \beta$)	cytoplasmic	141 + 146	574	$(\alpha\beta)_2$	108.0
bovine chymotrypsinogen	extracytoplasmic, enzymatic	245	245	α	104.8
bacteriorhodopsin from <i>Halobacterium halobium</i>	membrane-spanning	249	747	α_3	108.1
L-lactate dehydrogenase from <i>Squalus acanthias</i>	cytoplasmic, enzymatic	332	1328	$(\alpha_2)_2$	110.1
human immunoglobulin G Eu ($\alpha + \beta$)	extracytoplasmic	446 + 214	1320	$(\alpha\beta)_2$	108.9
human fibrinogen ($\alpha + \beta + \gamma$)	extracytoplasmic, fibrous	831 + 415 + 411	3314	$(\alpha\beta\gamma)_2$	111.8
human serum albumin	extracytoplasmic	585	585	α	113.6
glycogen phosphorylase <i>b</i> from <i>O. cuniculus</i>	cytoplasmic, enzymatic	841	1682	α_2	115.4
murine anion exchanger	membrane-spanning	929	1858	α_2	110.8
ovine Na ⁺ /K ⁺ -exchanging ATPase (α subunit)	membrane-spanning	1016	1319	$\alpha\beta$	110.4
embryonic skeletal myosin (α subunit) from <i>Rattus norvegicus</i>	cytoplasmic, fibrous	1940	4560	$\alpha\beta\gamma\alpha\beta\sigma$	115.4

mean molar mass of an amino acid = 110 ± 3

^aProteins composed of one or more polypeptides, the sequences of which were available in the Swissprot data bank, were chosen as examples. An attempt was made to include examples of all types of proteins, but extremely unusual proteins such as collagen were avoided. The constituent polypeptides the compositions of which were used are indicated. ^bThe lengths of the polypeptides chosen for analysis are presented in numbers of amino acids. ^cCalculated by dividing the molar mass of the protein portion of the polypeptide or polypeptides by the length or combined lengths, respectively.

ments has always relied ultimately upon the ability of the investigator to make an accurate **measurement of dry weight**. All values of molar mass can be traced back to such a determination. The difficulties involved in measurements of dry weight have been noted,¹⁸ and more than anything else, the accuracy of the values in Table 8-1 are a testimony to the careful measurements of this quantity. Accurate dry weight measurement, however, requires more protein than is usually available, and other measures of protein concentration have unfortunately but necessarily supplanted it. Ironically, when the amount of protein is in short supply, the most accurate method for assessing its concentration is quantitative amino acid analysis, which is a measure of moles of amino acids centimeter⁻³, and the use of A_{280} to follow protein in sedimentation equilibrium actually is a measure of the concentration of tyrosine and tryptophan in the solution.

One of the most peculiar manifestations of the abiding infatuation with molecular mass is the habit of naming proteins on the basis of estimates of the molecular mass performed by electrophoresis of their complexes with dodecyl sulfate. For example, protein p27 from simian retrovirus SRV-1 has an actual molecular mass of 24.73 kDa,⁷³ and protein p56 from murine lymphoma has an actual molecular mass of 57.82 kDa.⁷⁴ It is unclear what will be done when two-digit numbers run out.

Part of the description of a particular protein is an enumeration of the length of each of the polypeptides from which it is composed and the number of each subunit that it contains. At one time this information could be most conveniently learned by ascertaining both the molar mass of the entire protein and the molar mass of the isolated individual polypeptides. The history of this quest is interesting but beyond the scope of the present discussion. In two celebrated instances, that of aspartate carbamoyltransferase and that of fructose-bisphosphate aldolase,^{75,76} disagreements arose over the results from such measurements. These particular disagreements coincided with the development of the two techniques that have supplanted almost entirely the earlier methods of determining molar mass that were just described. These newer procedures are both based on the electrophoresis of complexes between polypeptides and dodecyl sulfate upon gels of polyacrylamide. In one procedure, sieving is used to display the different types of polypeptides in the protein and provide estimates of the length of each. In the other procedure, patterns of covalently cross-linked polypeptides separated by electrophoresis are used to count the number of each polypeptide present in the whole protein.

Suggested Reading

Schachman, H.K., & Edelstein, S.J. (1966) Ultracentrifuge studies with absorption optics. IV. Molecular weight determinations at the microgram level, *Biochemistry* 5, 2681-2705.

Becerra, S.P., Kumar, A., Lewis, M.S., Widen, S.G., Abbotts, J., Karawya, E.M., Hughes, S.H., Shiloach, J., Wilson, S.H., & Lewis, M.S. (1991) Protein-protein interactions of HIV-1 reverse transcriptase: implication of central and C-terminal regions in subunit binding, *Biochemistry* 30, 11707-11719.

Problem 8-1: Calculate the molar mass of bovine ribonuclease from its sequence.

Problem 8-2: Calculate the molar masses of these proteins that have the following incremental osmotic pressures at 25 °C.

protein	$\lim_{\gamma_{\text{prot}} \rightarrow 0} \frac{\Pi}{\gamma_{\text{prot}}}$
L-lactate dehydrogenase ⁵³	173 dyne cm ⁻² (g of protein) ⁻¹ L
β -lactoglobulin ⁷⁷	0.722 cm of H ₂ O (g of protein) ⁻¹ *
bovine serum albumin ⁸	3.57 N cm (g of protein) ⁻¹

* These are the centimeters that the level of the solution containing the protein rose above the level of the solution lacking the protein as a result of the expansion of the former at the expense of the latter. This additional layer of solution exerts a pressure because of the force of gravity. The units of centimeters are converted into newtons centimeters⁻² by multiplying by the density of the solution lacking the protein (grams centimeter⁻³) and the gravitational acceleration (980.6 cm s⁻² at sea level, 45° latitude) felt by the excess fluid on top of the solution of protein. The density of the solution has already been converted into the density of water.

Problem 8-3: The isoelectric pH of bovine serum albumin in a solution of 0.15 M NaCl is 5.37.⁷⁸ From the data in Figure 8-1, calculate the mean net charge number \bar{Z}_{SA} on the serum albumin at pH 7.0 by assuming that the difference in slope between the two lines is due entirely to the Donnan effect.

Problem 8-4: The second virial coefficient, B , for the osmotic pressure of lysozyme from *Gallus gallus* in solutions of (NH₄)₄SO₄ varies as a function of pH and ionic strength, I_c .³

pH	I_c (M)	second virial coefficient (cm ³ μ mol g ⁻²)
4	1	-140 \pm 23
7	1	-198 \pm 15
8	1	-307 \pm 21
4	3	-396 \pm 19
7	3	-423 \pm 34
8	3	-446 \pm 26

- What causes the second virial coefficient to become less negative as the pH of the solution is lowered?
- Why is the effect of pH smaller at the higher ionic strength?

- (C) What limit do these measurements place on the isoelectric point of lysozyme?
- (D) What limit do these measurements place on the second virial coefficient at the isoelectric pH for lysozyme?
- (E) Why does the second virial coefficient have the sign that it has at the isoelectric pH of lysozyme?

Problem 8-5: Human hemoglobin is a protein formed from two α polypeptides, two β polypeptides, and four hemes for a total molar mass of $64,450 \text{ g mol}^{-1}$. It is referred to as an $(\alpha\beta)_2$ tetramer to indicate that it is a heterotetramer formed from two $\alpha\beta$ heterodimers. The osmotic pressure at 20°C of a solution of hemoglobin,⁷⁹ at a concentration of 3 g L^{-1} , that had been flushed exhaustively with N_2 gas was $1750 \text{ dyne cm}^{-2}$. When the solution was flushed with O_2 gas, the osmotic pressure increased to $2500 \text{ dyne cm}^{-2}$ even though the concentration of the hemoglobin was unchanged.

- (A) Assume ideal behavior and that the virial coefficients are 0 and calculate the average molar mass for each circumstance.
- (B) Explain the values that you obtain.

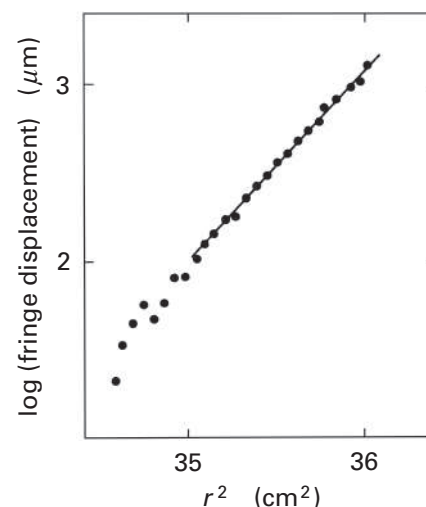
Problem 8-6: The enzyme glutamate–ammonia ligase, which catalyzes the ATP-dependent condensation of glutamate and ammonia, was isolated from *E. coli*. It was purified by $(\text{NH}_4)_2\text{SO}_4$ fractionation followed by ion-exchange chromatography. The final preparation of the enzyme was considered to be a pure protein because a single peak was obtained on repeated ion-exchange chromatography. The protein is a complex of 12 identical polypeptides each 468 amino acids in length. The length of its constituent polypeptides has been ascertained by sequencing its genomic DNA.

Glutamate–ammonia ligase was dissolved in 6 M guanidinium chloride, and the osmotic pressure of this solution was determined with a high-speed osmometer at 20°C . The following results were obtained

protein concentration (g L^{-1})	pressure (mmHg)
2	0.69
4	1.40
8	2.87
10	3.63
15	4.73
20	7.96

- (A) What is the molar mass of the protein in 6 M guanidinium chloride?
- (B) What has this salting-in solute done to the protein?

Problem 8-7: Calculate the molar mass of aspartate kinase I–homoserine dehydrogenase I from the data in this figure (adapted with permission from ref 80; copyright 1968 Springer-Verlag).



Sedimentation equilibrium of aspartate kinase I–homoserine dehydrogenase I from *E. coli*. The initial protein concentration was 0.6 mg mL^{-1} . Interference optics were used. In an interference scan, the logarithm of the distance in micrometers of the fringe displacement between blank corrected fringes at each point in the chamber and blank corrected fringes at zero level concentration is plotted against r^2 , the square of distance (centimeters²) from the center of the rotor to the point at which the measurement was made. It has been assumed that the concentration of the protein was zero at the top of the sample so that the fringe displacement (in micrometers) is directly proportional to the concentration of protein. Because the concentration of protein at the top of the sample is so small, the leftmost points are scattered.

Use the approximation of Equation 8-21, and assume $\rho_{\text{sol}} = \rho_{\text{H}_2\text{O}}$. The partial specific volume of the protein is $0.737 \text{ cm}^3 \text{ g}^{-1}$, the rotor was spinning at $11,272 \text{ revolutions min}^{-1}$ (1 revolution is 2π radians), and the temperature was 23°C .

Problem 8-8: The light scattering from a series of solutions of ovalbumin from *G. gallus* was measured at an angle θ of 90° .⁸¹ The apparatus sampled the light scattered from a volume of solution of 1.8 cm^3 with a photomultiplier tube at 10 cm from the center of the sample. The differences between the scattering of the buffer alone and the scattering of each solution were used to calculate the Rayleigh ratio, R_{90} , for each solution. The intensity of the incremental scattered light, i_θ , was less than 0.001% of the intensity of the incident light I_0 . The Rayleigh ratios for light of wavelength 436 nm in the vacuum were as follows:

$\gamma_{\text{prot}} (\text{g cm}^{-3})$	$R_{90} (\text{cm}^{-1})$
4.3×10^{-3}	1.13×10^{-4}
5.8×10^{-3}	1.52×10^{-4}
8.6×10^{-3}	2.22×10^{-4}
9.7×10^{-3}	2.54×10^{-4}
13.7×10^{-3}	3.66×10^{-4}

The refractive index \tilde{n}_0 of the solution without the protein was 1.333, the increment of the refractive index with concentration $[(\partial\tilde{n}/\partial\gamma_{\text{prot}})_{T,P,\mu}]$ for the protein at $\lambda_0 = 436 \text{ nm}$ was $0.1883 \text{ cm}^3 \text{ g}^{-1}$, and the temperature was 25°C .

- (A) Determine graphically the limit

$$\lim_{\gamma_{\text{prot}} \rightarrow 0} \frac{\gamma_{\text{prot}}}{R_{90}}$$

- (B) Assume that the Rayleigh ratio, R_θ , does not vary significantly with variation in θ for this small protein at this long wavelength and that

$$\lim_{\substack{\theta \rightarrow 0 \\ \gamma_{\text{prot}} \rightarrow 0}} \frac{\gamma_{\text{prot}}}{R_\theta} = \lim_{\gamma_{\text{prot}} \rightarrow 0} \frac{\gamma_{\text{prot}}}{R_{90}}$$

and use the value for this limit to estimate the molar mass of ovalbumin.

Problem 8–9: Calculate the mean molar mass of an amino acid from the amino acid composition of the set of proteins in Table 7–4.

Electrophoresis on Gels of Polyacrylamide Cast in Solutions of Dodecyl Sulfate

The sodium salt of dodecyl sulfate ($\text{H}_{25}\text{C}_{12}\text{OSO}_3^-$) is a detergent widely used commercially to dissolve nonpolar substances in water. It accomplishes this purpose by forming micelles. A **micelle of dodecyl sulfate** at moderate ionic strength (0.2 M) contains about 100 of the anions in an oblate ellipsoid that is 3 nm across at its minor axis.^{82,83} All of the anionic sulfonates are at the surface of the ellipsoid and the hydrocarbon is in the center. It is the hydrocarbon core of the micelle that dissolves individual molecules of a nonpolar substance, producing the detergent properties. Although dodecyl sulfate must be present at concentrations high enough to form micelles in order to interact with proteins, the complexes that result between the anions of dodecyl sulfate and polypeptides do not seem to involve discrete micelles.

When sodium dodecyl sulfate⁸⁴ is added to a solution of protein at a concentration greater than its critical micelle concentration* and at ratio greater than 2 g of dodecyl sulfate (g of protein)^{–1}, all of the polypeptides present in the solution are unfolded and separate from each other as they become coated with the dodecyl sulfate. The amount of dodecyl sulfate coating the **unfolded, separated polypeptides** at saturation is usually a function only of

their total length. Pitt-Rivers and Impiombato⁸⁵ observed that, within a series of globular, water-soluble proteins, each of the constituent polypeptides would bind 0.54 ± 0.01 molecule of dodecyl sulfate for every amino acid in its sequence. The important point, however, is not the numerical value of this ratio but the fact that it is constant (less than 2% variation) regardless of the protein examined, as long as it is of the usual water-soluble, globular variety and does not have significant segments of its sequence enriched in acidic amino acids and lacking basic amino acids.⁸⁶ This regularity in the binding of dodecyl sulfate, however, is observed only when all of the cystines in the proteins, if there were any, have been cleaved.⁸⁵ Usually this is done by disulfide interchange with a small thiol (Figure 3–20). The constant ratio between bound dodecyl sulfate and the number of amino acids presumably results from the fact that proteins displaying this behavior all have similar compositions of amino acids. Proteins with peculiar compositions,⁸⁷ an excess of charged side chains,⁸⁶ or an excess of hydrophobic side chains⁸⁸ behave anomalously.

The complexes that form between dodecyl sulfate and polypeptides are extended structures and have been variously described as cylindrical rods the length of which is directly proportional to the length of the polypeptide⁸⁹ or micellar pearls of dodecyl sulfate on a string of the flexible polypeptide.⁹⁰ No definitive description of their structure is available, but there is no evidence that the dodecyl sulfate in these complexes is present in discrete packets of 100 molecules of detergent, as would be expected if the micelles present in the absence of the protein were simply incorporated intact into a long string upon the unfolded polypeptide.

As with nucleic acids and presumably for the same reasons, the complexes between dodecyl sulfate and those polypeptides that bind a constant ratio of this strongly anionic detergent all display the same **free electrophoretic mobility**, $(-2.62 \pm 0.04) \times 10^{-4} \text{ cm}^2 \text{ s}^{-1} \text{ V}^{-1}$, regardless of the length of the polypeptide.⁹⁰ In the case of nucleic acids, the invariance of the free electrophoretic mobility with length results from the uniform distribution of negative charge along the regular polymer, and presumably this is also a necessary condition met by the complexes between dodecyl sulfate and polypeptides. With nucleic acids, however, this is a covalently conferred, intrinsic property of the phosphodiester of the backbone rather than the fortuitous and less reliable inclination of the polymer to bind a charge-conferring species uniformly along its length. As such, any polypeptide that binds dodecyl sulfate abnormally should have a different free electrophoretic mobility. When the amount of dodecyl sulfate bound to a series of polypeptides was purposely decreased, the magnitudes of their free electrophoretic mobilities also decreased.⁹⁰

As is the case with nucleic acids,⁹¹ native proteins (Figure 1–17), and other macromolecules submitted to electrophoresis on gels of polyacrylamide or other poly-

* The critical micelle concentration is the minimum concentration at which the detergent forms micelles.

meric supports, the electrophoretic mobilities of complexes between dodecyl sulfate and polypeptides (Figure 8-7)⁹² follow the relationship

$$u_i = u_i^\circ \exp(-K_{r,i} T_a) \quad (8-34)$$

where T_a is the concentration of acrylamide (in percent) from which the gel was cast and the **retardation coefficient**, $K_{r,i}$ is a constant unique to the particular polypeptide i . Because u_i° is the free electrophoretic mobility of the complex between dodecyl sulfate and polypeptide i and u° is the same for all complexes between dodecyl sulfate and well-behaved polypeptides, this relationship predicts that the lines in Figure 8-7 should intersect at the axis of the ordinate when T_a is equal to 0, which is almost the case. Because each complex has a unique retardation coefficient, electrophoresis on gels of polyacrylamide can be used to separate these complexes one from the other (Figure 8-8).⁹³

Systems for **stacking** complexes between dodecyl sulfate and polypeptides have been developed to improve the resolution of the separation. One system

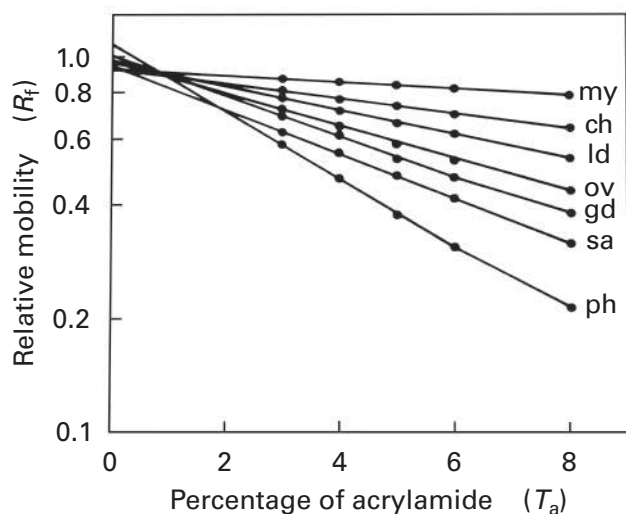


Figure 8-7: Relative electrophoretic mobilities, R_f , of complexes of polypeptides and dodecyl sulfate as a function of the concentration of acrylamide, T_a , used to cast the gel.⁹² Various proteins—myoglobin (my; 153 aa), chymotrypsinogen (ch; 245 aa), L-lactate dehydrogenase (ld; 331 aa), ovalbumin (ov; 385 aa), glutamate dehydrogenase (gd; 501 aa), bovine serum albumin (sa; 583 aa), and phosphorylase (ph; 842 aa)—were dissolved separately in a solution containing a concentration of dodecyl sulfate sufficient to saturate the polypeptides. Each was then submitted to electrophoresis on gels of polyacrylamide cast in a solution of 1% sodium dodecyl sulfate. A series of gels was used for each protein that differed in the percent acrylamide (T_a) from which they were cast. The gels were stained for protein, and the distance migrated by the protein was divided by the distance migrated by a dye, Pyronine-Y, of low molecular weight to obtain the relative electrophoretic mobility (R_f) of each polypeptide at each percent acrylamide. The assumption made was that the mobility of the Pyronine-Y would be unaffected by the percent acrylamide. Reprinted with permission from ref 92. Copyright 1972 *Journal of Biological Chemistry*.

releases the complexes from the descending boundary in which they are stacked by using an ascending boundary that increases the concentration of the neutral conjugate base of the cationic acid that is common to all of the solutions.⁹⁴⁻⁹⁶ The other releases them by using an ascending boundary that delivers the neutral conjugate base of a different cationic acid of much higher pK_a to jump the pH after the complexes have stacked.⁹⁷ The latter system is more effective at releasing the smaller polypeptides from the descending boundary than is the former. The former relies heavily on the increase in the concentration of polyacrylamide at the top of the running gel to accomplish the release and fails to do so when the concentration of polyacrylamide in the running gel is decreased below a certain level.

As with nucleic acids, the electrophoretic mobilities of complexes between dodecyl sulfate and polypeptides on gels of polyacrylamide are a regular function of the

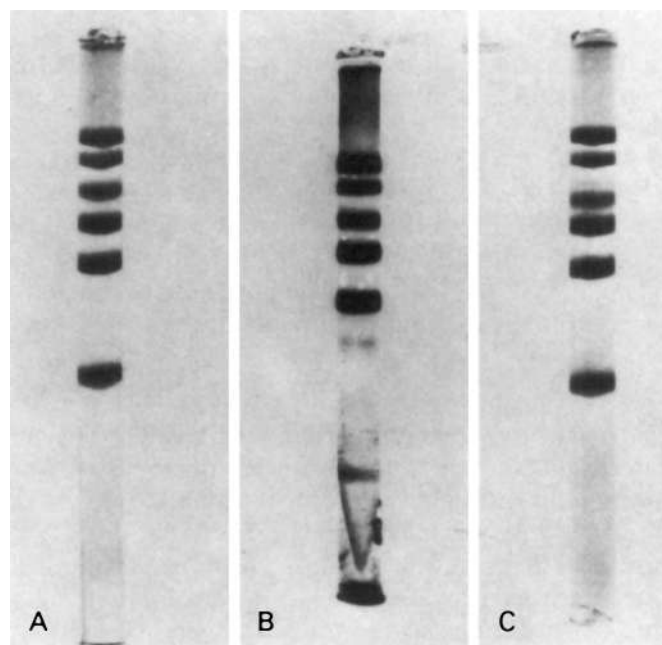


Figure 8-8: Separation of polypeptides by electrophoresis on gels of polyacrylamide cast in a solution of 0.2% sodium dodecyl sulfate.⁹³ Proteins containing the polypeptides were dissolved in solutions of sodium dodecyl sulfate sufficient to saturate them. They were submitted to electrophoresis on cylindrical gels (0.6 cm × 10 cm) cast from 10% acrylamide in 0.1% sodium dodecyl sulfate and 0.1 M sodium phosphate, pH 7.0. Following electrophoresis, the gels were stained for protein. The polypeptides that were run on gel A were those composing bovine catalase ($n_{aa} = 506$), the mitochondrial isoform of porcine fumarate hydratase ($n_{aa} = 466$), isoform A of fructose-bisphosphate aldolase from muscle of *Oryctolagus cuniculus* ($n_{aa} = 361$), glyceraldehyde-3-phosphate dehydrogenase from muscle of *O. cuniculus* ($n_{aa} = 332$), human carbonate dehydratase I ($n_{aa} = 260$), and equine cardiac myoglobin ($n_{aa} = 153$). The polypeptides run on gel B were the same as those run on gel A, but the myoglobin was omitted. The polypeptides run on gel C were catalase, fumarate hydratase, the E isoform of alcohol dehydrogenase from equine liver ($n_{aa} = 374$), glyceraldehyde-3-phosphate dehydrogenase, carbonate dehydratase, and myoglobin. Reprinted with permission from ref 93. Copyright 1969 *Journal of Biological Chemistry*.

length of the polypeptides, as long as they have a normal composition of amino acids⁸⁷ and bind the proper amount of dodecyl sulfate. To understand this property of the electrophoretic separations, the process known as sieving must be understood.

Suggested Reading

Weber, K., & Osborne, M. (1969) The reliability of molecular weight determinations by dodecyl sulfate–polyacrylamide gel electrophoresis, *J. Biol. Chem.* 244, 4406–4412.

Sieving

Sieving of macromolecules, for example, native proteins, nucleic acids, or complexes between dodecyl sulfate and polypeptides, occurs during both chromatography by molecular exclusion and electrophoresis on polymeric supports. **Sieving** is the discrimination between macromolecules on the basis of size that is accomplished by a random network of linear polymers. In chromatography by molecular exclusion, the network of polymer forms the beads among which the mobile phase percolates and is the sieve within the beads through which the macromolecule diffuses when it is inside of the stationary phase. In electrophoresis, the network of polymer forms an obstacle course through which the macromolecule must pass as it moves in the direction of the electric field.

Consider a geometric solid of any shape within a network of lines thrown at random through a volume of space completely containing the solid. An equation⁹⁸ for the probability that none of these lines intersects the solid, $P(\text{ni})$, was derived during the solution of an unrelated topological problem,⁹⁹ and

$$P(\text{ni}) = \exp(-lS/4) \quad (8-35)$$

where l is the density of the lines (centimeters centimeter⁻³) and S is the surface area of the solid (centimeters²). Assume that a macromolecule is a geometric solid and a network of chemical polymers is a network of lines. When a molecule of protein is submitted to **chromatography by molecular exclusion**, the fraction of the total volume available to macromolecule i , $K_{\text{av},i}$ (Equation 1-21), in the stationary phase of randomly arranged linear polymers should be that fraction of the total volume the occupation of which by the macromolecule does not cause any polymer to intersect the macromolecule. In this case,

$$K_{\text{av},i} = \exp(-bT_p S_{\text{app},i}) \quad (8-36)$$

where T_p is the concentration of polymer in percent [grams (100 cubic centimeters)⁻¹], b is a constant of proportionality to convert, among its other roles, the con-

centration of polymer [grams (100 cubic centimeters)⁻¹] into its linear density (centimeters centimeter⁻³), and $S_{\text{app},i}$ is the apparent surface area (centimeters²) of the macromolecule i .

Because the polymers are not lines but solids themselves, the **apparent surface area** of the macromolecule is not its real surface area. The apparent surface of macromolecule i , $S_{\text{app},i}$, lies outside its actual surface by a distance equal to the sum of the widths of any tight shells of hydration around either the macromolecule or the polymer and the width of the polymer itself. All of the dimensions that cause the polymer not to be a line and the macromolecule not to be a dry smooth solid object are incorporated into the dimensions of an apparent macromolecule that is larger than the actual macromolecule. When the actual macromolecule collides with the actual polymer, the apparent macromolecule collides with a line in the center of the polymer.

This model predicts that if a series of beaded stationary phases of increasing concentration of polymer is used to separate the same set of standard macromolecules by molecular exclusion chromatography, then

$$K_{\text{av},i} = \exp(-K_{r,i} T_p) \quad (8-37)$$

If this is so, then $\ln K_{\text{av},i}$ should be a linear function of T_p . It has been demonstrated that the relationship of Equation 8-37 describes the behavior of both proteins^{98,100,101} and polysaccharides¹⁰¹ during chromatography by molecular exclusion on gels of both polyacrylamide (Figure 8-9)^{98,100,102} and linear dextrans.¹⁰¹

If Equation 8-36 describes behavior during chromatography by molecular exclusion, then when T_p is fixed, $-\ln K_{\text{av},i}$ should be directly proportional to $S_{\text{app},i}$. Computer programs exist for calculating the accessible surface area of a molecule of a protein (6-12) by rolling a spherical probe over the surface of its crystallographic molecular model.¹⁰³ The accessible surface area is the surface area traced by the center of the probe. Unfortunately, there is no computer program that performs such a calculation for a cylindrical probe, which would not detect the smaller irregularities of the surface so readily as does a spherical probe (Figure 6-20). One can choose, however, a radius for the spherical probe that is large enough to include the radius of the polymer and the layers of hydration on the polymer and the protein as well as being large enough that the smaller irregularities of the surface of the protein that would not be detected by a cylinder are also not detected by the sphere. When a sphere of the appropriate radius is used as the probe, the values of $-\ln K_{\text{av},i}$ for a series of standard proteins that have been submitted to chromatography by molecular exclusion¹⁰⁴ upon cross-linked dextran are found to be directly proportional to the accessible surface areas¹⁰³ calculated from crystallographic molecular models of those same proteins (Figure 8-10A).

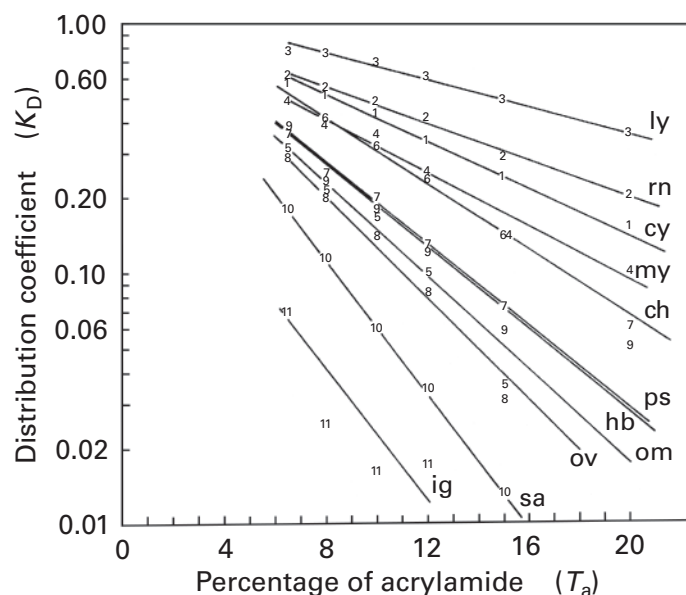


Figure 8-9: Distribution coefficients for a series of globular proteins submitted to chromatography by molecular exclusion on polyacrylamide gels of varying composition.^{98,100} A series of gels cast from different concentrations of acrylamide, T_a (percent), were separately fragmented to form suspensions of polyacrylamide granules of different porosities. Columns were made from these chromatographic media, and a set of standard globular proteins were submitted to chromatography by molecular exclusion on these columns and their respective elution volumes were used to calculate the respective distribution coefficients, K_D (Equation 1-22). The distribution coefficient K_D is directly proportional to the distribution coefficient K_{av} . The values of K_D are plotted on a logarithmic scale as a function of T_a . The proteins were, in order of their number of amino acids, (1) cytochrome c (cy; 104 aa), (2) ribonuclease (rn; 124 aa), (3) lysozyme (ly; 129 aa), (4) myoglobin (my; 153 aa), (5) ovomucoid (om; 186 aa), (6) chymotrypsinogen (ch; 245 aa), (7) pepsin A (ps; 327 aa), (8) ovalbumin (ov; 385 aa), (9) hemoglobin (hb; 574 aa), (10) serum albumin (sa; 583 aa), and (11) immunoglobulin G (ig; 1320 aa). Adapted with permission from ref 98. Copyright 1970 National Academy of Sciences.

Unfortunately, the accessible surface area of a molecule of protein is not one of its more interesting properties; usually sieving is used to estimate the **number of amino acids** a protein contains. It has been noted by Ogston¹⁰⁵ that if a set of macromolecules were all spheres of radius R_i and the polymers of the network were infinitely long right cylinders of radius r_p , then

$$S_{app,i} = 4\pi(r_p + R_i)^2 \quad (8-38)$$

Because the partial molar volume of a molecule of protein is a function only of its composition of amino acids^{18,106} and because the amino acid compositions of most proteins are similar, each of their partial molar volumes should be directly proportional to the number of amino acids each protein contains (Table 8-2). To the extent that a molecule of protein i is a sphere and has the normal composition of amino acids, the number of amino acids it contains, $n_{aa,i}$, should determine its radius by the relationship

$$R_i^o = \left[\frac{3}{4\pi} \left(\frac{n_{aa,i} \bar{V}_{aa}}{N_A} \right) \right]^{1/3} \quad (8-39)$$

where \bar{V}_{aa} is the mean partial molar volume of the amino acids in the usual protein ($82 \text{ cm}^3 \text{ mol}^{-1}$) and the superscript has been added to R to indicate that this is a sphere equivalent in volume to the volume of the protein, which is never exactly a sphere.

When Equations 8-36, 8-38, and 8-39 are combined

$$\left(\frac{-\ln K_{av,i}}{4\pi b T_p} \right)^{1/2} = r_p + \left[\frac{3}{4\pi} \left(\frac{n_{aa,i} \bar{V}_{aa}}{N_A} \right) \right]^{1/3} \quad (8-40)$$

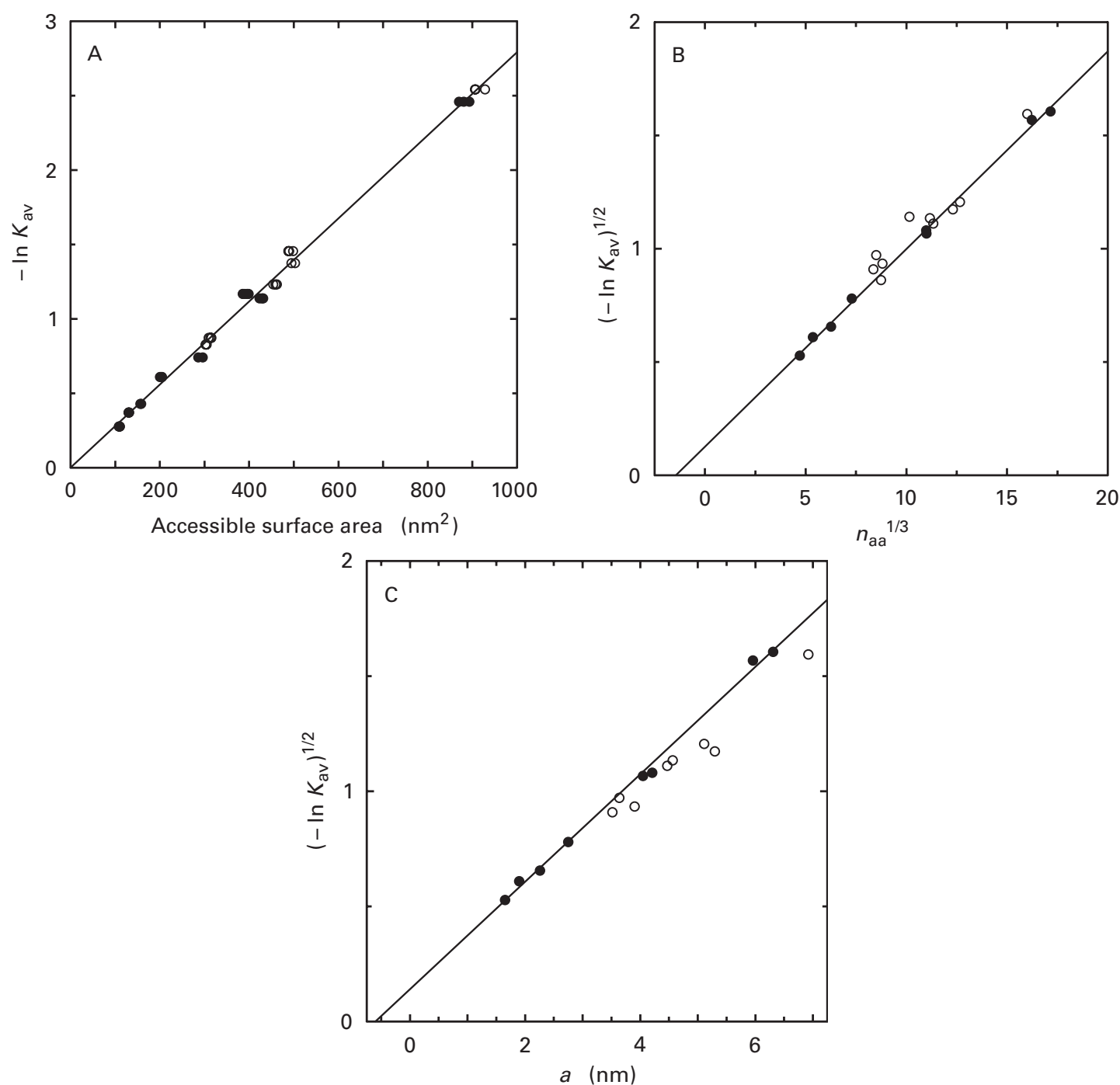
Figure 8-10: Sieving of globular proteins by molecular exclusion chromatography.¹⁰⁴ The proteins, dissolved in 0.1 M KCl at pH 7.5, were submitted to chromatography by molecular exclusion on a column (2.5 cm × 50 cm) of Sephadex G-200. The volumes at which the several proteins eluted from the column were tabulated. The distribution coefficient K_{av} for each was calculated (Equations 1-20 and 1-21) from its elution volume, V_e , and the void volume of the column, V_0 , and the included volume of the column, V_i (as determined by the volume at which sucrose eluted). It was assumed that $V_i = V_{H_2O}$, that $V_T = (1 - f_{poly})^{-1} V_{H_2O}$, and that $W_r = 20 \text{ mL g}^{-1}$. The proteins used by Andrews¹⁰⁴ were, in order of increasing total number of amino acids, equine cytochrome c ($n_{aa} = 104$), myoglobin from *Physeter catodon* ($n_{aa} = 153$), bovine chymotrypsinogen ($n_{aa} = 245$), ovalbumin from *G. gallus* ($n_{aa} = 385$), bovine serum albumin ($n_{aa} = 583$), bovine lactoperoxidase ($n_{aa} = 612$), the cytoplasmic isoform of malate dehydrogenase from porcine heart ($n_{aa} = 666$), bovine transferrin ($n_{aa} = 685$), glyceraldehyde-3-phosphate dehydrogenase from muscle of *O. cuniculus* ($n_{aa} = 1328$), the A isoform of L-lactate dehydrogenase from muscle of *O. cuniculus* ($n_{aa} = 1324$), alcohol dehydrogenase from *Saccharomyces cerevisiae* ($n_{aa} = 1388$), the A isoform of fructose-bisphosphate aldolase from muscle of *O. cuniculus* ($n_{aa} = 1452$), the mitochondrial isoform of porcine fumarate hydratase ($n_{aa} = 1864$), bovine catalase ($n_{aa} = 2024$), β -galactosidase from *E. coli* ($n_{aa} = 4092$), equine apoferritin ($n_{aa} = 4368$), and urease from *Canavalia ensiformis* ($n_{aa} = 5040$). (A) The quantity $-\ln K_{av}$ is plotted as a function of the accessible surface area (6-12; nanometers²) of those molecules of protein for which crystallographic molecular models were available (all of the proteins used by Andrews except lactoperoxidase, alcohol dehydrogenase, and urease). The accessible surface areas were calculated with a spherical probe of radius 1.1 nm (Figure 6-20) by use of the program of Lee and Richards¹⁰³ as adapted by Dr Ilya Shindyalov of the Protein Data Bank. Of necessity, crystallographic molecular models of proteins from species other than the species providing the protein for the molecular exclusion often had to be used. The access codes of the crystallographic molecular models in the Protein Data Bank that were chosen are 5CYT, 3CYT, 1CYC, 1CRC, 1HRC, 1ABS, 1DXD, 1HJT, 1SWM, 2MBW, 1MCY, 2CGA, 4CHA, 1OVA, 1BJ5, 1UOR, 1BKE, 1AO6, 4MDH, 5MDH, 1DOT, 1OVT, 1CB6, 1CE2, 1GPD, 4GPD, 3GPD, 1LDM, 2LDX, 3LDH, 5LDH, 9LDB, 9LDT, 6ALD, 4ALD, 2ALD, 1FUR, 1YFM, 1DGF, 1DGG, 4BLC, 7CAT, 8CAT, 1BGL, 1BGM, 1FHA, 1AEW, and 1DAT. The closed circles are those for proteins the frictional ratios of which are less than 1.20. (B) The quantity $(-\ln K_{av})^{1/2}$ is plotted as a function of the cube root of the number of amino acids in each protein. (C) The quantity $(-\ln K_{av})^{1/2}$ is plotted as a function of the Stokes radius, a , of each protein calculated from its diffusion coefficient by Equation 1-67. In panel A the line was fit to the averages of the surface areas for each protein, but in panels B and C it was fit only to the eight points (closed circles) for the proteins the frictional ratios of which are less than 1.2.

and a plot of $(-\ln K_{av,i})^{1/2}$ against $(n_{aa,i})^{1/3}$ should be a linear relationship. When the data of Andrews¹⁰⁴ are displayed in this fashion,⁹⁸ they are linearly related (Figure 8-10B). The intercept of the line with the abscissa is at a negative value as predicted by Equation 8-40, and this intercept yields an estimate of r_p , the mean radius of the polymers of dextran, of 0.46 nm, which is not unreasonable when it is considered that this parameter also includes the irregularities of the surface of the protein and the hydration of the molecule of protein and the molecules of polymer.

Several of the points in Figure 8-10B deviate from

the line that was drawn. One way to quantify the deviation of a molecule of protein from spherical behavior is to define a **frictional ratio**, $f(f_0)^{-1}$, which is simply the quotient between the measured frictional coefficient of the protein and the frictional coefficient it would have if its mass were distributed to form a hard sphere of the same partial specific volume. The measured frictional coefficient, f , is calculated from the diffusion coefficient (Equation 1-64) and the ideal frictional coefficient (Equation 1-66) is defined by

$$f_0 = 6\pi\eta R^0 \quad (8-41)$$



where R° is defined by Equation 8-39. For the most spherical of proteins, values of the frictional ratio of 1.1–1.2 are observed.⁵¹ The values of the frictional ratio are always greater than 1 because the water bound to a protein and the irregularities of its surface increase its actual frictional coefficient.

The solid circles in Figure 8-10B are those for all of the proteins chosen by Andrews that happen to have frictional ratios less than 1.2. They are, in ascending order of size (with the frictional ratios in parentheses), cytochrome *c* (1.09), myoglobin (1.16), chymotrypsinogen (1.12), ovalbumin (1.18), glyceraldehyde-3-phosphate dehydrogenase (phosphorylating) (1.16), L-lactate dehydrogenase (1.17), apoferritin (1.15), and urease (1.18).^{*} The line in Figure 8-10B was drawn through these points because they should be the most ideal examples. It can be seen that several of the points for proteins with larger frictional ratios indicate that they are behaving as if they were larger than they are, which makes sense if their behavior is a function only of their surface area. Further validating the conclusion that it is only the surface area of a molecule of protein that determines its distribution coefficient is the fact that the distribution coefficients of proteins with larger frictional ratios (open circles in Figure 8-10A) show no more deviation from linear behavior than do those of the proteins with frictional ratios less than 1.20 (solid circles in Figure 8-10A) when they are plotted as a function of surface area rather than volume.

It has been argued^{101,108,109} that, rather than the apparent surface area of a molecule of protein, the fundamental variable in describing its behavior when it is submitted to sieving on chromatography by molecular exclusion is its effective radius, or **Stokes radius**, a , calculated from its diffusion coefficient (Equation 1-67). When the same values of $(-\ln K_{av})^{1/2}$ displayed in Figure 8-10B are replotted against the effective radii for the various proteins (Figure 8-10C), no significant improvement is seen. Something can be learned, however, when a line is again drawn through the points for the eight most globular proteins listed above, the properties of which should be least affected by the change to effective radius from $n_{aa}^{1/3}$. It can be seen that the use of the effective radius significantly overcompensates for the deviations from the linear behavior displayed in Figure 8-10B for almost all of the proteins that have larger frictional ratios. It is hard to explain why proteins with irregular shapes would behave as if they were smaller than more spherical ones.

It is the correlation between K_{av} and n_{aa} that is exploited when data from chromatography by molecular exclusion are used to estimate the number of amino

acids contained within a protein of interest.¹⁰⁴ This estimation requires that the distribution coefficients, K_{av} , for a series of uncomplicated **standard proteins** of known number of amino acids be used to define the line for the chromatographic system chosen for the particular experiment. The estimate for the number of amino acids in the protein of interest is interpolated from the known values for the standards. A standard line for a particular chromatographic column must be established by running standards on that column, because the properties of each commercial batch of chromatographic medium are unique.¹⁰¹ It is also important to run the chromatographic system with a buffer of ionic strength 0.1–0.2 M to eliminate the effect of the dimensions of the ionic double layer around the charged macromolecules on the parameter K_{av} .¹¹⁰

As a macromolecule moves through a polymeric network during **electrophoresis**, it is also being sieved. In this case, it must travel through the network in a kinetic process, rather than equilibrating with the internal volume of a bead, but it appears that this distinction is inconsequential. It has been argued⁹⁸ that one can view the solid matrix of the polymerized gel as an array of screens through which the macromolecule must travel. A random cross section through a random three-dimensional network of lines will provide a distribution of points. The probability that none of these points lies within the randomly placed, random cross section of a geometric solid of any shape is still described by Equation 8-35.⁹⁸ If those points represent one of the screens in the gel, if the macromolecule can pass only through openings in that screen large enough so that no point forming that screen is found within the cross section of the macromolecule, and if the rate of its movement through the screen is proportional to the probability that openings of the proper size or larger will be encountered, then the mobility of a macromolecule through a gel during electrophoresis should be described by

$$u_i = u_i^\circ \exp(-b T_P S_{app,i}) \quad (8-42)$$

It has already been noted that the electrophoretic mobilities of proteins (Figure 1-17), nucleic acids,⁹¹ and complexes between dodecyl sulfate and polypeptides (Figure 8-7) satisfy Equation 8-34, and therefore their behavior is also consistent with Equation 8-42. It should also be the case that

$$K_{r,i} = b S_{app,i} \quad (8-43)$$

The values for the **retardation coefficients** K_r for a series of standard proteins in their native conformations submitted to electrophoresis on a series of polyacrylamide gels cast from increasing concentrations of acrylamide (Figure 1-17) are directly proportional to their accessible

* Values of frictional ratios were calculated from diffusion coefficients cited by Andrews¹⁰⁴ or in the tables of the *CRC Handbook of Biochemistry*¹⁰⁷ and the actual number of amino acids in each protein.

surface areas calculated from crystallographic molecular models of the same proteins (Figure 8–11A).^{40,111}

As already noted, however, the property of a molecule of protein that is usually of interest is not its accessible surface area but the **number of amino acids** it contains. If it is assumed that a series of proteins resembles a series of spheres and that Equations 8–38 and 8–39 are still valid approximations, then

$$\left(\frac{K_{r,i}}{4\pi b}\right)^{1/2} = r_p + R_i^{\circ} \quad (8-44)$$

and a plot of $(K_r)^{1/2}$ against $(n_{aa})^{1/3}$ for this series should yield a linear relationship.⁹⁸ When the data of Hedrick and Smith¹¹¹ and Bryan⁴⁰ for the retardation coefficients, $K_{r,i}$, of a series of native proteins that had been sieved by electrophoresis through gels cast from increasing concentrations of acrylamide, T_a , are plotted against the cube root of the number of amino acids they contain,

they do display linear behavior (Figure 8–11B).^{40,98,111} Again, the intercept with the abscissa is at a negative value as predicted by Equation 8–44, and this intercept yields a value of r_p , the mean radius of the polyacrylamide, of 0.9 nm.⁹⁸ Unlike the behavior of globular proteins on chromatography by molecular exclusion (Figure 8–10B), the proteins with higher frictional ratios (open symbols in Figure 8–11B) do not deviate systematically from linear behavior more significantly than those with frictional coefficients less than 1.20 (closed symbols in Figure 8–11B).

It has been suggested by Hedrick and Smith¹¹¹ that this linear correlation permits the number of amino acids in a native protein of unknown size to be estimated from its behavior on electrophoresis. This is particularly useful in situations in which only the electrophoretic mobility of the protein of interest can be measured.

Two types of macromolecules that are of interest in biochemistry are globular macromolecules, such as many of the proteins in their native state, and extended

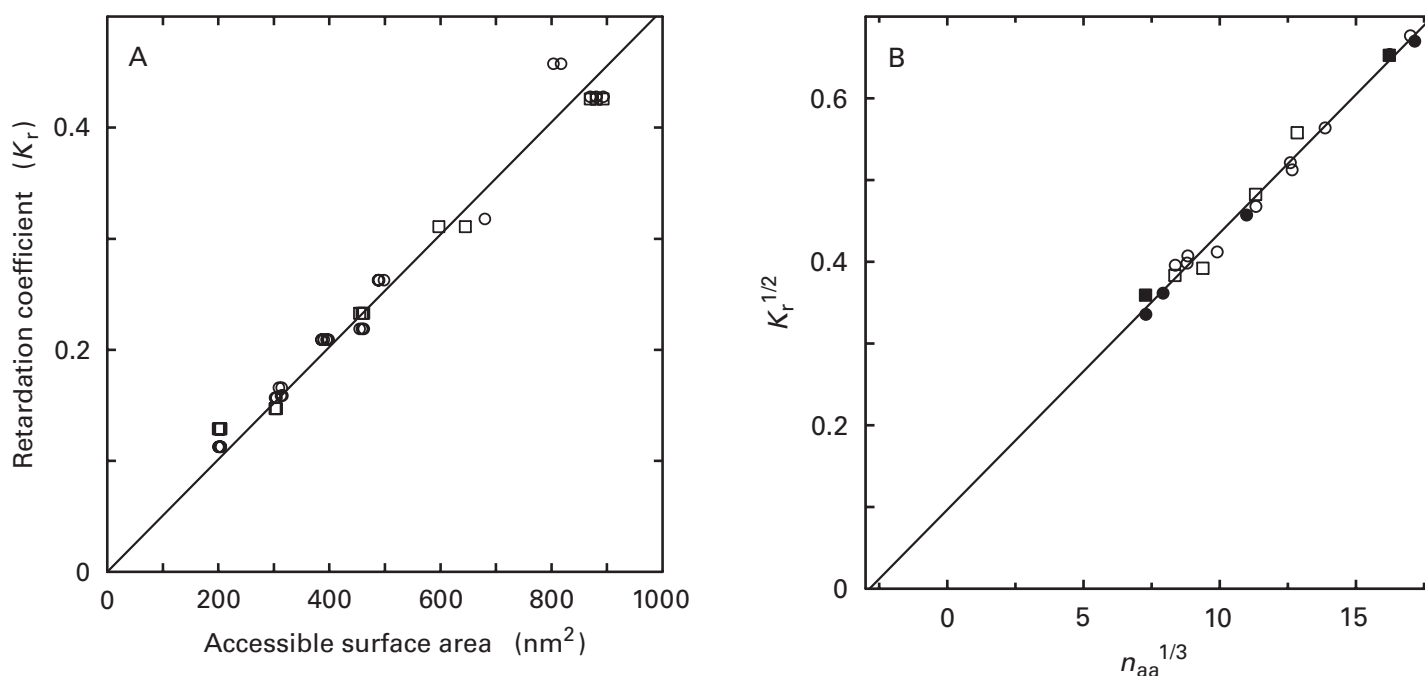


Figure 8–11: Relationship between the retardation coefficients K_r measured by electrophoresis and the surface areas or the numbers of amino acids n_{aa} for a set of proteins.^{40,111} A series of proteins were submitted to electrophoresis, each protein on a series of gels cast from solutions of increasing concentrations of acrylamide. The slopes of the lines from plots of the logarithm of the relative mobility against the percent of acrylamide were used to calculate K_r for each protein. The proteins used, in order of increasing number of amino acids, were ovalbumin from *G. gallus* ($n_{aa} = 385$), porcine α -amylase ($n_{aa} = 496$), bovine serum albumin ($n_{aa} = 583$), human transferrin ($n_{aa} = 679$), ovotransferrin from *G. gallus* ($n_{aa} = 686$), aspartate kinase–homoserine dehydrogenase from *Zea mays* ($n_{aa} = 828$), hexokinase from *S. cerevisiae* ($n_{aa} = 970$), the A isoform of L-lactate dehydrogenase from muscle of *O. cuniculus* ($n_{aa} = 1324$), the A isoform of fructose-bisphosphate aldolase from muscle of *O. cuniculus* ($n_{aa} = 1452$), β -amylase from *Ipomoea batatis* ($n_{aa} = 1992$), bovine catalase ($n_{aa} = 2024$), the M1 isoform of pyruvate kinase from *O. cuniculus* ($n_{aa} = 2120$), bovine xanthine oxidase ($n_{aa} = 2662$), equine apoferritin ($n_{aa} = 4272$), ribulose-bisphosphate carboxylase from *Chlamydomonas reinhardtii* ($n_{aa} = 4904$), and urease from *C. ensiformis* ($n_{aa} = 5040$). (A) The values of the retardation coefficients K_r are plotted as a function of the accessible surface areas (nanometers²) of the proteins calculated as described in Figure 8–10A with a spherical probe of 1.1 nm. Surface areas were calculated for ovalbumin, serum albumin, transferrin, ovotransferrin, L-lactate dehydrogenase, aldolase, pyruvate kinase (1PKM, 1A49), catalase, xanthine oxidase (1FO4), apoferritin, and ribulose-bisphosphate carboxylase (1AA1, 1RCO). (B) The square roots of the retardation coefficients for all of the proteins are plotted as a function of the cube roots of their number of amino acids. In both panels, circles are for the data of Hedrick and Smith¹¹¹ and squares are for the data of Bryan;⁴⁰ in panel B, solid symbols are for proteins with frictional ratios less than 1.20.

polymers, such as unfolded single-stranded DNA and unfolded polypeptides. **Extended polymers**, the shapes of which are unable to be approximated as spheres, nevertheless display regular behavior when they are submitted to sieving. The apparent surface areas, S_{app} , of extended, flexible polymers, such as unfolded polypeptides or unfolded single-stranded DNA, should increase linearly with their lengths, n_{aa} , because as each monomer is added, it increases S_{app} by the same increment, once the polymer is beyond a certain length. In this case,

$$S_{app,i} = c + d(n_{aa,i}) \quad (8-45)$$

where c incorporates all of the properties of homologous short polymers only a few segments in length.

When proteins are dissolved in solutions of guanidinium chloride, they unfold and their individual polypeptides become separated, random coils.¹¹² A series of these **randomly coiled polypeptides**, the lengths of which are now precisely known, were submitted to chromatography by molecular exclusion on beaded agarose, and the values of K_D (Equation 1-22) were reported.¹¹³ Combining Equations 1-21, 1-22, 8-36, and 8-45

$$-\ln K_{D,i} = \ln K_{av,R} + bT_P(c + dn_{aa,i}) \quad (8-46)$$

where $K_{av,R}$ is the distribution coefficient of the small reference solute R used to determine the apparent internal volume. This equation predicts that a plot of $\ln K_{D,i}$ against $n_{aa,i}$ should be linear, and it is (Figure 8-12).¹¹³ It has been proposed that the length of a polypeptide, the sequence of which is unavailable, could be estimated from its distribution coefficient by use of such a standard curve.

The regular behavior of **single-stranded nucleic acids** upon electrophoresis is crucial to the strategies for determining their sequences. The relative electrophoretic mobilities of the components in the ladder of single-stranded RNA displayed in Figure 3-13 can be measured from the photograph. Each relative mobility can in turn be related to the relative mobility of one of the components chosen as a standard, for example, the mobility of the one containing 30 nucleotides.¹¹⁴ If Equations 8-34, 8-43, and 8-45 are combined, and if it is remembered that u_i° for all unfolded single-stranded nucleic acids is the same, then

$$-\ln \left(\frac{u_i}{u_{30}} \right) = bT_a d(n_{b,i} - 30) \quad (8-47)$$

This predicts that a plot of $\ln(u_i/u_{30})$ against $n_{b,i}$ should be linear, and it is (Figure 8-13) with the exception of the compression at the discontinuity in the figure.

It could be ascertained, because sequencing was being performed,¹¹⁴ that the bands for the components

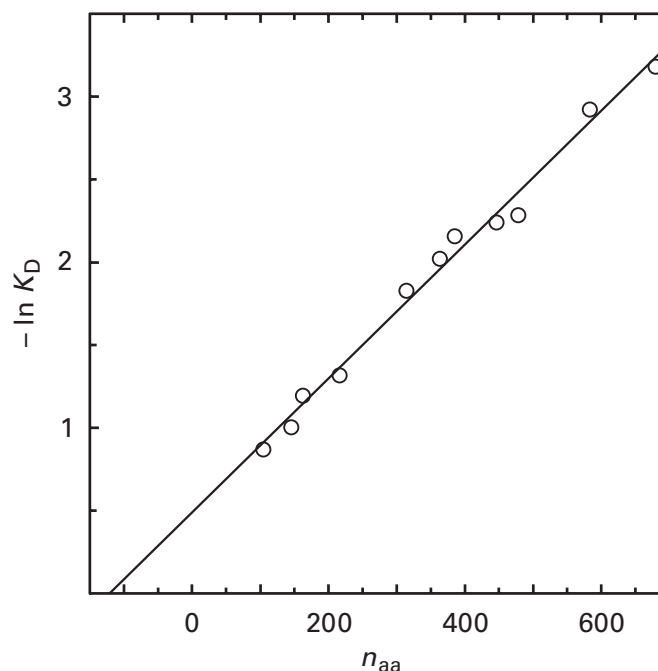


Figure 8-12: Sieving of unfolded, randomly coiled polypeptides on chromatography by molecular exclusion.¹¹³ Each of a series of proteins was dissolved in 6 M guanidinium chloride and 0.1 M 2-mercaptoethanol and submitted to chromatography on a column (1.5 cm × 90 cm) of beaded 6% agarose. The elution volume of each polypeptide was used to calculate its distribution coefficient K_D . The negative natural logarithm of K_D ($-\ln K_D$) is plotted as a function of the number of amino acids in the respective sequence, n_{aa} . The polypeptides chosen were those composing equine cytochrome c ($n_{aa} = 104$), bovine hemoglobin ($n_{aa} = 145$), bovine β -lactoglobulin ($n_{aa} = 162$), immunoglobulin G light chain from *O. cuniculus* ($n_{aa} = 220$), the mitochondrial isoform of malate dehydrogenase from liver of *Rattus norvegicus* ($n_{aa} = 314$), the A isoform of fructose-bisphosphate aldolase from *O. cuniculus* ($n_{aa} = 363$), ovalbumin from *G. gallus* ($n_{aa} = 385$), immunoglobulin G heavy chain from *O. cuniculus* ($n_{aa} = 450$), α -amylase A from *Aspergillus oryzae* ($n_{aa} = 478$), bovine serum albumin ($n_{aa} = 583$), and human transferrin ($n_{aa} = 679$).

in the ladder representing single-stranded ribonucleic acids of lengths 24–27 had all overlapped, producing this compression. It is usually assumed that a compression results from the ability of the 3' end of the single-stranded nucleic acid to double back upon itself and form a double-stranded hairpin as soon as the length of the nucleic acid becomes greater than a critical value in the expanding series. As the series approaches the discontinuity, it behaves regularly, because no hairpin is imminent. At the discontinuity and beyond it, the hairpin is present in each component, but it is eventually found far enough in the interior for the series to resume its linear behavior with the same slope it had previously but with a displacement. The displacement indicates that the polymer is behaving as if it were smaller than it actually is, presumably because the surface area of the double-helical hairpin in its interior is smaller than the surface area of the same number of nucleotides in a single-stranded state.

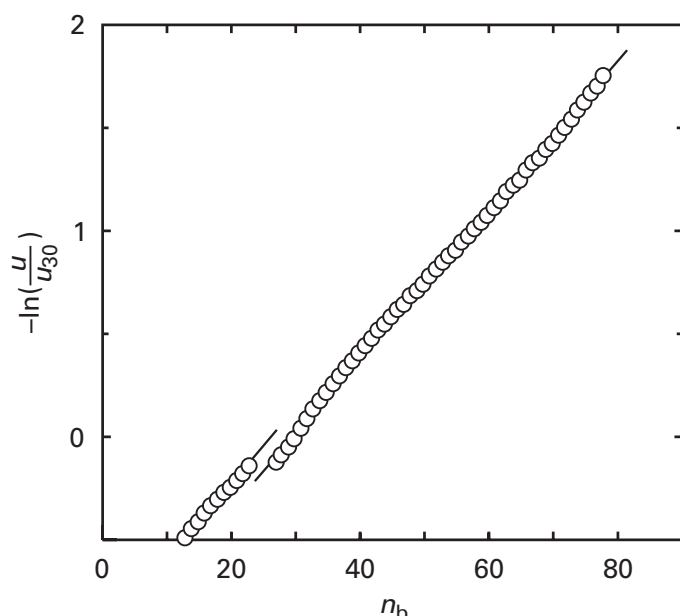


Figure 8-13: Sieving of single-stranded ribonucleic acid on electrophoresis in a gel of polyacrylamide. The distance between the origin and the final position of each band on the gel in Figure 3-13 was measured. These distances were each divided by the distance for the band corresponding to the ribonucleic acid 30 bases in length to obtain mobilities relative to this internal standard (u/u_{30}). The negative natural logarithms of these mobilities are plotted as a function of the lengths in bases, n_b , of each single-stranded ribonucleotide.

Unfolded polypeptides and unfolded single-stranded nucleic acids are examples of well-defined extended polymers. **Complexes between dodecyl sulfate and polypeptides**, because they are not chemically defined covalent polymers, are not so well understood. Nevertheless, both the behavior of polypeptides dissolved in solutions of guanidinium chloride (Figure 8-12) and the behavior of single-stranded nucleic acids (Figure 8-13) when they are respectively submitted to sieving suggest that the extended, unfolded complexes that form between dodecyl sulfate and polypeptides, which resemble the former in their unfolded state and the latter in both their distribution of negative charge and unfolded state, should display electrophoretic mobilities correlated with the length of the polypeptides. In fact, it was noted by Shapiro, Viñuela, and Maizel¹¹⁵ that this is the case. The electrophoretic mobility of the complex between dodecyl sulfate and polypeptide i is generally reported as a relative mobility:

$$R_{f,i} = \frac{u_i}{u_{\text{STD}}} \quad (8-48)$$

where u_{STD} is the mobility of a standard, either a small dye that can be readily followed visually or one of the obvious boundaries on a discontinuous gel. The advantage of the former point of reference is that because the

dye is dissolved in micelles of dodecyl sulfate, it marks the mobility of a micelle and hence the free electrophoretic mobility of the complexes between the proteins and the dodecyl sulfate (Figure 8-7). The advantage of the latter point of reference is that its absolute electrophoretic mobility can be calculated.

Equations 8-34 and 8-48 can be combined, and

$$\ln R_{f,i} = \ln \left(\frac{u_{\text{STD}}^\circ}{u_i^\circ} \right) - K_{r,\text{STD}} T_a + K_{r,i} T_a \quad (8-49)$$

This equation can be combined with Equations 8-43 and 8-45, and

$$\ln R_{f,i} = \ln \left(\frac{u_{\text{STD}}^\circ}{u_i^\circ} \right) - K_{r,\text{STD}} T_a + b T_a (c + d n_{aa,i}) \quad (8-50)$$

Because u_i° should be the same for all complexes between well-behaved polypeptides and dodecyl sulfate⁹⁰ and u_{STD}° , $K_{r,\text{STD}}$, and T_a are all constant, this equation predicts that a plot of $-\ln R_f$ against n_{aa} should be linear. When the natural logarithms of the relative mobilities measured by Weber and Osborn⁹³ are plotted as a function of the now known lengths of these polypeptides, n_{aa} , they conform to this expectation (Figure 8-14).^{93*}

At the present time, the method almost universally used to **estimate the length of a polypeptide**, the sequence of which is not yet known, is to determine the mobility of its complex with dodecyl sulfate upon electrophoresis on polyacrylamide gels. The mobility of the unknown is compared to the mobilities of complexes between dodecyl sulfate and standard polypeptides of known length, usually by plotting the data as in Figure 8-14. It should be realized, however, that the widespread reliance on this method is based on the assumption that the polypeptide of interest binds the same amount of dodecyl sulfate (amino acid)⁻¹ as the standards used. A comparison of Figure 8-13, which describes the behavior of a series of polymers in which the uniformity of the charge distribution is covalently dictated, with Figure 8-14, which describes the behavior of a series of polymers in which the uniformity of charge distribution depends only on a fortuitous consistency in its composition producing a fortuitous consistency in its ability to bind a small electrolyte, emphasizes the drawbacks of this assumption.

* When complexes between dodecyl sulfate and polypeptides are submitted to electrophoresis on polyacrylamide gels in which the complexes are stacked by moving discontinuities,⁹⁶ the relative mobilities of the standards do not fall on a line when they are plotted as in Figure 8-14. Nevertheless, their mobilities increase monotonically with their length, and the length of an unknown polypeptide can be estimated by interpolation.

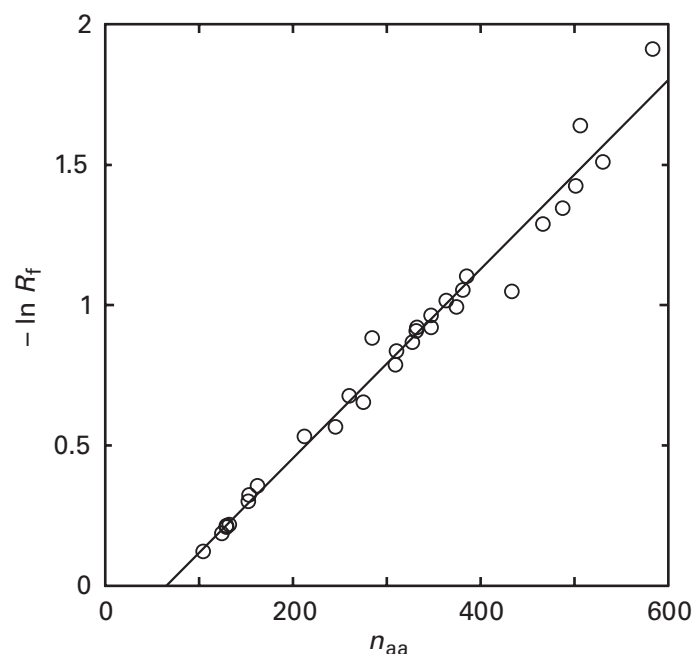


Figure 8-14: Sieving of complexes between dodecyl sulfate and unfolded polypeptides on electrophoresis in gels of polyacrylamide cast in solutions of sodium dodecyl sulfate.⁹³ A series of proteins (listed in Table I of ref 93) were submitted to electrophoresis as described in Figure 8-8. A marker dye was included in each sample, the mobility of which served as an internal standard. Its position was marked on each gel at the end of each run and the gels were stained. The distance migrated by each polypeptide was divided by the distance migrated by the marker dye to obtain a relative mobility R_f . The negative natural logarithms ($-\ln R_f$) of these values of relative mobility have been plotted as a function of the number of amino acids (n_{aa}) in the respective sequences of the polypeptides (www.expasy.org/sprot/).

Because they bind dodecyl sulfate with different stoichiometries, proteins that have peculiar compositions of amino acids,⁸⁷ that are excessively hydrophobic,⁸⁸ or that are significantly glycosylated do not have mobilities on polyacrylamide gels when they are coated with dodecyl sulfate that reflect only their lengths, and this procedure is unreliable in these instances.

A further problem that also should be realized is that beyond the ranges of relative mobilities displayed in Figure 8-14, the linear behavior of complexes between dodecyl sulfate and polypeptides often fails. This was originally pointed out by Weber and Osborn,⁹³ and it manifests itself in the tendency of complexes between dodecyl sulfate and very long polypeptides to travel faster than their lengths should permit. Because it is impossible to predict in what range of polymer lengths nonlinear behavior will become significant, a large collection of standard polypeptides the mobilities of which are close to and on either side of the mobility of the unknown should be chosen and the length of the unknown should be estimated by interpolation.

The failure of complexes between dodecyl sulfate and **long polypeptides** to behave as if they were geo-

metric solids is probably due to a change in the mechanism of sieving. It has been proposed that when the long dimension of a severely elongated macromolecule is significantly greater than the mean spacing between the fibers of polymer in a sieve, it will be hindered from reorienting significantly about any axes normal to that long dimension by the network itself. If so, it will be forced to move through the network as a worm in a randomly meandering burrow.¹¹⁶ At very low electric field gradients, the mobility of such a wormlike molecule should be proportional to n_{aa}^{-1} rather than $\exp(-n_{aa})$. At field gradients in the range normally used for electrophoresis, the burrow has a strong tendency to become aligned with the field, causing the mobility of the worm to become almost independent of n_{aa} .¹¹⁷ It is possible that the deviation of the retardation coefficients of complexes between dodecyl sulfate and longer polypeptides from linear behavior on gels of one polyacrylamide concentration^{88,93} results from a change in the mechanism of sieving that occurs as the longer dimension of the polymer becomes so long that it is forced to travel through the network as a worm rather than as a randomly reorienting, flexible geometric solid. The point at which this change in mechanism would set in would be a function not only of the length of the polymer but also of the spacing of the fibers in the network. If this is the case, it would explain the common observation that the behaviors of complexes between dodecyl sulfate and long polypeptides often become more linear when the concentration of polymer in the network is decreased.⁸⁸

Suggested Reading

Rodbard, D., & Chrambach, A. (1970) Unified theory for gel electrophoresis and gel filtration, *Proc. Natl. Acad. Sci. U.S.A.* 65, 970-977.

Problem 8-10: A series of standards was run to calibrate a cylindrical column of Sephadex G-75 for chromatography by molecular exclusion¹⁰⁴ so that it could be used to estimate the number of amino acids in bovine α -lactalbumin. The void volume of the column, V_0 , was 71 mL, and the total volume of the bed, V_T , was 226 mL. The following elution volumes were observed:

protein	n_{aa}	elution volume
cytochrome c	104	138 mL
ribonuclease	124	136 mL
myoglobin	153	127 mL
chymotrypsinogen	245	113 mL
α -lactalbumin		131 mL

(A) Calculate $K_{av} = (V_e - V_0)/(V_T - V_0)$ for every protein.

(B) Estimate the number of amino acids in α -lactalbumin.

Problem 8-11: Glycogen phosphorylase from *Oryctolagus cuniculus* was dissolved in a solution of sodium dodecyl sulfate sufficient to saturate the protein and submitted to electrophoresis on a polyacrylamide gel cast in a solution of sodium dodecyl sulfate. The relative mobilities of the glycogen phosphorylase and several standard proteins were measured.⁹³

protein	length of polypeptide (amino acids)	mobility relative to marker dye
myosin	1938	0.10
β -galactosidase	1023	0.16
serum albumin	583	0.33
catalase	506	0.37
glutamate dehydrogenase	501	0.43
fumarate hydratase	466	0.47
fructose-bisphosphate aldolase	363	0.56
glycogen phosphorylase		0.23

Estimate the length of the polypeptide composing glycogen phosphorylase.

Problem 8-12: Estimate the length of the polypeptide that composes porcine pepsin A from the following relative mobilities of complexes between the polypeptides and sodium dodecyl sulfate on electrophoresis on polyacrylamide gels cast in a solution of dodecyl sulfate.⁹³

polypeptide	n_{aa}	distance migrated (cm)
serum albumin	583	1.78
immunoglobulin G heavy chain	450	2.92
D-amino acid oxidase	347	4.80
glyceraldehyde-3-phosphate dehydrogenase	332	4.80
aspartate carbamoyltransferase, catalytic	310	5.22
polypeptide		
carboxypeptidase A	309	5.48
carbonate dehydratase I	260	6.12
pepsin A		5.06

Cataloguing Polypeptides

Rather than the intact oligomeric complexes of subunits formed from properly folded polypeptides that are separated during the electrophoresis of native proteins (Figure 1-19), the components separated when a mixture of different proteins is submitted to **electrophoresis in the presence of dodecyl sulfate** represent individual, unfolded, unassociated polypeptides. Consequently, such an electrophoretic separation is a **catalogue of the polypeptides** present in a sample⁹³ rather than a catalogue of the proteins. A graphic example of such a catalogue can be seen in Figure 1-22.

When a purified protein the homogeneity of which has been verified by electrophoresis in its native state is

submitted to electrophoresis in the presence of dodecyl sulfate, the pattern observed is a dissection of the protein into the different polypeptides of which it is composed. Usually a protein is composed of only one polypeptide, and that one polypeptide is usually present in the native protein in several copies. There are, however, many proteins that contain two or more different polypeptides, and a comprehensive description of the quaternary structure of such a protein requires that each of its constituent polypeptides be recognized as a unique component of the overall complex.

A catalogue of the polypeptides from which a protein is formed is reliable only if the **shortcomings** of electrophoresis in the presence of dodecyl sulfate have been recognized and eliminated. First, it has already been noted that, on discontinuous electrophoresis, components of high mobility often fail to escape the descending boundary. Complexes between dodecyl sulfate and short polypeptides are often trapped in this way and are unresolved. Second, it is also the case that, for reasons not well understood, all complexes between dodecyl sulfate and polypeptides less than 100 amino acids in length seem to have the same electrophoretic mobility,¹¹⁸ regardless of the concentration of polyacrylamide. This lower limit below which resolution fails can be lowered to about 25 amino acids in length by adding 8 M urea to the polyacrylamide gel.^{97,119} Third, because the cleavage of one peptide bond out of the hundreds present in an intact polypeptide always produces two new polypeptides that will be separated from each other and from their parent by electrophoresis in the presence of dodecyl sulfate, any degradation of the native protein by endopeptidases during or before its purification can artifactually multiply the apparent number of polypeptides without significantly altering the native protein or its own electrophoretic mobility. Fourth, endopeptidases often unfold more slowly than other proteins upon exposure to dodecyl sulfate, and they cleave their unfolded neighbors before they in turn succumb. If the purified protein is contaminated with even minute amounts of the endopeptidases that are always present in a homogenate, they can degrade the polypeptides during the preparation of the sample. Because these cleavages of the unfolded polypeptides are produced at random, as opposed to the unique cleavages usually produced during the degradation of a native protein by endopeptidases, they cause polypeptides in the sample to disappear into hundreds of fragments smeared over the field, each present in very low yield. Such an apparent disappearance of a polypeptide or polypeptides can also occur during the purification of the protein rather than during preparation of a sample for electrophoresis. For example, it was once thought that the stoichiometry of the subunits of nicotinic acetylcholine receptor from the electric eel was simpler than that from the electric ray until it was demonstrated that the missing polypeptides appeared

when indigenous endopeptidases, unavoidably present during the purification, were intentionally inactivated.¹²⁰ Yet the acetylcholine receptor originally purified, even though it had been cut up at random, was still biologically active. Finally, if only a fraction of the disulfides have been reduced, cross-linked, unreduced and un-cross-linked, reduced forms of the same polypeptide will appear as separate components. All of these and other artifacts must be recognized and eliminated¹²¹ before the pattern observed upon electrophoresis in the presence of dodecyl sulfate gives a reliable assessment of the different polypeptides present in a protein.

Each of the various components separated by electrophoresis in the presence of dodecyl sulfate may or may not represent a single polypeptide with a unique sequence of amino acids. The majority of the time they do, but it sometimes happens that one of the components represents two different polypeptides the lengths of which are so close to each other that they cannot be resolved. For example, the two subunits with unrelated amino acid sequences and unrelated functions composing the multienzyme complex from *Salmonella typhimurium* responsible for anthranilate synthase, glutamine amidotransferase, and anthranilate phosphoribosyltransferase happen to be 530 and 520 aa in length and are not resolved by electrophoresis in the presence of dodecyl sulfate. They are, however, cleanly resolved by electrophoresis in 8 M urea, a solution in which they are also unfolded but in which their differences in charge are not swamped by the binding of dodecyl sulfate.¹²² It is also possible that one or more of the components on the gel represent fragments of a larger component, also seen on the same gel. One way to resolve both of these ambiguities is to perform peptide maps.

A **peptide map** is a characteristic and reproducible display of the peptides produced when a polypeptide is digested with a specific endopeptidase. The display is usually produced by chromatographically or electrophoretically separating the digest in two dimensions to produce a characteristic pattern or map. Usually, the two dimensions are the respective orthogonal directions on a sheet of chromatographic paper or a thin layer of cellulose on a backing of plastic. Originally, electrophoresis was performed in the first dimension and chromatography in the second. Peptide maps are sensitive methods for assessing the similarity of two polypeptides, demonstrating that one polypeptide is a fragment of another, or revealing that one of the components on a polyacrylamide gel represents two polypeptides that fortuitously have the same electrophoretic mobility.

The most reliable maps are obtained from **tryptic digests** of a polypeptide because trypsin is the most specific and dependable of the endopeptidases. Polypeptides usually contain about 5 mol % arginine and 7 mol % lysine,¹²³ and for every 100 aa in length, about 11

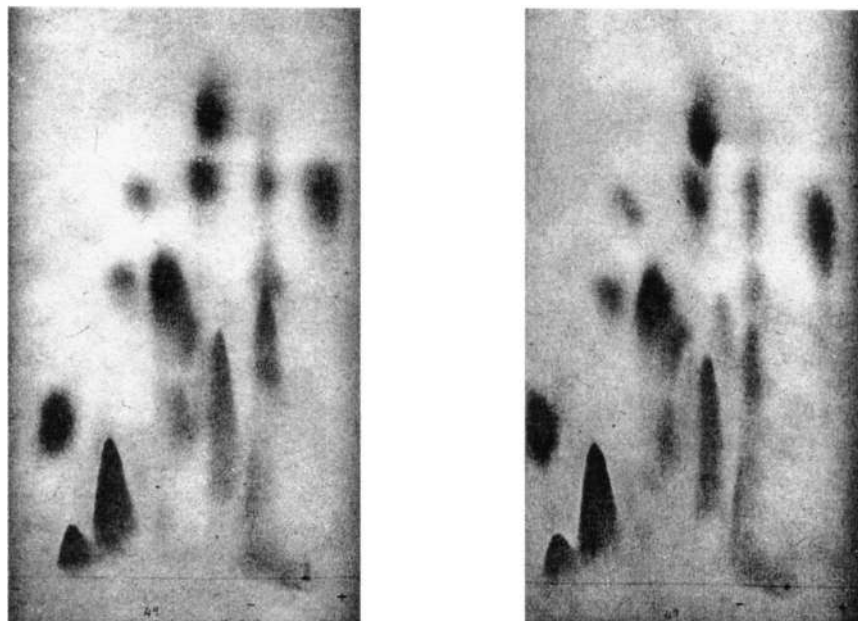
tryptic peptides should be present in the digest.* Each component on the map should represent a different tryptic peptide. Ideally, the resolution of the map should be high enough that every peptide in the digest appears as a separate, distinguishable component.

The initial triumph of analytical peptide mapping was in the examination of a mutant hemoglobin.¹²⁴ It had been proposed that the difference between normal hemoglobin, referred to as hemoglobin A, and hemoglobin S, a hemoglobin producing pathological distortions in erythrocytes, was due to a small difference in the amino acid sequence of the two proteins.¹²⁵ It was then shown that one and only one of the tryptic peptides on the respective peptide maps of the two proteins displayed an altered mobility (Figure 8–15).¹²⁶ It was concluded that all of the peptides the mobilities of which were the same between the two maps had identical sequences and the same relative locations in the two intact polypeptides, but that the one peptide the mobility of which was different had a sequence that differed between the two proteins by at least one amino acid. Because it is unlikely that the only two or three changes in the sequence of a polypeptide would occur in the same tryptic peptide, this result alone was substantial evidence that the two proteins differed from each other at only one location in their respective sequences. This was soon shown to be true by complete amino acid sequencing.

A similar strategy was used to evaluate the differences in the amino acid sequences of the different isoforms of actin.^{127,128} Each member of the set of the isoforms of actin chosen for the experiment, each of which had been isolated from a different species or a different tissue—a total of eight in all—was digested with trypsin, and each peptide map was compared to the peptide map of actin from skeletal muscle of *O. cuniculus*, the complete amino acid sequence of which was known. In all cases, the majority of the tryptic peptides were distributed over the map in the same pattern as the corresponding peptides on the map from the standard, and this permitted the various maps to be aligned with that of the standard. The peptides occupying the same positions in a pair of maps were assumed to be identical to each other in amino acid sequence. Amino acid analysis was used to verify these identities. Each unique peptide on the maps of the various unknowns was eluted and sequenced. Each of these amino acid sequences could be aligned with one of the tryptic peptides in the sequence of actin from muscle of *O. cuniculus*, and in this way the amino acid replacements in the sequences of the other actins could be readily established. This set of experiments relied on the fact that, aside from the first six amino acids in each sequence, all of the actins that were being compared show about 95% identity when their

* About 5% of the lysines and arginines in a protein are followed by proline, and trypsin is unable to cleave either a lysylproline or an arginylproline peptide bond.

Figure 8-15: Comparison of tryptic peptide maps of hemoglobin A (left panel) and hemoglobin S (right panel).¹²⁶ The respective hemoglobins were denatured at 90 °C for 4 min and the denatured proteins were digested with trypsin (1:50 trypsin/hemoglobin). The resulting digests were spotted on sheets of chromatographic paper, and the peptides were separated in the horizontal dimension by electrophoresis at pH 6.4 (negative pole to the left) and in the vertical direction by ascending chromatography in 1-butanol/acetic acid/water (3:1:1). The peptides were visualized with ninhydrin. A peptide in the middle of the right side of the left panel is replaced by a peptide in the center of the right panel. Reprinted with permission from ref 126. Copyright 1958 Elsevier Science Publishers.



amino acid sequences are aligned. This level of identity was what produced the underlying pattern that permitted the maps to be aligned and, in turn, permitted the ready identification of the peculiar peptides.

When the polypeptide is a long one, there may be so many components on a tryptic peptide map that they begin to overlap. One way to solve this problem is to modify the **tyrosine** side chains in the protein with radioactive iodine by electrophilic aromatic substitution. Because there are only 3–4 tyrosines for every 100 amino acids in a typical protein,¹²³ only about a third of the tryptic peptides become radioactive, and an autoradiogram* of the map is less cluttered than the entire map itself, but just as unique to the particular polypeptide. Such a map of tyrosine-containing chymotryptic peptides was used to show that the α polypeptides of Na^+/K^+ -exchanging ATPase from liver and kidney, respectively, both polypeptides now known to be 1018 amino acids in length with 24 mol of tyrosine (mol of polypeptide)⁻¹, were very similar if not identical to each other (Figure 8-16).¹²⁹ Another way of generating a peptide map from a long polypeptide is to digest the complex between it and dodecyl sulfate in a solution of dodecyl sulfate with an endopeptidase.¹³⁰ Under these conditions, the digestion is severely incomplete because the endopeptidase is rapidly inactivated by the dodecyl sulfate. Nevertheless, a reproducible set of large fragments of the polypeptide, characteristic of both it and the specificity of the endopeptidase used, is produced, and when these large fragments are separated by electrophoresis in the presence of dodecyl sulfate, the pattern of bands on the gel is a fingerprint unique to that polypeptide.

* An autoradiogram is a photographic image of the map on which only radioactive components are registered. It displays the distribution of radioactivity over the field.

Peptide mapping is sometimes performed by **adsorption chromatography** (Figure 3-7) because these separations are more rapidly accomplished,¹³¹ but this procedure is not so informative because it involves a separation in only one dimension. It is also possible to separate tryptic peptides from the digest of a polypeptide in one dimension by **mass spectrometry** following matrix-assisted laser desorption.^{132,133} The peptide map that results has the disadvantage that often only a portion of the tryptic peptides is present rather than a complete set so it functions mainly as a fingerprint of the particular polypeptide. Such a map, however has the advantages

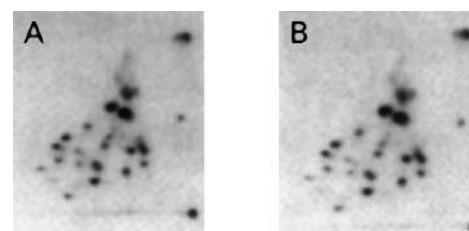


Figure 8-16: Peptide maps of tyrosine-containing chymotryptic peptides from the α polypeptide of Na^+/K^+ -exchanging ATPase.¹²⁹ After Na^+/K^+ -exchanging ATPase was purified from rat liver or rat kidney by immunoabsorption, its α polypeptide was isolated by electrophoresis on polyacrylamide gels in solutions of dodecyl sulfate. Each of the respective purified polypeptides was then chemically modified at its tyrosines by electrophilic aromatic substitution with ¹²⁵I. The radioactive polypeptides were then digested separately with chymotrypsin, and the digests were separated in two dimensions on thin layers of cellulose. Electrophoresis was performed from right to left followed by ascending chromatography with butanol/pyridine/water/acetic acid (65:50:40:10 v/v/v/v) from bottom to top. Peptides containing *o*-[¹²⁵I]iodotyrosine were identified by placing photographic film over the chromatogram. The images are those of the developed films. (A) Map from α polypeptide of kidney; (B) map from α polypeptide of liver. Adapted with permission from ref 129. Copyright 1986 American Chemical Society.

that the resolution is high so one dimension is sufficient, that the molecular mass of each peptide appearing on it is registered, and that the peptides can subsequently be sequenced (Figure 3–8). If the amino acid sequences of two closely related proteins are known, mass spectrometric peptide maps can often tell a sample of one from a sample of the other.¹³⁴ Another advantage of a mass spectroscopic peptide map is that it can be performed on a small amount of protein (1 pmol).

If two polypeptides yield peptide maps similar enough that they are judged to be related, the **percentage of identity** between their two sequences will be high; if they yield peptide maps that cannot be regarded as similar, they may still have clearly homologous sequences. In two-dimensional mapping of either all the peptides in a digest or just the tyrosine-containing peptides, or in peptide mapping by adsorption chromatography or mass spectrometry, the evaluation of the similarity of two polypeptides is based on comparison of the two patterns in which the peptides are displayed (Figure 8–15). Only if a significant fraction of the peptides on the two maps have the same relative positions on the field and produce a pattern that can be recognized is the judgment made that the two polypeptides are similar to each other. This implicit criterion of similarity requires that a significant fraction of the respective peptides be identical to each other in sequence for the decision to be made that the two polypeptides are related. One difference in the sequence of two otherwise identical peptides is usually sufficient to cause them to have different mobilities (Figure 8–15). Because the mean tryptic peptide is eight amino acids in length, differences between the sequences of two polypeptides at more than 20% of the positions will cause the two maps to be completely different, even though the two polypeptides are similar enough to be unambiguously judged homologous in amino acid sequence. Each of the four polypeptides of nicotinic acetylcholine receptor, although they are all homologous in sequence to each other (averaging 40% identity), yields a completely different peptide map.¹³⁵ The order in which the three ways of detecting homologies among polypeptides fail as the percentage of identity becomes smaller is peptide mapping before alignment of amino acid sequences and alignment of amino acid sequences before superposition of tertiary structures.

Whenever two or more polypeptides appear upon electrophoresis of a purified protein in the presence of dodecyl sulfate, the possibility that the smaller polypeptide or polypeptides are **fragments of the largest** should be examined. Such a relationship can be established by peptide mapping. The protein ankyrin from human erythrocytes is present in the cell under physiological conditions as the complete polypeptide and three progressively smaller fragments of that polypeptide. That these four polypeptides represent such a nested set derived by digestion of the largest by cellular endopeptidases could be demonstrated by producing peptide

maps of each.¹³⁶ This was done by separating these polypeptides on polyacrylamide gels in solutions of sodium dodecyl sulfate, iodinating their tyrosines, digesting them with trypsin, and producing tryptic peptide maps. These peptide maps all displayed the same pattern, but the maps from the smaller polypeptides lacked one or two of the peptides present in those from the next larger one. When collagen type XIV is isolated from epidermis of *Macaca fascicularis*, the purified protein contains two polypeptides that can be separated by electrophoresis in the presence of dodecyl sulfate. The complexes between each of these two polypeptides and dodecyl sulfate were digested separately with glutamyl endopeptidase. In the range of lengths less than that of the shorter of the two polypeptides, the two patterns of fragments that resulted from these partial digestions were identical to each other, a result demonstrating that the shorter of the two polypeptides was itself a fragment of the longer.¹³⁷

Peptide mapping can be used to determine whether an apparently unique component resolved by electrophoresis in the presence of dodecyl sulfate represents only one polypeptide or **two or more polypeptides** that fortuitously have the same electrophoretic mobility. The electrophoretic mobility of the complex between dodecyl sulfate and the polypeptide in question can be used to estimate its length. The mole percent of lysine and arginine in the protein can be either determined directly by total amino acid analysis or estimated from the fact that this number is usually about 11 mol %.¹²³ The mole percent of lysine and arginine and the length of the polypeptide can be used to estimate the number of tryptic peptides that should be produced if the component observed on the gel does represent only one unique polypeptide. If the number of peptides observed on the map agrees with this expectation, the component probably represents only one unique polypeptide. If there are about twice as many spots as expected, it must represent two different polypeptides. If the component does represent two or more polypeptides, it should be possible to separate them chromatographically or electrophoretically. Such a separation can usually be performed in 8 M urea, a solvent that unfolds polypeptides and separates them one from the other but that does not interfere with either chromatography by ion exchange or electrophoresis. The two or more separated polypeptides should each give unique peptide maps, the sum of which should be the peptide map of the original mixture.

When phosphoglycerate dehydrogenase was saturated with dodecyl sulfate and submitted to electrophoresis, one component was observed, the electrophoretic mobility of which was that of a polypeptide 360 aa in length. The content of lysine plus arginine in the protein was determined by amino acid analysis to be 9.6 mol %. A tryptic digest of the protein was separated by cation-exchange chromatography, and each of the pools from this first dimension was submitted to

electrophoresis on paper. This two-dimensional peptide map displayed 39–40 major peptides. The content of tryptophan in the protein was 1.0 mol %, and four of the peptides gave a positive test for tryptophan. If all of the polypeptides in this protein are identical, there should have been 36 tryptic peptides, four of which should have contained tryptophan. The **agreement between the observed numbers and the expected numbers** led to the conclusion that phosphoglycerate dehydrogenase was composed of identical polypeptides.¹³⁸

Glutamate-tRNA ligase was submitted to electrophoresis in the presence of dodecyl sulfate, and a single component was observed, the mobility of which was that of a polypeptide 500 aa in length. The protein had a content of lysine plus arginine of 12 mol %; tryptophan, 1.0 mol %; arginine, 6.3 mol %; and cysteine, 1.0 mol %. It could be concluded¹³⁹ that the component observed upon electrophoresis represented only one polypeptide because the tryptic peptide map of the protein displayed 55 peptides, 30 of which gave a positive test for arginine, five of which gave a positive test for tryptophan, and five of which became radioactive after the protein was reduced and carboxymethylated with [¹⁴C]iodoacetic acid (Equation 3–17).

Upon electrophoresis in dodecyl sulfate, the molybdenum-iron protein that is one of the components of nitrogenase gave two bands of stained material of very similar and often indistinguishable electrophoretic mobility, the apparent lengths of which were 540 amino acids. When the protein was reduced, carboxymethylated with [¹⁴C]iodoacetic acid, and submitted to amino acid analysis, its content of ([¹⁴C]carboxymethyl)cysteine was 1.7 mol %. Eleven of the tryptic peptides on a peptide map of the reduced and carboxymethylated protein were radioactive when nine were expected. Instead of passing this off as the result of incomplete digestion or inaccurate values for content of cysteine, the investigators proceeded to show that when the protein was dissolved in urea, to unfold its polypeptides, two polypeptides could be isolated by cation-exchange chromatography on (carboxymethyl)cellulose. Both were submitted to reduction, carboxymethylation, and peptide mapping. Four of the radioactive, cysteine-containing peptides from the map of the total protein were found on the map of one of the polypeptides, and the other seven radioactive peptides from the map of the total protein were found on the map of the other polypeptide, and no overlaps occurred between the two maps. It could be concluded that the molybdenum-iron protein had the subunit stoichiometry $\alpha\beta$.¹⁴⁰

A similar situation arose with methylmalonyl-CoA carboxytransferase. When this protein was submitted to electrophoresis in the presence of dodecyl sulfate, a component was present the apparent length of which was 550 amino acids, but under some conditions it would split into two bands of equal intensity. It was found that the native enzyme could be dissociated at pH 9.0 into two

proteins that could be separated from each other by molecular exclusion chromatography. Each of these proteins was composed of polypeptides the apparent lengths of which were 550 amino acids.¹⁴¹ Although their complexes with dodecyl sulfate were almost indistinguishable in electrophoretic mobility, the polypeptides in these separated proteins produced completely different tryptic peptide maps.¹⁴² It was later shown that they were polypeptides of 519 and 604 aa in length, respectively, with unrelated amino acid sequences.

The use of tryptic peptide mapping to provide evidence for the homogeneity of the polypeptides in a protein relies on the assumption that the trypsin has digested the polypeptide completely. This should be independently demonstrated. For example, initial tryptic digests of the polypeptides composing glucose-6-phosphate isomerase produced only two-thirds to three-fourths as many peptides as had been expected from the assumption that they were all identical. It was found that less base was consumed during the digestion than should have been, and this suggested that the digestion had been incomplete. When the protein was carbamylated on all of its lysines and then digested with trypsin, the quantity of base consumed during the digestion and the number of peptides observed on the map were those expected theoretically.¹⁴³ A more sensitive measure of complete tryptic digestion is to compare the total content of lysine and arginine in the digest to the amount of lysine and arginine released from the peptides in the digest when a sample is in turn digested with an appropriate carboxypeptidase.

Electron transfer flavoprotein is another protein composed of two polypeptides the lengths of which are very similar. It was originally believed to be a dimer of two identical polypeptides.¹⁴⁴ Under certain circumstances, however, two narrowly separated components would appear upon electrophoresis in the presence of dodecyl sulfate, and these were different enough to be separated by preparative electrophoresis. Each of the separated polypeptides was cleaved with cyanogen bromide, and the fragments produced were in turn saturated with dodecyl sulfate and separated in one dimension by electrophoresis in a solution of 0.1% dodecyl sulfate in 8 M urea (Figure 8–17).^{119,145} The maps produced in this way from each separated polypeptide were unique, and their sum was equal to the map of the intact protein. These results established the fact that the quaternary structure of the electron transfer flavoprotein is $\alpha\beta$ and explained the observation that there is only 1 mol of flavin (1.8 mol of polypeptide)⁻¹ in the protein.

Many proteins in addition to the molybdenum-iron protein of nitrogenase, methylmalonyl-CoA carboxytransferase, and electron transfer flavoprotein are composed of **two or more different polypeptides**. Often there is a functional basis for this arrangement. Many multienzyme complexes, rather than being composed of a string of enzymatic domains each formed from a dif-

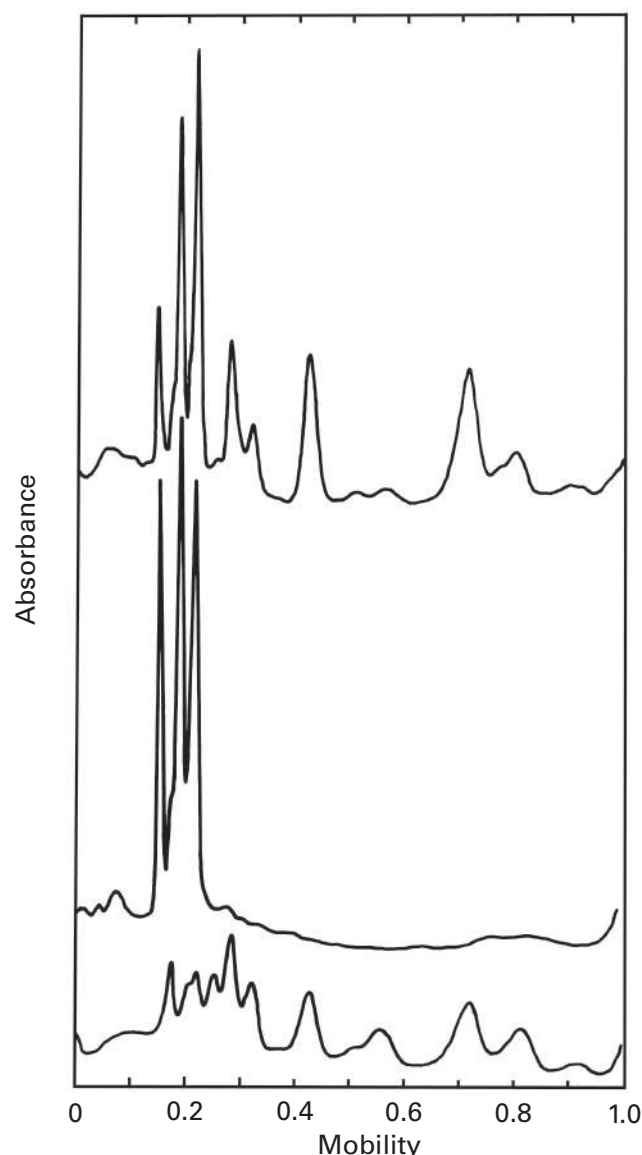
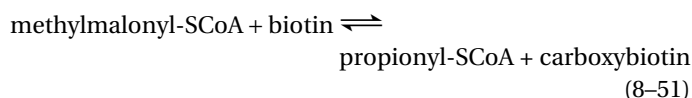


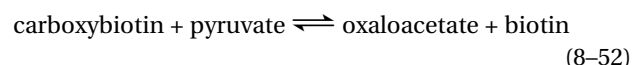
Figure 8-17: Peptide maps in one dimension performed by electrophoresis on a polyacrylamide gel of fragments derived from cleavage of electron-transferring flavoprotein from porcine liver with cyanogen bromide.¹⁴⁵ All proteins were reduced and alkylated with iodoacetamide before cleavage. Intact electron transfer flavoprotein (upper trace), its α polypeptide (middle trace), or its β polypeptide (lower trace), the latter two polypeptides isolated by preparative gel electrophoresis, were digested with cyanogen bromide (25 mM) in 88% formic acid and 0.03% sodium dodecyl sulfate for 24 h at 25 °C. The fragments produced were separated on gels cast from 13.4% acrylamide in 0.1% sodium dodecyl sulfate and 8 M urea¹¹⁹ and stained for protein. The gels were then scanned for absorbance as a function of length, and length was converted to mobility relative to the mobility of a marker dye. Reprinted with permission from ref 145. Copyright 1983 *Journal of Biological Chemistry*.

ferent region of the same polypeptide, are constructed from individual subunits gathered together in a larger complex. Functionally there is no distinction between these two types of **multienzyme complexes** because the important feature is that the different enzymes are gathered together, whether they are gathered as domains on

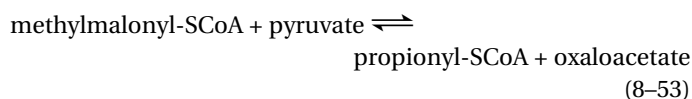
the same subunit or as different subunits. It has already been noted that methylmalonyl-CoA carboxytransferase is formed from different subunits. One of its three constituent polypeptides is folded to produce a protein that in isolation¹⁴⁶ can catalyze the reaction



and another, a protein that can catalyze¹⁴⁶ the reaction



The entire enzyme is composed of these two subunits and a third subunit formed from a single polypeptide bearing covalently attached biotin as a posttranslational modification of one of its lysines. The intact multienzyme complex catalyzes the overall reaction



In the case of the molybdenum-iron protein of nitrogenase, each of the two polypeptides of almost the same length (491 and 522 aa) forms one of its subunits. One of the two subunits contains the molybdenum and some of the iron, while the other subunit contains iron-sulfur clusters of a ferredoxin type. This assigns different functional roles to each of the subunits and explains the stoichiometries of the molybdenum and iron found in the intact protein.¹⁴⁰

Some of the proteins, however, thought to be composed of two different types of subunits, because they are isolated as complexes containing two different polypeptides, are actually the products of a **posttranslational cleavage**, either intentional or artifactual, of what was initially a single polypeptide. The internal modification of histidine decarboxylase from *Lactobacillus* producing the *N*-pyruvyl amino terminus (Equation 3-9) coincidentally produces two shorter polypeptides of lengths 81 and 229 amino acids from the originally intact precursor.^{147,148} Before the modification occurs, the protein is constructed from six identical copies of the intact polypeptide, each folded as an independent subunit. Many endopeptidases are normally stored safely as inactive precursors that are activated at the proper time by an internal cleavage or cleavages catalyzed by another molecule of endopeptidase. These cleavages occur in surface loops and have little effect on the overall structure of the protein, and the two resulting fragments are not separate subunits or even separate domains.¹⁴⁹ Another example of a posttranslational cleavage that produces two fragments of a constituent polypeptide occurs during the maturation of insulin receptor.¹⁵⁰

It was once believed that vertebrate acetyl-CoA carboxylase was constructed from two or three different polypeptides, each present in one or two copies.¹⁵¹ This belief was reinforced by the fact that the enzyme from *E. coli* is a multienzyme complex constructed from three different subunits,¹⁵² albeit folded polypeptides of lengths much shorter than the polypeptides found in the enzyme from vertebrates. When the vertebrate enzyme was purified by a rapid procedure employing affinity adsorption, however, it was found to be composed from only one polypeptide, the length of which (2345 amino acids) was more than twice the individual lengths of the separate polypeptides seen previously.¹⁵³ Only two folded copies of this longer polypeptide are present as subunits of the native protein. It was concluded that the smaller polypeptides seen previously were the products of artifactual digestion by endopeptidases.

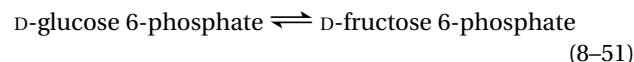
The distinction between a protein that contains two different polypeptides because of a posttranslational cleavage and one that has been assembled from two separately translated polypeptides is significant. The posttranslational cleavage of a protein usually occurs after the entire polypeptide has folded because it is only in the folded polypeptide that the cleavage can be directed to a precise location. Although significant changes may occur in the vicinity of the cleavage, the overall structure of the protein remains what it was before the cleavage. For all intents and purposes, a polypeptide cleaved either naturally or artifactualy after it has folded remains structurally a folded single polypeptide. On the other hand, when a protein is assembled from separately translated polypeptides, they first must fold independently before they can recognize each other and join together. Each subunit begins as and remains as a discrete entity and appears as such when a crystallographic molecular model of the protein is viewed. Yet when the polypeptides are resolved on a polyacrylamide gel, these two very different situations cannot be distinguished, and as with the confusion surrounding the connotations of the term domain, the urge arises to confer the structural integrities of separate subunits to a protein that may contain two or more polypeptides only by virtue of posttranslational modification.

It is possible that the number of proteins assembled from two or more different subunits containing polypeptides with different sequences and different lengths has been overestimated. Even if this is not so, such heterooligomers represent only a minority of the oligomeric proteins. Most oligomeric proteins are constructed from only one polypeptide that is present in two or more identical copies in the complete protein. The length of that polypeptide is estimated by electrophoresis in the presence of dodecyl sulfate. The number of copies present in the intact protein is counted.

Suggested Reading

Dowhan, W., & Snell, E.E. (1970) D-Serine dehydratase from *Escherichia coli*. II. Analytical studies and subunit structure, *J. Biol. Chem.* 245, 4618–4628.

Problem 8–13: Glucose-6-phosphate isomerase catalyzes the conversion



The enzyme has been purified from muscle of *O. cuniculus*. When it was dissolved in a solution of sodium dodecyl sulfate and submitted to electrophoresis, the following results were obtained:¹⁴³

protein	n_{aa}	R_f
ovalbumin	385	0.42
catalase	506	0.32
serum albumin	583	0.26
glucose-6-phosphate isomerase		0.30

The enzyme was unfolded in 6 M guanidinium chloride and modified with potassium cyanate, the reagents were removed by dialysis, and the protein was digested with trypsin. The following two maps were made of this tryptic digest. The only difference in these two maps is the pH at which the electrophoresis was performed.

The amino acid composition of the protein has been determined:

amino acid	moles (100,000 g of protein) ⁻¹	amino acid	moles (100,000 g of protein) ⁻¹
K	62	I	53
H	36	L	87
R	32	Y	19
D+N	88	F	46
T	62	G	67
S	58	A	64
E+Q	96	V	52
P	38	M	22

- What is the length of the polypeptides composing glucose-6-phosphate isomerase?
- How many different polypeptides does the protein contain?
- Which amino acid side chain are you certain the cyanate modified? Why?
- What conclusions did you draw from the peptide maps?

Problem 8–14: The molar mass of 2-dehydro-3-deoxyphosphogluconate aldolase (DDPG aldolase) from *Pseudomonas putida* has been estimated as 73,000 g

mol⁻¹ by sedimentation equilibrium. The enzyme was dissolved in a solution of sodium dodecyl sulfate and submitted to electrophoresis. Its mobility on the gel as well as the mobilities of several standards are tabulated.¹⁵⁴

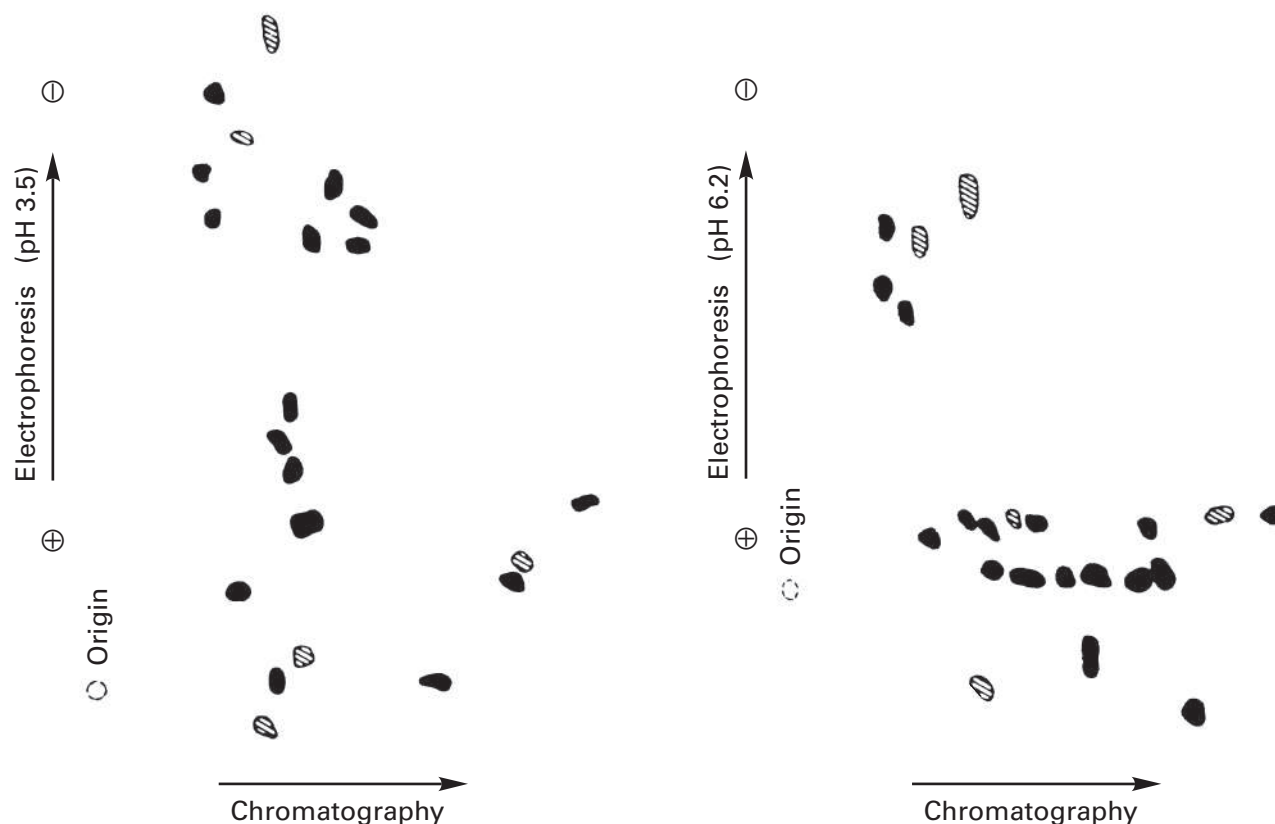
protein	n_{aa}	R_f
carbonate dehydratase	260	0.62
chymotrypsinogen	245	0.74
trypsin	223	0.76
myoglobin	153	1.00
DDPG aldolase		0.75

The protein was reduced, carboxymethylated, and digested with trypsin. There were 24 spots observed on a two-dimensional map of this tryptic digest; three were positive for tyrosine, as detected by Pauli stain, and 13 were positive for arginine. The amino acid composition of the protein was determined:

amino acid	mol (100 mol) ⁻¹	amino acid	mol (100 mol) ⁻¹	amino acid	mol (100 mol) ⁻¹
C ^a	1.7	G	9.2	Y	1.3
D+N	7.4	A	13.6	F	3.2
T ^b	4.5	V	6.5	K	3.1
S ^b	3.6	M	3.0	H	0.5
E+Q	8.9	I	8.2	R	6.6
P	6.9	L	9.5	W ^c	1.6

^aAs (carboxymethyl)cysteine at 24 h. ^bExtrapolated to zero hour. ^cDetermined spectroscopically.

The protein was dissolved in 8 M urea, 2-mercaptoethanol was added, and the mixture was incubated for 4 h. The reduced protein was then alkylated with [¹⁴C]iodoacetate, dialyzed, and digested with trypsin. The tryptic digest was run on an ion-exchange column, and four peaks of radioactivity were observed. Each radioactive peak was further purified to homogeneity. The four were shown by the composition of their amino acids to be unique and each contained one (carboxymethyl)cys-



Tryptic peptide maps of carbamylated glucose-6-phosphate isomerase from muscle of *O. cuniculus*.¹⁴³ Electrophoresis: right map, pyridine/acetic acid/water (520:1.4:1000 by volume), pH 6.2; left map, pyridine/acetic acid/water (7:66:1927 by volume), pH 3.5. Chromatography: descending chromatography with butanol/acetic acid/pyridine/water (15:10:3:12 by volume). Maps were performed on sheets (46 cm × 57 cm) of chromatographic paper, and peptides were located by ninhydrin.

teine. Their compositions were C₁D₂S₁E₃A₆I₂L₂, C₁D₁E₁A₁I₂K₁, C₁D₂T₁G₂A₁V₂F₁R₁, and C₁D₁T₁E₁G₁A₁V₁L₁R₁.

- (A) What are the lengths of the polypeptides composing this enzyme?
- (B) How many polypeptides are there in the protein?
- (C) How many different types of polypeptides are there in the protein?
- (D) What conclusions can you draw from the tryptic peptide map?
- (E) What conclusions can you draw from the tryptic peptides containing ([¹⁴C]carboxymethyl)cysteine?

Problem 8-15: A protein has been purified to homogeneity. It has the following properties.¹²⁶ When the protein is reduced with 2-mercaptoethanol and run on a sodium dodecyl sulfate gel in the presence of standards, the following results are obtained:

protein	<i>n</i> _{aa}	mobility
β-lactoglobulin	162	0.70
myoglobin	153	0.73
lysozyme	129	0.81
ribonuclease	124	0.82
cytochrome <i>c</i>	104	0.87
protein X		0.77

The amino acid composition of the protein is as follows:

amino acid	mol (100 mol) ⁻¹	amino acid	mol (100 mol) ⁻¹
G	6.9	Y	2.0
A	12.5	W	0.9
S	5.4	C	1.0
T	5.4	M	1.0
P	5.0	D	8.6
V	10.7	E	5.9
I	0.0	R	2.9
L	12.6	H	6.5
F	5.2	K	7.6

The protein was digested with trypsin, and a peptide map was prepared. It contained 26 well-defined peptides. The peptide map was stained for various amino acid side chains: five peptides were positive for arginine, 13 peptides were positive for histidine, four peptides were positive for methionine, three peptides were positive for tryptophan, and seven peptides were positive for tyrosine. The protein was carboxymethylated with [¹⁴C]iodoacetic acid and digested with trypsin. Three tryptic peptides containing radioactive (carboxymethyl)cysteine were isolated by ion-exchange

chromatography, and each of these was shown to have a unique composition of amino acids and to contain 1 mol of Cys (mol of peptide)⁻¹.

- (A) What is the length of a polypeptide composing this protein?
- (B) How many different types of polypeptides compose the protein?
- (C) Explain the peptide maps.

Cross-Linking

There arose a disagreement over the number of subunits contained in fructose-bisphosphate aldolase. The physical methods for estimating the molar mass of the native protein and the molar mass of its constituent polypeptides were unable to decide between three and four. It should be noted that everyone had an equal chance of being correct, so the point is not who turned out to be right but that the question could not be resolved simply by arguing over the numbers. What was needed instead was a different kind of experiment, and it was provided. When the fructose-bisphosphate aldolase in a homogenate from brain of *O. cuniculus* was submitted to electrophoresis in its native state, five evenly spaced components displaying enzymatic activity were observed (Figure 8-18).⁷⁶ Penhoet, Kochman, Valentine, and Rutter⁷⁶ decided that this must be due to the fact that, in the brain, two isoenzymatic polypeptides designated α and γ are translated from two different messenger RNAs continuously and coincidentally. These polypeptides fold separately to form monomeric subunits that then combine at random with subunits of their own kind or of the other isoenzymatic type to produce **hybrids** of the stoichiometries α₄, α₃γ, α₂γ₂, αγ₃, and γ₄, designated A, I, II, III, and C in Figure 8-18.* The two different subunits, α and γ, differ in the sequences of their polypeptides and hence in their charge. Each hybrid in turn has a different electrophoretic mobility because each has a different mean charge number *Z*. The hybrids are capable of forming in the first place because the two different polypeptides are homologous in their sequences, have superposable tertiary structures in their folded state, share a common ancestor, and have not diverged sufficiently from that common ancestor to have lost the ability to combine with each other in the same way that they are required to do with subunits identical to themselves. If this explanation is correct, the number of subunits in any molecule of aldolase can be determined by simply counting the components on the electrophoretic separation. There must be four. A similar hybridization was used to verify that chloramphenicol *O*-acetyltransferase is a trimer.¹⁵⁵

* The proteins with quaternary structures α₄ and γ₄ have been designated the A isoform and the C isoform, respectively, of fructose-bisphosphate aldolase.

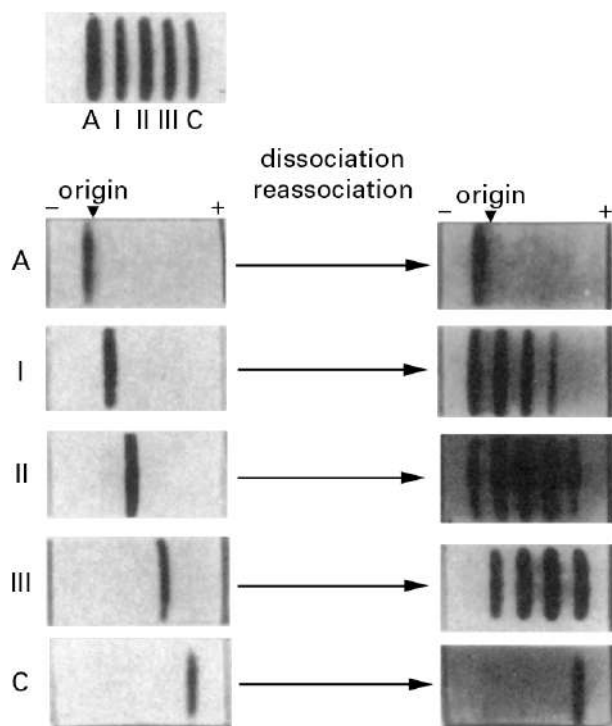


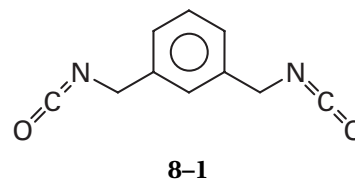
Figure 8-18: Dissociation and reassociation of oligomeric hybrids of fructose-bisphosphate aldolase.⁷⁶ A clarified homogenate from brain of *O. cuniculus* was submitted to electrophoresis on cellulose acetate at pH 8.6, and the strip was then stained for the activity of fructose-bisphosphate aldolase. Five evenly spaced bands of enzymatic activity were observed (top pattern). On the basis of their calculated points of zero net charge, the most anionic component, labelled C in the figure, could be identified as isoform C of fructose-bisphosphate aldolase; and the most cationic, labelled A in the figure, as isoform A of fructose-bisphosphate aldolase. The homogenate was then submitted to substrate elution from cellulose phosphate followed by anion-exchange chromatography on (diethylaminoethyl)cellulose (Figure 1-2). The five components that differed in electrophoretic mobility could be identified on these chromatograms by their enzymatic activity and could be cleanly separated from each other in this way. Each was submitted to electrophoresis separately at pH 8.6 and only one respective component with enzymatic activity was found in each (the five patterns on the lower left). Each of these single components was then exposed to 0.33 M H_3PO_4 at pH 2.0 for 30 min at 0 °C and then brought back to pH 7.5. The low pH served to dissociate the subunits of each hybrid, and the neutralization reassociated them but at random. Each of the dissociated and reassociated mixtures of hybrids was then resubmitted to electrophoresis, and the strips of cellulose acetate were stained for enzymatic activity (the five patterns on the right). Reprinted with permission from ref 76. Copyright 1967 American Chemical Society.

In the experiments with fructose-bisphosphate aldolase, the investigators proved that the components observed upon electrophoresis were the hybrids that they proposed by isolating each of them, dissociating each to subunits, reassociating the subunits, and demonstrating that each of these reassociated samples contained a new set of hybrids consistent with the random rearrangement of the stoichiometry assigned to the parent (Figure 8-18). For example, α_4 gave back only α_4 and γ_4 gave back only γ_4 , but $\alpha_2\gamma_2$ gave all five hybrids in

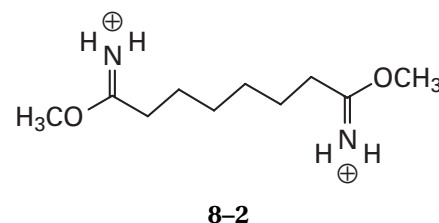
approximately binomial ratios. It was also shown that an equal mixture of α_4 and γ_4 when dissociated and reassociated reproduced all five hybrids.

It is possible to increase the difference in charge between two isoenzymatic polypeptides, and hence the spacing between the bands on electrophoresis, or to produce a charge variant of a single polypeptide by mutating lysines to glutamates.¹⁵⁵ It is also possible to make **electrophoretic variants** by simply succinylating any native protein of interest,¹⁵⁶ rather than relying on isoforms coincidentally provided by nature. The pattern of functional heterogeneity, rather than the pattern of electrophoretic heterogeneity, of a hybrid mixture formed from subunits with differences in function, rather than differences in charge, can also be used to count the subunits in an oligomer.¹⁵⁷ Nevertheless, the number of subunits in a native protein is rarely counted by hybridization. The point, however, is not that hybridization solves the problem of determining the stoichiometries of subunits but that an experiment can be designed so that the number of subunits can be counted directly instead of calculated from physical measurements. This new way of looking at the problem stimulated the development of a technique that could be used to count the number of subunits in most soluble proteins. This technique relies upon cross-linking.

A reagent capable of covalently **cross-linking** two polypeptides is a chemical compound that contains at least two electrophilic functional groups attached to each other by the remainder of the molecule. Aromatic diisocyanates, such as *m*-xylylene diisocyanate¹⁵⁸

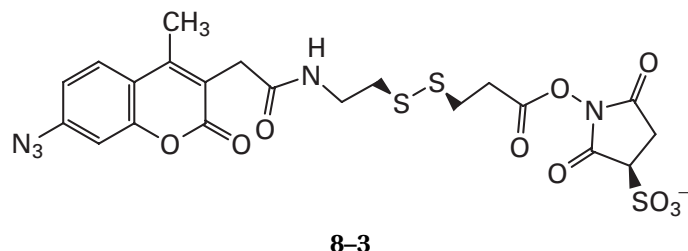


were among the first such reagents to be used for this purpose, but they suffer from problems of low solubility. A widely employed cross-linking agent, and the one originally used to count subunits,¹⁵⁹ is dimethyl suberimide



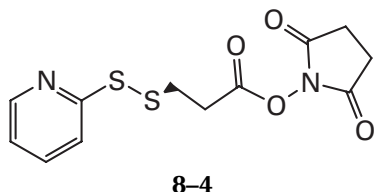
By far the most prevalent and accessible nucleophiles on the surface of a molecule of protein are the primary amines of its **lysines**, and the majority of the reagents, such as the diisocyanates and the diimidoesters, are directed to lysines.

There have been a large array of reagents synthesized over the years for cross-linking proteins. They have been designed to react with a broader array of amino acids than just lysine. They have also been designed to form cross-links that can be reversed.¹⁶⁰ A reagent that illustrates all of the various elements in the design of such **cross-linking reagents** is sulfosuccinimidyl 2-(7-azido-4-methylcoumarin-3-acetamido)-ethyl-1,3'-dithiopropionate:¹⁶¹



The ester of *N*-hydroxysulfosuccinimide is an electrophile that reacts with lysines. The *N*-hydroxysulfosuccinimide is used as a leaving group, rather than *N*-hydroxysuccinimide, to increase solubility. The azide is photoactivated to the nitrene that has a broader spectrum of reactivity than a more common electrophile. The two electrophiles react respectively with two nucleophiles on the protein to cross-link them. The disulfide can be reductively cleaved to reverse the cross-linking. The coumarin makes the reagent and hence the products of the cross-linking fluorescent. At the other end of the scale of complexity is the cross-linking of two proteins by the oxidative formation of covalent dimers between a tyrosine on one of the proteins and a tyrosine on the other.¹⁶²

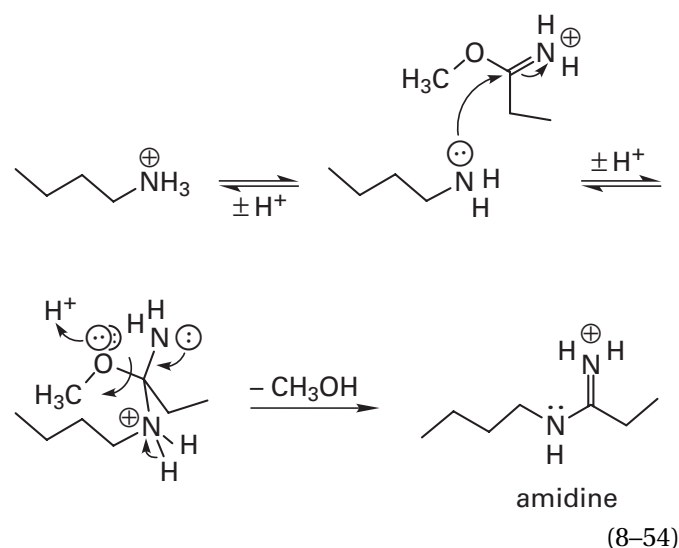
Cross-linking reagents that contain an activated disulfide that reacts efficiently with cysteines on the surface of a protein are also widely used. A mixed disulfide with 2-thiopyridine, an excellent leaving group, can be used to attach an electrophile to a cysteine on the first protein. For example, *N*-succinimidyl 3-(2-pyridyldithio)propionate¹⁶³



reacts by disulfide interchange (Figure 3-20) with the thiol of a cysteine to form a mixed disulfide between that cysteine and the *N*-succinimidyl 3-thiopropionate. The reagent can be directed to a particular location on the first protein by inserting a cysteine at a particular position in its sequence by site-directed mutation.¹⁶⁴ The ester of *N*-hydroxysuccinimide can then react with a lysine on a second protein, cross-linking it to the first protein. The cross-link is reversible because the mixed disulfide can be reduced by disulfide interchange with a

mercaptan such as dithiothreitol (Equation 3-18). This reaction generates a thiol at the location on the second protein that participated in the cross-link. This unmasked thiol can then be used to isolate a peptide from the second protein containing the amino acid that was cross-linked.¹⁶⁴ Because the position in the amino acid sequence of the first protein at which the reagent was attached was determined by the site-directed mutation and the position in the second protein with which the electrophile reacted can be identified by isolating a peptide containing the modified amino acid, the cross-linking reaction can be used to identify adjacent amino acids in the complex that forms between the two proteins.

The most commonly used cross-linking reagents are **bisimidates** and bisaldehydes. An imidoester such as the ones in **8-2** reacts with a primary amine to form an amidine:



When a molecule of dimethyl suberimidate happens to react at one of its ends with a lysine on one subunit and at its other end with a lysine from another subunit, those two subunits become covalently cross-linked and their polypeptides migrate together upon electrophoresis in the presence of dodecyl sulfate with the mobility of a polypeptide the length of which is equal to the sum of the lengths of the two polypeptides so joined.

The intramolecular cross-linking of an oligomeric protein provides a **count of the number of subunits** it contains. At concentrations of protein below 10 μM , no significant intermolecular cross-linking between separate molecules of protein occurs when a solution of protein is mixed with a cross-linking agent such as dimethyl suberimidate. Instead, what is observed are the products that result from intramolecular cross-linking among the fixed number of subunits of which the protein is composed.¹⁵⁹ When the products of such intramolecular cross-linking are separated on a polyacrylamide gel, a ladder of bands, vaguely reminiscent of the ladders seen with randomly cleaved nucleic acids, is observed (Figure

8–19).¹⁶⁵ These ladders, however, end abruptly because no more polypeptides can be cross-linked than are present in the complete protein. A count of the rungs in the ladder provides the number of subunits in the protein. In the case of glycerol kinase, the protein used for the analyses presented in Figure 8–19, it could be concluded that it is composed of four and only four subunits.¹⁶⁵ Peptide maps showed that all four of the polypeptides composing the subunits are identical to each other.¹⁶⁵

There are three **reassurances** that should be provided. First, the relative mobility of each of the bands in

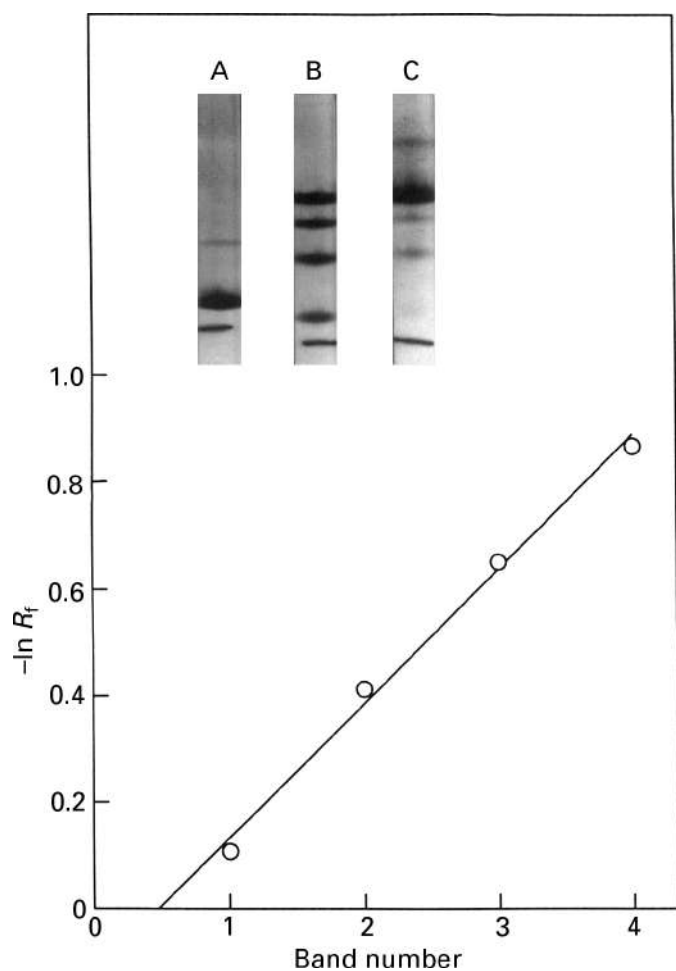
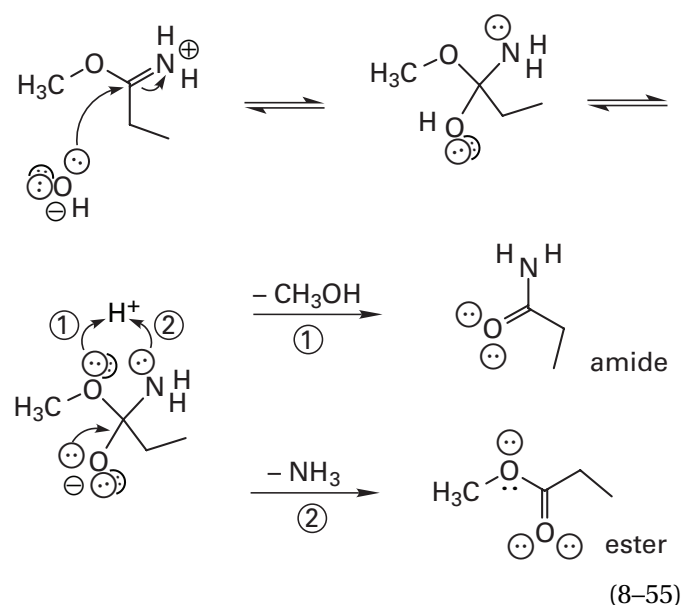


Figure 8–19: Definition of the number of subunits in glycerol kinase.¹⁶⁵ Three samples of a solution of glycerol kinase were mixed with final concentrations of 10 mg mL⁻¹ dimethyl suberimidate (inset, gel C), 0.25 mg mL⁻¹ dimethyl suberimidate (inset, gel B), or no additions (inset, gel A). After 4 h at pH 8.5, the samples were dissolved in a solution of sodium dodecyl sulfate and subjected to electrophoresis on polyacrylamide gels cast in 0.1% sodium dodecyl sulfate. The gels were stained for protein. The irregular line at the bottom of each gel is a line of India ink used to mark the position of the marker dye at the end of the electrophoresis. The band of stain just above this mark on gels A and B is the un-cross-linked polypeptide of glycerol kinase. The mobilities of the four components on gel B relative to the marker dye were calculated from the photograph, and the negative natural logarithm of each of these mobilities ($-\ln R_f$) is plotted against a scale of successive integers. Adapted with permission from ref 165. Copyright 1971 *Journal of Biological Chemistry*.

the ladder should be shown to be a regular function of its number (graph in Figure 8–19).¹⁶⁶ This is a reassurance that one of the members of the set is not missing from the pattern. Second, the cross-linking reaction should be forced to completion so that only the highest oligomer is seen (gel C, inset to Figure 8–19). This result provides the reassurance that this oligomer does represent a true limit to the reaction and that the solution contains only one unique multimer of the subunits rather than a mixture of multimers each containing a different number of subunits. Third, the possibility that intermolecular cross-linking is occurring should be ruled out by varying the concentration of protein. The amounts of the components arising from intramolecular cross-linking will not vary with the concentration of protein, but those that arise from intermolecular cross-linking will. Random intermolecular cross-linking also yields a distribution that gradually declines in its amplitude with band number rather than displaying an abrupt limit at a certain unique polymer size. It is this discontinuous behavior that is the logical basis for believing the results.¹⁵⁹

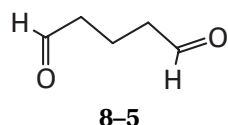
It is in fulfilling the requirement that the reaction be forced to completion that dimethyl suberimidate usually fails. The reaction of an imidoester with a primary amine in aqueous solution is a competition between formation of the desired amidine and hydrolysis of the imidoester to amide and ester:



There is no way to avoid this competition by some informed adjustment of the conditions; it can only be made worse. It complicates the reaction because when one end of the diimidoester has attached to a lysine, there is a significant probability that its other end has either already been hydrolyzed or will hydrolyze before it can react with another lysine. A high percentage of the lysines end up with defunct reagent after all of the lysines have become amidines and no further reaction can occur

regardless of how much reagent is added. Because of this, it is often the case that not all of the subunits can be cross-linked among themselves. The logical chemical solution to this problem would be to use an electrophile at the two ends of the cross-linking reagent that is more selective for a primary amine relative to a hydroxide anion, but this has not been explored systematically. Rather, it has been inadvertently accomplished.

The most versatile cross-linking agent is **glutaraldehyde**:



Unfortunately, the chemistry of its reactions with proteins has never been elucidated. Presumably as an aliphatic aldehyde it engages in all of the same chemical reactions with lysine that occur after aliphatic aldehydes are produced in collagen by protein-lysine 6-oxidase (Figure 3-19). It is part of the lore surrounding glutaraldehyde that the freshly distilled reagent is far less active; this is meant to suggest that compounds derived from glutaraldehyde itself, such as its aldol and dehydrated aldol, are important to the cross-linking. The dehydrated aldol as well as many of the products resulting from the imine formed upon the reaction of glutaraldehyde with a lysine are $\alpha\beta$ unsaturated aldehydes or imines. It is the greater preference of these $\alpha\beta$ unsaturated aldehydes and imines for reaction with the weak base lysine rather than the strong base hydroxide that produces a higher yield of cross-linked product when glutaraldehyde is used. In situations when the cross-linking reaction with dimethyl suberimidate cannot be forced to completion, cross-linking with glutaraldehyde will usually produce the fully cross-linked protein in high yield. The efficiency of glutaraldehyde has permitted it to be used to provide a quantitative catalogue of the various oligomeric complexes present in a heterodisperse solution of a single, pure protein.

Quantitative cross-linking is cross-linking carried to an extent sufficient to connect covalently every subunit in a macromolecular complex to every other subunit in the same macromolecular complex, either directly or indirectly, but not to any subunit in another macromolecular complex. In order for glutaraldehyde to perform quantitative cross-linking, the formation of intermolecular covalent connections between independent, unassociated oligomeric complexes in the solution must be negligible because every covalent complex must represent only the product of intramolecular cross-linking, and every multimeric complex in the solution must be completely cross-linked within itself so that every one of its constituent subunits is covalently attached to all of the others. Cross-linking with appropriately high concentrations of glutaraldehyde often fulfills these requirements.¹⁶⁷⁻¹⁷⁰

This fulfillment can be illustrated with two experiments. In the first experiment, a monodisperse solution containing a homogeneous population of the oligomeric protein L-lactate dehydrogenase, a protein known to be a tetramer composed of four identical subunits, was mixed with glutaraldehyde.¹⁶⁷ The reaction was permitted to proceed a short period of time, and the products were examined by electrophoresis on polyacrylamide gels in the presence of dodecyl sulfate. At high enough concentrations of glutaraldehyde, the only component that was observed contained the number of polypeptides, four, known to be present in the protein, all covalently connected to each other (Figure 8-20).¹⁶⁷ No larger products, which would have resulted from intermolecular cross-linking, and no smaller products, which would have resulted from incomplete intramolecular cross-linking, were observed. In the second experiment, the receptor for epidermal growth factor was cross-linked. This protein forms an α_2 dimer from two α monomers upon the addition of epidermal growth factor to the solution. Before the epidermal growth factor was added, only the monomer of the receptor was seen on the polyacrylamide gel after the protein had been cross-linked with glutaraldehyde and then dissolved in a solution of dodecyl sulfate. When epidermal growth factor was added and samples were removed at different times and then cross-linked, unfolded in a solution of dodecyl sulfate, and submitted to electrophoresis, the un-cross-linked monomer was seen to be gradually but completely replaced by the covalent cross-linked dimer.¹⁷⁰ If any intermolecular cross-linking had been occurring after the glutaraldehyde was added, covalent dimer should have been observed in the absence of epidermal growth factor, but it was not. If incomplete intramolecular cross-linking were occurring after the glutaraldehyde had been added, some un-cross-linked monomer should have remained after completion of the dimerization, but it did not.

There are two possible **problems** that can affect the outcome of quantitative cross-linking. First, if the protein is cross-linked at a concentration that is significantly lower than its physiological concentration, a naturally occurring oligomer may dissociate; or if it is cross-linked at a concentration significantly higher than its physiological concentration, artifactual aggregation may occur. Second, if the electrophilic cross-linking reagent reacts with nucleophilic side chains on the protein that are involved in its function and if its intact function is required for maintenance of its proper oligomerization, the protein could dissociate or artifactually associate because of its chemical inactivation before it becomes cross-linked. For example, glutaraldehyde inactivates the ATPase of the chaperone protein GroEL within 2 s of its application, and the ATPase activity is required to maintain the correct oligomerization of the protein. Controls were devised to demonstrate that the oligomeric state of the protein nevertheless did not change in the time required for quantitative cross-linking to be accomplished.¹⁶⁶

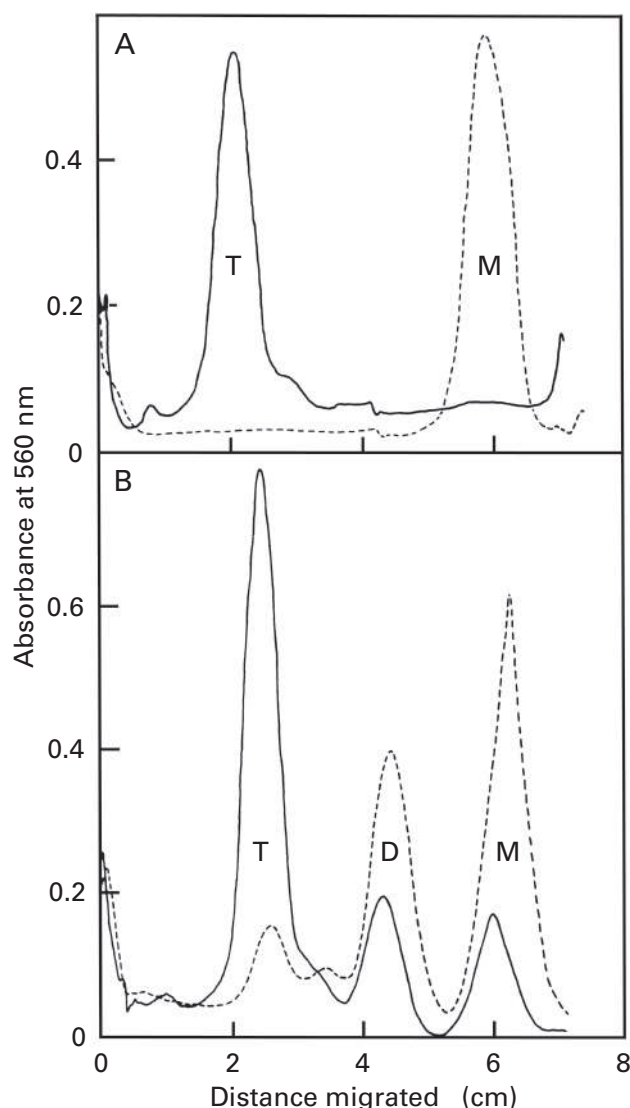


Figure 8-20: Cross-linking of L-lactate dehydrogenase.¹⁶⁷ Samples of L-lactate dehydrogenase were submitted to cross-linking for 2 min with 0.4 M glutaraldehyde at 20 °C and the reaction was terminated by adding sodium dodecyl sulfate to 20 μ M. The samples were then submitted to electrophoresis on polyacrylamide gels, the gels were stained for protein, and the absorbance at 560 nm as a function of the distance migrated is presented. (A) Cross-linking in a solution of native tetramers (20 nM). Equivalent samples were either cross-linked (solid line) or not cross-linked (dashed line). In the former, only fully cross-linked tetramers (T) are observed; in the latter, only un-cross-linked monomers (M). (B) Cross-linking of reassembling L-lactate dehydrogenase after it had been dissociated into subunits. L-Lactate dehydrogenase was dissociated into subunits at pH 2.3 and then diluted into a solution at pH 7.6 (final concentration of subunits 340 nM). After 210 s (dashed line) and 1 h (solid line), samples were removed and complexes (M, monomer; D, dimer; T, tetramer) were catalogued by cross-linking them for 2 min with 0.4 M glutaraldehyde. Reprinted with permission from ref 167. Copyright 1979 *Nature*.

If all of these requirements have been satisfied, quantitative cross-linking can be used to define the **quaternary structure** of a protein. For example, it has been used to show that 1-aminocyclopropane-1-carboxylate synthase is a dimer of two identical subunits.¹⁷¹ The

(α_7)₂ tetradecamer of GroEL, the (α_7)₂ β_7 heterooligomer of GroEL and GroES, and the β_7 (α_7)₂ β_7 heterooligomer of GroEL and GroES are each also quantitatively cross-linked by glutaraldehyde.¹⁷²

While the preceding examples illustrate how glutaraldehyde can be used to define the oligomer present in a monodisperse solution of a protein, it can also be used to provide a catalogue of the **oligomers in a heterodisperse solution**. When Na⁺/K⁺-exchanging ATPase is dissolved in a solution of nonionic detergent, a mixture is formed of various oligomers, ($\alpha\beta$)_{*n*}, which are combinations of the two different subunits, α and β , from which this protein is composed. The fraction of the total protein engaged in each of these oligomers, respectively, could be rapidly and accurately determined by quantitative cross-linking (Figure 8-21).^{169,173} From the pattern displayed in the figure, it could be calculated that before the glutaraldehyde had been added to the solution, the fractions of the protein present as the ($\alpha\beta$)₄ tetramer, the ($\alpha\beta$)₃ trimer, the ($\alpha\beta$)₂ dimer, and the $\alpha\beta$ monomer were 0.25, 0.19, 0.28, and 0.31, respectively.

As was done with the receptor for epidermal growth factor, cross-linking with glutaraldehyde can also be

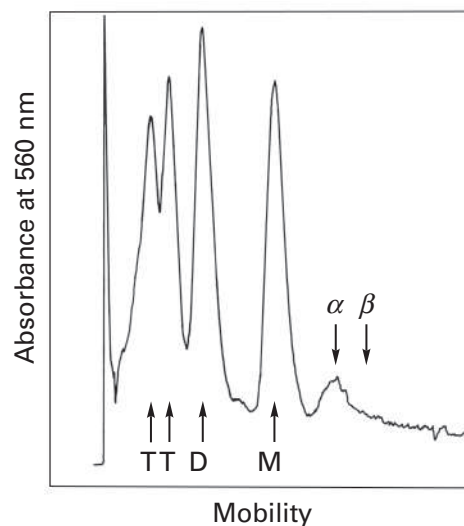


Figure 8-21: Cross-linking of the oligomers in a heterodisperse solution of Na⁺/K⁺-exchanging ATPase.¹⁶⁹ A suspension of biological membranes containing only Na⁺/K⁺-exchanging ATPase was dissolved in a 5 mM solution of the nonionic detergent octa(ethylene glycol) dodecyl ether. After the solution was clarified by centrifugation, glutaraldehyde was added to 8 mM. After 45 min at 22 °C, sodium dodecyl sulfate was added, and the sample was submitted to electrophoresis on a gel cast from 3.6% acrylamide and 0.1% dodecyl sulfate. After the gel was stained, it was scanned for absorbance at 560 nm as a function of the distance migrated. By their mobilities the components were identified as covalently cross-linked $\alpha\beta$ monomer (M), ($\alpha\beta$)₂ dimer (D), ($\alpha\beta$)₃ trimer (T), and ($\alpha\beta$)₄ tetramer (T) of the enzyme, which in its native form is a monomer ($\alpha\beta$) of one α subunit and one β subunit in noncovalent association.¹⁷³ Because the enzyme is a complex of an α subunit and a β subunit, almost no un-cross-linked α polypeptide (α) or β polypeptide (β) remained following the cross-linking. Reprinted with permission from ref 169. Copyright 1982 American Chemical Society.

used to follow the **kinetics of the assembly** of a protein. When L-lactate dehydrogenase is transferred to a solution at pH 2.3, it dissociates into monomers consisting of single subunits, and when it is transferred back to pH 7.6, it reassociates over an hour to its normal tetrameric state. The concentration of monomer, dimer, and tetramer can be ascertained at any given minute by removing a sample and cross-linking it with glutaraldehyde (Figure 8–20B).¹⁶⁷ An extensive study of the kinetics of this stepwise association could be performed in this way.¹⁶⁸ The rate and mechanism of the association of the subunits of the catalytic trimer of aspartate carbamoyltransferase were also monitored by quantitative cross-linking,¹⁷⁴ as well as the conversion of the $(\alpha_7)_2$ tetradecamer of GroEL into the $(\alpha_7)_2\beta_7$ and $\beta_7(\alpha_7)_2\beta_7$ heterooligomers of GroEL and GroES caused by MgATP.¹⁷²

Cross-linking has also been used to determine the **stoichiometric ratio** between two dissimilar subunits. For example, the fact that the α and β polypeptides of succinate-CoA ligase (ADP-forming) from *E. coli* disappear in concert almost entirely during the formation of a covalent $\alpha\beta$ heterodimer and covalent $(\alpha\beta)_2$ heterotetramer (Figure 8–22)¹⁷⁵ means that they must be present in equimolar ratio in the protein. A similar result was observed with the two α and β polypeptides composing Na^+/K^+ -exchanging ATPase.¹⁷⁶ In either of these cases, if the polypeptides had not been present in equimolar ratio, either one polypeptide would have disappeared from its position on the polyacrylamide gel more rapidly than the other during the formation of a covalent $\alpha\beta$ heterodimer or significant amounts of covalent products of the form $\alpha_2\beta$ or $\alpha\beta_2$ would have appeared in addition to the covalent $\alpha\beta$ heterodimer.

The patterns seen in the ladders from partially cross-linked samples of a protein (Figure 8–19) can provide information about the **arrangement of the subunits** in the oligomer. Succinate-CoA ligase (ADP-forming) from *E. coli* is a protein composed of two different subunits, 388 and 288 amino acids in length, each present in two copies. These subunits could have been arranged in at least two ways to produce stoichiometries of either $(\alpha\beta)_2$ or $\alpha_2\beta_2$. The former designation implies that the association between an α subunit and a β subunit is more intimate than that between either two α subunits or two β subunits; the latter implies the reverse. Examples illustrating this distinction would be either a dimer of two identical subunits, each of which was posttranslationally cleaved to produce a pair of subunits, each an entwined α polypeptide and β polypeptide, or a heterotetramer assembled from a fully folded α_2 dimer and a fully folded β_2 dimer, respectively. Succinate-CoA ligase (ADP-forming) was reacted with enough dimethyl suberimidate to cross-link completely the α and β polypeptides among themselves but not enough to produce a high yield of the completely cross-linked product.¹⁷⁵ The covalent $\alpha\beta$ heterodimer was by far the major product of this incomplete reaction. A reasonable yield of the covalent heterote-

trimer, $(\alpha\beta)_2$, was also produced during the reaction, but little or no covalent trimer, covalent $\alpha\alpha$ homodimer, or covalent $\beta\beta$ homodimer was seen (Figure 8–22).

This result demonstrates that the formation of a cross-link between one folded α polypeptide and one folded β polypeptide in native succinate-CoA ligase (ADP-forming) is a far more likely event than the formation of a cross-link between either two α polypeptides or two β polypeptides, and it suggests that α and β polypeptides are more intimately associated with each other than are α polypeptides with α polypeptides or β polypeptides with β polypeptides. This conclusion has been validated by the crystallographic molecular model.¹⁷⁷ Therefore, the proper designation for the arrangement of the subunits in the native, un-cross-linked protein is $(\alpha\beta)_2$.

The yield of heterotrimer on the gel in Figure 8–22 is significantly lower than the yield of heterotetramer. A similar disparity among the products of a partial cross-linking reaction was seen when several tetrameric proteins, each composed of four identical subunits, were examined.¹⁷⁸ L-Lactate dehydrogenase, pyruvate kinase, fructose-bisphosphate aldolase, fumarate hydratase, and catalase were each partially cross-linked with various dimethyl bisimides. In each case, the yield of the covalent trimer was 2–6-fold lower than that of either the covalent dimer or the covalent tetramer. This result is reminiscent of the one seen with succinate-CoA ligase (ADP-forming), and its explanation is the same. Within

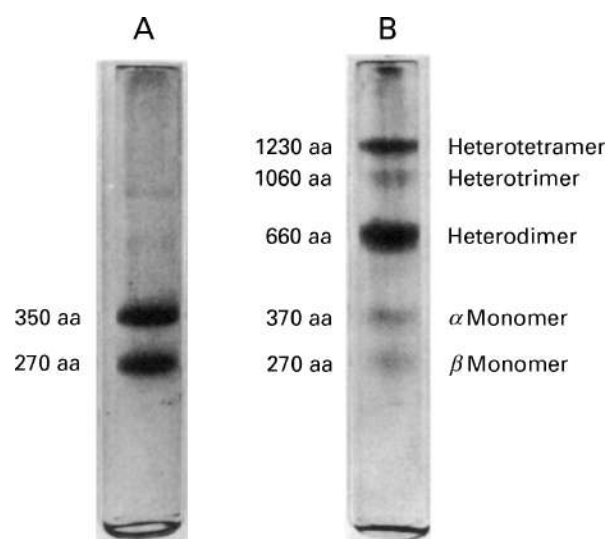


Figure 8–22: Cross-linking of succinate-CoA ligase (ADP-forming) from *E. coli*.¹⁷⁵ A solution of succinate-CoA ligase (1.0 mg mL^{-1}) was cross-linked with dimethyl suberimidate (2.0 mg mL^{-1}) for 30 min. The resulting covalent complexes were dissolved in a solution of sodium dodecyl sulfate and submitted to electrophoresis. (A) Un-cross-linked control showing the α and β polypeptides from which the enzyme is composed. (B) Cross-linked product. The components observed were assigned as the covalent $\alpha\beta$ heterodimer, the covalent heterotrimer, and the covalent $(\alpha\beta)_2$ heterotetramer on the basis of their apparent lengths (numbers to the left of each gel) determined by their mobilities on electrophoresis relative to a set of polypeptides of known length. Reprinted with permission from ref 175. Copyright 1975 *Journal of Biological Chemistry*.

each of these molecules of protein there must be associations between some pairs of α subunits that are more intimate than the associations between other pairs of α subunits. The results suggest that the proper designation for the arrangement of the subunits in each of these proteins is $(\alpha_2)_2$. To understand why this is so, the rules governing the evolution of oligomeric proteins must be understood. These rules are based on rotational axes of symmetry.

Suggested Reading

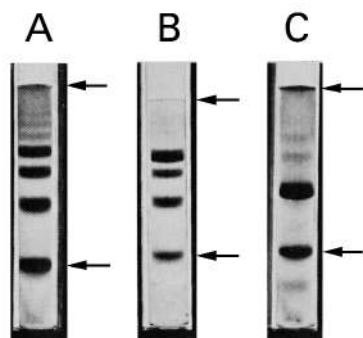
Davies, G.E., & Stark, G.R. (1970) Use of dimethyl suberimide, a cross-linking agent, in studying the subunit structure of oligomeric proteins, *Proc. Natl. Acad. Sci. U.S.A.* 66, 651–656.

Day, P.J., Murray, I.A., & Shaw, W.V. (1995) Properties of hybrid active sites in oligomeric proteins: kinetic and ligand binding studies with chloramphenicol acetyltransferase trimers, *Biochemistry* 34, 6416–6422.

Problem 8–16: Assume that each pure hybrid of fructose-bisphosphate aldolase shown in Figure 8–18⁷⁶ was dissociated completely and reassociated at random during the experiment described in the figure. Refer to the two types of subunits as α and γ .

- What is the respective subunit structure of each of the five purified hybrids?
- Assume the subunits reassemble at random by the binomial formula $a^4 + 4a^3c + 6a^2c^2 + 4ac^3 + c^4$, where a is the fraction of the dissociated subunits that are α and c is the fraction that are γ and predict the ratio of components expected from the reassociation of each of the five dissociated hybrids.

Problem 8–17: The following pictures are of polyacrylamide gels cast in 0.1% sodium dodecyl sulfate on which proteins unfolded with dodecyl sulfate were submitted to electrophoresis.¹⁵⁹



In all three experiments, the native proteins (at concentrations of 0.03–0.2 mM in subunit) were first treated with dimethyl suberimide (at concentrations of 3–7 mM) before they were unfolded with the dodecyl sulfate. In each experiment, the upper arrow marks the top

of the gel, and the lower arrow marks the stained band corresponding to the single polypeptide observed when treatment with dimethyl suberimide was omitted.

- By drawing a graph for each of the gels, A, B, and C, show that none of the products of the reaction between the respective protein and dimethyl suberimide were overlooked.
- What is the stoichiometry of the subunits of the protein run on gel A?
- What is the stoichiometry of the subunits of the protein run on gel B?
- What is the stoichiometry of the subunits of the protein run on gel C?

Assume in making all of these assignments that proteins A, B, and C are homooligomers.

- In making these assignments, you have ignored the minor bands of lower mobility seen in gels A and C. If you are correct in ignoring these bands, what should have happened to these bands if the proteins had been more dilute in concentration when they were treated with the same concentrations of dimethyl suberimide? Why?

References

- Tanford, C. (1961) *Physical Chemistry of Macromolecules*, pp 180–316, Wiley, New York.
- Eisenberg, H. (1976) *Biological Macromolecules and Polyelectrolytes in Solution*, Vol. 6, pp 100–153, Clarendon Press, Oxford, England.
- Moon, Y.U., Anderson, C.O., Blanch, H.W., & Prausnitz, J.M. (2000) *Fluid Phase Equilib.* 168, 229–239.
- McMillan, W.G., & Mayer, J.E. (1945) *J. Chem. Phys.* 13, 276–305.
- Edelhoch, H. (1967) *Biochemistry* 6, 1948–1954.
- Gornall, A.G., Bardawill, C.J., & David, M.M. (1948) *J. Biol. Chem.* 177, 751–766.
- Moczydlowski, E.G., & Fortes, P.A. (1981) *J. Biol. Chem.* 256, 2346–2356.
- Scatchard, G., Batchelder, A.C., & Brown, A. (1946) *J. Am. Chem. Soc.* 68, 2320–2331.
- Van Holde, K.E. (1985) *Physical Biochemistry*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ.
- Eisenberg, H. (1981) *Q. Rev. Biophys.* 14, 141–172.
- Eisenberg, H. (1994) *Biophys. Chem.* 53, 57–68.
- Schachman, H.K., & Edelstein, S.J. (1966) *Biochemistry* 5, 2681–2705.
- Cohen, G., & Eisenberg, H. (1968) *Biopolymers* 6, 1077–1100.
- Reisler, E., & Eisenberg, H. (1969) *Biochemistry* 8, 4572–4578.
- Becerra, S.P., Kumar, A., Lewis, M.S., Widen, S.G., Abbotts, J., Karawya, E.M., Hughes, S.H., Shiloach, J., Wilson, S.H., & Lewis, M.S. (1991) *Biochemistry* 30, 11707–11719.
- Eisenberg, H. (2000) *Biophys. Chem.* 88, 1–9.
- Kharakoz, D.P. (1997) *Biochemistry* 36, 10276–10285.

18. Cohn, E.J., & Edsall, J.T. (1943) *Proteins, Amino Acids and Peptides as Ions and Dipolar Ions*, pp 374–381, Reinhold Publishing Corporation, New York.
19. Brown, J.R. (1976) *Fed. Proc.* 35, 2141–2144.
20. Reynolds, C.M., & Poole, L.B. (2001) *Biochemistry* 40, 3912–3919.
21. Velten, M., Villoutreix, B.O., & Ladjimi, M.M. (2000) *Biochemistry* 39, 307–315.
22. Golbik, R., Naumann, M., Otto, A., Muller, E., Behlke, J., Reuter, R., Hubner, G., & Kriegel, T.M. (2001) *Biochemistry* 40, 1083–1090.
23. LaRonde-LeBlanc, N., & Wolberger, C. (2000) *Biochemistry* 39, 11593–11601.
24. Ebel, C., Eisenberg, H., & Ghirlando, R. (2000) *Biophys. J.* 78, 385–393.
25. Beernink, H.T., & Morrical, S.W. (1998) *Biochemistry* 37, 5673–5681.
26. Xia, J., Sinclair, J.F., Baldwin, T.O., & Lindahl, P.A. (1996) *Biochemistry* 35, 1965–1971.
27. Zondlo, J., Fisher, K.E., Lin, Z., Ducote, K.R., & Eisenstein, E. (1995) *Biochemistry* 34, 10334–10339.
28. Robertson, J.G. (1995) *Biochemistry* 34, 7533–7541.
29. Mechanic, L.E., Hall, M.C., & Matson, S.W. (1999) *J. Biol. Chem.* 274, 12488–12498.
30. Ratcliff, G.C., & Erie, D.A. (2001) *J. Am. Chem. Soc.* 123, 5632–5635.
31. Titus, G.P., Mueller, H.A., Burgner, J., Rodriguez De Cordoba, S., Penalva, M.A., & Timm, D.E. (2000) *Nat. Struct. Biol.* 7, 542–546.
32. Lerman, J.C., Robblee, J., Fairman, R., & Hughson, F.M. (2000) *Biochemistry* 39, 8470–8479.
33. Tennyson, R.B., & Lindsley, J.E. (1997) *Biochemistry* 36, 6107–6114.
34. Musatov, A., & Robinson, N.C. (1994) *Biochemistry* 33, 13005–13012.
35. Debye, P. (1947) *J. Phys. Chem.* 51, 18–32.
36. Einstein, A. (1910) *Ann. Phys. (Berlin)* 33, 1275–1298.
37. Casassa, E.F., & Eisenberg, H. (1964) *Adv. Protein Chem.* 19, 287–395.
38. Zimm, B.H.D. (1948) *J. Chem. Phys.* 16, 1099–1116.
39. Dandliker, W.B. (1954) *J. Am. Chem. Soc.* 76, 6036–6039.
40. Bryan, J.K. (1977) *Anal. Biochem.* 78, 513–519.
41. Edsall, J.T., Edelhoch, H., Lontie, R., & Morrison, P.R. (1950) *J. Am. Chem. Soc.* 72, 4641–4656.
42. Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F., & Whitehouse, C.M. (1989) *Science* 246, 64–71.
43. Kathmann, E.C., Naylor, S., & Lipsky, J.J. (2000) *Biochemistry* 39, 11170–11176.
44. Ronk, M., Shively, J.E., Shute, E.A., & Blake, R.C. (1991) *Biochemistry* 30, 9435–9442.
45. Ullah, J.H., Walsh, T.R., Taylor, I.A., Emery, D.C., Verma, C.S., Gamblin, S.J., & Spencer, J. (1998) *J. Mol. Biol.* 284, 125–136.
46. Musatov, A., & Robinson, N.C. (1994) *Biochemistry* 33, 10561–10567.
47. Yang, G., Sandalova, T., Lohman, K., Lindqvist, Y., & Rendina, A.R. (1997) *Biochemistry* 36, 4751–4760.
48. Fitzgerald, M.C., Chernushevich, I., Standing, K.G., Whitman, C.P., & Kent, S.B. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 6851–6856.
49. Whitelegge, J.P., le Coutre, J., Lee, J.C., Engel, C.K., Prive, G.G., Faull, K.F., & Kaback, H.R. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 10695–10698.
50. Dayhoff, M.O. (1978) *Atlas of Protein Sequence and Structure*, Vol. 5, Suppl. 3, National Biomedical Research Foundation, Silver Spring, MD.
51. Edsall, J.T. (1953) in *The Proteins, Chemistry, Biological Activity, and Methods* (Neurath, H., & Bailey, K. E., Eds.) Vol. I, pp 549–726, Academic Press, New York.
52. Dayhoff, M.O. (1972) *Atlas of Protein Sequence and Structure*, Vol. 5, National Biomedical Research Foundation, Silver Spring, MD.
53. Jaenicke, R., & Knof, S. (1968) *Eur. J. Biochem.* 4, 157–163.
54. Titani, K., Koide, A., Ericsson, L.H., Kumar, S., Hermann, J., Wade, R.D., Walsh, K.A., Neurath, H., & Fischer, E.H. (1978) *Biochemistry* 17, 5680–5693.
55. Seery, V.L., Fischer, E.H., & Teller, D.C. (1967) *Biochemistry* 6, 3315–3327.
56. Dandliker, W.B., & Fox, J.B. (1955) *J. Biol. Chem.* 214, 275–283.
57. Schroeder, W.A., Shelton, J.R., Shelton, J.B., Robberson, B., & Apell, G. (1969) *Arch. Biochem. Biophys.* 131, 653–655.
58. Samejima, T., & Yang, J.T. (1963) *J. Biol. Chem.* 238, 3256–3261.
59. Murthy, M.R., Reid, T.J., Sicignano, A., Tanaka, N., & Rossmann, M.G. (1981) *J. Mol. Biol.* 152, 465–499.
60. Stellwagen, E., & Schachman, H.K. (1962) *Biochemistry* 1, 1056–1068.
61. Kawahara, K., & Tanford, C. (1966) *Biochemistry* 5, 1578–1584.
62. Richter, G.W., & Walker, G.F. (1967) *Biochemistry* 6, 2871–2880.
63. Boyd, D., Vecoli, C., Belcher, D.M., Jain, S.K., & Drysdale, J.W. (1985) *J. Biol. Chem.* 260, 11755–11761.
64. Gerhart, J.C., & Schachman, H.K. (1965) *Biochemistry* 4, 1054–1062.
65. Hoover, T.A., Roof, W.D., Foltermann, K.F., O'Donovan, G.A., Bencini, D.A., & Wild, J.R. (1983) *Proc. Natl. Acad. Sci. U.S.A.* 80, 2462–2466.
66. Hammerstedt, R.H., Meohler, H., Decker, K.A., & Wood, W.A. (1971) *J. Biol. Chem.* 246, 2069–2074.
67. Falcoz-Kelly, F., Janin, J., Saari, J.C., Vaeron, M., Truffa-Bachi, P., & Cohen, G.N. (1972) *Eur. J. Biochem.* 28, 507–519.
68. Katinka, M., Cossart, P., Sibilli, L., Saint-Girons, I., Chalvignac, M.A., Le Bras, G., Cohen, G.N., & Yaniv, M. (1980) *Proc. Natl. Acad. Sci. U.S.A.* 77, 5730–5733.
69. Cassman, M., & Schachman, H.K. (1971) *Biochemistry* 10, 1015–1024.
70. Moon, K., & Smith, E.L. (1973) *J. Biol. Chem.* 248, 3082–3088.
71. Eisenberg, H., & Tomkins, G.M. (1968) *J. Mol. Biol.* 31, 37–49.
72. Johnson, M.S., & Overington, J.P. (1993) *J. Mol. Biol.* 233, 716–738.
73. Power, M.D., Marx, P.A., Bryant, M.L., Gardner, M.B., Barr, P.J., & Luciw, P.A. (1986) *Science* 231, 1567–1572.
74. Marth, J.D., Peet, R., Krebs, E.G., & Perlmutter, R.M. (1985) *Cell* 43, 393–404.
75. Rosenbusch, J.P., & Weber, K. (1971) *J. Biol. Chem.* 246, 1644–1657.

76. Penhoet, E., Kochman, M., Valentine, R., & Rutter, W.J. (1967) *Biochemistry* 6, 2940–2949.
77. Bull, H.B., & Currie, B.T. (1946) *J. Am. Chem. Soc.* 68, 742–745.
78. Tanford, C., Swanson, S.A., & Shore, W.S. (1955) *J. Am. Chem. Soc.* 77, 6414–6421.
79. Guidotti, G. (1967) *J. Biol. Chem.* 242, 3694–3703.
80. Truffa-Bachi, P., Van Rapenbusch, R., Janin, J., Gros, C., & Cohen, G.N. (1968) *Eur. J. Biochem.* 5, 73–80.
81. Halwer, M., Nutting, G.C., & Brice, B.A. (1951) *J. Am. Chem. Soc.* 73, 2786–2790.
82. Tanford, C. (1980) *The Hydrophobic Effect: Formation of Micelles and Biological Membranes*, 2nd ed., Wiley-Interscience, New York.
83. Mysels, K.J., & Princen, L.H. (1959) *J. Phys. Chem.* 63, 1696–1700.
84. Burgess, R.R. (1969) *J. Biol. Chem.* 244, 6168–6176.
85. Pitt-Rivers, R., & Impiombato, F.S. (1968) *Biochem. J.* 109, 825–830.
86. Montserret, R., McLeish, M.J., Bockmann, A., Geourjon, C., & Penin, F. (2000) *Biochemistry* 39, 8362–8373.
87. Weinreb, P.H., Zhen, W., Poon, A.W., Conway, K.A., & Lansbury, P.T., Jr. (1996) *Biochemistry* 35, 13709–13715.
88. Peterson, G.L., & Hokin, L.E. (1981) *J. Biol. Chem.* 256, 3751–3761.
89. Reynolds, J.A., & Tanford, C. (1970) *J. Biol. Chem.* 245, 5161–5165.
90. Shirahama, K., Tsujii, K., & Takagi, T. (1974) *J. Biochem. (Tokyo)* 75, 309–319.
91. Fisher, M.P., & Dingman, C.W. (1971) *Biochemistry* 10, 1895–1899.
92. Banker, G.A., & Cotman, C.W. (1972) *J. Biol. Chem.* 247, 5856–5861.
93. Weber, K., & Osborn, M. (1969) *J. Biol. Chem.* 244, 4406–4412.
94. Davis, B.J. (1964) *Ann. N.Y. Acad. Sci.* 121, 404–427.
95. Ornstein, L. (1964) *Ann. N.Y. Acad. Sci.* 121, 321–349.
96. Laemmli, U.K. (1970) *Nature* 227, 680–685.
97. Kyte, J., & Rodriguez, H. (1983) *Anal. Biochem.* 133, 515–522.
98. Rodbard, D., & Chrambach, A. (1970) *Proc. Natl. Acad. Sci. U.S.A.* 65, 970–977.
99. Cornfield, J., & Chalkley, H.W. (1951) *J. Wash. Acad. Sci.* 41, 226–229.
100. Fawcett, J.S., & Morris, C.J.O.R. (1966) *Sep. Sci.* 1, 9–26.
101. Laurent, T., & Killander, J. (1964) *J. Chromatogr.* 14, 317–330.
102. Morris, C.J.O.R. (1966) *Protides Biol. Fluids* 14, 543–551.
103. Lee, B., & Richards, F.M. (1971) *J. Mol. Biol.* 55, 379–400.
104. Andrews, P. (1965) *Biochem. J.* 96, 595–606.
105. Ogston, A.G. (1958) *Trans. Faraday Soc.* 54, 1754–1757.
106. Cohn, E.J., McMeekin, T.L., Edsall, J.T., & Blanchard, M.H. (1934) *J. Am. Chem. Soc.* 56, 784–794.
107. Sober, H.A. (1970) *CRC Handbook of Biochemistry*, CRC Press, Cleveland, OH.
108. Ackers, G.K. (1964) *Biochemistry* 3, 723–730.
109. Siegel, L.M., & Monty, K.J. (1966) *Biochim. Biophys. Acta* 112, 346–362.
110. Potschka, M., Nave, R., Weber, K., & Geisler, N. (1990) *Eur. J. Biochem.* 190, 503–508.
111. Hedrick, J.L., & Smith, A.J. (1968) *Arch. Biochem. Biophys.* 126, 155–164.
112. Tanford, C. (1968) *Adv. Protein Chem.* 23, 121–282.
113. Fish, W.W., Mann, K.G., & Tanford, C. (1969) *J. Biol. Chem.* 244, 4989–4994.
114. Kumagai, I., Pieler, T., Subramanian, A.R., & Erdmann, V.A. (1982) *J. Biol. Chem.* 257, 12924–12928.
115. Shapiro, A.L., Viñuela, E., & Maizel, J.V. (1967) *Biochem. Biophys. Res. Commun.* 28, 815–820.
116. Lumpkin, O.J. (1982) *Biopolymers* 21, 2315–2316.
117. Lumpkin, O.J., Daejardin, P., & Zimm, B.H. (1985) *Biopolymers* 24, 1573–1593.
118. Williams, J.G., & Gratzer, W.B. (1971) *J. Chromatogr.* 57, 121–125.
119. Swank, R.T., & Munkres, K.D. (1971) *Anal. Biochem.* 39, 462–477.
120. Lindstrom, J., Cooper, J., & Tzartos, S. (1980) *Biochemistry* 19, 1454–1458.
121. Weber, K., Pringle, J.R., & Osborn, M. (1972) *Methods Enzymol.* 26 (Part C), 3–27.
122. Henderson, E.J., & Zalkin, H. (1971) *J. Biol. Chem.* 246, 6891–6898.
123. Doolittle, R.F. (1981) *Science* 214, 149–159.
124. Ingram, V.M. (1957) *Nature* 180, 326–328.
125. Pauling, L., Itano, H.A., Singer, S.J., & Wells, I.C. (1949) *Science* 110, 543–548.
126. Ingram, V.M. (1958) *Biochim. Biophys. Acta* 28, 539–545.
127. Vandekerckhove, J., & Weber, K. (1978) *Proc. Natl. Acad. Sci. U.S.A.* 75, 1106–1110.
128. Vandekerckhove, J., & Weber, K. (1978) *Eur. J. Biochem.* 90, 451–462.
129. Hubert, J.J., Schenk, D.B., Skelly, H., & Leffert, H.L. (1986) *Biochemistry* 25, 4156–4163.
130. Cleveland, D.W., Fischer, S.G., Kirschner, M.W., & Laemmli, U.K. (1977) *J. Biol. Chem.* 252, 1102–1106.
131. Fullmer, C.S., & Wasserman, R.H. (1979) *J. Biol. Chem.* 254, 7208–7212.
132. Shevchenko, A., Wilm, M., Vorm, O., & Mann, M. (1996) *Anal. Chem.* 68, 850–858.
133. Zhang, X., Herring, C.J., Romano, P.R., Szczepanowska, J., Brzeska, H., Hinnebusch, A.G., & Qin, J. (1998) *Anal. Chem.* 70, 2050–2059.
134. Dreger, M., Otto, H., Neubauer, G., Mann, M., & Hucho, F. (1999) *Biochemistry* 38, 9426–9434.
135. Lindstrom, J., Merlie, J., & Yogeewaran, G. (1979) *Biochemistry* 18, 4465–4470.
136. Luna, E.J., Kidd, G.H., & Branton, D. (1979) *J. Biol. Chem.* 254, 2526–2532.
137. Schuppan, D., Cantaluppi, M.C., Becker, J., Veit, A., Bunte, T., Troyer, D., Schuppan, F., Schmid, M., Ackermann, R., & Hahn, E.G. (1990) *J. Biol. Chem.* 265, 8823–8832.
138. Grant, G.A., & Bradshaw, R.A. (1978) *J. Biol. Chem.* 253, 2727–2731.
139. Kern, D., Potier, S., Boulanger, Y., & Lapointe, J. (1979) *J. Biol. Chem.* 254, 518–524.
140. Lundell, D.J., & Howard, J.B. (1978) *J. Biol. Chem.* 253, 3422–3426.
141. Green, N.M., Valentine, R.C., Wrigley, N.G., Ahmad, F., Jacobson, B., & Wood, H.G. (1972) *J. Biol. Chem.* 247, 6284–6298.

142. Zwolinski, G.K., Bowien, B.U., Harmon, F., & Wood, H.G. (1977) *Biochemistry* 16, 4627–4637.
143. James, G.T., & Notmann, E.A. (1973) *J. Biol. Chem.* 248, 730–737.
144. Hall, C.L., & Kamin, H. (1975) *J. Biol. Chem.* 250, 3476–3486.
145. McKean, M.C., Beckmann, J.D., & Frerman, F.E. (1983) *J. Biol. Chem.* 258, 1866–1870.
146. Chuang, M., Ahmad, F., Jacobson, B., & Wood, H.G. (1975) *Biochemistry* 14, 1611–1619.
147. Recsei, P.A., Huynh, Q.K., & Snell, E.E. (1983) *Proc. Natl. Acad. Sci. U.S.A.* 80, 973–977.
148. van Poelje, P.D., & Snell, E.E. (1990) *Biochemistry* 29, 132–139.
149. Baldwin, E.T., Bhat, T.N., Gulnik, S., Hosur, M.V., Sowder, R.C.n., Cachau, R.E., Collins, J., Silva, A.M., & Erickson, J.W. (1993) *Proc. Natl. Acad. Sci. U.S.A.* 90, 6796–6800.
150. Olson, T.S., Bamberger, M.J., & Lane, M.D. (1988) *J. Biol. Chem.* 263, 7342–7351.
151. Guchhait, R.B., Zwergel, E.E., & Lane, M.D. (1974) *J. Biol. Chem.* 249, 4776–4780.
152. Guchhait, R.B., Polakis, S.E., Dimroth, P., Stoll, E., Moss, J., & Lane, M.D. (1974) *J. Biol. Chem.* 249, 6633–6645.
153. Song, C.S., & Kim, K.H. (1981) *J. Biol. Chem.* 256, 7786–7788.
154. Robertson, D.C., Hammerstedt, R.H., & Wood, W.A. (1971) *J. Biol. Chem.* 246, 2073–2081.
155. Day, P.J., Murray, I.A., & Shaw, W.V. (1995) *Biochemistry* 34, 6416–6422.
156. Meighen, E.A., & Schachman, H.K. (1970) *Biochemistry* 9, 1163–1176.
157. Cooper, E., Couturier, S., & Ballivet, M. (1991) *Nature* 350, 235–238.
158. Singer, S.J. (1959) *Nature* 183, 1523–1524.
159. Davies, G.E., & Stark, G.R. (1970) *Proc. Natl. Acad. Sci. U.S.A.* 66, 651–656.
160. Lambert, J.M., Jue, R., & Traut, R.R. (1978) *Biochemistry* 17, 5406–5416.
161. Thevenin, B.J., Shahrokh, Z., Williard, R.L., Fujimoto, E.K., Kang, J.J., Ikemoto, N., & Shohet, S.B. (1992) *Eur. J. Biochem.* 206, 471–477.
162. Brown, K.C., Yu, Z., Burlingame, A.L., & Craik, C.S. (1998) *Biochemistry* 37, 4397–4406.
163. Carlsson, J., Drevin, H., & Axen, R. (1978) *Biochem. J.* 173, 723–737.
164. Itoh, Y., Cai, K., & Khorana, H.G. (2001) *Proc. Natl. Acad. Sci. U.S.A.* 98, 4883–4887.
165. Thorner, J.W., & Paulus, H. (1971) *J. Biol. Chem.* 246, 3885–3894.
166. Azem, A., Weiss, C., & Goloubinoff, P. (1998) *Methods Enzymol.* 290, 253–268.
167. Hermann, R., Rubolph, R., & Jaenicke, R. (1979) *Nature* 277, 243–245.
168. Hermann, R., Jaenicke, R., & Rudolph, R. (1981) *Biochemistry* 20, 5195–5201.
169. Craig, W.S. (1982) *Biochemistry* 21, 2667–2674.
170. Canals, F. (1992) *Biochemistry* 31, 4493–4501.
171. White, M.F., Vasquez, J., Yang, S.F., & Kirsch, J.F. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 12428–12432.
172. Azem, A., Kessel, M., & Goloubinoff, P. (1994) *Science* 265, 653–656.
173. Craig, W.S. (1982) *Biochemistry* 21, 5707–5717.
174. Burns, D.L., & Schachman, H.K. (1982) *J. Biol. Chem.* 257, 8638–8647.
175. Teherani, J.A., & Nishimura, J.S. (1975) *J. Biol. Chem.* 250, 3883–3890.
176. Craig, W.S., & Kyte, J. (1980) *J. Biol. Chem.* 255, 6262–6269.
177. Fraser, M.E., James, M.N., Bridger, W.A., & Wolodko, W.T. (1999) *J. Mol. Biol.* 285, 1633–1653.
178. Hucho, F., Meullner, H., & Sund, H. (1975) *Eur. J. Biochem.* 59, 79–87.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Chapter 9

Symmetry

The arrangement in space of the subunits in a multimeric protein is its **quaternary structure**. Many multimeric proteins are composed of multiple copies of only one particular subunit. These subunits, each necessarily identical to the others when free in solution, combine together to form the final molecule of protein. In such homomultimeric proteins, each of the subunits in the crystallographic molecular model can be formally designated a protomer¹ of the final overall structure. A **protomer** of a multimeric protein is the smallest portion of that protein from copies of which its entire quaternary structure is created. Consequently, all of the protomers in the protein must be the same.

Some multimeric proteins, like the isoenzymatic hybrids of fructose-bisphosphate aldolase (Figure 8–18), are composed of two or more distinct subunits that are, nevertheless, each the offspring of the same common ancestor. Although different in amino acid sequence and in the atomic details of their tertiary structure, these subunits are still related closely enough to participate together to form the complete heteromultimeric protein, much as identical protomers would participate in a homomultimeric protein. In such instances, each of the individual subunits, although actually different, can be considered at low resolution to be one of the indistinguishable protomers forming the overall structure. In contrast to the hybrids formed by the isoenzymatic subunits of fructose-bisphosphate aldolase, other proteins built from such similar but distinct subunits usually incorporate those subunits in unvarying ratios. For example, hemoglobin is always an $(\alpha\beta)_2$ tetramer and acetylcholine receptor is always an $\alpha_2\beta\gamma\delta$ pentamer, even though the α and β subunits of hemoglobin or the α , β , γ , and δ subunits of acetylcholine receptor are respectively homologous to each other and necessarily superposable. These exclusive stoichiometries are established by the distinct atomic interactions between the subunits that take place as the multimer assembles.

Some multimeric proteins contain two or more dissimilar, unrelated subunits. For example, aspartate carbamoyltransferase contains α subunits and β subunits that are unrelated to each other. Even though the subunits composing such a protein are completely different, the final structure produced, when observed as a crystallographic molecular model, can often be divided formally into identical protomers, each containing one copy of each of the different subunits. A particular number of

these heterooligomeric protomers arranged in an array unique to that protein makes up its quaternary structure. Because every protomer is identical to all the others, the arrangement of the protomers of such a heteromultimeric protein is formally equivalent to the arrangement of the subunits of a homomultimeric protein.

Each of the multimeric proteins discussed so far can be divided formally into a set of identical protomers or almost identical subunits. Aside from a few peculiar exceptions, the rules that govern the way these protomers or these subunits are arranged in space to produce the complete molecular structure of the entire protein seem to be the same. In an oligomeric protein formed from a fixed number of identical protomers, those protomers are arrayed around rotational axes of symmetry. In an oligomeric protein formed from a fixed number of homologous but nonidentical subunits, those subunits are arrayed around rotational axes of pseudosymmetry. In a polymeric protein with an indefinite number of identical protomers or homologous subunits, those protomers or those subunits are arrayed around a screw axis of symmetry or a screw axis of pseudosymmetry, respectively.

There is also a set of multimeric proteins composed of nonidentical subunits that are assembled haphazardly with no regard, or sometimes a slight regard, for symmetry. In these proteins, the various subunits are associated with each other by interfaces that, other than their lack of symmetry, are almost indistinguishable in their details from those holding together symmetric proteins. Such asymmetric, heteromultimeric proteins are much more likely to assemble and disassemble under different situations, and such alternations in quaternary structure are often involved intimately in their function. There are also members of this set of heteromultimers, however, that have stable quaternary structures. Unlike symmetric homomultimeric proteins and symmetric heteromultimeric proteins, asymmetric heteromultimeric proteins seem to be cobbled together in the absence of any set of rules.

Rotational and Screw Axes of Symmetry

The **fundamental symmetry operations** that are available to asymmetric objects such as the protomers of a protein when they assemble into a homomultimeric

structure are those around rotational axes and screw axes. If a protein is constructed with rotational symmetry, only a finite number of protomers produce the final structure. If a protein is assembled with screw symmetry, a potentially infinite number of protomers usually can combine to produce a polymer of indefinite length. In a few isolated instances observed so far, a protein, although it is assembled from subunits arranged with screw symmetry, is nevertheless forced to have only a finite number of protomers in the final structure. Such finite structures assembled with screw symmetry are thought to be very rare.

Consider two proteins that illustrate the observation that multimeric proteins constructed from identical protomers are assembled around rotational or screw axes of symmetry. The proteins are malate dehydrogenase from *Aquaspirillum arcticum* (Figure 9-1A),² a dimer built from two identical protomers that are each folded polypeptides 329 amino acids in length, and actin (Figure 9-1B,C),³⁻⁶ a protein that forms helical polymeric fibers of indefinite length with each fiber built from many identical protomers, each of which is the same folded polypeptide 375 amino acids in length.⁴

A **rotational axis of symmetry** is a line about which a structure can be rotated by $360^\circ/n$, where n is an integer larger than 1, to superpose upon itself. An exact 2-fold rotational axis of symmetry runs through the center of the α_2 dimer in the crystallographic molecular model of malate dehydrogenase. If the two subunits of malate dehydrogenase, still held together by the same interface, had been distorted intramolecularly by the structure of the protein itself into significantly different conformations or if they had had different sequences but were nevertheless homologous to each other, a rotational axis of pseudosymmetry would superpose one of them upon the other. A **rotational axis of pseudosymmetry** is a line about which the structure of a protein can be rotated by $360^\circ/n$ to superpose upon each other subunits with identical sequence but significantly different conformation, subunits of different sequence but the same common fold, or distinct but homologous domains.

The value of n is the **fold of the symmetry**. The 2-fold rotational axis of symmetry in the crystallographic molecular model of malate dehydrogenase is a line perpendicular to the plane of the page in Figure 9-1A passing through the center of the molecule of protein. If the image of the upper protomer in the figure is rotated 180° about this axis, it superposes exactly on the lower protomer. Because of this rotational axis of symmetry, the two protomers in the protein are indistinguishable.

Through the center of the indefinitely long **helical polymer** of actin, of which only a segment is drawn in Figure 9-1B, runs a screw axis of symmetry. A **screw axis of symmetry** is a line, passing through a structure, about which the structure can be rotated by an angle between -180° and 180° and along which the structure can be simultaneously translated to superpose upon itself. In

the molecular model of the actin filament, the screw axis of symmetry is a vertical line parallel to the plane of the page. If the image of any protomer is transposed by being rotated -166° (left-handed) around this axis while it is lifted 2.8 nm in a direction parallel to it, it superposes on the next protomer in the helical polymer. Because this operation can be repeated indefinitely, all of the protomers are indistinguishable from each other except by their place in line.

A rotational axis of symmetry (Figure 9-1A) is necessarily 2-fold, 3-fold, 4-fold, and so forth. This requirement arises from the fact that as one of the images of a protomer is being rotated to superpose it on its neighbor, all of the images of its neighbors are also simultaneously being rotated (Figure 9-1A). When the rotation is completed, each of the images must superpose on its respective partner. This can be accomplished only if the protomers are arrayed about the axis at angles to each other that are **integral quotients** of 360° ($360^\circ/n$). The integer defines the number of times the rotation can be accomplished before returning to the beginning. The rotational axis of symmetry within malate dehydrogenase is a 2-fold rotational axis of symmetry. After two superpositions, the original locations are regained. For the rotation to superpose the images of all protomers simultaneously each time, no translation along the axis can occur.

Screw axes of symmetry are defined by a rotation through any angle and a translation of any distance, and they are designated as left-handed or right-handed by the same rules as apply to α helices. If the translation is not zero, a screw axis of symmetry produces a helical array. By the principle that the majority rules, right-handed screws are given a positive sign and left-handed screws are given a negative sign. For example, one of the screw axes of symmetry in the actin polymer, the left-handed one, has a rotational angle of -166° . Trivially, it is also possible to generate an actin polymer with a right-handed screw axis of $+194^\circ$ but, less trivially, by two coaxial right-handed screw axes of $+28^\circ$. A helical array always comes with such a set of different coaxial screw axes. A rotational axis of symmetry is simply a special case of a screw axis of symmetry where the translation is zero and the rotational angle is required to assume values that are integral quotients of 360° .

A screw axis of symmetry, other than the special case of a rotational axis of symmetry, in which translation cannot occur, does not require the angular steps to be integral quotients of 360° . As the image of one of the protomers is rotating and rising, the point of superposition need not occur at any particular angular disposition along the helix produced by the screw axis. There is, however, one requirement that limits the angles and translations permitted a screw axis of symmetry. Its operation is most readily observed by comparing the crystallographic molecular models of the protomer of protocatechuate 3,4-dioxygenase from *Pseudomonas putida* (Figure

Figure 9-1: α -Carbon diagrams drawn from crystallographic molecular models of malate dehydrogenase from *A. arcticum*² and actin from *Oryctolagus cuniculus*.³ (A) Malate dehydrogenase is constructed from two subunits, one drawn with thicker line segments than the other. The view is down a crystallographic 2-fold rotational axis of symmetry in the space group $P2_12_12$ of the array. This drawing was produced with MolScript.⁴⁸⁵ (B) A model of the actin filament was constructed by placing individual actin monomers, a crystallographic molecular model for which is available,⁴ in positions and orientations indicated by the map of electron density calculated from an X-ray fiber diffraction pattern of a gel of oriented actin filaments.³ The orientation and position of the actin monomer that were assigned in this way are in agreement with the orientation and position of the protein encoded by the *mreB* gene of *Escherichia coli*, a homologue of actin, within its filament, which serendipitously happens to be present in a crystal of this protein.⁶ There are eight actin monomers in the segment of the filament displayed in this figure. The circles within each monomer designate the location of a bound Ca^{2+} ion. This drawing was produced with MolScript.⁴⁸⁵ The atomic coordinates on which the drawing is based were provided by Ken Holmes. (C) A low-resolution molecular model of the actin filament, calculated by image reconstruction of electron micrographs of ordered actin filaments,⁵ is included, at the same scale, to illustrate more clearly the packing of the monomers along the helix. Reprinted with permission from ref 5. Copyright 1983 Academic Press.

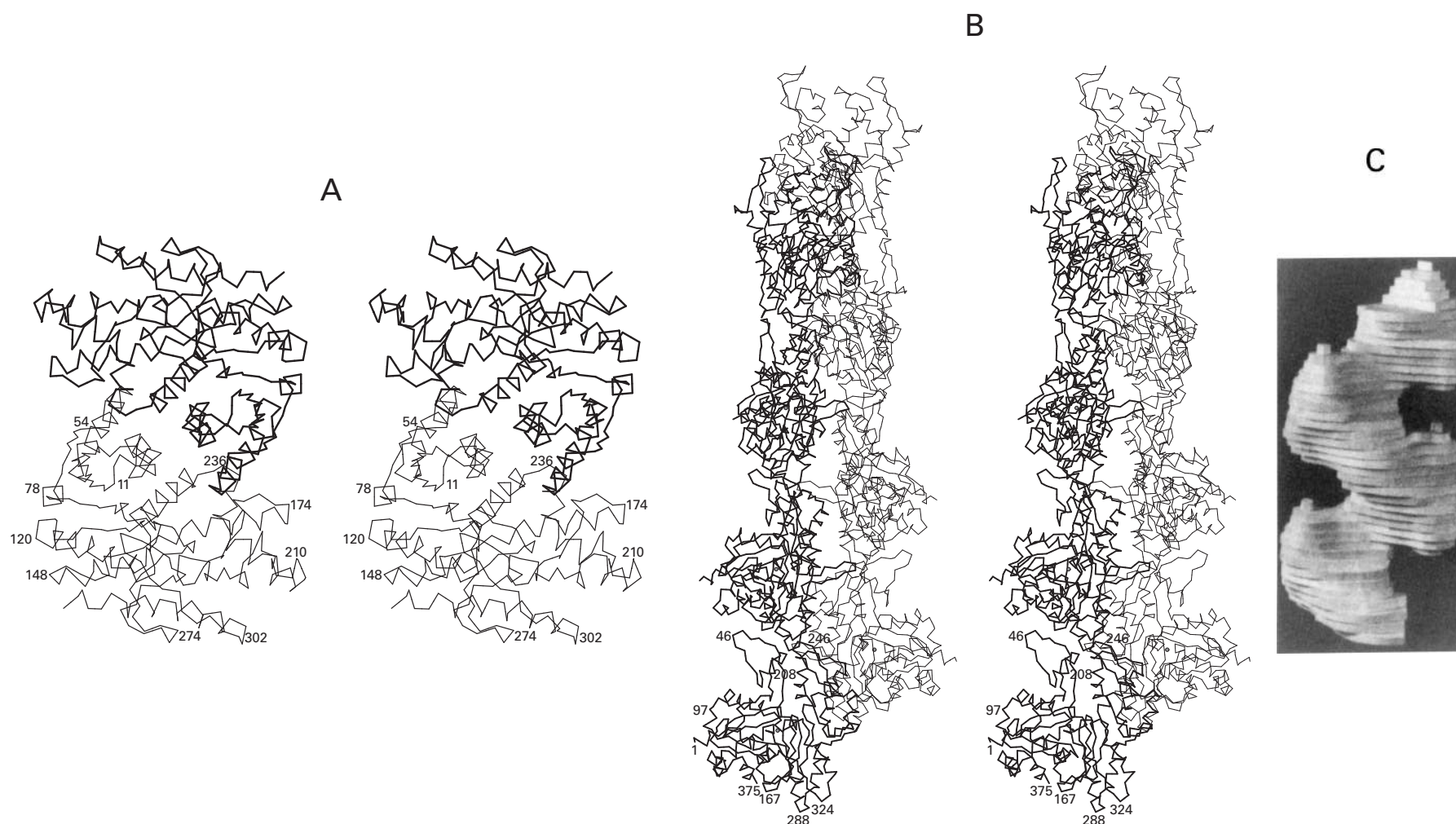
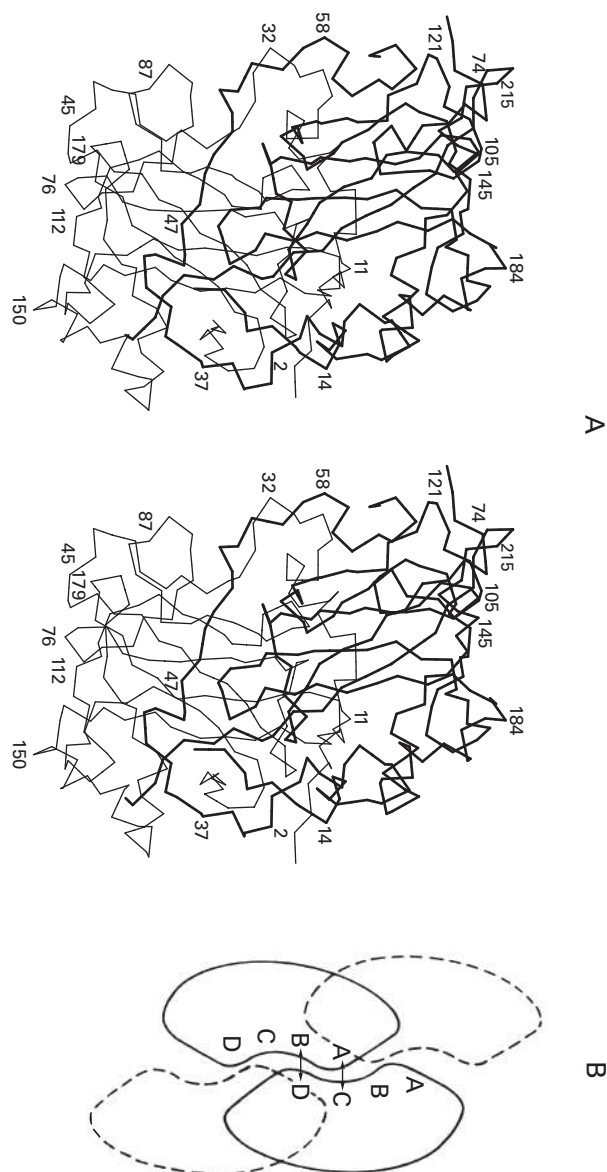


Figure 9-2: (A) α -Carbon diagram drawn from the crystallographic molecular model of the protomer of protococatechuate 3,4-dioxygenase from *P. putida*.^{7,8} The thickness of the line segments connecting the positions of the α -carbons differs between the two homologous subunits that are arrayed around the screw axis of pseudosymmetry. The α subunit is drawn with the thin line; the β subunit, with the thick line. The screw axis of pseudosymmetry is horizontal, in the plane of the page passing through the center of the protomer. The two different subunits have different sequences but homologous folds, except for an extension at the amino terminus in the β subunit of about 35 aa. Homologous positions in the respective subunits are numbered, with the equivalent numbers for the β subunit always about 35 greater than those for the α subunit. This drawing was produced with MolScript.⁴⁶⁵ (B) Schematic arrangement of the array of intersubunit interactions in a closed structure built around a screw axis of symmetry where there is translation. The dashed outlines indicate where adjacent protomers would add but for the steric effect. Reprinted with permission from ref 9.



9-2A)^{7,8} and actin (Figure 9-1B). The protomer of protococatechuate 3,4-dioxygenase is composed of two polypeptides of different sequence, but they are clearly homologous to each other because they share a common fold. Through the center of the $\alpha\beta$ heterodimer of protococatechuate 3,4-dioxygenase (Figure 9-2A) there runs a screw axis of pseudosymmetry. It is a horizontal line parallel to the plane of the page. If the image of the upper, β subunit is transposed by being rotated $+169.2^\circ$ around this axis as it is simultaneously shifted to the right 0.674 nm in a direction parallel to the axis, it superposes on the lower, α subunit.⁸

A helical polymer of actin can be constructed by placing one protomer upon an origin, properly oriented with respect to the axis of symmetry; transposing its image around the axis by -166° and along the axis by 2.8 nm; placing another protomer at this next location; and repeating this process indefinitely. Suppose this were attempted with the common subunit for protococatechuate 3,4-dioxygenase. Place an α subunit upon an origin, properly oriented with respect to the axis. Move the image of this subunit $+169.2^\circ$ around the axis and 0.674 nm along the axis. Place a β subunit at this next location. Move the image of this β subunit $+169.2^\circ$ around the axis and 0.674 nm along the axis and try to place an α subunit at this third location. This third subunit would have to overlap the first. The problem here is that the translation of the screw axis in protococatechuate 3,4-dioxygenase is insufficient to move the image of the subunit far enough along to clear the protomer next to it as the helical path completes the turn (Figure 9-2B).⁹ Therefore, two subunits can be combined to form a heterodimer, but three or more cannot be combined to form a polymer. The parameters of the helix generating actin do not cause such collisions. Actin forms a helical polymer of indefinite length, while protococatechuate 3,4-dioxygenase is limited to a dimer. In an additional example that reinforces the concept, two identical subunits of hexokinase are arranged in crystals of the space group $P2_12_12_1$ along a screw axis of pseudosymmetry the translation of which, 1.38 nm, is so short and the rotation of which, 156° , is so large that the protein is also limited to a dimer.⁹

In theory, the same requirement restricting protococatechuate 3,4-dioxygenase and hexokinase to be dimers also restricts malate dehydrogenase to being a dimer. Because no translation occurs along the axis in malate dehydrogenase, the third protomer would completely intersect the first if the game just played with protococatechuate 3,4-dioxygenase were repeated with malate dehydrogenase. It is the inescapable and obvious rules of this game that cause protococatechuate 3,4-dioxygenase and malate dehydrogenase to be closed structures¹ and actin to be an open structure. A **closed structure** is a structure that contains a finite number of protomers distributed by rotational axes of symmetry or a screw axis of symmetry and to which the addition of further protomers by the

one or more of these symmetry operations is precluded. An **open structure** is a structure built upon a screw axis of symmetry to which protomers can be infinitely added by that symmetry operation. An **oligomeric protein** is a multimeric protein with a closed structure of a fixed number of protomers, and a **polymeric protein** is a multimeric protein with an open structure of an indefinite number of protomers and of an indefinite length.

Each multimeric protein composed of identical protomers, even a helical polymer of indefinite length, can be considered to be the manifestation of a set of interfaces between those protomers. Each of these **interfaces** includes all of the points of contact that lie between two protomers in the structure, and each is formed by the association of two **complementary faces**, one from each of the two protomers. These faces are particular regions on the respective surfaces of the two associating protomers. Because the structures of the protomers are identical to each other, each necessarily possesses on its own surface all of the unique faces forming the interfaces found in the complete molecule. The interface between the two subunits of malate dehydrogenase (Figure 9-1A) is formed from two identical faces, one on the surface of each of its subunits. Note the particularly intimate contact across the interface as the secondary structures, which mimic each other across the axis of symmetry, interdigitate.

Following its biosynthesis, a polypeptide folds to form a structure capable of recognizing and being recognized by other folded polypeptides. If it is to combine with its twins to form a multimeric protein constructed from identical subunits, it must do so in a series of individual steps, and each step must involve the formation of an interface from two complementary faces. The atomic contacts within each of these consecutively formed interfaces are as specific as the atomic contacts throughout the protein. For the same reasons that a folded polypeptide assumes a precise and unique atomic structure, the interface between two subunits has a precise and unique atomic structure.

If, as the result of evolution, a face appears anywhere on the surface of a folded polypeptide and a face complementary to the first appears anywhere else on the surface of the same folded polypeptide, the face on one copy of that folded polypeptide will associate with its complement on another copy of the same folded polypeptide. Any such **random association** between any two identical asymmetric objects always defines a unique screw axis, an angle of rotation about that screw axis, and a translation along that screw axis that will superpose the image of one of the asymmetric objects upon the other. Either these three parameters are consistent with an open structure or they are consistent with a closed structure. If they are consistent with an open structure, the very fact that the one interface can form means that many others will subsequently form. A series of such interfaces is a helical polymer.

The interfaces that are repeated throughout a multimer composed of identical protomers are the origin of the geometry of the final structure. Because they are created by **evolution**, their appearance is determined by a completely random process. As time passes, variation in the identity of the amino acids on the surface of a given monomeric protein occurs. At some point, in some organism, in some species, a constellation of amino acids appears on the surface that permits a stable interface to form between two of these identical monomers.

Natural selection operates at this point. It takes little imagination to realize that if the vast majority of multimeric proteins were not closed structures, the cell would rapidly fill with helical polymers and become a solid, inflexible object incapable of the pliability essential to life. The difficulties encountered with helical polymers of hemoglobin S in sickled erythrocytes dramatically illustrate this problem.¹⁰ An example of a polymerization that is undesirable for a different reason is that of a tRNA-intron endonuclease from *Archaeoglobus fulgidus*, which can form a helical polymer that is enzymatically inactive because the active site is sterically blocked by neighboring monomers in the polymer.¹¹ If a monomeric protein were to sustain a series of mutations dictating that it combine with its twins in such a way that a polymeric fiber necessarily results, this set of mutations would probably be eliminated by natural selection. Mutations leading to closed structures, however, aside from lowering the osmotic pressure of the cytoplasm, may be neutral initially, but oligomeric proteins have potentials denied to monomeric proteins, and the appearance of an oligomeric protein during evolution is the first step in the eventual exploitation of these potentials. Nevertheless, if the interface is compatible with a closed structure, it can initially be fixed by genetic drift as a neutral variation. With its fixation within that species, the protein has become an oligomer of identical protomers.

There remains one perplexing fact. As in the case of malate dehydrogenase (Figure 9-1A), the vast majority of homomultimeric proteins that have been examined conclusively are built around rotational axes of symmetry. Often, as in the case of malate dehydrogenase, these rotational axes of symmetry can be proven to be exact; they always seem to be exact. In fact, protocatechuate 3,4-dioxygenase (Figure 9-2) is one of the few exceptions to this rule, and it is not even a homodimer. If rotational axes of symmetry are no more than severely restricted cases of screw axes of symmetry, and if a screw axis of symmetry is compatible with a closed structure, why are multimeric proteins almost always rotationally symmetric?

The main difference between a dimer like malate dehydrogenase and a dimer like protocatechuate 3,4-dioxygenase lies in the respective interfaces defining these structures. In a rotationally symmetric dimer such as malate dehydrogenase, individual interactions between the two protomers come in **sets of identical**

pairs. The best way to see this is to consider the α helices containing Isoleucines 59 on either side of the rotational axis of symmetry (Figure 9–1A). These α helices and their carboxy-terminal loops of random meander insert into pockets in the opposite protomers. Consider a specific position such as Isoleucine 59 in the segment of the sequence of the lower protomer forming this insertion. The amino acid at this location is making several contacts with amino acids in the pocket in the upper protomer. Suppose a mutation occurred at position 59 that strengthened these interactions by a certain increment. Because the upper protomer was read from the same gene, at sequence position 59 in its α helix the same favorable change would occur automatically so the increment for the increase in stability for the whole interface would be twice that of each individual increment. The same argument could be made for each location in the interface.

In a protein built on a screw axis of symmetry, however, the amino acids at the same sequence positions in the two protomers never interact with the same amino acids from the other protomer across the screw axis of symmetry (Figure 9–2B). A mutation, occurring anywhere in the interface, that adds an increment of stability to the dimer is not duplicated automatically. It follows that as variation proceeds during evolution, the incremental changes that occur within an interface built around a 2-fold rotational axis of symmetry are amplified 2-fold relative to those that occur within each interface around a screw axis. This conclusion is valid whether one of these interfaces has already appeared during evolution or is merely incipient.

The formation of an interface between two identical monomeric proteins, which is the evolutionary event that precedes the appearance of a multimeric protein, is not an all or none phenomenon. The chemical reaction in question is



Associated with this reaction is a change in standard free energy, and it is this change in standard free energy that determines the extent of the reaction. The numerical value of this change in standard free energy is dictated by the particular interactions that occur among the amino acids within the interface. The particular interactions that occur are the product of evolution by natural selection. Each explicit variation in one of these interactions adds or subtracts an increment of free energy to the overall change. If the increments are automatically doubled, overall decreases in the standard free energy change for the reaction proceed more rapidly over evolutionary time.

Incremental decreases in the standard free energy change, however, are also doubled. Although rotationally symmetric dimers should appear more frequently, they should also disappear more frequently, unless they rep-

resent **advantageous variations**. Improvements in a certain protein are retained by natural selection, and their retention is unaffected by the frequency with which retrograde changes arise. Mutations turning the dimer back into a monomer, such as those that can be performed experimentally,^{12,13} would be eliminated by natural selection if they were disadvantageous. It is possible that oligomerization of a protein has an immediate advantageous effect. If it did, the fact that most oligomeric proteins are built around 2-fold rotational axes of symmetry would be a reflection of the fact that such oligomers arise with a high frequency and of the fact that they are fixed by natural selection because oligomers are advantageous.

Because events were discussed in the opposite order of the normal progression, a summary of the historical and logical sequence seems appropriate. As a result of genetic variation among the individuals in a given species, a constellation of amino acids appears on the surface of a monomeric protein within one of those individuals. The constellation causes molecules of that previously monomeric protein to associate with each other. This association necessarily creates an interface. This interface necessarily forces the two protomers it brings together to be related to each other by a unique screw axis of symmetry. The association between the two or more protomers created by this screw axis of symmetry is tested by natural selection. Occasionally, a helical polymer, which necessarily results from a screw axis that is not closed, is advantageous and is retained. Most of the time, however, the survivors of natural selection are closed oligomeric proteins the interfaces of which dictate rotational axes of symmetry.

Suggested Reading

Monod, J., Wyman, J., & Changeux, J.P. (1965) On the nature of allosteric transitions: a plausible model, *J. Mol. Biol.* 12, 88–118.

Problem 9–1: Using as your three examples malate dehydrogenase, protocatechuate 3,4-dioxygenase, and filamentous actin, discuss the topics of rotational axes of symmetry, screw axes, open structures, closed structures, interfaces, and helical polymers.

Space Groups

Screw and rotational operations around axes of symmetry occur within crystals of proteins in addition to the translational operations relating the unit cells. These axes of symmetry are the fundamental operations that define the space groups. A **space group** of identical enantiomeric objects is a potentially infinite array of those objects, the positions and orientations of which are related to each other by screw axes of symmetry, rotational axes of symmetry, and translational operations.

For reasons mainly associated with the phenomenon of diffraction, a unit cell is defined solely in terms of translation. A unit cell is the smallest unit from exact copies of which, distributed only by simple translational movements, the entire crystal is created. In contrast, the **crystallographic asymmetric unit** is the smallest unit from exact copies of which, distributed both by translation and by rotation around axes of symmetry, the entire crystal is created. Crystallographic asymmetric units are usually delineated to include one or more intact subunits of a protein or one or more intact molecules of a protein.

If crystallographic asymmetric units containing one or more subunits or one or more molecules of a protein were always distributed in a crystal so that each of those asymmetric units had exactly the same rotational orientation, all unit cells would be of the same type, *P1*, and each unit cell would contain only one crystallographic asymmetric unit. **Packing** the same asymmetric units in different rotational orientations to form a crystal, however, is not forbidden, and strangely shaped enantiomeric objects, such as asymmetric units containing protein, usually are packed with greater efficiency when they can assume different rotational orientations. If these rotational orientations are to be compatible with the infinite regular array that is a crystal, they must be related by particular symmetry operations. Dismissing mirror symmetry, which is irrelevant to enantiomeric objects such as proteins, one is left with axial symmetry.

In a diagram of the simple **space group *P2*** (Figure 9-3), the array of unit cells portrayed represents one of the layers in a three-dimensional crystal. The array of the enantiomeric objects is created by distributing them about rotational axes of symmetry and by translational operations. Within each unit cell, the two identical enantiomeric objects are related by a central 2-fold rotational

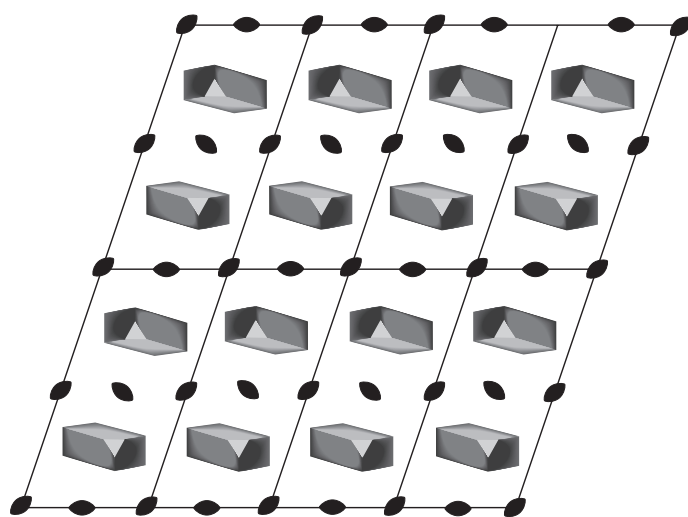


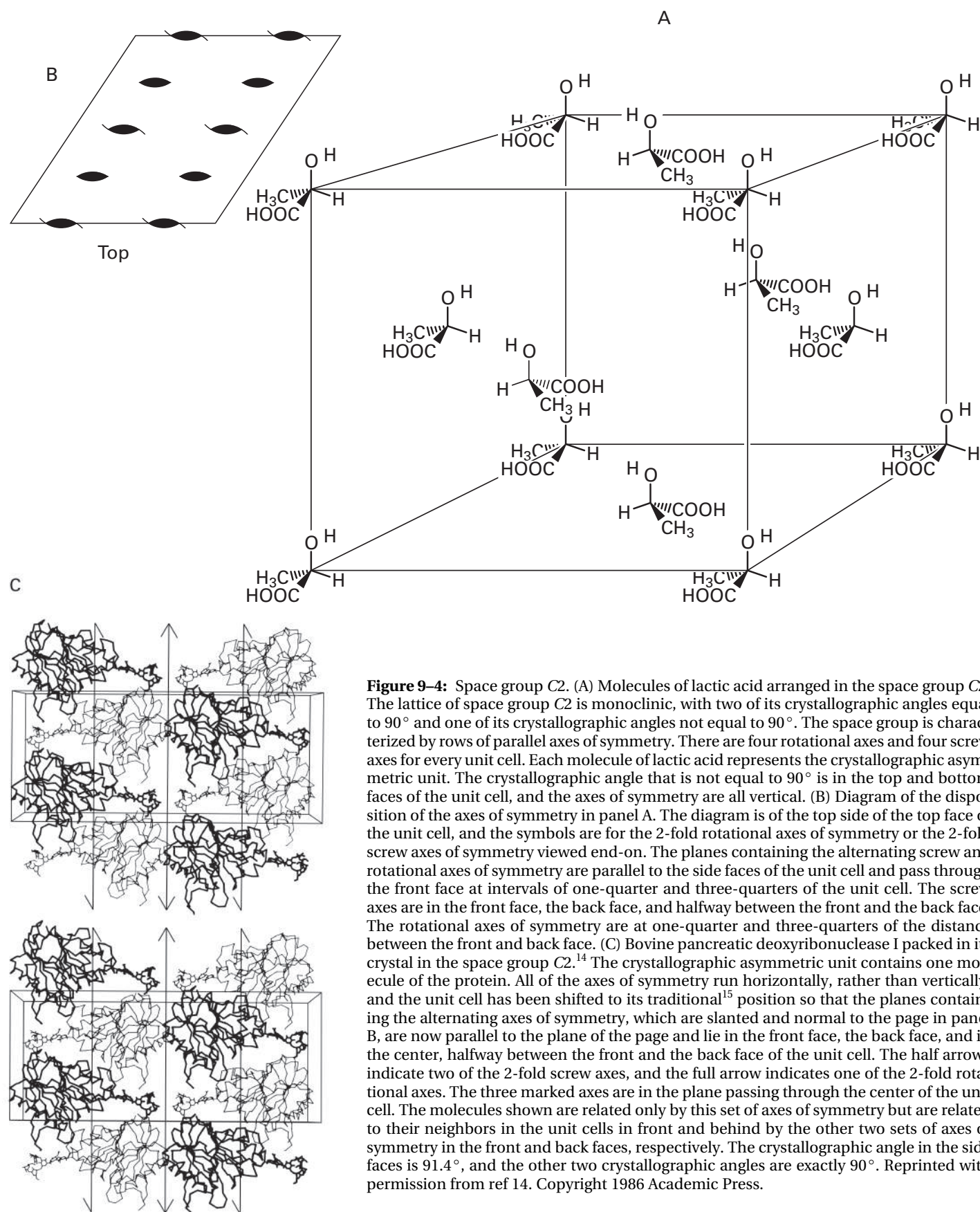
Figure 9-3: Packing of identical asymmetric objects, which represent the crystallographic asymmetric units from which the array is formed, in the space group *P2*, in which they alternately assume two different rotational orientations. The symbol ● indicates a 2-fold rotational axis of symmetry perpendicular to the page.

axis of symmetry perpendicular to the page. Because of the two rotational dispositions, which arise only because of the increased efficiency of packing, the unit cell ends up containing two of the enantiomeric objects rather than one. Each of the 2-fold rotational axes in the center of each of these unit cells is itself a rotational axis of symmetry for the entire array if it is assumed that the array is infinitely propagated in three dimensions. There are also three distinct sets of 2-fold rotational axes of symmetry between the unit cells. It is a feature of axes of symmetry in space groups that more than one distinct set appears at a time, and together these **sets of axes of symmetry** define the space group.

In crystals, as opposed to individual oligomers and polymers, screw axes of symmetry are required to have rotational angles that are integral quotients of 360° . This arises from the fact that these screw axes of symmetry operate on the entire array in the crystal. As the image of one of the crystallographic asymmetric units from which the crystal is composed is rotating and rising around the screw axis, the images of every other asymmetric unit in the array are rotating and rising around the same axis. At the completion of the operation, all of the images in the array must superpose on identical partners. This can occur only if the rotations of both the screw axes and the rotational axes of symmetry in a space group are 2-fold, 3-fold, 4-fold, or 6-fold. No other rotational multiplicities are compatible with an infinite array of asymmetric objects. Space groups never have 5-fold rotational or screw axes of symmetry because an infinite repetitive array of pentagons cannot be formed. Any translational distance, as long as it is compatible with an unclosed screw axis of symmetry, is compatible with an infinite array. Technically no crystal is infinite, but it always has the potential to be so, and this potential is all that matters.

The arrangement of asymmetric units in a space group can be displayed by distributing drawings of an enantiomeric object representing a crystallographic asymmetric unit of protein around appropriately positioned axes of symmetry. It is convenient to use an enantiomeric object familiar to all chemists, as well as one that is easy to draw, namely, a small enantiomeric molecule. Lactic acid is one of the smallest enantiomeric molecules. Drawings of distributions of lactic acid in the space groups *C2*, *P2₁2₁2₁*, and *P3₁21* (Figures 9-4, 9-5, and 9-6, respectively)¹⁴⁻¹⁸ illustrate certain properties of space groups, their rotational axes of symmetry, and the unit cells they create.

Although there are four distinct sets of 2-fold rotational axes in the space group *P2* of Figure 9-3, it is considered to be a simple 2-fold array, designated by the single integer 2, because all of its axes are parallel to each other and no set can exist without all of the others. The presence of a rotational axis of symmetry in a space group is indicated by an unadorned integer: 2, 3, 4, or 6. The presence of a screw axis of symmetry is indicated by



an integer, 2, 3, 4, or 6, followed by another integer in subscript, for example, the 3_1 screw axes of symmetry in the space group $P3_121$ (Figure 9–6). The main integer, n , is the integer by which 360° is divided to obtain the rotational angle of the steps. The integer in subscript, m , determines the fraction, m/n , of the unit cell over which the translation occurs with each rotational step. The translation is always right-handed to the rotation. Consequently, by this convention a 3_1 screw and a 4_1 screw are right-handed, and a 3_2 screw and a 4_3 screw are left-handed.

The **designation of the space group** in a crystalline array takes the form of a capital letter* followed by one or more numbers. An example would be $P2_12_12_1$, which would mean a primitive lattice made of rectangular parallelepipeds intersected by three orthogonal sets of 2-fold screw axes of symmetry (Figure 9–5). The arrangements of the axes in a particular space group can be learned only by consulting a diagram of that space group.¹⁵

The crystallographic asymmetric unit has a volume that is always an integral quotient of the unit cell and thus its volume is equal to or smaller than that of the unit cell. An integral number of asymmetric units, but not necessarily of intact asymmetric units, composes a **unit cell**. For example, one whole, two halves, four fourths, and eight eighths gathered from 15 different asymmetric units, each represented by one intact molecule of lactic acid, can together create a unit cell containing a total of four asymmetric units in the space group $P2_12_12_1$ (Figure 9–5). The crystallographic asymmetric unit in Figure 9–4C is one molecule of deoxyribonuclease I, and the unit cell contains a total of four asymmetric units but not four intact molecules of the protein.

The space group imposes certain **constraints** on the structure of the unit cell. For example, in the space group $P2$ that produces the array of Figure 9–3, the 2-fold rotational axes of symmetry must be normal of the plane of Figure 9–3 or the superposition cannot occur. Therefore, each unpictured asymmetric unit above and below the plane of the page in the lattice must be perpendicularly aligned with one of the asymmetric units in the plane of the page. This requires that the two angles of the fundamental unit cell aligning the axis normal to the page be precisely 90° . A lattice where two of the angles must be 90° is monoclinic (caption to Figure 4–2). In the space group $P2_12_12_1$ displayed in Figure 9–5, the three necessarily orthogonal sets of screw axes force the unit cell to be a rectangular parallelepiped and the lattice to be orthorhombic. Each space group other than the most

primitive, $P1$, which lacks axial symmetry entirely, enforces one or more of the angles of the unit cell to be 90° or 120° or enforces one or more of the axes of the unit cell to be the same length or enforces requirements on both angles and lengths. These are not coincidental identities but required identities. They are dictated by the symmetry operations and are thus exact quantities. If the angle in the monoclinic crystal of Figure 9–3 were not exactly 90° , the crystal would be filled with fractures and would not be a crystal.

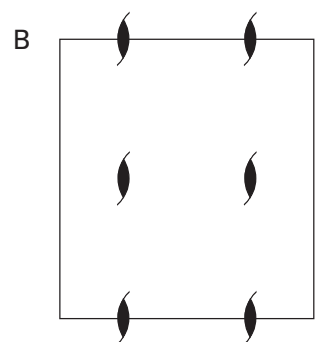
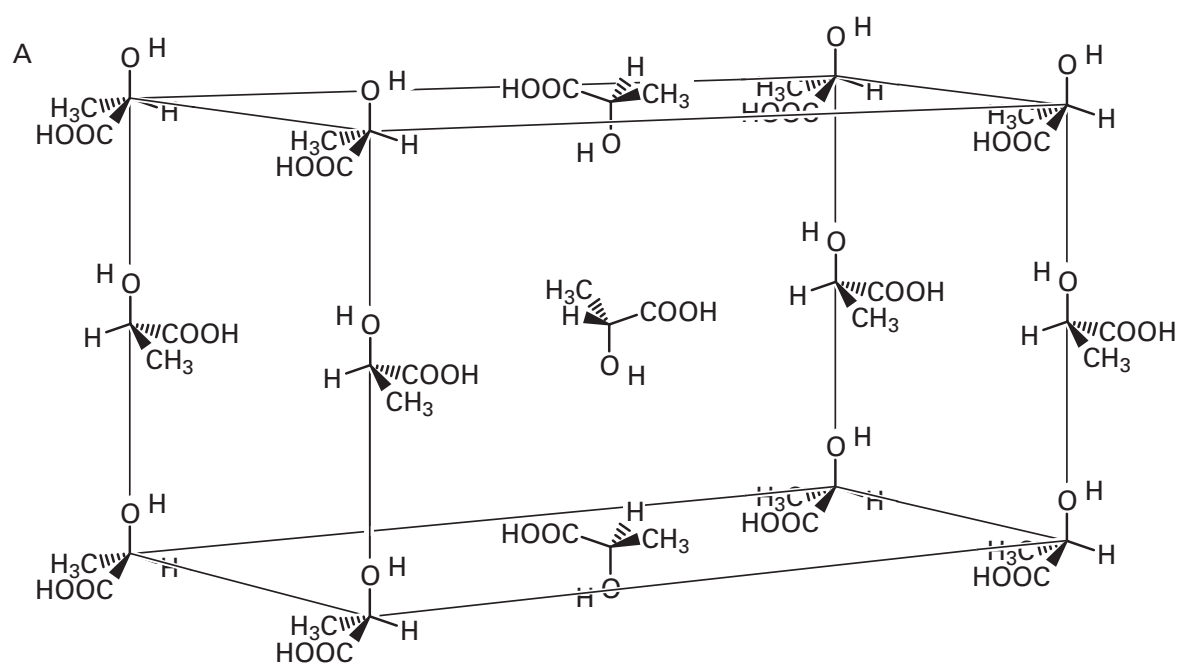
Reality seems to take place in Cartesian space, and there are only 71 space groups into which an infinite array of identical crystallographic asymmetric units, each containing one or more subunits of protein, can be arranged in Cartesian space to produce a crystal. Every crystal of protein has its crystallographic asymmetric units arrayed in one of these space groups. The space group is established as the crystal nucleates and grows in the dish; it cannot be dictated by the investigator. She can only try to change the conditions of crystallization in the hope that another space group will be generated by the process. This is often attempted because the identity of the space group determines how difficult it will be to calculate a map of electron density.

The 71 space groups available to a crystallizing protein are distinguished one from the other by the arrangement in space of their respective screw axes and rotational axes of symmetry. In turn, the space group of the particular crystal of protein that is formed is identified by the investigator from a characteristic pattern created in the data set by its particular arrangement of axes of symmetry. These are patterns in which identities occur in the amplitudes of the reflections. For example, in the oscillation photograph in Figure 4–1B, the fact that the patterns of the intensities of the reflections above and below the equator are mirror images of each other is consistent with the existence of a rotational axis of symmetry in the crystal parallel to the axis of oscillation. The particular **pattern of identities** within the entire data set identifies the axes of symmetry in the crystal and their arrangement in space, and hence the space group of the crystal.

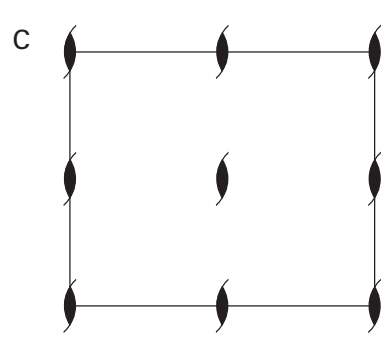
The packing of deoxyribonuclease I in the space group $C2$ (Figure 9–4C),¹⁴ that of the lectin from *Pisum sativum* in the space group $P2_12_12_1$ (Figure 9–5E),^{16,17} that of telokin from *Meleagris gallopavo* in the space group $P3_221$ (Figure 9–6C),¹⁸ that of porin from *Rhodobacter capsulatus* in the space group $R3$ (Figure 9–7),¹⁹ and that of ferredoxin from *Aphanothece sacrum* in the space group $P4_1$ (Figure 9–8)²⁰ illustrate the accommodations of the molecules of proteins to the axes of symmetry defining these five space groups.

There are no rotational axes of symmetry, only screw axes of symmetry, relating the asymmetric units in the space groups $P2_12_12_1$ and $P4_1$, but in the space groups $C2$, $P3_121$, and $R3$, pairs of asymmetric units are disposed around 2-fold rotational axes or triplets of asymmetric

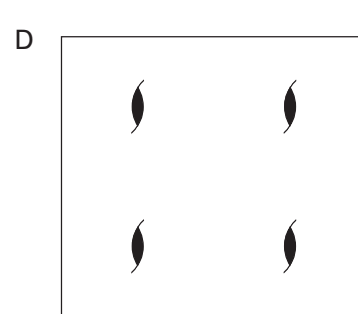
* The capital letter refers to the particular relationship between the underlying lattice and the unit cell for the space group of interest. These relationships are primitive (*P*), C-face centered (*C*), A-face centered (*A*), B-face centered (*B*), all-face centered (*F*), body centered (*R*), or hexagonally centered (*H*).



Left



Top



Front

E

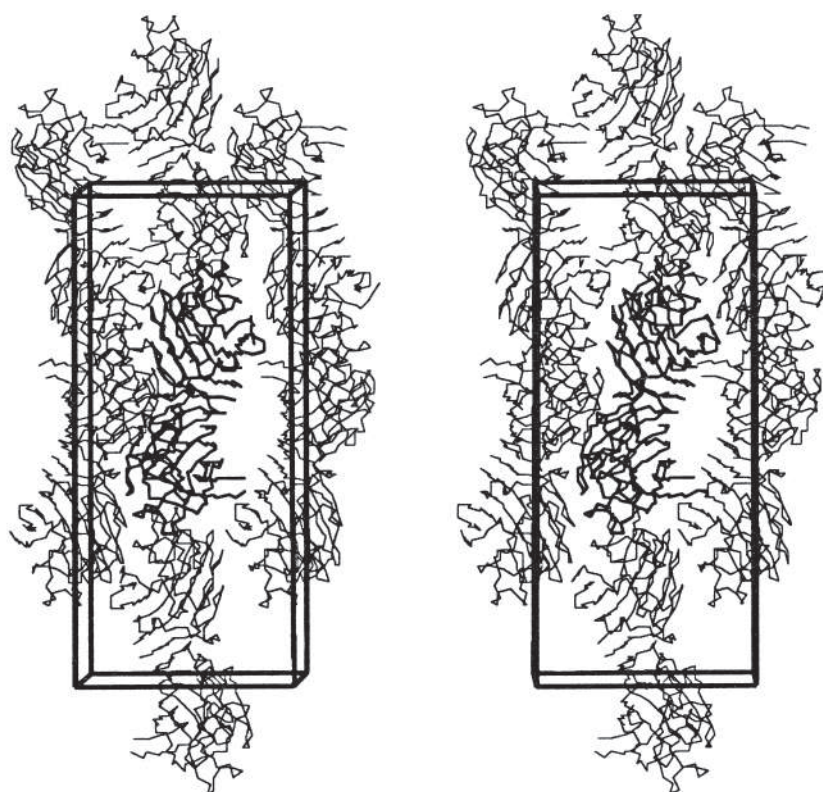


Figure 9-5: Space group $P2_12_12_1$. (A) Molecules of lactic acid arranged in the space group $P2_12_12_1$. The lattice of space group $P2_12_12_1$ is orthorhombic, with all three crystallographic angles equal to 90° , and the unit cell is rectangular. This results from the fact that there are three sets of 2-fold screw axes of symmetry, each set is necessarily orthogonal to the others, and in each set the screw axes are parallel to one of the crystallographic axes. There are four screw axes for each unit cell in each of the three sets. The molecule of lactic acid represents a crystallographic asymmetric unit, and from any one molecule of lactic acid the entire array can be created by performing the operations of 2-fold screw symmetry in the three orthogonal directions. (B) The four parallel horizontal screw axes of symmetry passing through the left face of the unit cell in panel A from front left to back right. These axes are in the top face, bottom face, and halfway between the top face and bottom face, one-quarter and three-quarters of the way across the unit cell. (C) Vertical 2-fold screw axes of symmetry passing through the top face of the unit cell in panel A. Half of these vertical axes coincide with each vertical column of lactic acids, of which there are two for each unit cell. The other half of these vertical screw axes of symmetry coincide with vertical lines in the center of each vertical face. When rotated 180° about any one of these vertical axes while simultaneously rising half of a unit cell, the lattice superposes on itself. (D) The four parallel horizontal screw axes of symmetry passing through the front face of the unit cell in panel A from front right to back left. These axes are at one-quarter and three-quarters of the distance between top and bottom face and one-quarter and three-quarters of the distance between the side faces. The symbols in panels B–D are the 2-fold screw axes of symmetry seen end-on. (E) The lectin from *P. sativum* packed in its crystal in the space group $P2_12_12_1$.^{16,17} The asymmetric unit contains one complete molecule of the protein, which is a homodimer of identical subunits related by a local noncrystallographic 2-fold rotational axis of symmetry. The unit cell is positioned in the traditional¹⁵ location with respect to the three orthogonal sets of 2-fold screw axes of symmetry so that the top is shifted one quarter of its width forward and the front is shifted one quarter of its width downward relative to their positions in panels C and D, respectively. The central, complete molecule of protein is represented in thicker lines. It is related to the two molecules drawn with thinner lines above and below it by one of the vertical screw axes of symmetry halfway across the unit cell and one quarter of the way forward from the back face. It is related to the two molecules to its upper right and upper left, respectively, by one of the horizontal screw axes of symmetry parallel to the plane of the page, one quarter of the way down the unit cell, and halfway between the front and back faces. And it is related to the two molecules to its lower right and lower left, respectively, by the two screw axes of symmetry normal to the plane of the page half of the way up from the bottom face and each one quarter of the way in from one of the sides. Reprinted with permission from ref 17. Copyright 1990 Elsevier B.V.

units are disposed around 3-fold rotational axes of symmetry that are inherent in the space groups. In the space groups $C2$ and $P3_121$, each and every asymmetric unit is related to at least one of its adjacent twins by a particular 2-fold rotational axis of symmetry. This unique relationship establishes a particular pair of twins. This pair is exceptional because the whole lattice can be divided into an array of these pairs, and in each of these pairs the orientation of the two twins to each other is the same. In the space group $C2$ pictured in Figure 9-4A, such a pair of twins includes any two lactic acid molecules that have their carboxylic acid functional groups opposite the hydroxyls of their neighbors. Every lactic acid molecule in the crystal participates in one and only one such symmetric relationship.

The particular 2-fold rotational axes of symmetry connecting the noted pairs of rotationally symmetric twins are crystallographic axes. A **crystallographic axis of symmetry** is any one of the axes of symmetry that defines the space group of the crystal. It exists only when the protein is in a crystal. The 2-fold rotational axis of symmetry running through the center of malate dehydrogenase (Figure 9-1) and connecting its twin subunits is a molecular axis of symmetry. A **molecular axis of symmetry** is an axis of symmetry that exists in the molecule of a protein regardless of whether or not that molecule is in solution or in a crystal. Crystallographic axes and molecular axes arise under different circumstances. The molecular axes of symmetry are created as the oligomeric protein assembles in the cell, and the crystallographic axes of symmetry are created as the crystal grows in a dish. These two types of axes of symmetry are independent properties. They can, however, but they are never required to, coincide. If a molecular rotational axis of symmetry coincides with a crystallographic rotational axis of symmetry, it can be stated that the molecular axis

when it is within the crystal is an exact rotational axis of symmetry because a crystallographic axis of symmetry is necessarily an exact rotational axis of symmetry. If a crystallographic axis of symmetry were not exact, crystal growth could not continue because the small equal deviations between the actual rotational operation and an exact rotational operation would add up across the crystal and eventually produce an interruption in the lattice. An **exact rotational axis of symmetry** in a protein is a rotational axis which, in a crystallographic molecular model of that protein, coincides with a crystallographic axis of symmetry.

The molecular 2-fold rotational axis of symmetry in the middle of each dimer of malate dehydrogenase from *A. arcticum* (Figure 9-1) coincides with one of the crystallographic 2-fold rotational axes of symmetry in the space group $P2_12_12$ in which it crystallizes.² Consequently, the molecular axis is exact. The molecular 3-fold rotational axis of symmetry in the middle of each trimer of porin from *R. capsulatus* coincides with one of the crystallographic 3-fold rotational axes of symmetry in the space group $R3$ in which it crystallizes (Figure 9-7)¹⁹ and is exact.

There is one crystal that contains an educational exception to this rule that a molecular axis of symmetry coinciding with a crystallographic axis of symmetry is exact. Each $\alpha\beta$ protomer of the $(\alpha\beta)_3$ trimeric portion of rat mitochondrial H^+ -transporting two-sector ATPase²¹ is found in its own asymmetric unit when the protein crystallizes in the space group $R32$. Each $(\alpha\beta)_3$ trimer, however, has one and only one γ subunit associated with it. Consequently, only one of the asymmetric units in each triplet of asymmetric units containing the entire $(\alpha\beta)_3$ trimer can contain a particular segment of a γ subunit. In fact, the γ subunits are distributed at random among the three so the map of electron density contains

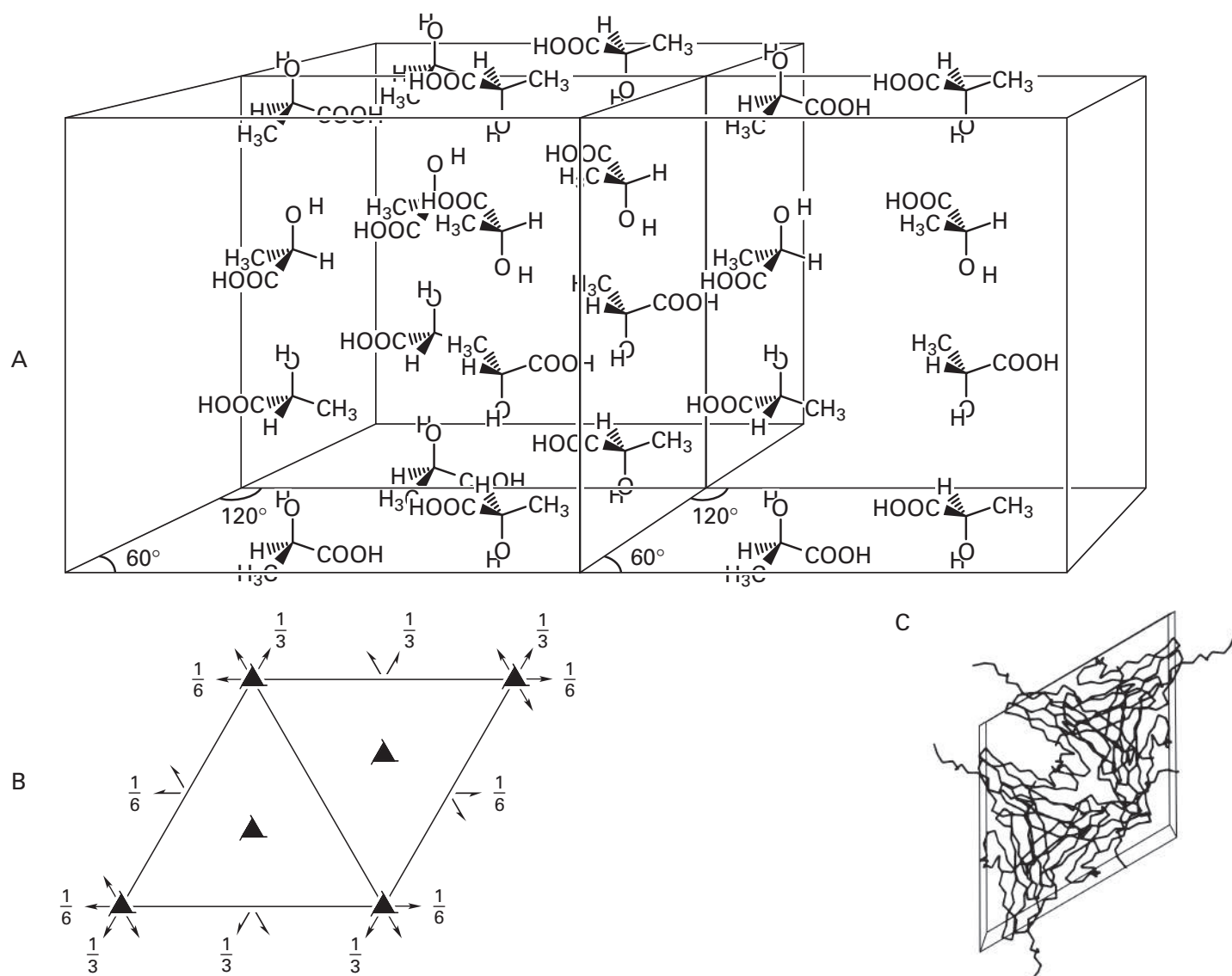


Figure 9-6: Space group $P3_121$. (A) Molecules of lactic acid arranged in the space group $P3_121$. The lattice of space group $P3_121$ is trigonal, with two crystallographic angles of 90° and one crystallographic angle of 60° . The angle of 60° is in the bottom faces of the three unit cells that are drawn. The space group $P3_121$ has a set of right-handed 3-fold screw axes of symmetry. In the drawing, two of the 3-fold screw axes of symmetry of each unit cell coincide with the vertical columns of lactic acid molecules. The third 3-fold screw axis of symmetry of each unit cell coincides with the vertical edge of each unit cell. Because the other crystallographic angle in the bottom face is 120° , counterclockwise rotation of 120° about this latter axis while simultaneously rising one-third of a unit cell causes the three upward columns of lactic acid around this axis to superpose on themselves and the three downward columns of lactic acid around this axis to superpose on themselves. Normal to each of the 3-fold screw axes of symmetry running along each vertical edge of the unit cells and intersecting perpendicularly with these latter vertical axes are 2-fold rotational axes of symmetry arrayed at 60° angles to each other. In the bottom face of the unit cell, the first 2-fold rotational axis of symmetry is the diagonal bisecting the 120° angle. Each successive 2-fold rotational axis of symmetry is 60° counterclockwise to the one below it and one-sixth of the distance up the axis of the unit cell. (B) Diagram¹⁵ of the arrangement of all of the axes of symmetry in space group $P3_121$. The view is looking down onto the bottom face of the unit cell containing the 60° angle. The flared triangles denote 3-fold screw axes normal to the page. The full arrows are 2-fold rotational axes of symmetry, and the half arrows are 2-fold screw axes of symmetry parallel to the plane of the page. The fractions indicate how far up the vertical edge or vertical face of the unit cell the axes parallel to the plane of the page are found. Reprinted with permission from ref 15. Copyright 1983 D. Reidel. (C) Telokin from the gizzard of *M. gallopavo* packed in its crystal¹⁸ in the closely related space group $P3_221$. The protein is a monomer of a single folded polypeptide, and the asymmetric unit contains only one of these monomers. The view is from the top so the 3-fold screw axes of symmetry in this view are normal to the plane of the page rather than vertical. They are in the same locations in the unit cell as those in the space group $P3_121$ (panel B) but are left-handed rather than right-handed screw axes (hence the 3_2 instead of 3_1). This difference causes the 2-fold screw axes of symmetry and 2-fold rotational axes of symmetry parallel to the plane of the page (panel B) to be encountered in a clockwise succession rather than a counterclockwise succession but at the same locations, angles, and depths. Reprinted with permission from ref 18. Copyright 1992 Elsevier B.V.

three overlapping, symmetrically displayed copies of the γ subunit each copy having one third the expected electron density. The actual individual asymmetric units in each molecule of the protein in the crystal are not symmetric within themselves, but the crystal is 3-fold symmetric. The individual molecular rotational axes of pseudosymmetry relating the three $\alpha\beta$ protomers is precisely 3-fold, but the symmetry of each molecule is perturbed by the presence of the necessarily asymmetric γ subunit.

The results from several crystallographic experiments serve to illustrate the **distinction between crystal symmetry and molecular symmetry** and the consequences of their coincidence.

Triose-phosphate isomerase, a dimer composed of two identical subunits, crystallizes in the space group $P2_12_12_1$. The crystallographic asymmetric unit is the α_2 dimer, and the 2-fold molecular rotational axis of symmetry within the dimer cannot coincide with a crystallographic rotational axis of symmetry because there is none.²² Glyceraldehyde-3-phosphate dehydrogenase (phosphorylating), an $(\alpha_2)_2$ tetramer composed of four identical subunits, also crystallizes in the space group $P2_12_12_1$, and the asymmetric unit is necessarily the entire tetramer. Glutathione peroxidase, however, also an $(\alpha_2)_2$ tetramer composed of four identical subunits, crystallizes in the space group $C2$, and the asymmetric unit is the α_2 dimer.²³ Consequently, in this instance one of the molecular axes of symmetry coincides with a crystallographic axis of symmetry, and the two α_2 dimers composing the $(\alpha_2)_2$ tetramer must be related to each other by an exact 2-fold molecular rotational axis of symmetry. The other two molecular 2-fold rotational axes of symmetry orthogonal to the one that coincides cannot also coincide with orthogonal crystallographic 2-fold rotational axes of symmetry because there is none.

Phosphorylase *b*, a dimer composed of two identical subunits, crystallizes in the space group $P4_32_12_1$, the asymmetric unit is one subunit,²⁴ and the dimer must be constructed upon an exact 2-fold molecular rotational axis of symmetry. Alcohol dehydrogenase, a dimer composed of two identical subunits, crystallizes in the space group $C22_12_1$, and the molecular 2-fold rotational axis of symmetry coincides with a crystallographic 2-fold rotational axis of symmetry and must be exact.²⁵ The α_3 trimer of chloramphenicol *O*-acetyltransferase from *E. coli* crystallizes in the space group $R32$,²⁶ and the α_3 trimer of bovine purine-nucleoside phosphorylase crystallizes in the space group $P2_13$,²⁷ and in both crystals molecular and crystallographic 3-fold rotational axes of symmetry coincide, and the molecular axes must be exact. L-Lactate dehydrogenase, a tetramer composed of four identical subunits, crystallizes in the space group $F422$, and one single subunit is the asymmetric unit.²⁸ Consequently, the tetramer is constructed around three exact orthogonal 2-fold molecular rotational axes of symmetry that coincide with three precisely orthogonal crys-

tallographic 2-fold rotational axes of symmetry. (S)-2-Hydroxy-acid oxidase crystallizes in the space group $I422$. As a result, its molecular 4-fold rotational axes of symmetry and its four molecular 2-fold axes of symmetry coincide with crystallographic rotational axes of symmetry²⁹ and must be exact. Dihydrolipoyllysine-residue acetyltransferase from *Azotobacter vinelandii*³⁰ and dihydrolipoyllysine-residue succinyltransferase from *E. coli*,³¹ related oligomers each composed of 24 identical subunits, both crystallize in the space group $F432$ in which the asymmetric unit is a single subunit, and the three molecular 4-fold rotational axes of symmetry, the

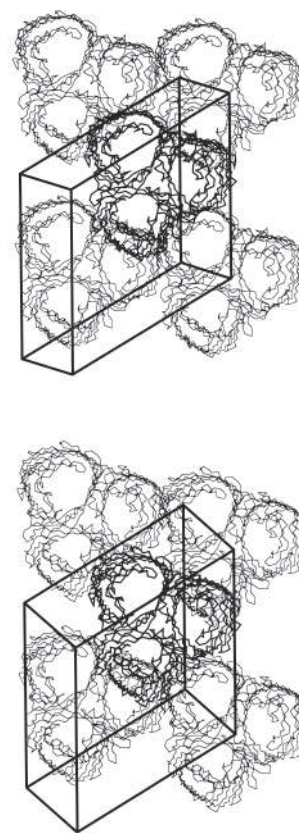
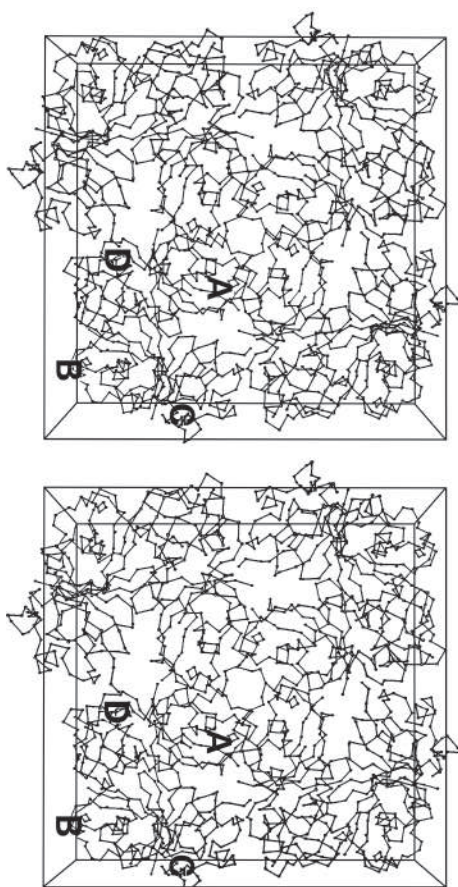


Figure 9-7: Molecules of porin from *R. capsulatus* packed in the space group $R3$ of its crystal.¹⁹ The lattice of space group $R3$ is trigonal with two crystallographic angles of 90° and one of 60° . Three-fold rotational axes of symmetry and 3-fold screw axes of symmetry normal to the front face pass along its edges and through its interior. The protein is a trimer of three identical folded polypeptides, and the asymmetric unit contains only one single folded polypeptide of this trimer. Each molecule of the protein has one of the crystallographic rotational axes of symmetry running through its center so that each of its subunits is exactly the same as the others and related to them by an exact 3-fold rotational axis of symmetry. The 3-fold screw axes of symmetry running between the trimers all superpose trimers on other trimers. Reprinted with permission from ref 19. Copyright 1992 Elsevier B.V.

Figure 9-8: Molecules of ferredoxin I from *A. sacrum* packed in the space group $P4_1$ of its crystal.²⁰ The lattice of space group $P4_1$ is tetragonal, with all crystallographic angles equal to 90° , and the top face of the unit cell is necessarily a square. Four-fold screw axes of symmetry run normal to the page at the four corners of the square and in its center. Consequently, 2-fold screw axes of symmetry normal to the page are located in the middle of each edge of the square. The protein is a monomer of a single folded polypeptide, and the asymmetric unit contains four copies, labeled A through D in the lower right hand asymmetric unit. Each A molecule participates in an interface with an adjacent A molecule around the central 4-fold screw axis of symmetry. The same interface joins B molecules around the 4-fold screw axes of symmetry at the corners and alternating C and D molecules around the 2-fold screw axes of symmetry. Within the asymmetric unit the only symmetry is a local 2-fold screw axis of symmetry relating an A molecule to a B molecule. Reprinted with permission from ref 20. Copyright 1990 Elsevier B.V.



four molecular 3-fold rotational axes of symmetry, and the four molecular 2-fold rotational axes of symmetry all coincide with crystallographic axes of symmetry and must be exact. Because both the 2-fold molecular axis of symmetry of the decamer of DNA with the sequence ATGACGTCAT and the 2-fold molecular axis of symmetry running through the coiled coil of α helices in the α_2 dimer of the crystallographic molecular model of a

portion of the general control protein GCN4 coincide with the same crystallographic axis of symmetry, the molecular axis of the DNA must be exact and the molecular axis of the coiled coil must be exact and precisely perpendicular to the central helical axis of the DNA.³²

Each of the last eight oligomeric proteins has all of its subunits arranged with a perfect rotational symmetry that can be conclusively proven. The coincidences of their molecular axes of symmetry and the crystallographic axes of symmetry that permitted these proofs, however, were by chance, and it seems reasonable to assume that in solution, the first three proteins, that just happened to crystallize in space groups incompatible with one or more of their molecular rotational axes of symmetry, are no less symmetric than the last eight. Consequently, while it is sometimes possible to deduce the whole symmetry of the protein from the crystallographic symmetry, the absence of the appropriate coincidences permitting this deduction does not mean that the molecule of protein lacks the missing symmetry.

Often it is assumed that molecular axes of symmetry within the asymmetric unit of a crystal are exact so that the electron density of the protomer can be enhanced by averaging around those molecular axes. Cystathionine γ -synthase from *Nicotiana tabacum*, an α_4 tetramer of folded polypeptides 445 aa in length, crystallizes in the space group $P2_12_12_1$ with two $(\alpha_2)_2$ tetramers in each of the four asymmetric units. Data could only be gathered to Bragg spacing of 0.29 nm, but when the electron densities of the eight protomers in the asymmetric unit were superposed about the respective molecular axes of symmetry and averaged, the map of electron density that resulted was a remarkable improvement over any one of the unaveraged maps.³³

Once **averaging around noncrystallographic rotational axes of symmetry** has been used to obtain a map of electron density accurate enough to build a convincing molecular model of the protomer, copies of the individual protomers are placed in their locations in the asymmetric unit, and the refinement of the molecular model is performed without rotational averaging. Only by refining the individual structures without rotational averaging do the differences among the protomers in the asymmetric unit, often dramatic ones,³⁴ become apparent. Often these differences, even though significant, can be dismissed as being due to distortions of an otherwise symmetric structure resulting from the asymmetric demands of packing it into the crystal.

When molecular axes of symmetry fail to coincide with crystallographic axes of symmetry, other criteria are used to evaluate the precision of the molecular rotational symmetry. For example, in the case of the $(\alpha_2)_2$ tetramer of glyceraldehyde-3-phosphate dehydrogenase (phosphorylating) in the space group $P2_12_12_1$, the spacing of the heavy metal atoms in the isomorphous replacement³⁵ and the final map of electron density³⁶ were both consistent with the tetramer being

constructed around three orthogonal apparently exact 2-fold molecular rotational axes of symmetry. The molecular 2-fold rotational axis of symmetry of the dimeric lectin from *P. sativum* does not coincide with a crystallographic rotational axis of symmetry because there is none in its space group $P2_12_12_1$ (Figure 9–5), but when the one folded polypeptide in the crystallographic molecular model is rotated around the molecular axis of symmetry, its α carbons superpose on those of its twin with a root mean square deviation of 0.06 nm.¹⁶ The rotational angle around the molecular axis of symmetry producing the smallest root mean square deviation (0.019 nm) of the **superposed α carbons** of the two subunits in the crystallographic molecular model of formate dehydrogenase from *Pseudomonas* was 179.9° .³⁷ In similar superpositions, the α carbons of the two subunits of triose-phosphate isomerase from yeast coincided with a root mean square deviation of less than 0.04 nm,³⁸ and those of the two subunits of transketolase from yeast, by 0.024 nm.³⁹ The error in the coordinates for the crystallographic molecular model of inorganic diphosphatase from yeast was estimated to be 0.037 nm; upon superposition of the one of its subunits on the other by rotation around the molecular axis of symmetry, the root mean square deviation of all of the atoms in the two polyamide backbones was only 0.038 nm.⁴⁰ Although there were functional indications that the subunits of ribulose-bisphosphate carboxylase were in different environments in solution, when the different folded polypeptides in the crystallographic molecular model were superposed by rotation around the molecular axes of symmetry, their α carbons coincided with a root mean square deviation of less than 0.02 nm, well within the accuracy of the coordinates themselves, and it was concluded that there was no structural evidence for asymmetry.⁴¹

When such superpositions are performed about molecular rotational axes of symmetry that do not coincide with crystallographic rotational axes of symmetry, it is often found that flexible regions of the crystallographic molecular model do not coincide as well as more rigid regions because they respond readily to variations in their surroundings resulting from **differences in crystal packing**.⁴² For example, most of the α carbons of the five identical β subunits of heat-labile enterotoxin from *E. coli* superpose to within 0.04 nm upon successive rotations about the molecular 5-fold rotational axis of symmetry, but the positions of the α carbons in the flexible loop between Glycine 54 and Serine 60 deviate by 0.1–0.2 nm from each other.⁴³ Unlike malate dehydrogenase from *A. arcticum*, cytoplasmic malate dehydrogenase from *Sus scrofa* crystallizes in the space group $P2_12_12_1$ with its α_2 dimer in each asymmetric unit.⁴⁴ The rotational angle around the molecular axis of symmetry producing the superposition with the minimum root mean square deviation is 174° instead of 180° . It was, however, concluded that this observation was mislead-

ing because it had been shown previously that crystallization of the protein causes several of its enzymatic properties, which are uncomplicated in solution, to become asymmetric. Consequently, it was concluded that crystal packing forces caused an otherwise symmetric protein to become remarkably asymmetric.

There are, however, a few observations suggesting that there may be subtle **asymmetries** in some homooligomers. For example, when the subunits of the crystallographic molecular model of the $(\alpha_2)_2$ tetramer of L-2-hydroxyisocaproate dehydrogenase were superposed intramolecularly, it was found that two superposed well on each other (root mean square deviation of 0.02 nm) and the other two did also (root mean square deviation of 0.03 nm) but that neither member of one of these pairs superposed well on either member of the other pair (root mean square deviation of 0.13 nm).⁴⁵ This asymmetry seemed too large to be due to crystal packing, and it is possible that this protein may be asymmetric in solution.

It is also possible to perform a **self-rotation function** on the asymmetric unit in the space group of a crystal to detect molecular symmetry before phases are available. The 11-fold molecular rotational axis of symmetry in the *trp* RNA-binding attenuation protein from *Bacillus subtilis*, which cannot coincide with any of the permissible crystallographic rotational axes of symmetry, was readily detected within the asymmetric unit of its space group C2 by performing a self-rotation function.⁴⁶

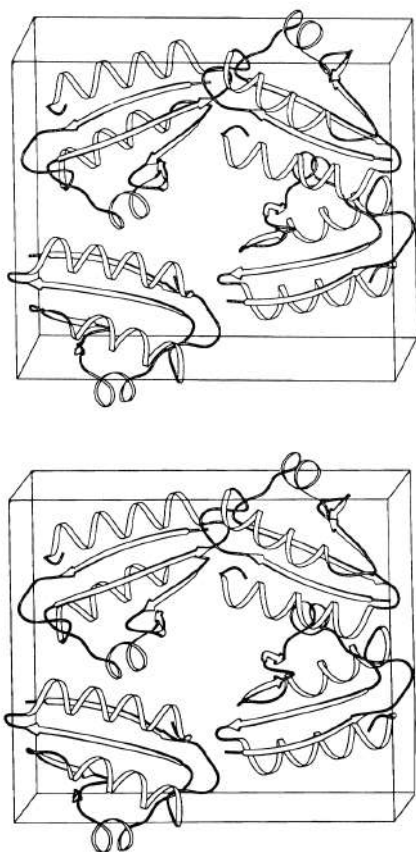
Suggested Reading

Hahn, T., Ed. (1983) *International Tables for Crystallography, Vol. A: Space-Group Symmetry*, D. Reidel, Dordrecht, The Netherlands.

Problem 9–2: In the space group $P3_121$ portrayed in Figure 9–6, every lactic acid molecule is related to one of its neighbors by the same 2-fold rotational axis of symmetry. Draw two lactic acid molecules arranged around that specific rotational axis of symmetry.

Problem 9–3: The stereodiagram on the next page is based on the crystallographic molecular model of the four individual molecules of the protein HPr from the monosaccharide transport system of *Streptococcus faecalis* arranged in the tetragonal unit cell.⁴⁷ Reprinted with permission from ref 47. Copyright 1994 Elsevier B.V.

On three rectangles representing the front, the side, and the bottom, respectively, of the unit cell in the orientation of the figure, indicate the positions of any screw axes of symmetry or rotational axes of symmetry passing through the unit cell. Use symbols for rotational or screw axes of symmetry like those in the diagrams in Figures 9–3 to 9–8.¹⁵



Problem 9-3

Oligomeric Proteins

A protomer is the smallest portion of a protein from copies of which its entire quaternary structure is created. The protomers of a homooligomeric protein are arranged around rotational axes of symmetry, and the number of protomers and the way in which they are arranged designates the point group to which the oligomer belongs. A **point group** is the distribution in which a particular number of protomers are arranged about one or more particular rotational axes of symmetry oriented at particular angles to one another and intersecting at a common origin, in which all of the centers of mass of the protomers are equidistant from this origin, and in which all of the symmetrically related positions are occupied. The finite and specific number of the protomers, their equidistance from the origin, and the common intersection of all of the rotational axes of symmetry distinguishes the point groups from the space groups as well as from the linear groups, which designate open linear multimers such as helical polymers. Oligomeric proteins have exploited all of the available point groups lacking mirror planes.

A **cyclic point group** arranges protomers in a circle about only one rotational axis of symmetry, and the different cyclic point groups are distinguished by the fold of the axis. In the simplest of these cyclic point groups,

point group $2(C_2)$,* two protomers are arranged around a 2-fold rotational axis of symmetry to form a dimer. Half of all homooligomeric proteins are dimers (Table 9-1) the protomers of which, with few exceptions (Figure 9-2), are arranged with the symmetry of point group $2(C_2)$. Malate dehydrogenase (Figure 9-1A) and κ bungarotoxin from *Bungarus multicinctus* (Figure 9-9)⁴⁹ are examples of oligomers of point group $2(C_2)$.

κ Bungarotoxin crystallizes in the space group $P6$ with an α_2 dimer in the crystallographic asymmetric unit. In the crystal, one protomer superposes on the other upon a 178.5° rotation about the molecular axis of symmetry. With the exception of the flexible loops between Cysteine 27 and Proline 36 and between Proline 15 and Glutamine 18, which adjust malleably to the constraints of crystal packing, the α carbons superpose to a root mean square deviation of 0.05 nm around the molecular rotational axis of symmetry, which probably differs from 180° also because of crystal packing. The interdigitations of the side chains forming the interface mimic each other across the axis of symmetry, for example, the hydrophobic cluster containing Isoleucine 20, Cystine 46/58, and Valine 60 from one protomer and Phenylalanine 49 from the other. The axis of symmetry itself runs through the hydrogen bond between the two Glutamines 48. The structure is unquestionably closed; every amino acid in the common sequence of the two identical polypeptides that is enclosed within the interface from one of the protomers is also enclosed from the other.

In larger proteins with more than one domain, it is often the case that the interface forming the dimer connects only one of the domains to its twin in the other pro-

Table 9-1: Frequency of Homooligomers

number of subunits	percent observed ^a
2	50
3	5
4	35
6	10
8	3
10	1
12	2

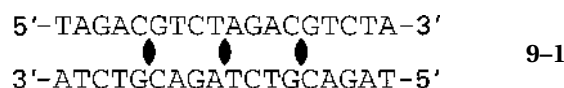
^aThe table of oligomeric stoichiometries published by Darnell and Klotz⁴⁸ was used to calculate these frequencies. Because this is a selected and incomplete list, some numbers were rounded to the nearest 5%.

* There are two notations currently in use to identify the individual point groups. Crystallographers use the Hermann-Mauguin notation, 2, 3, 4, ..., 222, 322, 422, ..., 23, 432, and 532, which will be the notation used here out of parentheses. Spectroscopists use the Schönflies notation, C_2 , C_3 , C_4 , ..., D_2 , D_3 , D_4 , ..., T , O , and I , which will be the notation used here within parentheses. Chemists other than crystallographers and spectroscopists will use one or the other of these notations depending on who taught them point groups or what book they happened to open.

tomers, but this limited interface nevertheless dictates the symmetry of point group $2(C_2)$ for the whole dimer.^{50,51} The domain (117 aa of the 450 aa in the intact protein) forming the entire interface holding together the dimer of glutathione-disulfide reductase⁵² has been detached genetically and shown to form by itself a dimer.⁵³ In the α_2 dimer of human hexokinase, each of the two subunits contains two internally duplicated domains, each superposable on a complete subunit of hexokinase from yeast, and they are connected by an α -helical segment of six turns. The amino-terminal domain of one subunit forms an interface with the carboxy-terminal domain of the other subunit and vice versa to form a dimer with two well-separated but identical interfaces on either side of the molecular 2-fold rotational axis of symmetry.⁵⁴

Proteins that associate with **double-helical DNA** are often dimeric, and such a dimer uses its own 2-fold rotational axis of symmetry to recognize a local 2-fold rotational axis of pseudosymmetry in the double helix of the DNA. Regardless of its sequence, between the two bases in any one of the pairs of bases in a molecule of DNA and running perpendicular to the hydrogen bonds between them is a local 2-fold axis of pseudosymmetry (Figure 3–9). A **local rotational axis** is an axis of rotation around which superpose upon one another only structural units immediately adjacent to that axis. Because a real molecule of DNA is rarely straight, if the whole molecule of DNA is rotated around one of these local axes by 180° , it roughly superposes on itself in the immediate vicinity of the axis but does not superpose beyond the immediate vicinity because of its curvature. Because the two bases in the central base pair are never the same, the base pairs on either side of the central base pair are usually not the same, and the DNA is usually curved, this local axis is always pseudosymmetric. Halfway between any consecutive two of these 2-fold rotational axes of pseudosymmetry and at an angle halfway between their two angles, there is also a local 2-fold rotational axis of pseudosymmetry.

The existence of this second set of local 2-fold rotational axes of pseudosymmetry means that **palindromic sequences*** such as



in which the two individual strands have the same sequence and the sequence of the duplex inverts at its center, have local 2-fold rotational axes of symmetry (indicated by \bullet). Rotation around any one of these axes by 180° superposes identical bases, within the segment.

* A **palindromic sequence**, for example, “able was i ere i saw elba”, is a sequence that superposes upon itself when rotated around an axis running through its center.

The existence of the local 2-fold rotational axes of pseudosymmetry running through the base pairs means that for the sequence



which contains a split palindrome, rotation about the local 2-fold axis of pseudosymmetry through the central

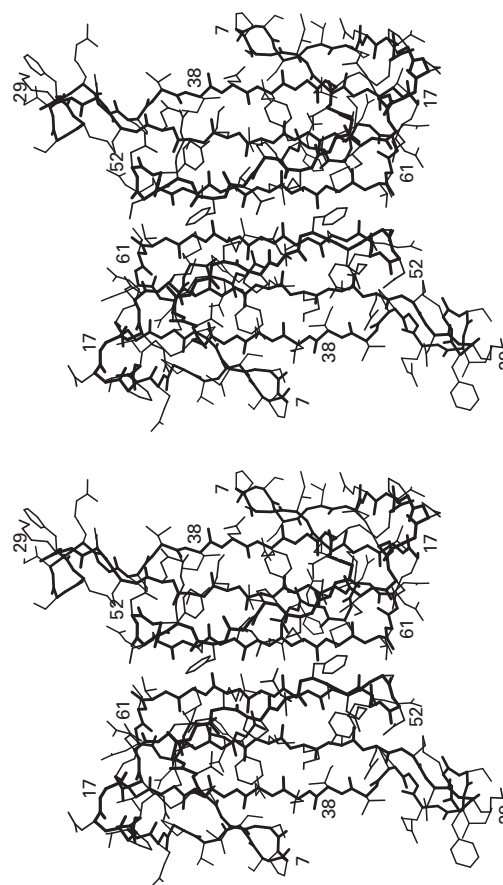
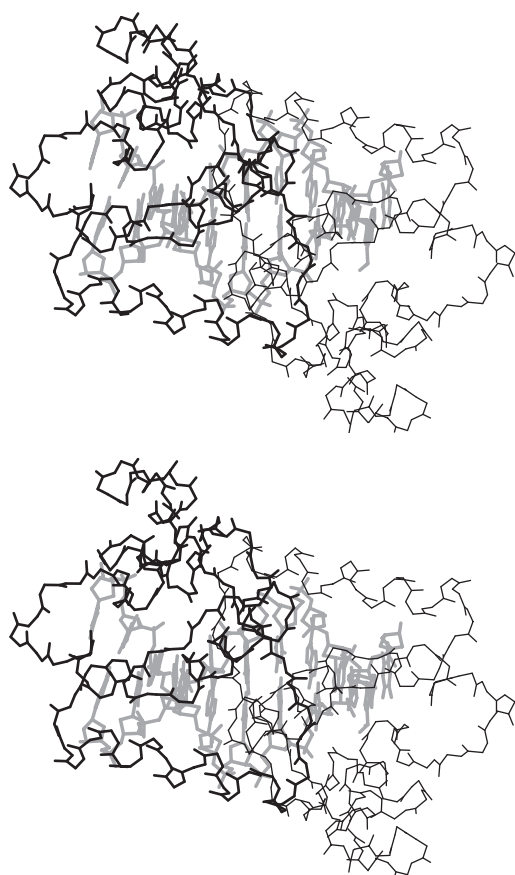


Figure 9-9: Two-fold rotational symmetry of point group $2(C_2)$. The crystallographic molecular model of the α_2 dimer of κ bungarotoxin from *B. multivittatus*⁴⁹ is formed from two identical folded polypeptides 66 aa in length arranged around a 2-fold rotational axis of symmetry running through the center of the molecule normal to the page. This drawing was produced with MolScript.⁴⁸⁵

Figure 9–10: Alignment of the 2-fold rotational axis of symmetry of the dimer of *met* repressor from *E. coli* with the local 2-fold axis of pseudosymmetry of the palindromic double-stranded DNA that it recognizes.⁵⁵ The portion of the crystallographic molecular model displayed is of an α_2 dimer of *met* repressor bound to a segment of chemically synthesized double-stranded DNA with the self-complementary sequences TTAGAGTCT and AGACGTCTA. The view is down the aligned molecular 2-fold rotational axis of symmetry of the dimer of protein and the local 2-fold rotational axis of pseudosymmetry of the DNA. Because the DNA is not completely palindromic, neither of these local, molecular axes of symmetry is able to coincide with one of the crystallographic 2-fold rotational axes of symmetry in the space group 6_222 . The double-stranded DNA is in the rear, and the protein is in the front. The DNA is drawn with thick gray lines, and the two subunits of the α_2 dimer are drawn with black lines of different thickness. Only the polyaniline backbone of the protein is represented. An antiparallel β sheet of two identical strands, one from each subunit, arrayed around the molecular 2-fold rotational axis of symmetry passes through the major groove of the DNA. The amino-terminal four amino acids assume different conformations in the two subunits. This drawing was produced with MolScript.^{48b}



G–C pair superposes the palindromic sequences in the boxes.

Many of the sequences of DNA that particular proteins are required to recognize are palindromic. A protein built from two copies of a folded polypeptide in

which the two copies are associated with each other around a molecular 2-fold rotational axis of symmetry can present, one on each copy, two identical faces complementary to faces on the surface of the palindromic segments of DNA. One of these surfaces can bind to one half of the palindrome and the other to the other half of the palindrome simultaneously if those surfaces are positioned relative to their own 2-fold rotational axis of symmetry as the two palindromes are to theirs. Evolution by natural selection has discovered that, by this simple strategy, the standard free energy of association between the DNA and the protein is automatically doubled and the specificity of the interaction is more than squared.

The *met* repressor of *E. coli* is a **dimeric protein** built from two copies of a folded polypeptide 104 aa in length that are associated with each other around a molecular 2-fold rotational axis of symmetry such that each copy presents a face complementary to a face on one half of the palindrome **9–1**. When the repressor is bound to the DNA, the molecular 2-fold rotational axis of symmetry of the protein coincides with the local palindromic 2-fold rotational axis of symmetry of the DNA (Figure 9–10).⁵⁵

The *arc* repressor of *E. coli* is a similarly symmetric protein that binds to the split palindrome in **9–2**.⁵⁶ When the palindrome is split, the protein can recognize both its symmetry and the length of the separation between its two halves in the segment of DNA. In the crystallographic molecular model⁵⁷ of the dimeric E2 DNA-binding domain of bovine papillomavirus-1, each of the identical subunits binds to one half of a palindrome, the two halves of which are split by four base pairs rather than five. In this complex, the molecular 2-fold rotational axis of symmetry of the protein and the molecular 2-fold rotational axis of symmetry of the DNA are coaxial and both coincide with a crystallographic rotational axis of symmetry.

In the cyclic **point group 3**(C_3), the three protomers of a homooligomeric protein are arranged around a 3-fold rotational axis of symmetry. Chloramphenicol *O*-acetyltransferase from *E. coli* is a trimer with symmetry of point group 3(C_3) (Figure 9–11).⁵⁸ It crystallizes in the space group $R32$ with one of its three identical folded polypeptides as the crystallographic asymmetric unit. Consequently, the molecular 3-fold rotational axis of symmetry is exact. The individual subunits are compact globular structures, and the three interfaces connecting them together are formed by the association of two relatively flat faces. The interfaces that produce the trimer are all identical to each other, and rotations of 120° about the axis of symmetry superpose them upon each other. Each subunit has on its surface one copy of each of the two distinct but complementary faces that form the interface. Three identical β strands, one from each subunit, direct their side chains into the center of the trimer at the axis of symmetry.

The frequency with which trimers with symmetry of

point group $3(C_3)$ arise during evolution by natural selection is about 10 times lower than the frequency with which dimers with symmetry of point group $2(C_2)$ arise (Table 9-1). In fact, at one time it was thought that such trimeric proteins did not exist, but there are now crystallographic molecular models for a number of them.^{19,59-66} It is by considering the problem of assembling a trimer relative to that of assembling a dimer that the reason for the scarcity of trimers becomes apparent.

The interface is the feature that evolves and not the oligomer. A dimer built around a 2-fold rotational axis of symmetry is held together by one more or less continuous **interface centered on the rotational axis of symmetry**. The axis divides the complete interface into two identical halves (Figures 9-1A and 9-9). Each half is the formal equivalent to one of the three identical interfaces distributed around the 3-fold rotational axis of symmetry in a trimer (Figure 9-11). The incremental decreases of free energy associated with favorable mutations are not automatically doubled during the evolution of an interface in a trimer as they are in the evolution of an interface in a dimer. Because termolecular collisions rarely occur, the assembly of an oligomeric protein must proceed through a series of bimolecular steps. A bimolecular collision producing a dimer automatically involves the simultaneous formation of the two halves of its interface and incorporates the free energies of formation of both halves into the immediate product. The first step in the assembly of a trimer, however, is the collision of two monomers to form only one of its three interfaces. This first interface, standing alone, must exist long enough or form often enough for the third protomer to complete the ring, yet it is the evolution of this initial interface that does not benefit from symmetry as does the evolution of the initial and final interface of the dimer. Therefore, trimers should appear less frequently than dimers during evolution. In favor of the symmetric trimer, however, is the fact that, as with the two halves of the interface in a dimer, its three identical interfaces can evolve simultaneously, a fact that magnifies the incremental decrease in free energy change in the overall formation of the complete oligomer for each favorable mutation.

If the 3-fold axis of symmetry of chloramphenicol O-acetyltransferase were not an exact rotational axis of symmetry but a closed screw axis, one of the interfaces could not be equivalent to the other two because the ring could not be completed. It is most likely that this peculiar one of the three interfaces would not fit together properly because it would be formed from the same two faces now required to associate in a different way from the way they associated at the other two interfaces. Such a structure would be significantly weaker than a rotationally symmetric trimer because of the one misaligned interface.

α_4 Tetramers with cyclic symmetry of point group 4 (C_4), such as L-lactate dehydrogenase (cytochrome) from *Saccharomyces cerevisiae* (Figure 9-12),⁶⁷ L-fucose-

phosphate aldolase from *E. coli*,⁶⁸ L-ribulose-phosphate 4-epimerase from *E. coli*,⁶⁹ and mammalian⁷⁰ and bacterial⁷¹ IMP dehydrogenase; α_5 pentamers with cyclic symmetry of point group 5 (C_5), such as acetylcholine-binding protein from *Lymnaea stagnalis*,⁷² the B subunit of heat-labile enterotoxin from *E. coli*,⁴³ and human serum amyloid P component;⁷³ α_6 hexamers with cyclic symmetry of point group 6 (C_6), such as transitional endoplasmic reticulum ATPase from *Mus musculus*⁷⁴ and the replicative DNA helicase encoded by the bacterial plasmid RSF1010;⁷⁵ and α_7 heptamers with cyclic symmetry of point group 7 (C_7), such as transcriptional

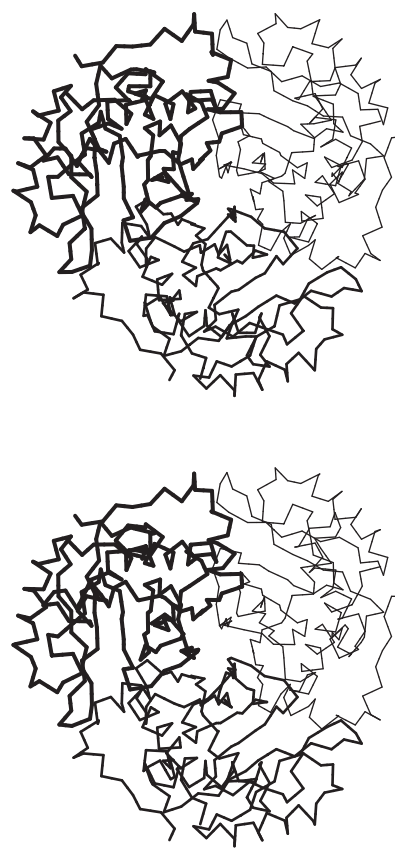
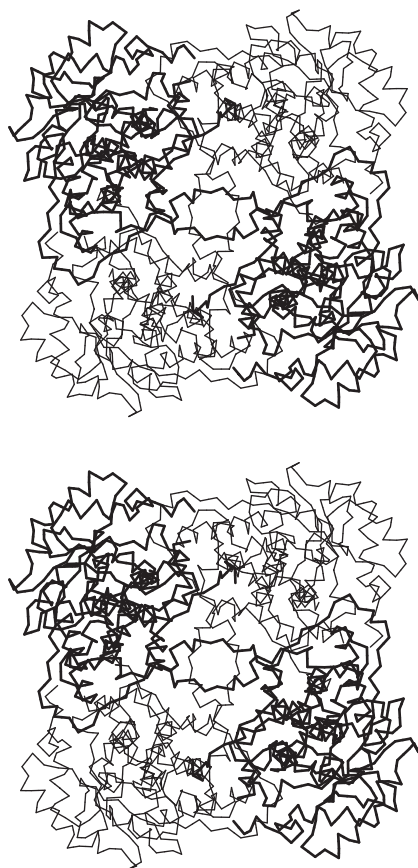


Figure 9-11: Three-fold rotational symmetry of point group $3(C_3)$. An α -carbon diagram of chloramphenicol O-acetyltransferase from *E. coli* is drawn from the crystallographic molecular model of this trimeric protein.³⁸ The three identical subunits are distinguished by the different widths of the line segments. This drawing was produced with MolScript.⁴⁸⁵

Figure 9–12: Four-fold rotational symmetry of point group 4(C_4). An α -carbon diagram of L-lactate dehydrogenase (cytochrome) from *S. cerevisiae* is drawn from its crystallographic molecular model.⁶⁷ The α_4 tetramer crystallized in the space group $P3_221$ with two subunits in the crystallographic asymmetric unit. The molecular 4-fold axis of symmetry coincided with a crystallographic 2-fold axis of symmetry, and both run through the center of the molecule normal to the plane of the page. The first 100 amino acids of the folded polypeptide form a separate domain, which is ordered and hence visible in two of the four subunits and disordered and hence invisible in the other two. To enhance the symmetry, these amino-terminal domains have been omitted. The four subunits have been drawn with lines that alternate in thickness around the circle. The carboxy-terminal 25 amino acids of each polypeptide cross over each other to form a central entwined cone around the 4-fold axis of symmetry. This drawing was produced with MolScript.⁴⁸⁵



activator NTRC1 from *Aquifex aeolicus*⁷⁶ and the small nuclear ribonucleoprotein from *Pyrobaculum aerophilum*,⁷⁷ are even more rare than trimers with cyclic symmetry of point group 3(C_3). Amazingly, however, there is a protein that is an α_{11} undecamer with symmetry of cyclic point group 11(C_{11}).⁷⁸

Tetramers are the second most common type of oligomeric protein (Table 9–1). Almost all tetrameric proteins have their protomers arranged in the symmetry of dihedral point group 222(D_2) instead of the symmetry of cyclic point group 4(C_4). A **dihedral point group** is a point

group in which the protomers are arranged on two circles of equal radius around a central n -fold rotational axis of symmetry and n 2-fold rotational axes of symmetry perpendicular to that central axis. In the **dihedral point group 222(D_2)**, the central axis is a 2-fold rotational axis of symmetry, there are two 2-fold rotational axes of symmetry orthogonal to it and to each other, and all of the axes intersect at the same point. If a sphere is placed at each of the four vertices of an equilateral tetrahedron, and if each of the four spheres has a diameter equal to the length of a side of the tetrahedron, the spheres will contact each other at six points (Figure 9–13A). If the spheres are asymmetric objects, for example, the subunits of a tetramer, then the six points of contact, for example, the six possible interfaces between pairs of subunits, are three identical twins, designated 1, 2, and 3 in the diagram, each different from the other two. Each of the six interfaces contains within itself one of the 2-fold rotational axes of symmetry. Each interface is superposed on its twin when rotation occurs about either of the two axes of symmetry orthogonal to the one passing through it, but no rotational axes of symmetry can superpose an interface 1 on an interface 2 or an interface 3 or superpose an interface 2 on an interface 3.

Ideally, a tetrameric protein with dihedral symmetry of point group 222(D_2) should have **three different pairs of interfaces**, but usually one pair does not form or is almost nonexistent because the tetrahedron is squashed in one dimension. For example, if the tetrahedral arrangement of spheres in Figure 9–13A were squashed from above the plane of the page, the interfaces 3 would pull apart. The tetramer of symmetry of point group 222(D_2) that is the crystallographic molecular model of 2,2-dialkylglycine decarboxylase (pyruvate) from *Burkholderia cepacia* is such a squashed tetrahedron (Figure 9–13B).⁷⁹ The two identical vertical interfaces, normal to the plane of the page through which runs the exact vertical molecular 2-fold rotational axis of symmetry in the plane of the page (the interfaces 1 in panel A), are the most extensive (75 nm² interface⁻¹)* and were designated by the crystallographers as the interfaces connecting the monomers that form the two dimers of the structure. In these two roughly flat interfaces, there are loops of polypeptide that interpenetrate the two subunits. The two identical horizontal interfaces,

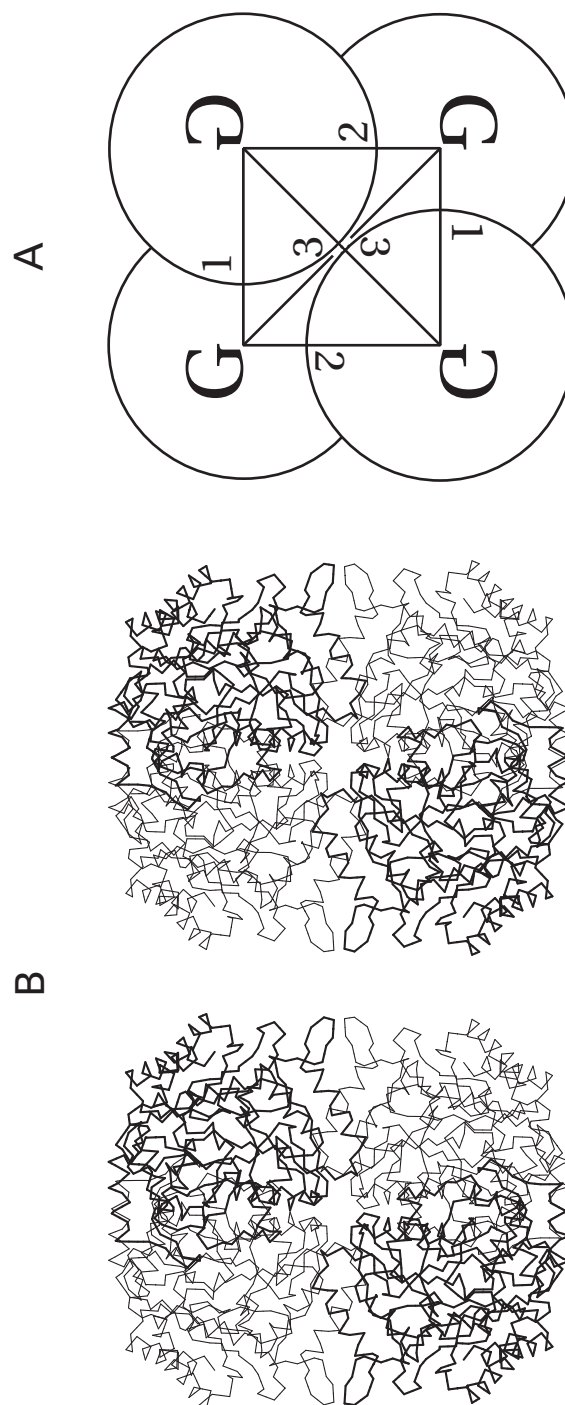
* The size of an interface will be presented as the total accessible surface area from its two participants that is buried upon its formation. Consequently, each interface is formally defined as the adhesive interaction between any two subunits that holds only those two subunits together in the complex. This definition ignores any cooperativity that arises in interfaces in which n subunits are held together by a total of n interfaces around an n -fold rotational axis of symmetry when n is greater than 2, such as the stability gained when the third subunit is added to complete a cyclic trimer or the stability realized in the entwined cone of four carboxy termini (Figure 9–12) in the four formal interfaces around the 4-fold rotational axis of symmetry in L-lactate dehydrogenase (cytochrome).

Figure 9-13: A dimer of dimers with dihedral symmetry of point group $222(D_2)$. (A) Point group $222(D_2)$. Four spheres are placed at the vertices of an equilateral tetrahedron. The spheres are made asymmetric by placing the letter G inside of each. The letters G are superposed on each other by rotation around the three axes of symmetry, one vertical and one horizontal in the plane of the page and the third normal to the page. The three different pairs of interfaces between the spheres, labeled 1, 2, and 3, are each distinct: interfaces 1 connecting the open sides of the letters G, interfaces 2 connecting the bottoms of the letters G, and interfaces 3 connecting the T segments of the letters G. (B) α -Carbon diagram of the $(\alpha_2)_2$ tetramer of 2,2-dialkylglycine decarboxylase (pyruvate) from *B. cepacia* drawn from its crystallographic molecular model.⁷⁹ Two of the subunits have been drawn with thicker line segments than the other two. The protein crystallizes in the space group $P6_322$ with a single subunit in the crystallographic asymmetric unit, and all three molecular 2-fold rotational axes of symmetry coincide with crystallographic rotational axes of symmetry. This drawing was produced with MolScript.⁴⁸³

which are tilted somewhat from being normal to the plane of the page and through which runs the exact horizontal molecular 2-fold rotational axis in the plane of the page (the interfaces 2 in panel A), are less extensive ($24 \text{ nm}^2 \text{ interface}^{-1}$) and were designated as the interfaces holding together the two dimers to form the tetramer. The interfaces along the exact 2-fold rotational axis of symmetry normal to the plane of the page (interfaces 3 in panel A) are almost nonexistent because of the squashing of the tetrahedron.

Almost all tetramers of dihedral symmetry of point group $222(D_2)$ are constructed along these lines, with an almost nonexistent or a completely **nonexistent pair of interfaces** and with one of the remaining two pairs of interfaces being more extensive than the other. Consequently, tetramers of dihedral symmetry, which include almost all tetrameric proteins, are **dimers of dimers** and the complete structure of such a dimer of dimers can be designated as that of an $(\alpha_2)_2$ tetramer to distinguish it from an α_4 tetramer of cyclic symmetry of point group $4(C_4)$. The same points noted in the description of the evolution of a rotationally symmetric dimer from an ancestral monomeric protein are of equal validity in describing the evolution of a rotationally symmetric dimer of dimers from an ancestral dimeric protein. Also, for the same reasons, a dimer of dimers is a closed structure.

Because each of the three pairs of interfaces in a tetramer with dihedral symmetry is completely different, they each have different strengths, reflected in their **free energies of association**. One of the two or three pairs of interfaces must be the strongest, and if there is any dissociation of an $(\alpha_2)_2$ tetramer into its constituent α_2 dimers, it is usually this strongest pair of interfaces that will be retained, one in each of the two dimers. For example, the uncomplicated, **reversible dissociation** of $(\alpha_2)_2$ homotetrameric, enzymatically active bacterial 6-phosphofructokinase into enzymatically inactive α_2 dimers results from the sundering of the less extensive



of the two distinct types of interfaces that produce the structure of dihedral symmetry.⁸⁰⁻⁸²

Hemoglobin is an honorary homotetramer. It is built from two different polypeptides, α and β , each present in two copies to provide the four protomers. The two polypeptides are homologous in sequence⁸³ and their tertiary structures are superposable.⁸⁴ In the heterotetramer, the four protomers of hemoglobin occupy the same arrangement as the four protomers of 2,2-dialkylglycine decarboxylase (pyruvate), but two of the rotational axes are rotational axes of pseudosymmetry. The $\alpha\alpha$ in-

terface and the $\beta\beta$ interface, through which runs the molecular 2-fold axis of symmetry, are the rudimentary pair, so the structure can be represented as $(\alpha\beta)_2$.

The oxygenated form of hemoglobin in solution participates in the dissociation



and the value of the dissociation constant

$$K_d = \frac{[\alpha\beta]^2}{[(\alpha\beta)_2]} \quad (9-3)$$

is $1 \mu\text{M}$.⁸⁵ That the same interface always remains in the dimer follows from the fact that no hybrid dimers of the type $\alpha'\beta$ or $\alpha\beta'$ are formed under conditions where the reaction of Equation 9-2 is rapidly interconverting tetramers and dimers in a mixture of two hemoglobins, $(\alpha\beta)_2$ and $(\alpha'\beta')_2$, from two species, dog and human, respectively, even though **hybrids** of the type $(\alpha\beta)(\alpha'\beta')$ did form readily.⁸⁶ In this experiment, hemoglobins from the two species were used to permit isoelectrophoretic separation of the hybrids (as in Figure 8-18). As a control, it was shown that the hybrid dimers $\alpha'\beta$ and $\alpha\beta'$ could be artificially formed and easily separated isoelectrophoretically, from each other and from the parent dimers $\alpha\beta$ and $\alpha'\beta'$. The $\alpha'\beta$ and $\alpha\beta'$ hybrids also could not be shuffled by the reaction of Equation 9-2. Therefore one of the two different pairs of interfaces between α and β subunits in the hemoglobin tetramer⁸⁴ must be much stronger than the other.

A number of other oligomeric proteins also participate in dissociations the equilibrium constants for which are large enough to be measured. In fact, the dimer of interleukin-8 that is observed crystallographically has a dissociation constant so large that the protein is actually a monomer at physiological concentrations.⁸⁷ Most oligomeric proteins, however, have interfaces so strong that their dissociation does not occur within normal ranges of concentration.

A **molecular asymmetric unit** is the smallest unit of the structure of an oligomeric molecule that, when submitted to the appropriate symmetry operations, creates the entire structure. The individual subunits of the cyclic oligomers in Figures 9-9 to 9-12 are the molecular asymmetric units of their respective molecules. In the $(\alpha_2)_2$ tetramer of 2,2-dialkylglycine decarboxylase (pyruvate) (Figure 9-13B), the molecular asymmetric unit is, by inspection, the single folded polypeptide. If the one polypeptide is positioned in space, its image is rotated 180° around any one of the 2-fold rotational axes of symmetry, and another identical folded polypeptide is placed where the image of the first has thus been positioned, one of the three possible dimers in the tetramer is created. If the image of this dimer is then rotated about

another of the 2-fold rotational axes of symmetry and another identical dimer is placed where the image of the first has been positioned, then the entire tetramer is created.

If the tetrahedron of Figure 9-13A is squashed flat, the structure becomes a **ring of four protomers** that vaguely resembles a ring with cyclic symmetry of point group $4(C_4)$, often with a large hole in the middle.^{88,89} Nevertheless, it is easy to distinguish the former from the latter because its ring has dihedral symmetry of point group $222(D_2)$. The orientation of the protomers in an oligomer with dihedral symmetry alternates up-down-up-down around the ring, and the two orthogonal 2-fold rotational axes of symmetry in the plane of the ring between every pair of subunits, which do not exist in a structure with cyclic symmetry, remain.

As with protocatechuate 3,4-dioxygenase (Figure 9-2), which is a rare example of a dimer the subunits of which are arranged around a screw axis of symmetry, *lac* repressor from *E. coli*⁹⁰ and the lectin from *Arachis hypogaea*⁹¹ are tetramers in each of which a pair of rotationally symmetric dimers of identical subunits is arranged around a screw axis of symmetry.

Ribulose-phosphate 3-epimerase from chloroplasts of *Solanum tuberosum* (Figure 9-14A)⁹² is an $(\alpha_2)_3$ hexamer with the symmetry of **dihedral point group 322(D_3)** (Figure 9-14B), phosphoribulokinase from *Rhodobacter sphaeroides* (Figure 9-15A)⁹³ is an $(\alpha_2)_4$ octamer with the symmetry of **dihedral point group 422(D_4)** (Figure 9-15B), and peroxiredoxin from *Crithidia fasciculata* (Figure 9-16)⁹⁴ is an $(\alpha_2)_5$ decamer with the symmetry of **dihedral point group 522(D_5)**. The crystallographic molecular model of the $(\alpha_2)_3$ hexamer of ribulose-phosphate 3-epimerase has a central molecular 3-fold rotational axis of symmetry and three molecular 2-fold rotational axes of symmetry at angles of 60° to each other, each of them orthogonal to the central axis and one of them exact. The crystallographic molecular model of the $(\alpha_2)_4$ octamer of phosphoribulokinase has a central, exact 4-fold rotational axis of symmetry and four exact 2-fold rotational axes of symmetry at angles of 45° to each other and all of them orthogonal to the central axis. The crystallographic molecular model of the $(\alpha_2)_5$ decamer of peroxiredoxin has five molecular 2-fold rotational axes of symmetry at angles of 36° to each other, all of them orthogonal to the central molecular 5-fold rotational axis of symmetry.

In the dihedral point groups of odd fold (3-fold, 5-fold, and 7-fold), the **interfaces found at the two ends of each 2-fold rotational axes of symmetry**, although rotationally symmetric about the axis, are different from each other (Figure 9-14B); in the dihedral point groups of even fold (4-fold and 6-fold), the interfaces found at the two ends of each 2-fold rotational axis of symmetry are the same, but there are two different kinds of 2-fold rotational axes of symmetry that alternate around the central axis (Figure 9-15B). Nevertheless, in both instances, odd

and even, there are only two types of interfaces associated with the 2-fold rotational axes of symmetry regardless of how large the fold (hence the notation $n22$). A third type of n -fold symmetric interface can occur across the central n -fold axis.

In oligomeric proteins with dihedral symmetry there are usually a set of n strong, identical interfaces distributed around the central axis that connect pairs of subunits into dimers. Examples are the three interfaces approximately normal to the plane of the page each connecting an upper subunit with a lower subunit in Figure 9-14A and the five interfaces at about 2 o'clock, at about 5 o'clock, at 7 o'clock, at about 10 o'clock and at about 12 o'clock in Figure 9-16. As is the case with those in isolated dimers with cyclic symmetry, each of these interfaces has a 2-fold rotational axis of symmetry running through its center. In ribulose-phosphate 3-epimerase, phosphoribulokinase, and peroxiredoxin, these interfaces forming the dimers are more extensive than the interfaces joining the dimers into the hexamer, the octamer, or the decamer, respectively, so the proteins are a **trimer of dimers**, a **tetramer of dimers**, and a **pentamer of dimers**, respectively, all with dihedral symmetry.

There are three configurations in which such dimers can be assembled into rings with dihedral symmetry: eclipsed, staggered, and splayed (Figure 9-17). The dimers in ribulose-phosphate 3-epimerase (Figure 9-14A) are **eclipsed**; those in ribulokinase (Figure 9-15A) are **staggered**; and those in peroxiredoxin (Figure 9-16) are **splayed**. In ribulose-phosphate 3-epimerase the only interfaces holding the three dimers together are the six identical ones, three between the subunits in the upper ring and three between the subunits in the lower ring,⁹² and in peroxiredoxin the only interfaces holding five dimers together are the five identical ones, each between

an upper subunit and a lower subunit in neighboring dimers.⁹⁴ In phosphoribulokinase, however, the six identical interfaces connecting the three upper subunits to each other and the three lower subunits to each other are as extensive as the five identical interfaces between a lower subunit of one dimer and an upper subunit of a neighboring dimer, but the interfaces holding the dimers

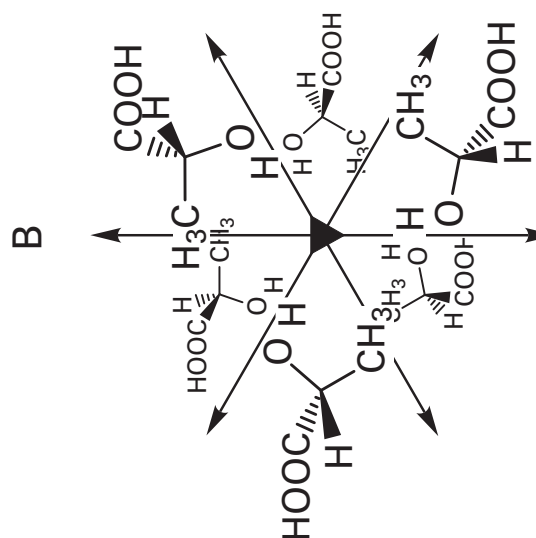
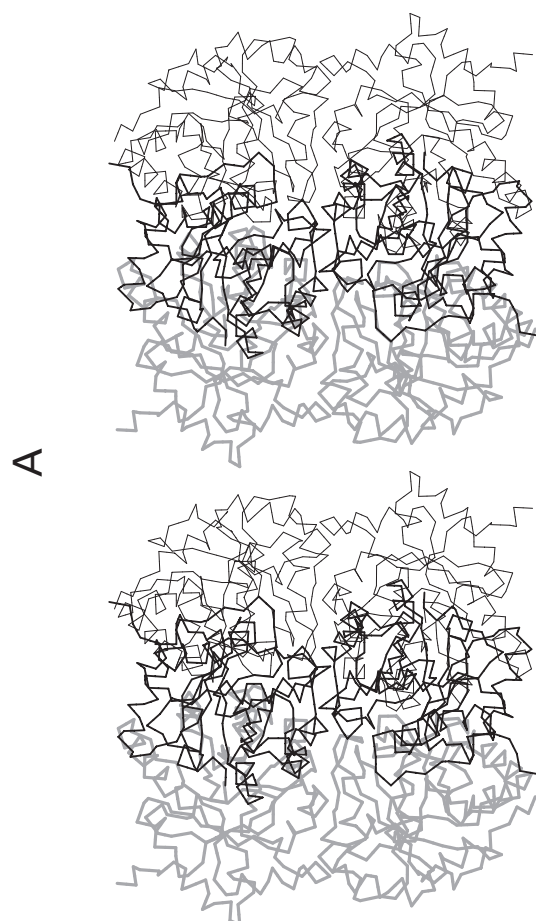
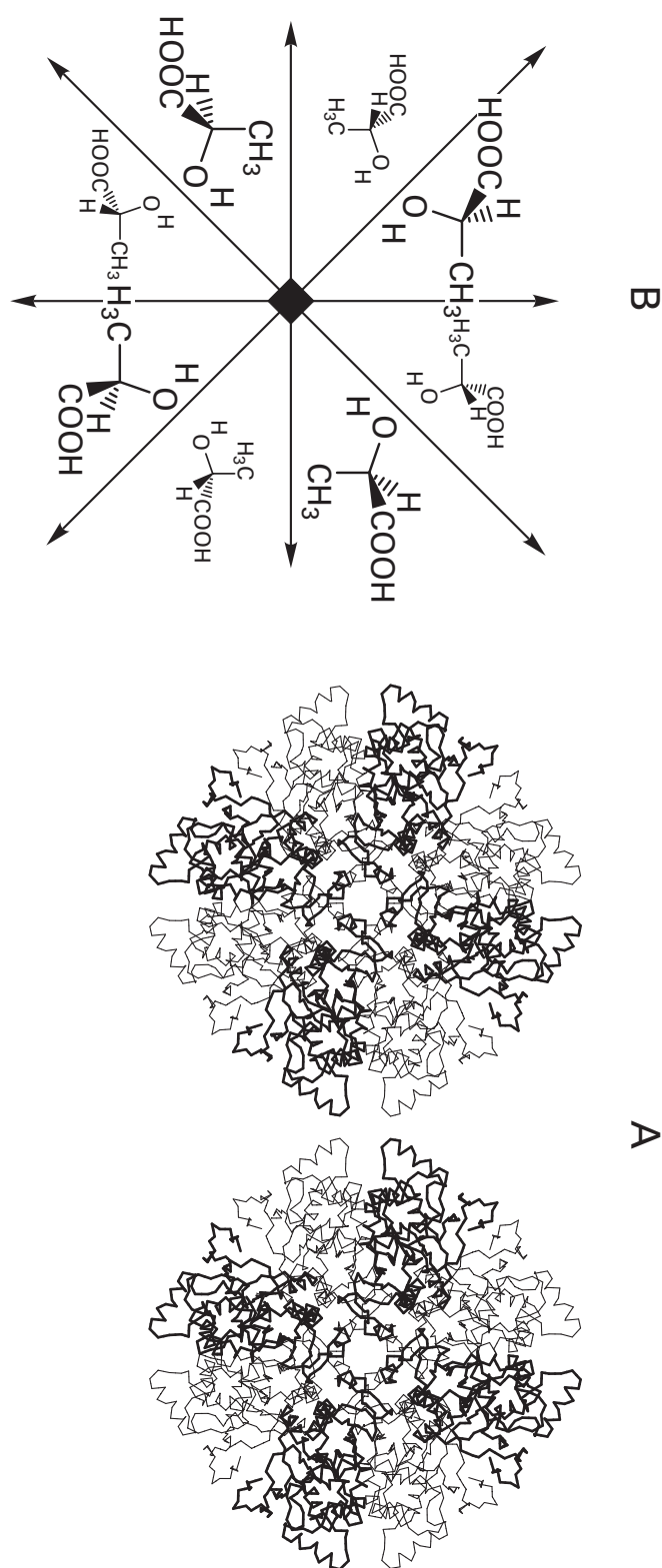


Figure 9-14: A trimer of eclipsed dimers with dihedral symmetry of point group $322(D_3)$. (A) α -Carbon diagram of the homohexamer of ribulose-phosphate 3-epimerase from the chloroplasts of *S. tuberosum*, drawn from the crystallographic molecular model (space group $P3_121$) of the enzyme.⁹² The view is along one of the dihedral 2-fold rotational axes of symmetry, which coincides with a crystallographic rotational axis of symmetry. The central 3-fold noncrystallographic molecular rotational axis of symmetry is vertical in the plane of the page. The three eclipsed dimers are drawn with lines of different widths and different shading. There are no interactions between an upper subunit and any lower subunit other than the one with which it forms a dimer. The accessible surface area buried in the interface within each dimer is $21 \text{ nm}^2 \text{ interface}^{-1}$, and that within an interface connecting the dimers around the central 3-fold rotational axis of symmetry is $15 \text{ nm}^2 \text{ interface}^{-1}$. (B) Six molecules of lactic acid arranged with dihedral symmetry of point group $322(D_3)$. The molecules of lactic acid at 1, 5, and 9 o'clock are above the plane of the page, and those at 3, 7, and 11 o'clock are below the plane of the page. The three 2-fold rotational axes of the dihedral point group (arrows with full heads) are in the plane of the page, and the solid triangle denotes a 3-fold rotational axis of symmetry normal to the plane of the page. This drawing was produced with MolScript.⁴⁸⁵



together are three times more extensive than either of these types.⁹³ In the staggered hexamer of sulfate adenylyltransferase from *Penicillium chrysogenum*, which also has the symmetry of point group $322(D_3)$, the interfaces connecting each subunit in one trimer with one subunit in the other trimer and with another subunit in the other

Figure 9-15: A tetramer of staggered dimers with dihedral molecular symmetry of point group $422(D_4)$. (A) α -Carbon diagram of phosphoribulokinase from *R. sphaeroides*, drawn from the crystallographic molecular model (space group $P4_32$) of the enzyme.⁹³ The view is along the central 4-fold rotational axis of symmetry. All molecular axes of symmetry coincide with crystallographic axes of symmetry. The four subunits to the front are drawn with thin lines; the four to the back, with thick lines. The accessible surface area buried in the interface within each of the four fundamental dimers centered on the vertical and horizontal 2-fold rotational axes of symmetry is $35 \text{ nm}^2 \text{ interface}^{-1}$; that within an interface connecting a subunit to the front or a subunit to the back within one of those dimers to a subunit to the front or a subunit to the back, respectively, in a neighboring dimer is $6.4 \text{ nm}^2 \text{ interface}^{-1}$; and that within an interface connecting a subunit to the front or a subunit to the back within one of those dimers to a subunit to the back or a subunit to the front, respectively, in a neighboring dimer is $9.0 \text{ nm}^2 \text{ interface}^{-1}$. This drawing was produced with MolScript.⁴⁶⁵ (B) A drawing of eight molecules of lactic acid arrayed with dihedral symmetry of point group $422(D_4)$. The molecules at 2, 5, 8, and 11 o'clock are above the plane of the page, and the other four are below the plane of the page. The four 2-fold rotational axes of symmetry of the dihedral point group are in the plane of the page, and the solid square denotes a 4-fold rotational axis of symmetry normal to the plane of the page. The atomic coordinates on which this drawing is based were provided by David H. T. Harrison.

trimer, respectively, are formed by two different domains in that subunit.⁹⁵ In an eclipsed structure, a loop⁹⁶ or terminal segment⁹⁷ of polypeptide sometimes extends across the central plane of the structure to connect the upper subunit of one dimer to the lower subunit of a neighboring dimer and vice versa.

There are other examples of eclipsed trimers of dimers,^{98,99} staggered tetramers of dimers,^{100,101} and splayed pentamers of dimers¹⁰² as well as splayed trimers of dimers,^{103,104} a splayed tetramer of dimers,^{105,106} an eclipsed pentamer of dimers,⁹⁷ and an eclipsed hexamer of dimers.¹⁰⁷ In all of these proteins with dihedral symmetry, the fact that the most extensive interfaces in the structure are those holding the dimers together suggests that evolution assembled these oligomers from preexisting symmetric dimers. This conclusion would make sense because dimeric proteins are far more common than either trimeric proteins with cyclic symmetry, tetrameric proteins with cyclic symmetry, or pentameric proteins with cyclic symmetry.

One also comes to the conclusion that the dimer is the most common **fundamental unit** in the assembly of oligomeric proteins because there are examples of proteins that have different dihedral quaternary structures even though the α_2 dimers from which they are constructed are obviously related and are themselves held together by homologous interfaces. Human nucleoside-diphosphate kinase is a hexamer with dihedral symmetry of point group $322(D_3)$, but nucleoside-diphosphate kinase from *Myxococcus xanthus* is a tetramer with dihedral symmetry of point group $222(D_2)$, even though the α_2 dimer from human protein is readily superposable on the α_2 dimer of the protein from *M. xanthus*.¹⁰⁸ Within the

family of superoxide dismutases, a fundamental α_2 dimer, which is superposable among the proteins from different species, either stands alone or is arranged in several different ways to form distinct $(\alpha_2)_2$ tetramers.¹⁰⁹ The cytoplasmic ribulose-phosphate 3-epimerases from animals and plants are α_2 dimers rather than $(\alpha_2)_3$ hexamers, as are those from chloroplasts (Figure 9-14A).^{92,110} In fact, the dimeric ribulose-phosphate 3-epimerases from the cytoplasms of fungi and animals are more closely related to the dimeric ribulose-phosphate 3-epimerase from a given plant than is the hexameric ribulose-phosphate 3-epimerase from its own chloroplasts. The dihydrodipicolinate synthases from *Nicotiana sylvestris* and *E. coli* are tetramers with dihedral symmetry assembled from superposable α_2 dimers, but the dimers face each other in opposite directions in the two proteins.¹¹¹ Consequently, it seems that most oligomers with dihedral symmetry are $(\alpha_2)_2$ dimers of dimers, $(\alpha_2)_3$ trimers of dimers, $(\alpha_2)_4$ tetramers of dimers, $(\alpha_2)_5$ pentamers of dimers, and $(\alpha_2)_6$ hexamers of dimers.

There are of course exceptions to the observation that most proteins with dihedral symmetry are assembled from symmetric dimers. Histidine decarboxylase from *Lactobacillus* is a **dimer of trimers**,¹¹² glutamate-ammonia ligase from *Salmonella typhimurium* is a **dimer of hexamers**,¹¹³ and human serum amyloid P component is a **dimer of pentamers**.¹¹⁴ There are also two proteins, a chaperonin¹¹⁵ and an intracellular, multifunctional endopeptidase¹¹⁶ that are $[(\alpha\beta)_7]_2$ dimers of heptamers. The fact that the octamer of IMP dehydrogenase from *Tritrichomonas foetus* with dihedral symmetry of point group $422(D_4)$ dissociates into two tetramers with a dissociation constant of $1 \mu\text{M}$ demonstrates that it is a dimer of two tetramers, each of cyclic symmetry.¹¹⁷

In proteins with cyclic symmetry such as dimers¹¹⁸⁻¹²⁰ or trimers, tetramers, and pentamers¹²¹ as well as proteins with dihedral symmetry,¹⁰⁶ related superposable monomers have been assembled by evolution around different interfaces or into different quaternary structures. The malleability of the arrangements of protomers with both dihedral and cyclic symmetry within sets of oligomers of the same family of proteins or even the same species of proteins¹²⁰ leads to the conclusion that the quaternary structure of a protein provides little information about its **evolutionary relationships**.

If, however, the quaternary structure is retained in the same protein from widely different species of organisms, the evolutionary constraints within the interfaces producing that quaternary structure are as stringent as those in the interior of the protein. For example, in the heterodimer of isoform 1 and isoform 2 of the R2 protein of ribonucleoside-diphosphate reductase from *S. cerevisiae*, the interface closely resembles those in the homodimers of the same protein from *E. coli* and *M. musculus* even though the folded polypeptides of isoforms 1 and 2 from *S. cerevisiae* differ significantly from

their folded homologues from *E. coli* and *M. musculus* at their peripheries.¹²²

When monomers are transformed by evolution into a homooligomeric ring with cyclic symmetry (Figures 9-9, 9-11, and 9-12) or when homodimers are transformed by evolution into a homooligomeric ring with dihedral symmetry (Figures 9-14A, 9-15A, and 9-16), a face and the complement to that face must be created on the surface of the monomer or the monomer in the dimer, respectively. The face and its complement must be positioned,

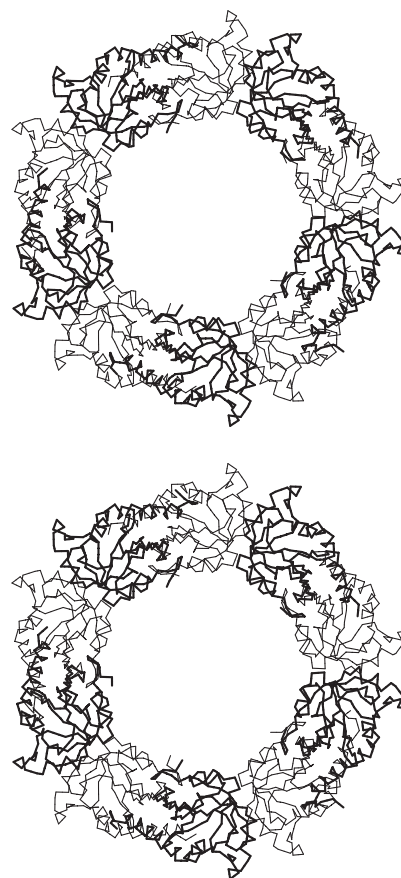


Figure 9-16: A pentamer of played dimers with dihedral symmetry of point group $522(D_5)$. An α -carbon diagram of peroxidase from *C. fasciculata*⁹⁸ is viewed down its central noncrystallographic, molecular 5-fold rotational axis of symmetry. Because the space group of the crystal is $P2_1$, none of the 2-fold molecular axes of symmetry coincides with a crystallographic axis of symmetry. The subunits are drawn with lines of different thickness. Of the total accessible surface area of each subunit, 13% is buried in one interface and 8.5% in the other interface. This drawing was produced with MolScript.⁴⁸⁵

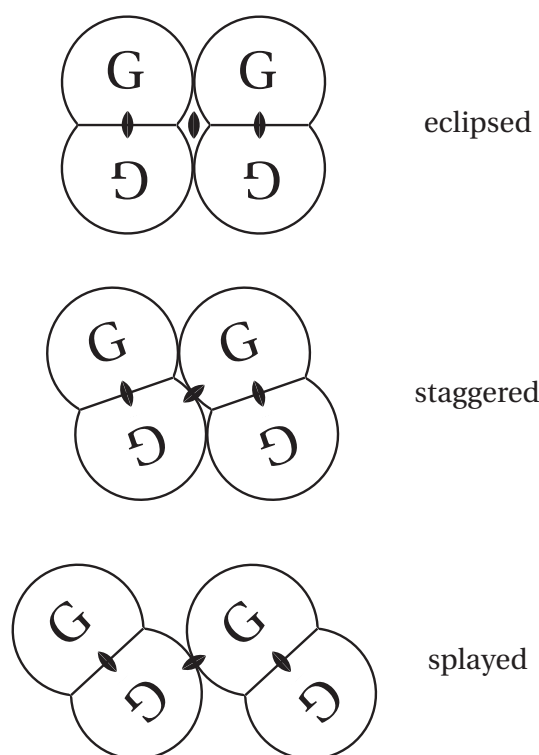


Figure 9-17: Interfaces between dimers in rings of dihedral symmetry. When dimers stand vertically around the ring, they are **eclipsed** when viewed along the central axis, and the interfaces connecting those dimers to each other are all identical and symmetrically displayed around dihedral 2-fold rotational axes of symmetry (Figure 9-14A). When dimers tilt sufficiently so that both the pair of interfaces symmetrically displayed around each of their central 2-fold rotational axes of symmetry and an interface along the adjacent 2-fold rotational axis of symmetry between an upper subunit and a lower subunit in a neighboring dimer are present simultaneously, the dimers are **staggered** when viewed along the central axis (Figure 9-15A). When the dimers tilt so much that only the interface between an upper subunit and a lower subunit in a neighboring dimer can form along alternate 2-fold rotational axes of symmetry, the dimers are **splayed** when viewed along the central axis (Figure 9-16). The transformation from staggered to splayed is formally equivalent to squashing the tetrahedron of spheres and losing interfaces 3 in Figure 9-13.

respectively, on the surface of the monomer or the monomer in the dimer so that they are directed relative to each other at angles of 60° , 90° , 108° , or 120° , consistent with the angles relating the orientations of the interfaces in each monomer in a triangular, square, pentagonal, or hexagonal ring, respectively. Only in this way can every interface lock fully into place in the complete ring. Because it is only when all of the interfaces are fully locked that the full standard free energy of formation is realized, most of the existing **rings are continuous and symmetric**.¹²³ In HNRNP arginine methyltransferase from *S. cerevisiae*, however, the face and its complement on each dimer that produces the trimer of dimers have evolved so they are directed outward at a little more than 60° to each other. As a result, one of the three interfaces cannot quite lock into place, and there is a gap in the ring.¹²⁴

The homooctamer of the repressor from bacteriophage λ , which has dihedral symmetry, dissociates neither into four cyclic dimers nor into two cyclic tetramers as do most homooctamers of dihedral symmetry. Instead, it splits apart along a cleavage plane parallel to the 4-fold rotational axis of symmetry and one of the 2-fold rotational axes of symmetry, equivalent to a horizontal plane normal to the plane of the page in Figure 9-15B, to produce two tetramers that each retain an exact 2-fold rotational axis of symmetry, equivalent to the vertical 2-fold rotational axis of symmetry in Figure 9-15B.^{125,126} The other two now local 2-fold rotational axes of symmetry, however, in each of these tetramers end up at an 80° angle to each other instead of the 90° angle they were forced to assume in the octamer. If two of these relaxed tetramers were to form an octamer, there would necessarily be a gap of 20° at one of the 2-fold rotational axes of symmetry. Unlike what happens in arginine methyltransferase, when the octamer of the repressor forms, the tetramers distort to close the gap and produce a structure with complete dihedral symmetry of point group $422(D_4)$. Because the strain of this distortion is relieved on dissociation, the octamer dissociates along an unexpected set of interfaces.

In many proteins, the folded polypeptides of which contain **internal duplications**, the two superposable, duplicated domains are related by a 2-fold rotational axis of pseudosymmetry. Thiosulfate sulfurtransferase is an example of such an arrangement (Figure 9-18).¹²⁷ The amino acid sequences of its two domains are not significantly related to each other (11% identity with no gaps upon structural alignment). Nevertheless, one domain superposes upon the other with a root mean square deviation of 0.2 nm (117 out of 146 α carbons) upon a 179° rotation about the axis of pseudosymmetry normal to the page in Figure 9-18 and a translation along that axis of less than 0.1 nm. As with other domains of this level of kinship, the central portions coincide well, but loops connecting elements of secondary structure differ, often dramatically. The two internally duplicated domains of methionyl aminopeptidase from *E. coli* (Figure 7-7B) superpose on each other upon a rotation of 174° and a translation of 0.06 nm along a screw axis of pseudosymmetry between them¹²⁸ and those of porcine pepsin superpose upon a rotation of 173° .¹²⁹ Other examples of proteins formed by internal duplication in which the two domains are related by an approximate 2-fold rotational axis of symmetry are arabinose binding protein,¹³⁰ chymotrypsinogen,¹³¹ and sulfite reductase.¹³²

Presumably, these 2-fold **rotational axes of pseudosymmetry** are the remains of the 2-fold rotational axes of symmetry in the dimers of two identical protomers that were the ancestors of each of these proteins before the gene duplication occurred. The duplicated polypeptide incorporated the original rotational axis of symmetry, but following the duplication the two halves began to evolve separately and diverge. Because the sequences have diverged, the superposition is not between struc-

tures with the same amino acid sequence, and the axis has become a 2-fold rotational axis of pseudosymmetry. There is at least one example of a protein in which a single folded polypeptide has an internal 3-fold rotational axis of pseudosymmetry^{133,134} and one with an internal 5-fold rotational axis of pseudosymmetry,¹³⁵ presumably the remains of an ancestral trimer and an ancestral pentamer, respectively. The first and the third of the three domains in the single folded polypeptide composing pyruvate oxidase from *Lactobacillus plantarum* superpose on each other with a root mean square deviation of 0.19 nm upon rotation of 190° around a rotational axis of pseudosymmetry between them.¹³⁶ These two halves of the duplicated gene, when unduplicated, encoded the two identical subunits of an α_2 dimer. There is, however, another unrelated domain of about the same size inserted into the polypeptide between the two that nevertheless remain symmetrically arrayed.

There are also examples of proteins in which an early gene duplication, which produced two domains now related by a 2-fold axis of pseudosymmetry bearing witness to that duplication, was then followed by a later gene duplication. This later duplication produced another 2-fold axis of pseudosymmetry relating two copies of the product of the early duplication.^{137,138} Because the three 2-fold axes of pseudosymmetry in each of these proteins, the two early ones and the one later one, are almost parallel to each other, the original protein must have been an α_2 dimer and after its two identical subunits were fused, the product of the fusion then evolved to become an α_2 dimer, the identical subunits of which were then fused. One of these proteins is now again an α_2 dimer of two subunits, each with four internally duplicated, symmetrically arrayed domains,¹³⁷ perhaps on its way to yet another duplication.

Not all folded polypeptides with internally repeating domains display such rotational axes of pseudosymmetry, and there is no direct correspondence between internally repeating domains and rotational axes of symmetry. Immunoglobulin G (Figure 7–13) contains two different polypeptides, one long and one short, that are both composed of internally repeating domains. Within neither of the polypeptides are any of the adjacent internally repeating domains related by a rotational axis of pseudosymmetry. Only proteins that were symmetric oligomers before the replication occurred can retain vestiges of their former rotational axes of symmetry.

Phaseolin from *Phaseolus vulgaris* contains an α_3 trimer with cyclic symmetry.¹³⁹ The subunit of phaseolin is the product of a gene duplication, and its two superposable domains are related by a 2-fold rotational axis of pseudosymmetry. In the trimer, these three 2-fold rotational axes intersect the central 3-fold rotational axis of symmetry and are normal to it. Consequently, the ancestral protein must have been a hexamer with dihedral symmetry, the constituent α_2 dimers of which were fused by gene duplication. It is certain that such an event

occurred to the ancestor of 5-carboxymethyl-2-hydroxymuconate Δ -isomerase. It is also a trimer with subunits containing internally duplicated domains around local 2-fold rotational axes of pseudosymmetry orthogonal to the central 3-fold rotational axis of symmetry. 4-Oxalocrotonate tautomerase is an enzymatically related protein, the monomer of which is homologous to both of the duplicated halves of the monomer of 5-carboxymethyl-2-hydroxymuconate Δ -isomerase. 4-Oxalocrotonate tautomerase is an α_6 hexamer with dihedral symmetry, the entire structure of which super-

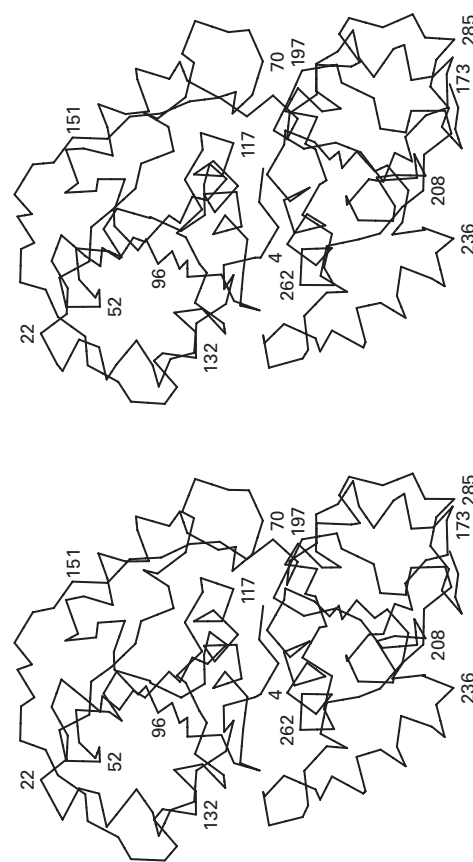


Figure 9-18: α -Carbon diagram drawn from the crystallographic molecular model of the thiosulfate sulfurtransferase.¹²⁷ The amino acid sequence of the protein does not contain significant internal homology, but the internal duplication is manifest in the crystallographic molecular model of the folded polypeptide that forms this monomeric protein. The two domains, above and below a horizontal plane passing through its center, are related to each other by a 2-fold rotational axis of pseudosymmetry normal to the plane of the page running through the center of the protein. This drawing was produced with MolScript.⁴⁸⁵

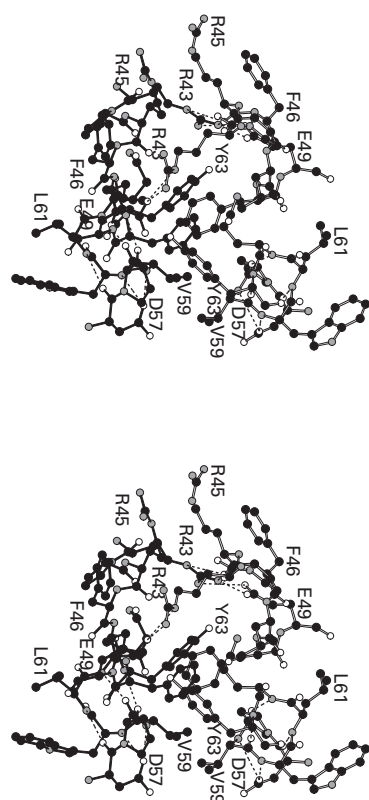


Figure 9-19: An interface between the two subunits of an α_2 dimer. A portion of the crystallographic molecular model of protein Cro from bacteriophage 434¹⁴⁵ is represented. The backbone and side chains from the upper subunit are drawn with white sticks; those from the lower, with black sticks. The molecular 2-fold rotational axis of symmetry runs horizontally through the center of the interface. Because there are no crystallographic 2-fold rotational axes of symmetry in the space group $P2_1$ of the crystal, the 2-fold axis of symmetry is a molecular axis, and at the periphery of the interface, side chains differ noticeably from their twins in conformation (for example, the two Arginines 43 and the two Arginines 45).⁴⁶⁵ This drawing was produced with MolScript.⁴⁶⁵

poses on the trimer of 5-carboxymethyl-2-hydroxymuconate Δ -isomerase.¹⁴⁰ The tautomerase retains the quaternary structure of the common ancestor of itself and the isomerase, while the isomerase is an internal duplication of its ancestor.

The **interfaces between folded polypeptides** in an oligomeric protein are almost indistinguishable in their **hydropathy** from the interfaces between secondary structures within a folded polypeptide.¹⁴¹ In the interfaces between subunits, $65\% \pm 4\%$ of the accessible surface area of the complementary faces is nonpolar and $22\% \pm 7\%$ is polar but uncharged. These percentages are indistinguishable from those for interfaces between ele-

ments of secondary structure within a subunit ($70\% \pm 5\%$ and $24\% \pm 6\%$, respectively). The average interface between subunits, however, is enriched in its percentage of charged accessible surface area ($13\% \pm 5\%$) relative to the percentage of charged accessible surface area ($6\% \pm 2\%$) in the average interface between secondary structures within a subunit. Most of this **elevation in charged accessible surface area** is attributable to the guanidinium functional groups of arginines,¹⁴¹ and arginines are the donors in 33% of all of the hydrogen bonds located within interfaces between subunits. Otherwise, the amino acid composition of an interface is about the same as that of the buried interior of a folded polypeptide.

A representative set of ionized and un-ionized **hydrogen bonds** can be found in the interface between the two subunits in the dimeric carboxylesterase ESTA from *A. fulgidus* (Table 9-2). The compositions of the amino acids within interfaces, however, can vary widely. For example, each of the two identical symmetrically displayed interfaces creating the dimer of 4- α -glucanotransferase from *Thermotoga maritima* is formed from 11 side chains that are completely hydrocarbon (valines, leucines, isoleucines, and phenylalanines),¹⁴³ while the interface between the two subunits of phosphopyruvate hydratase from *E. coli*, through which the molecular 2-fold rotational axis of symmetry passes, has more than twice as many charged side chains as the average.¹⁴⁴

The interfaces between subunits are probably elevated in their composition of charged side chains because the constituent subunits, once they have folded, have to remain soluble until they can find a complementary partner. The charged groups prevent the folded subunits from aggregating nonspecifically with other proteins while they are searching for a partner.

A typical interface between two subunits of a dimer is that in the crystallographic molecular model of Cro protein from bacteriophage 434, through which the molecular 2-fold axis of symmetry runs (Figure 9-19).¹⁴⁵ To the right, there is a hydrophobic cluster of symmetrically arranged prolines, valines, leucines, phenylalanines, and tyrosines; to the left, a hydrogen-bonded cluster of glutamates, arginines, and the phenolic oxygen-hydrogens of the tyrosines. A significant portion of this polar half of the interface on the left, however, is formed from a hydrophobic cluster of the methylenes from the four arginines sandwiched between the two phenyl rings of two phenylalanines. **Arginine** is probably the most common charged amino acid in such interfaces because it provides charge to keep the folded subunit in solution but also has three methylenes that provide hydrophobic hydrogen-carbon bonds.

Because neither hydrogen bonds nor ion pairs can provide favorable standard free energy of formation to an interface between two folded polypeptides, the standard free energy of formation must be provided by the hydrophobic effect. Although there are clusters of

Table 9-2: Hydrogen Bonds in the Interface between the Subunits in the Dimer of Carboxylesterase ESTA from *A. fulgidus*^a

subunit A ^b	subunit B	length (nm)
Asp7 OD1	Arg269 NH2	0.286
Arg269 NH2	Asp7 OD1	0.295
Ala247 O	Lys295 NZ	0.315
Lys295 NZ	Ala247 O	0.306
Glu274 OE1	Arg298 NH2	0.300
Arg298 NH2	Glu274 OE1	0.29
Ser276 OG	Asp299 OD2	0.258
Asp299 OD2	Ser276 OG	0.248
Ile277 O	Arg281 N	0.287
Arg281 N	Ile277 O	0.286
Arg279 N	Arg279 O	0.285
Arg279 O	Arg279 N	0.289
Tyr280 OH	Gln303 OE1	0.336
Gln303 OE1	Tyr280 OH	0.344

^aThe donor or acceptor in one subunit (subunit A) and its acceptor or donor, respectively, in the other subunit (subunit B) across the interface in the crystallographic molecular model¹⁴² of the homodimer of the carboxylesterase from *A. fulgidus* are tabulated. ^bThe atom performing the donation or the acceptance in each amino acid is in the notation of Figure 4–14. The crystallographic abbreviations for an acyl oxygen of the backbone and an amido nitrogen of the backbone are O and N, respectively. Two dimers together form the crystallographic asymmetric unit so the symmetrically arrayed hydrogen bonds have different lengths.

hydrogen–carbon bonds throughout the interface in the Cro protein (Figure 9–19), such clustering is not necessary for the expression of the **hydrophobic effect**. All that is required is that hydrogen–carbon bonds be removed from the aqueous phase and sequestered within the interface during its formation. What they end up next to is inconsequential.

Almost all of the interfaces between subunits in oligomeric proteins are those that form symmetric dimers. The 2-fold rotational axis of symmetry passes through the center of such an interface, and the detailed atomic contacts (Figures 9–9 and 9–19) or the interdigitations of secondary structure (Figure 9–1A) are **symmetrically duplicated**, except at the periphery where side chains in contact with the water are more flexible (for example, the hydrogen bonding of the Arginines 43 in Figure 9–19). The second most common interface is that forming a trimer with cyclic symmetry. Examples of the **symmetrically triplicated atomic contacts** and hydrogen bonds around a 3-fold rotational axis of symmetry are found around the exact rotational axes in the crystallographic molecular models of dihydrolipoyllysine residue acetyltransferase from *A. vinelandii* (Figure 9–20)³⁰ and the bacterial porins.¹⁴⁶ At the center of a tetramer with dihedral symmetry, where the three orthogonal 2-fold rotational axes of symmetry intersect, the atomic contacts and hydrogen bonds are also arrayed with the same dihedral symmetry about the point of intersection.^{147,148}

When the side chains of the twins across a rotational

axis of symmetry from each other intersect that axis, they usually assume alternate conformations in each of which one of the side chains sits on the axis and the other is pushed to one side and vice versa (Figure 6–29).¹⁴⁹ In the interface between the two subunits in the crystallographic molecular model of isoenzyme 3–3 of glutathione transferase from *Rattus norvegicus*, however, the stack of π molecular orbitals from the two stacked guanidiniums of the Arginines 77 is intersected through its center by the molecular 2-fold rotational axis of symmetry.¹⁵⁰ There are several examples in which the central sulfur–sulfur bond of a cystine sits upon a molecular 2-fold rotational axis of

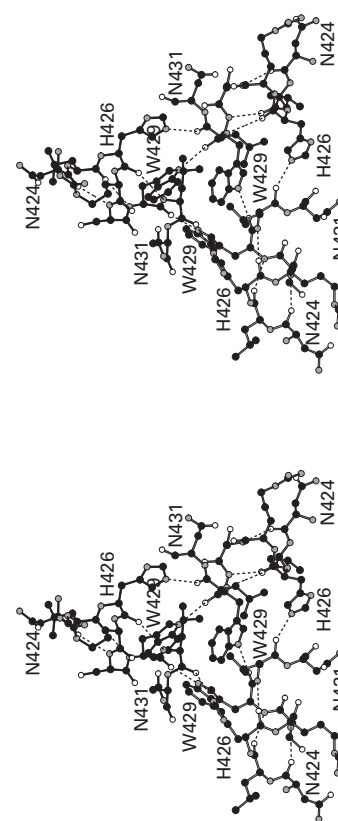


Figure 9-20: An interface between three subunits arranged around a 3-fold rotational axis of symmetry. Portions of three of the 24 subunits in the crystallographic molecular model of the catalytic core (amino acids 382–637) of dihydrolipoyllysine-residue acetyltransferase from *A. vinelandii*³⁰ are presented. The exact molecular 3-fold rotational axis of symmetry, coinciding with one of the 3-fold rotational axes of symmetry in the space group *F*₄₃₂ of the crystal, runs back left to front right through the center of the interface. This drawing was produced with MolScript.⁴⁸⁵

symmetry.^{151–153} In the crystallographic molecular model of the α_2 dimer of human nitric-oxide synthase, a Zn^{2+} cation sits upon the molecular 2-fold rotational axis, symmetrically bound by the two Cysteines 110 and the two Cysteines 115;¹⁵⁴ and in the crystallographic molecular model of the $(\alpha_2)_2$ tetramer of methionine adenosyltransferase from *R. norvegicus*, a K^+ cation sits upon one of the molecular 2-fold rotational axes of symmetry symmetrically bound by the four amido oxygens of the backbone from the two positions 264 and the two positions 265.¹⁵⁵

Most of the interfaces between subunits in oligomeric proteins are formed from two complementary faces, each of which is a portion of the surface of its globular, folded polypeptide, and this portion is no more irregular than the usual exposed surface of the usual globular, folded polypeptide. In some instances, for example, glucose oxidase from *Aspergillus niger*,¹⁵⁶ superoxide dismutase from *Pseudomonas ovalis*,¹⁵⁷ chloramphenicol *O*-acetyltransferase (Figure 9–11), and ribulose-phosphate 3-epimerase (Figure 9–14A), the two faces are almost flat and the resulting interface is almost planar. Often, however, segments of secondary structure or loops between secondary structures will penetrate superficially the subunit across the interface (Figures 9–1A and 9–13B).

In contrast to such classical interfaces, there are interfaces that are formed from **regular arrays of secondary structure**. The most common examples of this type of interface are those in oligomers that are held together by coiled coils of α helices, as are general control protein GCN4 (Figure 6–29) and methyl-accepting chemotaxis protein II (Figure 6–30). The interface between the two subunits in the α_2 dimer of the variant surface glycoprotein from *Trypanosoma brucei* is an antiparallel coiled coil of four α helices, each about 50 aa long;¹⁵⁸ the interface between the two subunits of translocated intimin receptor is an antiparallel coiled coil of four α helices, each about 20 aa long;¹⁵⁹ the interface connecting the three subunits of human mannose-binding protein is a parallel coiled coil of three α helices, each 20 aa long;¹⁶⁰ and the interface between the two α_2 dimers of the tetrameric *lac* repressor from *E. coli* is an antiparallel coiled coil of four α helices, one from each subunit.⁹⁰ The central cores holding together the four subunits of the $(\alpha_2)_2$ tetramers of fumarate hydratase II from *E. coli*,¹⁶¹ histidine ammonia-lyase from *P. putida*,¹⁶² and adenylosuccinate lyase from *P. aerophilum*¹⁶³ are bundles of 20 antiparallel α helices, five from each subunit and each about 30 aa long.

Regular arrays of β structure also are used to connect subunits of oligomeric proteins together. In the α_3 trimer of UDP-*N*-acetylglucosamine diphosphorylase, the interfaces are formed by three identical β helices, one from each subunit that run parallel to each other around the molecular 3-fold rotational axis of symmetry, which coincides with a crystallographic axis.¹⁶⁴ Continuous β -pleated sheets run from one subunit into the other in

the α_2 dimers of concanavalin A,¹⁶⁵ alcohol dehydrogenase (Figure 6–9), κ bungarotoxin (Figure 9–9), and heme-binding protein 23 from *R. norvegicus*.¹⁶⁶ A β sheet of six strands is orthogonally (Figures 6–33 and 6–34) and symmetrically packed against an identical β sheet of six strands from the other subunit in the α_2 homodimer of the lectin from *A. hypogaea*,¹⁶⁷ and a β sheet of six strands is packed in parallel and symmetrically against an identical β sheet of six strands from the other subunit of the α_2 dimer of glucose-fructose oxidoreductase from *Zymomonas mobilis*.¹⁶⁸ In the α_4 cyclic tetramer of each half of the dihedral octamer of dihydroneopterin aldolase from *Staphylococcus aureus*¹⁶⁹ and in the α_2 dimer of each half of the dihedral tetramer of urate oxidase from *Aspergillus flavus*,¹²¹ each subunit contributes four β strands or eight β strands, respectively, to the dramatic antiparallel β barrel of 16 strands in the center of these oligomers. In the pentameric rings within the coat protein of rhinovirus 14, each of the identical subunits contributes one β strand to the parallel β barrel of five strands in the center of the oligomer.¹⁷⁰

Subunits in some homooligomeric proteins are joined to each other by **structural swapping**.¹⁷¹ When the subunit in such an oligomer is in its monomeric form, it is a compact, globular structure. When that monomer combines with another identical monomer, however, one or more of its elements of secondary structure, for example, an amino-terminal α helix,¹⁷² a β hairpin,¹⁷³ two α helices and two strands of β structure,¹⁷⁴ or a structural domain,¹⁷¹ takes the place of its twin on the other subunit, and its twin takes its place on its own subunit. Because the two elements of structure that have swapped are identical to each other, each can fit precisely into the cavity vacated by the other. The resulting α_2 dimer is held together by the respective strands of polypeptide connecting each swapped segment with the rest of its subunit. Often it is only these strands of random meander that hold the two subunits together and no formal interface is formed between them.¹⁷⁵ Conclusive proof of structural swapping requires a crystallographic molecular model of the unswapped monomer and the swapped dimer so that it can be shown that the swapped segments occupy in the same orientation the same locations in the dimer that were occupied by the unswapped segments in their respective monomers.^{171,172,176–178} In the coat protein of bacteriophage MS2, however, two of the three subunits in a homotrimeric substructure have swapped a β hairpin but in the third subunit that β hairpin occupies the same location on its own subunit occupied by the swapped β hairpins on the other two subunits.¹⁷³

The requirements for structural swapping have been examined by site-directed mutation.^{177,179} Site-directed mutation has also been used to convert an otherwise monomeric protein into a structurally swapped dimer.¹⁸⁰

Conclusive evidence of structural swapping is available for only a few proteins, but there are a number of oligomeric proteins in which one or more segments of the

polypeptide forming one subunit reaches over to embrace its neighboring subunit just as the same segments from its neighbor symmetrically embrace it. One example of this is the α_2 dimer of glucose-6-phosphate isomerase from *O. cuniculus*;¹⁸¹ another is the α_3 trimer of 4-chlorobenzoyl-CoA dehalogenase from *Pseudomonas*.⁶² In the symmetric α_2 dimer of human interleukin-5, the carboxy-terminal 24 amino acids of each subunit run symmetrically down one side of the other in random meander and then turn to run across the surface of the other in an α helix.¹⁸² In the α_2 dimer of ADP-ribose diphosphatase from *E. coli*, the first 57 amino acids of each subunit form a three-stranded antiparallel β sheet that lies upon the surface of the remaining globularly folded 153 amino acids of the other subunit,¹⁸³ and these 57 amino-terminal amino acids are missing in monomers from the same family. In the $(\alpha_2)_2$ tetramer of catalase from *Penicillium vitale*, the first 13 amino-terminal amino acids of one subunit are threaded through a large loop of 39 amino acids bulging out of the surface of its symmetric twin and vice versa to form a dimer in which each subunit is hooked to the other.¹⁸⁴

The ultimate expression of cyclic and dihedral symmetry is erythrocrucorin. Erythrocrucorin from *Lumbricus terrestris* is an $\{[(\alpha\beta)(\gamma\delta)]_3\}_{12}\epsilon_{36}$ oligomer of 180 folded polypeptides with a total of 29,916 aa arranged with dihedral symmetry of point group $622(D_6)$.¹⁸⁵ The core of the protein, which confers the dihedral symmetry, is a hexamer of dimers of trimers. Each of the 36 ϵ subunits (240 aa) of this core is first assembled into a homotrimer with cyclic symmetry. Along the 3-fold rotational axis of symmetry of the trimer, the amino-terminal 50 aa of each subunit forms a parallel coiled coil of three α helices holding the three subunits together. These coiled coils then combine in pairs to form dimers of trimers. Six of these dimers of trimers assemble around a central 6-fold rotational axis of dihedral symmetry in a splayed array, the interfaces of which are formed between the coiled coils of α helices. This structure displays the 12 globular

trimers of the carboxy-terminal domains (200 aa) of the ϵ subunits directed outward in upper and lower rings of six.

The globin subunits of the protein, each containing a heme, come in four isoforms (α , 151 aa; β , 145 aa; γ , 153 aa; and δ , 142 aa). They first pair as $\alpha\beta$ dimers and $\gamma\delta$ dimers with cyclic pseudosymmetry of point group $2(C_2)$. An $\alpha\beta$ dimer and a $\gamma\delta$ dimer then assemble around a molecular 2-fold rotational axis of pseudosymmetry, but their own local 2-fold rotational axes of pseudosymmetry intersect the central 2-fold rotational axis of pseudosymmetry of this tetramer at angles of 54° ¹⁸⁶ instead of 90° , as they would in a tetramer of dihedral pseudosymmetry. Nevertheless, the structure is closed not because of steric exclusion but because the subunits are not identical. The δ subunits of three copies of this asymmetric tetramer then associate with each other through three identical interfaces around a 3-fold rotational axis of symmetry to produce a symmetric $[(\alpha\beta)(\gamma\delta)]_3$ trimer of asymmetric tetramers. One of these trimers of tetramers then attaches to each of the 12 globular trimers directed outward from the dihedral core of ϵ subunits to produce the final molecule containing 36 ϵ subunits and 144 globin subunits.

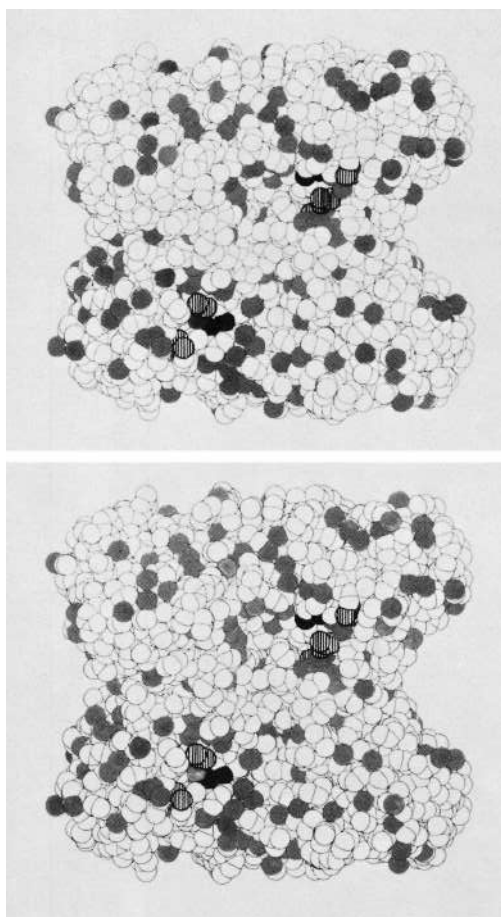
The strategy employed to assemble erythrocrucorin provides a way of assembling more than 100 folded polypeptides into an enormous structure by a hierarchy of symmetries. It is also possible to accomplish the same goal even more dramatically with hexagonally expanded icosahedral symmetry.

Suggested Reading

- Buehner, M., Ford, G.C., Moras, D., Olsen, K.W., & Rossmann, M.G. (1974) Three-dimensional structure of D-glyceraldehyde-3-phosphate dehydrogenase, *J. Mol. Biol.* 90, 25–49.
- Royer, W.E., Jr., Heard, K.S., Harrington, D.J., & Chiancone, E. (1995) The 2.0 Å crystal structure of *Scapharca* tetrameric hemoglobin: cooperative dimers within an allosteric tetramer, *J. Mol. Biol.* 253, 168–186.

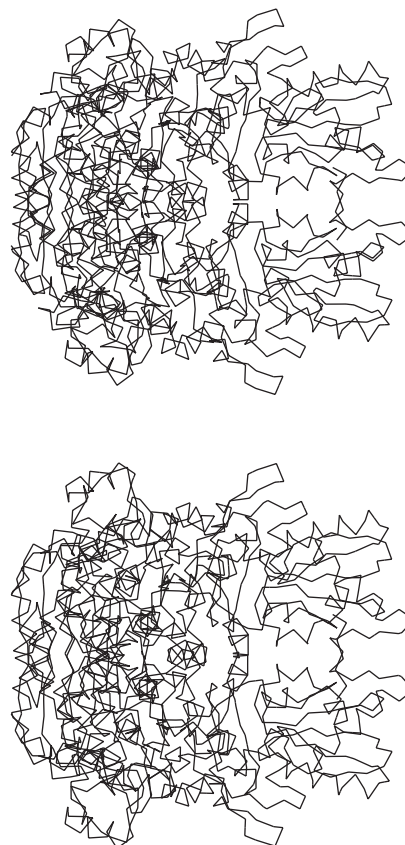
Problem 9-4: Make a xerographic copy of the following figure, reprinted with permission from ref 23, copyright 1983 *European Journal of Biochemistry*.

Using a ruler, draw all of the rotational axes of symmetry on one of the two members of the stereo pairs. Use the abbreviations for rotational axes found in the *International Tables for Crystallography*.¹⁵



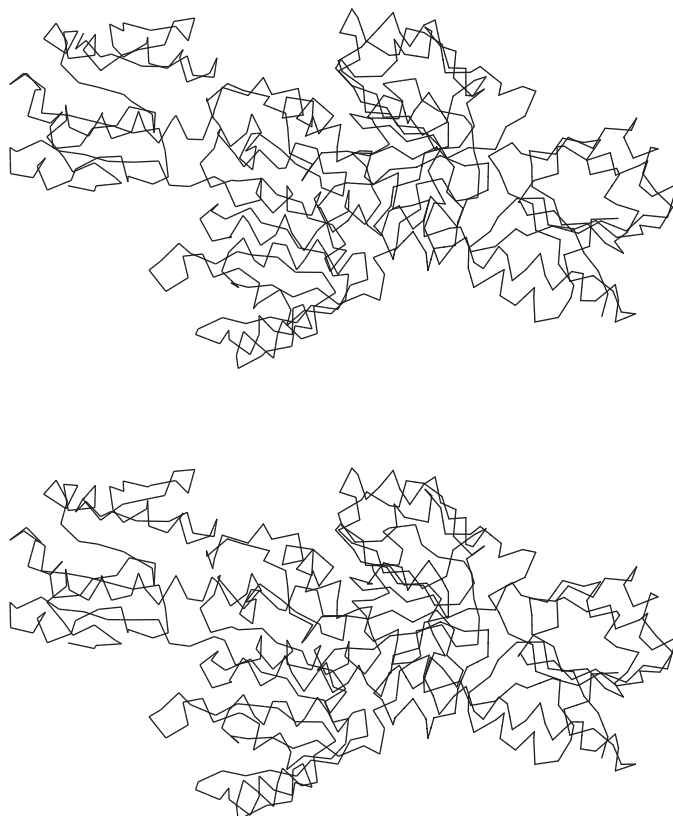
Problem 9-5: The following diagram is based on the crystallographic molecular model of transketolase.³⁹ This drawing was produced with MolScript.⁴⁸⁵

- (A) How many protomers does the protein contain, what types of axes of symmetry does the structure contain, and what are their locations in the structure?



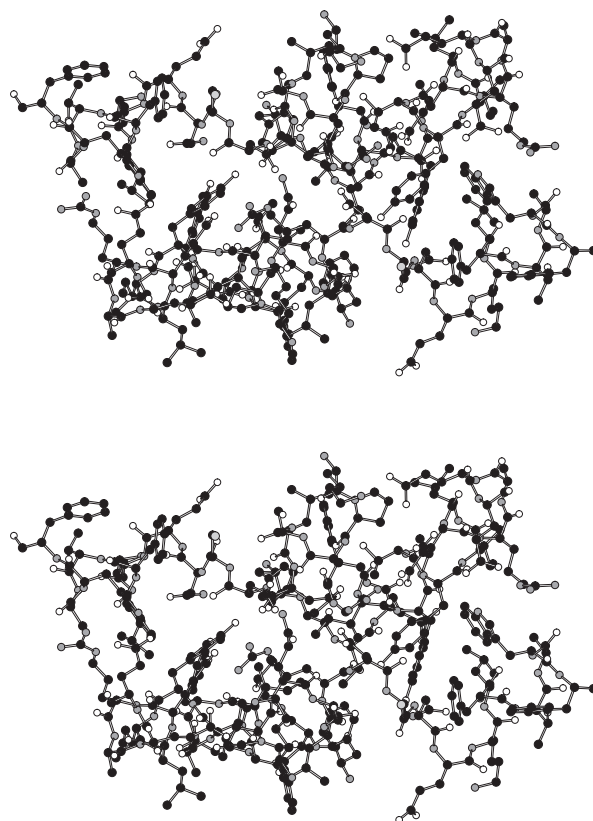
The following diagram is based on the crystallographic molecular model of glycerate dehydrogenase.¹⁸⁷ This drawing was produced with MolScript.⁴⁸⁵

- (B) How many protomers does the protein contain, what types of axes of symmetry does the structure contain, and what are their locations in the structure?

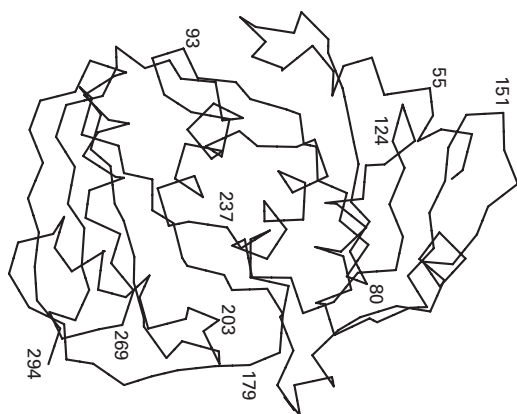
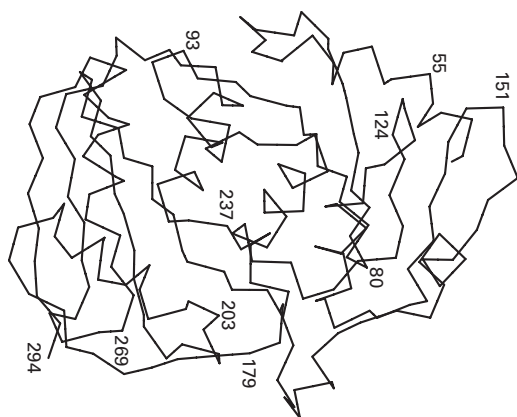


Problem 9-6: The following figure is a drawing of the region of a crystallographic molecular model of glutathione synthase¹⁸⁸ that includes a portion of one of the interfaces between the protomers. This drawing was produced with MolScript.⁴⁸⁵

- (A) What type of axis of symmetry runs through the figure?
- (B) Describe in detail the location of the axis of symmetry in the portion of the structure presented in the figure by naming the three amino acid side chains in each protomer that are immediately adjacent to that axis of symmetry



Problem 9-7: A crystallographic molecular model of cytidine deaminase from *E. coli* has been constructed. The protein is a homodimer formed from two identical folded polypeptide chains, each 294 aa in length. The following figure is a tracing of the α carbons from Glutamate 49 to Alanine 294 in one of the two folded polypeptides in the crystallographic molecular model.¹⁸⁹ This drawing of the crystallographic molecular model was produced with MolScript.⁴⁸⁵



- (A) How many domains are there in the portion of the crystallographic molecular model shown in the figure?
- (B) What criterion did you use to decide how many domains there are?
- (C) By what pseudosymmetry operation are the domains related to each other, and where is the axis of pseudosymmetry located in the figure?
- (D) How did this structure arise during evolution?

Below is an alignment of the segment of the sequence of cytidine deaminase from *E. coli* between Glutamate 49 and Leucine 177 with the segment of the sequence from Glycine 183 to Alanine 294.

```

49      EDALAFALLPLAAACARTPLSNFNVGAIARGVSG
183  GYALTGDALSQAAIAAANRSHMPYSKSPSGVALECKDG
TWYFGANMEFIGATMQQTVHAEQSAISHAWLSGEK--ALAAI
RIFSGSYAENA--AFNPTLPPLQGALILLNLKGYDYPDIQRA
TVN---YTPCG--HCRQFMNELNSGLDLRIHLPGREAHALRD
VLAEKADAPLIQWDATSATLKALGC---HSIDRVLLA 294
YLPDAFGPKDLEIKTLL 177

```

- (E) This alignment is based on the structure shown in the figure above. How was this alignment performed?
- (F) Over the aligned region (Glutamate 49 to Histidine 155 aligned with Threonine 187 to Alanine 294), what is the percentage of identity and how many gaps are there? In calculating the percentage of identity, assume that the length of the aligned region is the average of the lengths of the two aligned sequences.

The figure on the next page is a tracing of the α carbons of the two subunits of cytidine deaminase from *E. coli* in the crystallographic molecular model. Again, in each of the two subunits the structure of the first 48 amino acids has been left out of the tracing. Each subunit starts at Glutamate 49 and ends at Alanine 294, and they have identical sequences. This drawing was produced with MolScript.⁴⁸⁵

- (G) There are three axes in the figure: horizontal, vertical, and normal. Designate correctly each of these three axes as 2-, 3-, 4-, or 6-fold axes of symmetry or axes of pseudosymmetry. One of these axes is a crystallographic axis of symmetry. Which is it?
- (H) What is the point group of the pseudosymmetry and what type of oligomer usually has this type of symmetry?

The amino acid sequence of cytidine deaminase from *B. subtilis* is

```

MNRQELITEALKARDMAYAPYSKFQVGAALLTKDGKVYRGCNIE
NAAYSMCNCAERTALFKAVSEGDTFQMLAVAADTPGPVSPCGA
CRQVISELCTKDVIVVLTLNLQGQIKEMTVEELLPGAFSSDLHD
ERKL

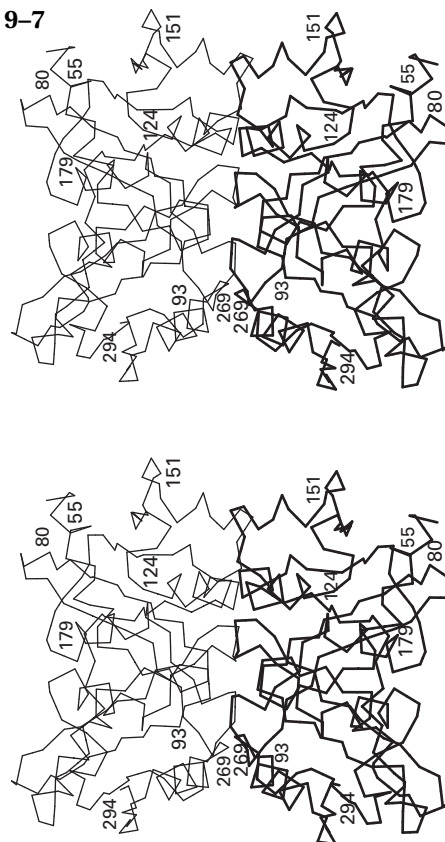
```

- (I) Align this sequence with the sequence of amino acids 49–177 of the protein from *E. coli*. The most conserved region in these two sequences is PCGX-CRQ, which contains amino acids from the active site of cytidine deaminase. Aligning these two regions from the two proteins will give you a start in the alignment. Put in gaps and try to get the best alignment.
- (J) For your alignment, what is the percentage of identity and how many gaps are there? In calculating the percentage of identity, assume that the length of the aligned region is the average of the lengths of the two aligned sequences.
- (K) Which are more closely related, the two sequences from *E. coli* or the sequence from *E. coli* and the sequence of cytidine deaminase from *B. subtilis*?
- (L) Does cytidine deaminase from *B. subtilis* contain a segment homologous to the segment from Alanine 154 to Leucine 176 in cytidine deaminase from *E. coli*? Why is this of interest in deciding how these two proteins evolved?
- (M) What would you guess is the quaternary structure of cytidine deaminase from *B. subtilis*?

Problem 9-8: Hemocyanin from *Sepia officinalis* is a decamer with dihedral symmetry of point group $522(D_5)$.¹⁹⁰ The ten subunits form a ring that can be divided into five segments. As required by this point group, each of these segments is identical to the other four and has a 2-fold rotational axis of symmetry running through its center, and each segment is formed from two of the subunits of the protein. The portion of each subunit forming one of these segments of the cylinder is a folded polypeptide with six internally repeating domains. Consequently, the ancestral segment of the cylinder must have been formed from 12 identical sub-

units. Arrange 12 identical subunits around local molecular rotational axes of symmetry, including the global 2-fold rotational axis of symmetry of the dihedral point group, to produce a segment composing one fifth of a cylinder.

Problem 9-7



Isometric Oligomeric Proteins

Aside from tetramers with symmetry of point group $222(D_2)$, protomers arranged with circular and dihedral symmetry form structures in which they are arranged cylindrically. There are three other point groups, however, in which asymmetric protomers can be arranged to form oligomers that are **isometric structures** (Table 9-3). These three point groups are the only remaining ones in which asymmetric objects can be arranged. They are the tetrahedral point group $23(T)$, the octahedral point group $432(O)$, and the icosahedral point group $532(I)$. In these three point groups, the centers of mass of the protomers are arrayed systematically over the surface of a sphere centered on the point of intersection of the rotational axes of symmetry around which the protomers are positioned.

As with proteins the subunits of which are arranged with either cyclic symmetry or dihedral symmetry, it is only the rotational axes of symmetry and their relative

Table 9–3: Isometric Arrangements of Identical Asymmetric Objects¹⁹¹

symmetry	Hermann-Mauguin	Shönflies	rotational axes of symmetry ^a	angles between axes (degrees) ^b	number of asymmetric units
tetrahedral	23	<i>T</i>	three 2-fold four 3-fold	90 70.53	12
octahedral	432	<i>O</i>	six 2-fold four 3-fold three 4-fold	63.43 70.53 90	24
icosahedral	532	<i>I</i>	15 2-fold 10 3-fold six 5-fold	36 41.81 63.43	60

^aA rotational axis or symmetry is a line passing through the center of the oligomeric structure. Because it is a line, it extends in both directions from the center. As a result, each axis of symmetry passes out of the oligomeric structure at two opposite points, and at each of these two points there is a symmetric arrangement of asymmetric objects on the surface of the structure. ^bThese are the angles between the immediately adjacent axes of the same fold.

orientations, not the ultimate shape of a particular oligomer, that define its point group. Nevertheless, there are three **regular polyhedra**¹⁹² that have, respectively, tetrahedral, octahedral, and icosahedral symmetries and that illustrate the types of rotational axes of symmetry in each of these three point groups and their orientations in space. These are Kepler's rhombic dodecahedron (Figure 9–21A), the triangular expansion of Kepler's rhombic dodecahedron (Figure 9–21B), and the triangular expansion of Kepler's rhombic triacontahedron (Figure 9–21C).^{*} The Platonic solids that are the namesakes for each of these three point groups are the tetrahedron, the octahedron, and the icosahedron, respectively. The other two Platonic solids, the cube and the dodecahedron, have octahedral and icosahedral symmetry, respectively. None of the five Platonic solids, however, is an adequate representative of its point group because each of them is formed from rotationally symmetric faces (equilateral triangles, squares, or regular pentagons) that are centered on axes of symmetry. In Kepler's three polygons all of the rotational axes of symmetry lie between the faces, just as in an oligomeric protein all of the rotational axes of symmetry must pass between its necessarily asymmetric protomers.

In the **tetrahedral point group 23(*T*)**, 12 identical protomers are arranged about four isometrically spaced 3-fold rotational axes of symmetry (at 70.53° to each other) and three orthogonal 2-fold rotational axes of symmetry that all intersect at a common origin (Figure 9–21A). When 12 protomers of protein are assembled in the tetrahedral point group 23(*T*), they do not fit into the neat geometrical boundaries of any polyhedron.

^{*} The unexpanded rhombic triacontahedron of Kepler, although constructed from 30 faces arrayed around rotational axes of symmetry, cannot accommodate 30 asymmetric objects. Consequently, it does not represent a point group.

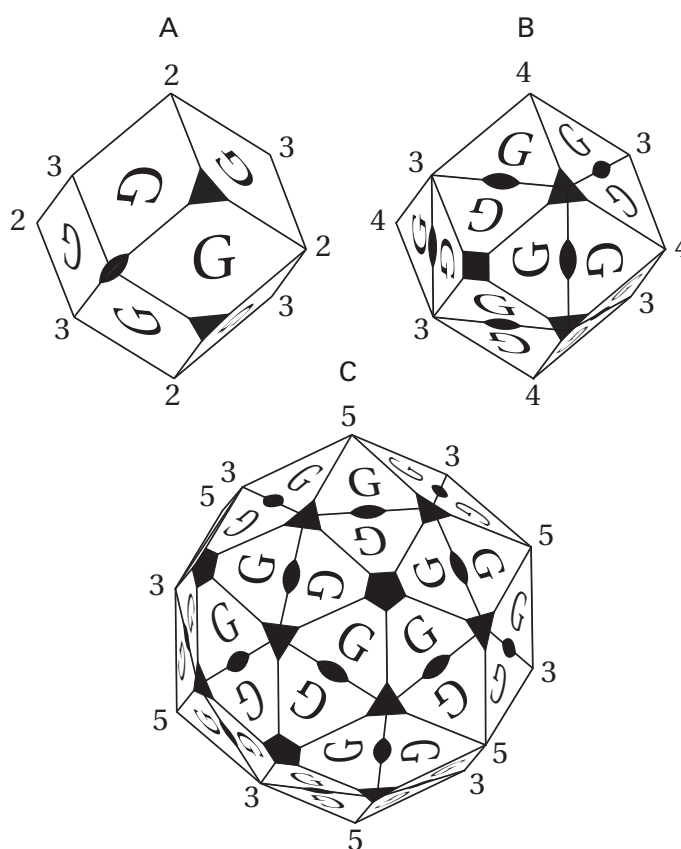


Figure 9–21: Regular polyhedra that have isometric symmetries:¹⁹² (A) the rhombic dodecahedron with tetrahedral symmetry of point group 23(*T*), (B) the tetracosahedron that is the triangular expansion of the rhombic dodecahedron and that represents octahedral symmetry of point group 432(*O*), and (C) the hexacontahedron that is the triangular expansion of a rhombic triacontahedron and that has icosahedral symmetry of point group 532(*I*). Kepler derived the rhombic triacontahedron from the intersection of the dodecahedron and octahedron by connecting the vertices at the 5-fold and 3-fold axes of symmetry with lines. In each of the figures, rotational axes on the circumference are labeled with the number of their fold.

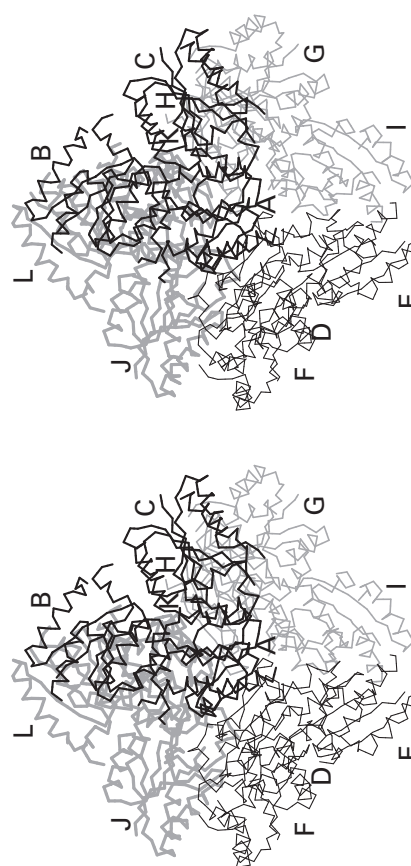
Nevertheless, they are arrayed around the proper 2-fold and 3-fold rotational axes of symmetry. 3-Dehydroquinase dehydratase from *Mycobacterium tuberculosis* (Figure 9-22)¹⁹³ is a protein in which the 12 folded polypeptides, each 146 aa long, are arranged with tetrahedral symmetry. In this particular protein, the most extensive interfaces are those producing the four trimers, each of which sits on its own 3-fold rotational axis of symmetry. As with the 2-fold rotational axes of symmetry in the dihedral point groups of odd fold, the two ends of each of the 3-fold axis of symmetry in point group $23(T)$ are different. Each of the four trimers in 3-dehydroquinase dehydratase is at one end of a 3-fold rotational axes of symmetry, while the other end of each 3-fold rotational axis of symmetry is surrounded by the other three trimers. The monomers in a trimer are superposed by one end of a 3-fold rotational axis of symmetry, while the other three trimers are superposed upon each other by the other end of that same axis.

Several of the proteins with tetrahedral symmetry are such **tetramers of trimers**. In dilute solutions of guanidinium chloride, 3-dehydroquinase dehydratase dissociates into its constituent trimers.¹⁹⁴ Phaseolin from *P. vulgaris* is a trimeric protein¹³⁹ that associates to form a dodecamer¹⁹⁵ with tetrahedral symmetry¹³⁹ below pH 4.5. The self-rotation function of the asymmetric unit of crystals of catabolic ornithine carbamoyltransferase from *Pseudomonas aeruginosa* has the maxima consistent with an oligomer with tetrahedral symmetry,¹⁹⁶ and when the protein is cross-linked the major covalent species are trimer, hexamer, nonamer, and dodecamer, a result consistent with the trimer being the fundamental unit.¹⁹⁷

In contrast to these three dodecamers, the dodecamer of protocatechuate 3,4-dioxygenase from *P. aeruginosa* is a structure in which six dimers, each formed by an extensive interface centered on a 2-fold rotational axis of symmetry, are arrayed around the 3-fold axes of symmetry of its tetrahedral point group by less extensive interfaces.⁸ The interfaces forming the constituent dimers of bromoperoxidase from *Corallina officinalis* are also more extensive (105 nm² interface⁻¹) than those forming the tetrahedral dodecamer (32 nm² interface⁻¹) from six of those dimers.¹⁹⁸ Consequently, there are both tetramers of trimers and **hexamers of dimers** with tetrahedral symmetry.

In the **octahedral point group 432(O)**, the 24 identical asymmetric objects of the tetracosamer are arranged about three orthogonal 4-fold rotational axes of symmetry, four isometrically spaced 3-fold rotational axes of symmetry at 70.53° to each other, and six isometrically spaced 2-fold rotational axes of symmetry at 63.43° to each other (Figure 9-21B). Again, the 24 protomers of a protein with octahedral symmetry only have to be arranged about the rotational axes of symmetry; they do not have to assume any particular geometric shape. As long as the interfaces among the protomers

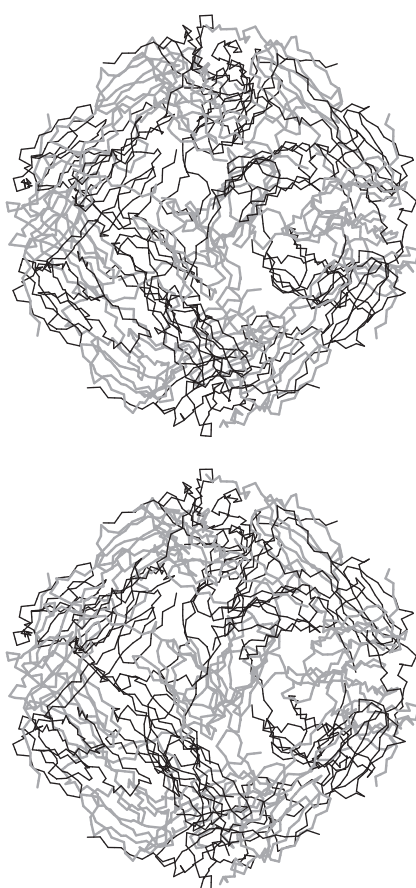
Figure 9-22: Tetrahedral symmetry of point group $23(T)$. An α -carbon diagram of 3-dehydroquinase dehydratase from *M. tuberculosis* is drawn from the crystallographic molecular model.¹⁹³ The protein crystallized in the space group F23, and an individual subunit of 146 aa is the crystallographic asymmetric unit. Consequently, all rotational axes of symmetry are exact. Each of the four trimers is drawn with line segments of different thickness or shading, and each subunit is labeled with a letter. The molecule is drawn in the orientation of the rhombic dodecahedron in Figure 9-21A. This drawing was produced with MolScript.⁴⁸⁵



position them in space around those rotational axes of symmetry, the oligomer will be a closed, octahedral, isometric structure. In point group $432(O)$, unlike in point group $23(T)$, identical sets of interfaces are found at the two ends of each of the rotational axes of symmetry.

Heat shock protein 16.5 from *Methanococcus jannaschii* (Figure 9-23)¹⁹⁹ is a tetracosameric protein in which subunits of 147 aa are arranged with octahedral symmetry to form a hollow spherical shell that has sizeable holes at the 3-fold rotational axes of symmetry and at the 4-fold rotational axes of symmetry. The most

Figure 9-23: Octahedral symmetry of point group 432(*O*). An α -carbon diagram of the heat shock protein 16.5 from *M. jannaschii* is drawn from the crystallographic molecular model.¹⁹⁹ The protein crystallized in the space group *H*3 with one of the molecular 3-fold rotational axes of symmetry coinciding with a crystallographic 3-fold rotational axis of symmetry and eight subunits in the crystallographic asymmetric unit. The two subunits in each dimer are given different shading. Each dimer is centered on one of the 2-fold rotational axes of symmetry. The molecule is drawn in the orientation of the regular tetracosahedron in Figure 9-21B.



extensive interfaces are those between monomers paired as dimers around the 2-fold rotational axes of symmetry. The second most extensive set of interfaces are the groups of four dimers arrayed around each of the 4-fold rotational axes of symmetry, so the structure is a trimer of tetramers of dimers. The four symmetrically arrayed interfaces among the four monomers around a 4-fold rotational axis of symmetry, one from each dimer, incline those four dimers relative to each other to form a jagged cup with the proper curvature so that three of these cups

fit together to form the final spherical structure. Coincidentally, the protein crystallizes in the space group *H*3 with one of these octameric cups as the asymmetric unit.

In the **icosahedral point group 532(*I*)**, 60 identical asymmetric objects are arranged about 31 rotational axes of symmetry (Figure 9-21C; Table 9-3) to produce the oligomer. Each of the 31 rotational axes of symmetry intersects all of the others at the center of the structure. In every icosahedral arrangement, the relative angular dispositions of these axes is always the same. Although there are 31 rotational axes of symmetry in this point group, it can be generated from one protomer by four successive rotations around specified axes in a given sequence. 6,7-Dimethyl-8-ribityllumazine synthase from *B. subtilis*²⁰⁰ and the protein coat of satellite panicum mosaic virus (Figure 9-24)²⁰¹ are oligomeric proteins in each of which 60 identical subunits are arranged with icosahedral symmetry of point group 532(*I*) to form a spherical shell.

The **protein coat of a virus** such as satellite panicum mosaic virus is a thick, continuous, spherical layer of protein that serves the purpose of enclosing and protecting the viral nucleic acid. This nucleic acid encodes the genetic information necessary for the virus to control the host parasitically and divert the purpose of the host from its own growth and replication to the growth and replication of the virus. These requirements can be satisfied only by a fairly large molecule of nucleic acid, and it all must fit within the protein coat so that it can be protected from the environment. A viral protein coat is made from 60 identical, or almost identical,^{202,203} protomers. In a spherical virus, these protomers are arranged about the icosahedral rotational axes of symmetry to produce spherical shells that can enclose the nucleic acid. In the case of satellite panicum virus, the protomer is a single folded polypeptide.

Again, it is not any regular polygon that defines a protein built from 60 protomers arranged with icosahedral symmetry but the rotational axes of symmetry. Regardless of how intertwined or encroaching the protomers in such a protein become, the number and relative angular dispositions of the rotational axes of symmetry that dictate the positions of those protomers are permanent features of the structure. If 60 identical folded polypeptides are arranged around these rotational axes of symmetry and their shapes are so constructed as to mesh symmetrically at each of the boundaries among themselves, they will necessarily form a tightly sealed icosahedral shell.

The **interfaces among the protomers** are the fundamental determinants of the multimeric structure. In 6,7-dimethyl-8-ribityllumazine synthase, the pentamer appears to have been the unit assembled by evolution into the oligomer of 60 subunits.²⁰⁴ At each of the five equivalent outer edges of this pentamer an interface evolved, connecting the pentamer to another identical

pentamer in the usual way that two identical proteins are associated, which is around a 2-fold rotational axis of symmetry. This particular 2-fold rotational axis of symmetry, however, happened to incline the two pentamers with respect to each other so that their respective 5-fold rotational axes of symmetry both intersected the 2-fold axis of symmetry and formed the angle required to exist between the 5-fold rotational axes of symmetry in an icosahedron, which is 63° . If two interfaces, the one defining the pentamer and the other at the 2-fold rotational axis defining an angle of 63° , are built into faces on a protomer, 60 such protomers will automatically assemble into an icosahedral shell.

In the protein coat of satellite panicum mosaic virus, the interfaces holding the trimers together around the 3-fold axes of symmetry ($15.7 \text{ nm}^2 \text{ interface}^{-1}$) are more extensive than those holding together the dimers around the 2-fold axes ($12.8 \text{ nm}^2 \text{ interface}^{-1}$) or the pentamers around the 5-fold axes ($10.5 \text{ nm}^2 \text{ interface}^{-1}$).²⁰¹ If the trimer was the fundamental unit from which the protein coat arose, the evolution of two complementary faces on the trimer that formed a dimer of trimers inclining the two 3-fold rotational axes of symmetry at 42° would also automatically create the entire icosahedrally symmetric oligomer of 60 subunits.

Vertebrate ferritin is a tetracosamer the identical subunits of which are arranged with octahedral symmetry (Figure 9-25A),^{205,206} but ferritin from *Listeria innocua* is a dodecamer the identical subunits of which are arranged with tetrahedral symmetry.²⁰⁷ These are examples of **two different quaternary structures for the same species of protein**. The subunits of vertebrate ferritin and the ferritin from *L. innocua* are both antiparallel coiled coils of four α helices (Figure 9-25B) that are superposable on each other and consequently homologous to each other. In both proteins, these coiled coils of α helices associate side by side in opposite directions to form dimers around 2-fold rotational axes of symmetry, and the interfaces forming the dimers from the two proteins are homologous to each other. The dimers, in turn, in each protein combine in triplets around 3-fold rotational axes of symmetry in which the three identical interfaces around each axis are formed by the two respective ends of the monomers in one dimer butting up against the side of one of the monomers in a neighboring dimer (Figure 9-25C). These interfaces around the 3-fold axes in the two proteins are homologous to each other. Vertebrate ferritin is formed from four of these trimers of dimers; the ferritin from *L. innocua*, from two. As with any oligomer with tetrahedral symmetry (Figure 9-22), there are two different types of interfaces at the two ends of each 3-fold axis in the tetrahedral dodecamer from *L. innocua*. One of these types is the set homologous to the set of interfaces around the 3-fold axes in vertebrate ferritin (Figure 9-25C).

If one of the dimers around a 4-fold rotational axis of symmetry in vertebrate ferritin were removed (for

example, the one containing monomer V) and the creases between the remaining three dimers were made more acute, the space previously occupied by the missing dimer would close up and a new interface could form (one between dimer I/II and the side of monomer IV) now identical to the other two around the once 4-fold but now 3-fold axis of symmetry. If this transformation were performed at each end of each 4-fold axis in vertebrate ferritin, six dimers would be removed, the structure would be converted from a tetracosamer with octahedral

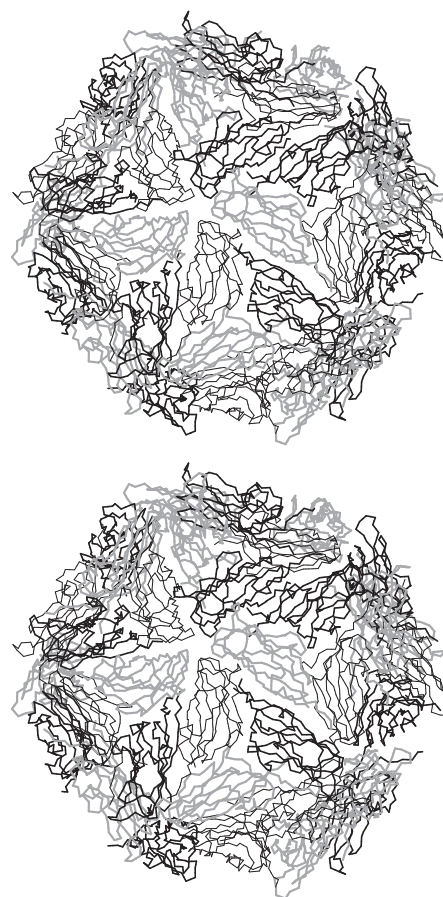


Figure 9-24: Icosahedral symmetry of point group 532(I). An α -carbon diagram of one hemisphere of the protein coat of satellite panicum mosaic virus is drawn from the crystallographic molecular model of the protein.²⁰¹ The protein crystallized in the space group $P4_332$ with a pentamer of subunits, each of 157 aa, in the crystallographic asymmetric unit, the smallest asymmetric unit available to icosahedral symmetry. Individual subunits have been drawn with line segments with one of three different thicknesses or shadings. Only 34 of the 60 subunits are drawn to make the structure easier to visualize. These 34 subunits form the front of the sphere that is the protein coat. The portion of the molecule drawn is in the orientation of the hexacontahedron of Figure 9-21C. This drawing was produced with MolScript.⁴⁸⁵

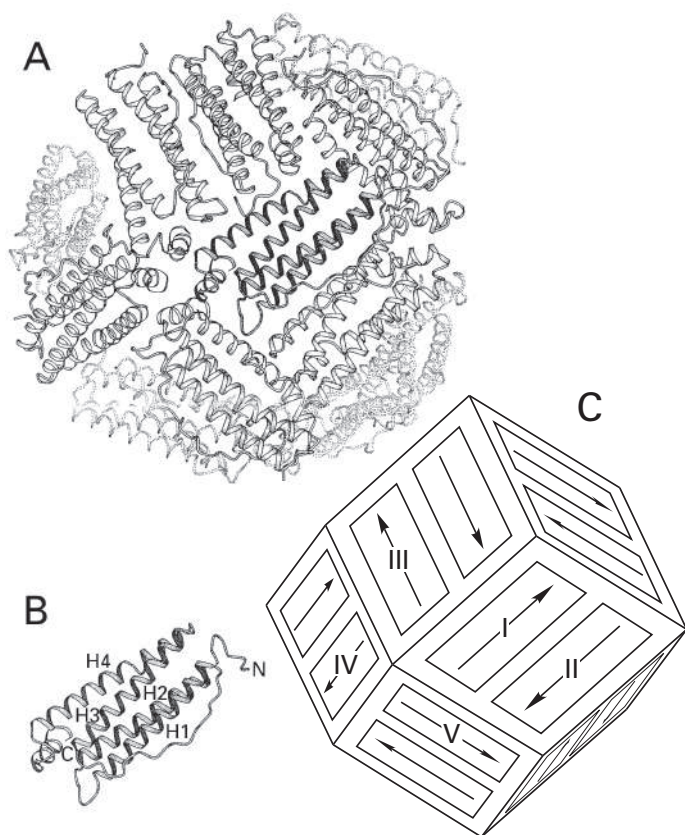


Figure 9-25: Representations of the octahedral symmetry of mammalian ferritin.^{205,206} (A) Ribbon diagram of only the front half of the crystallographic molecular model of the entire spherical molecule showing 12 of the 24 subunits. Reprinted with permission from ref 206. Copyright 1997 Elsevier B.V. (B) Ribbon diagram of just the central of the 12 subunits portrayed in panel A. Reprinted with permission from ref 206. Copyright 1997 Elsevier B.V. (C) Representation of the arrangement of the dimers of the subunits of ferritin on the faces of a rhombic dodecahedron (Figure 9-21A). Arrows are drawn to indicate the antiparallel orientations of the two subunits around the 2-fold rotational axes of symmetry in the center of each face. The 2-fold, 3-fold, and 4-fold rotational axes of symmetry are located as they are in Figure 9-21B.

symmetry into a dodecamer with tetrahedral symmetry, and this new dodecamer would be structurally homologous to the ferritin from *L. innocua*. The two different ends of each of the new 3-fold rotational axes of symmetry in the new tetrahedral dodecamer would be the end not affected by the transformation and the end that had been at a 4-fold rotational axis of symmetry before the transformation. Consequently, in these two different quaternary structures of ferritin, the octahedral and the tetrahedral, the homologous complementary faces on two different, but homologous, dimers must accommodate in the one case a 4-fold rotational axis of symmetry and in the other case a 3-fold rotational axis of symmetry.

Among the small heat shock proteins, there are also two different quaternary structures for the same species of protein in different species of organisms. The small

heat shock protein from *M. jannaschii* (Figure 9-23) is a tetracosamer with octahedral symmetry of point group $432(O)$, while the same protein from *Triticum aestivum* is a dodecamer with dihedral symmetry of point group $322(D_3)$.²⁰⁸ In both structures, a dimer is the fundamental unit, and the respective dimers have homologous tertiary and quaternary structures. In the protein from *M. jannaschii*, the molecular 2-fold rotational axis of symmetry of each dimer is coincident with one of the 2-fold rotational axes of the symmetry of point group $432(O)$, while in the protein from *T. aestivum*, the entire dimer is the asymmetric unit for the symmetry of point group $322(D_3)$ and its 2-fold rotational axis is local and pseudosymmetric. Nevertheless, the way in which the three dimers are arranged around the single, central 3-fold rotational axis of symmetry in the latter is the same as that in which three dimers are arranged about one of the four 3-fold rotational axes of symmetry in the former (Figure 9-23), and the fundamental unit in the two different structures is this homologous hexamer. Two of these hexamers lie back to back in the dodecamer, and four of these hexamers are each centered on the respective 3-fold rotational axes of symmetry in the tetracosamer.²⁰⁸ Presumably, because this hexamer is the fundamental unit, the dodecamer is one of symmetry of point group $322(D_3)$ rather than that of point group $622(D_6)$ as are most other dodecamers.

Dihydrolipoyllysine residue acetyltransferase, dihydrolipoyllysine residue succinyltransferase, and dihydrolipoyllysine residue (2-methylpropanoyl)transferase are closely related proteins, the sequences of which can be readily aligned with each other with greater than 30% identity.²⁰⁹ All of the dihydrolipoyllysine residue succinyltransferases and dihydrolipoyllysine residue (2-methylpropanoyl)transferases and dihydrolipoyllysine residue acetyltransferases from Gram-negative bacteria are tetracosamers of 24 identical subunits with octahedral symmetry,²¹⁰⁻²¹² while the dihydrolipoyllysine-residue acetyltransferases from Gram-positive bacteria and eukaryotes are hexacontamers of 60 identical subunits with icosahedral symmetry.^{210,213}

In both the octahedral oligomers²¹² and the icosahedral oligomers,²¹⁰ the homologous interfaces forming the trimers centered on the respective 3-fold rotational axes of symmetry (Figure 9-21A,B) are the most extensive in the structures. In the octahedral oligomers, the eight trimers at the eight 3-fold vertices are connected by 12 interfaces centered on the 2-fold rotational axes of symmetry, so that the structure is a cube with wide openings in each face at the 4-fold rotational axes of symmetry.³¹ In the icosahedral oligomers, the trimers are again joined at the 2-fold rotational axes of symmetry but by 30 interfaces, and the structure is a dodecahedron with wide openings in each face at the 5-fold rotational axes of symmetry. In the octahedral oligomers, four trimers associate with each other around 4-fold rotational axes of symmetry, while in the icosahedral oligomers, five

trimers associate with each other around 5-fold rotational axes of symmetry. The respective interfaces at the 2-fold rotational axes of symmetry between these homologous trimers that are arranged around these two different rotational axes of symmetry adapt flexibly to the dispositions of the trimers that are required by those axes.²¹⁰ In these two different quaternary structures of these homologous acetyltransferases, the octahedral and the icosahedral, the homologous complementary faces on two different, but homologous, trimers must accommodate in the one case a 4-fold rotational axis of symmetry and in the other case a 5-fold rotational axis of symmetry.

Ferritin, small heat shock protein, and dihydrolipoyllysine-residue acetyltransferase, because they each are examples of the same protein having different quaternary structures, reinforce the conclusion that quaternary structure contains no information relevant to the evolution of proteins. The examples of ferritin and dihydrolipoyllysine-residue acetyltransferase also illustrate the ability of two different subunits, formed respectively from closely related polypeptides, to assume **similar arrangements around rotational axes of symmetry of different fold**. It is this ability of interfaces to adjust flexibly to different rotational axes of symmetry that has been exploited by evolution to increase the size of viral protein coats.

Although there are a few other viruses like satellite panicum mosaic virus that have protein coats assembled from only 60 subunits arranged with 532(*I*) symmetry,^{203,214,215} the problem with such protein coats is that they are too small. In order to enclose enough DNA to accomplish a successful subversion of the host, the protein coats usually must be larger. In two instances, this problem has been solved by using an elongated homodimer as a protomer and having these homodimers arrayed as spokes around the 5-fold rotational axes of symmetry. One monomer of the dimer forms the near end of the spoke adjacent to the 5-fold axis; and the other monomer, the far end of the spoke. A local 2-fold rotational axis of pseudosymmetry relating the two elongated monomers is located in the center of the spoke. The interdigitation of these long spokes forms the protein coat. Although each is formed from copies of the same polypeptide, a monomer at the 5-fold hub is required to assume a significantly different shape from a monomer at the periphery of the spoke in order for the interdigitation to succeed and the global 3-fold and 2-fold rotational axes of icosahedral symmetry to be satisfied.^{216,217} Most viral protein coats, however, are built with a different strategy that takes advantage of quasi-equivalence and pseudosymmetry to provide a general solution to the problem of expanding the size of an icosahedral shell.^{218,219}

Quasi-equivalence²¹⁸ is the manifestation of the ability of either two or more copies of the same subunit or two or more homologous subunits to adapt flexibly in

different situations to the requirements of two rotational axes of symmetry of different fold. The respective homologous subunits in the two different quaternary structures of ferritin or in the two different quaternary structures of dihydrolipoyllysine-residue acetyltransferase are quasi-equivalent to each other, but because in each case they have different amino acid sequences, it is the differences in amino acid sequence that might explain their abilities to assume the different dispositions. Furthermore, the two rotational axes of symmetry of different fold to which they adapt are in different oligomers. In many viral protein coats, however, the several quasi-equivalent subunits are formed from folded polypeptides of the same sequence, and the quasi-equivalent subunits are found together in the same icosahedral shell.

To understand the relationships of such quasi-equivalent subunits to each other in such an oligomer, the local rotational axes within a protomer must be distinguished from the global rotational axes of icosahedral symmetry governing the entire structure. A **global rotational axis of symmetry** is an axis of symmetry around which a rotation of $360^\circ/n$ causes the entire oligomer to superpose upon itself. A global rotational axis is thus distinguished from a local rotational axis that operates only on structural units in its immediate vicinity.

The triangle from which the expanded rhombic triacontahedron (Figure 9-21C) is constructed, although formally an asymmetric object because two of its vertices lie at global 3-fold rotational axes of symmetry and one vertex lies at a global 5-fold rotational axes of symmetry and only one of its three edges lies on a global 2-fold rotational axis of symmetry, is nevertheless an equilateral triangle and locally symmetric. Were the equivalent mass of three identical folded polypeptides, related to each other by this **local 3-fold rotational axis of pseudosymmetry**, to fill this triangle, it would have three times more area than if the equivalent mass of only one folded polypeptide filled it, and the shell could then contain 3.5 times more nucleic acid. This solution requires (Figure 9-21C) that this folded polypeptide be capable of quasi-equivalence because those subunits forming the interfaces around one vertex of the triangle would have to be arrayed around a global 5-fold rotational axis of symmetry (72° for each step; complementary faces at 108°), while those subunits forming the interfaces around the other two vertices would have to be arrayed around local 6-fold rotational axes of pseudosymmetry (60° for each step; complementary faces at 120°) that each coincide with a global 3-fold rotational axis of symmetry. The requirements of this quasi-equivalence would force each of the three subunits arrayed around the rotational axis in the center of each triangle to assume a significantly different conformation, so the local 3-fold rotational axis around which they are arrayed is one of pseudosymmetry.

Quasi-equivalent subunits arrayed around rotational axes of pseudosymmetry cannot each be individ-

ual protomers, but the set of all of the quasi-equivalent subunits arrayed around a rotational axis of pseudosymmetry or several rotational axes of pseudosymmetry can be a protomer of the overall quaternary structure. Consequently, it is the three quasi-equivalent subunits arrayed around the local 3-fold rotational axis of pseudosymmetry that would form the protomer of the icosahedral array. The global rotational axes of this icosahedral array would remain true global rotational axes of symmetry for the entire structure because the asymmetry would be confined entirely within each of the protomers of the point group.

Tomato bushy stunt virus has a protein coat with just such an arrangement of subunits (Figure 9–26).²²⁰ Each of its 60 identical **pseudosymmetric, trimeric protomers** is formed from three folded polypeptides, subunits A, B, and C, each of the same sequence 386 aa in length and the tertiary structures of which, when they are in the viral protein coat, are homologous and superposable. The differences in their respective conformations permit each of them to play its required quasi-equivalent

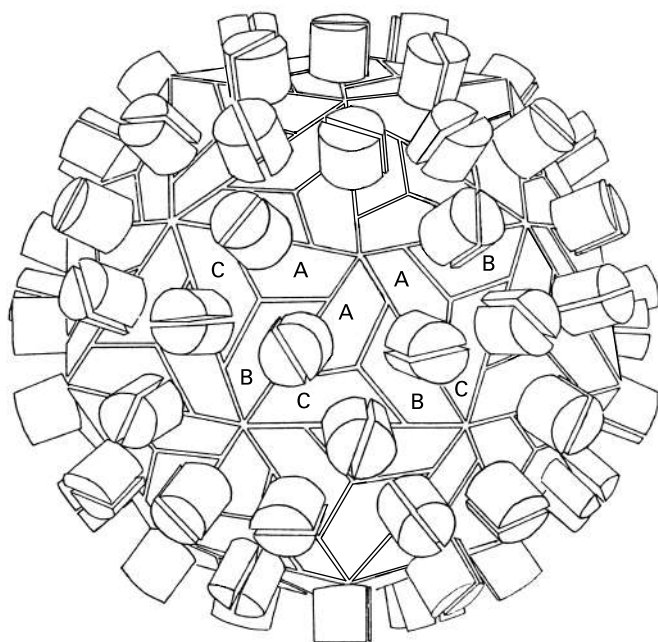


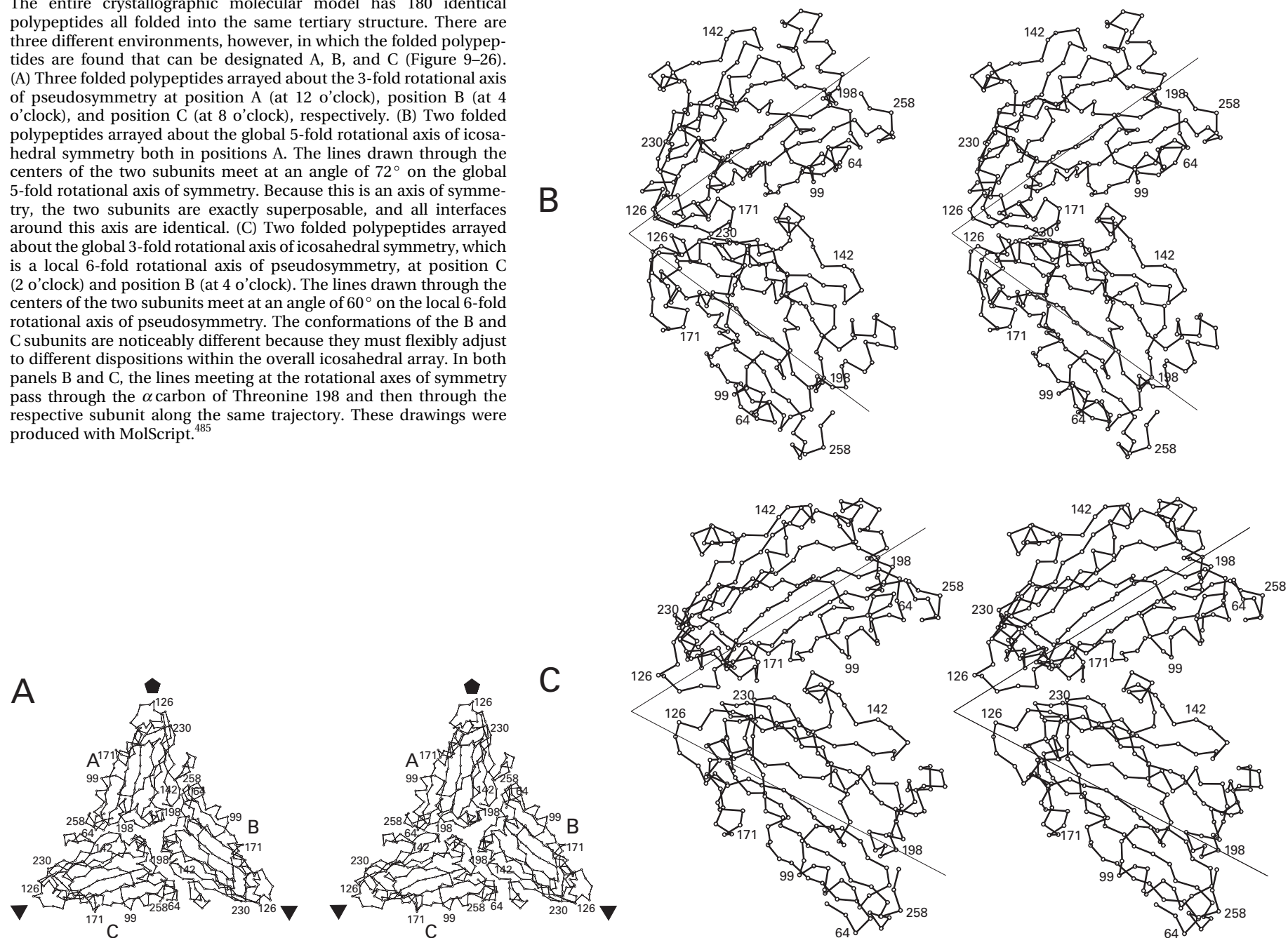
Figure 9–26: Arrangement of the 180 subunits in the protein coat of tomato bushy stunt virus.²²⁰ Each tile is a single folded polypeptide, and all of the polypeptides are identical to each other in amino acid sequence. Three folded polypeptides, designated A, B, and C, are arrayed around a local 3-fold axis of pseudosymmetry to produce the trimeric protomer of the icosahedral array. The vertex occupied by the A subunit lies at a global icosahedral 5-fold rotational axis of symmetry, and the axes occupied by the B and C subunits lie at global icosahedral 3-fold rotational axes of symmetry that are also local 6-fold rotational axes of pseudosymmetry. Global 2-fold rotational axes of symmetry relate C subunits, and local 2-fold rotational axes of pseudosymmetry relate A and B subunits. Each subunit has a protrusion that runs up its associated 2-fold rotational axis. The diagram was adapted from the crystallographic molecular model of this viral protein coat. Reprinted with permission from ref 220. Copyright 1983 Academic Press.

role. For example, only subunits C are adjacent to the global 2-fold rotational axes of symmetry. Subunits B and C must alternate around the global 3-fold rotational axes of symmetry to produce local 6-fold rotational axes of pseudosymmetry, while subunits A are distributed around the global 5-fold rotational axis of symmetry. These quasi-equivalent situations produce alterations in the structures of these folded polypeptides that are most obvious at the interfaces among the homotrimers; it is here that the strain of requiring the same protein to adapt to the different rotational axes of symmetry is the strongest.

The **packing at the quasi-equivalent interfaces** has been described in detail for the protein coat of southern bean mosaic virus,²²¹ which is closely related to tomato bushy stunt virus. The three identical but quasi-equivalently folded polypeptides, each 260 aa in length, are arranged around a local 3-fold rotational axis of pseudosymmetry (Figure 9–27A)²²¹ to create the homotrimeric protomer in which the three subunits adapt to their respective quasi-equivalent environments. The A subunits are arranged around the global icosahedral 5-fold rotational axes of symmetry (Figures 9–21C, 9–26, and 9–27B). Each of the B and C subunits uses the same vertex to form the local 6-fold rotational axis of pseudosymmetry (Figures 9–26 and 9–27C) as that used by the A subunit to conform to the global 5-fold rotational axis of symmetry. Careful inspection of Figure 9–27B,C shows that the two unique defining interfaces are similar but significantly adjusted to accommodate the differences in the angular requirements around these two axes. These adjustments, in turn, create conformational changes throughout each of the individual subunits, causing their overall structures to differ. These differences are most obvious in the angular orientations of both the pleats within the β sheets and the β sheets themselves when the three different conformations are compared.

The protein coats of tomato bushy stunt virus, southern bean mosaic virus, turnip yellow mosaic virus,²²² black beetle nodavirus,²²³ and primate calicivirus,²²⁴ among others, are all constructed from 180 identical subunits distributed among 60 identical homotrimers. Because, however, the three subunits in a trimer (A, B, and C in Figure 9–26) are not in identical environments, they need not be identical in amino acid sequence. In some icosahedral protein coats with trimeric protomers, **gene triplication** of the nucleic acid encoding the amino acid sequence of the protein forming the coat has occurred to produce three genes. The protein coats of the comoviruses, of which those of cowpea mosaic virus and beanpod mottle virus are examples, are an interesting intermediate case in this process.²²⁵ In the protein coat of these viruses, two of the subunits in the heterotrimeric protomer are internally repeating domains on the same polypeptide. This suggests that the general sequence of events has been a gene duplication

Figure 9-27: α -Carbon diagrams of the folded polypeptides composing the protein coat of southern bean mosaic virus drawn from the crystallographic molecular model of the entire viral protein coat.²²¹ The entire crystallographic molecular model has 180 identical polypeptides all folded into the same tertiary structure. There are three different environments, however, in which the folded polypeptides are found that can be designated A, B, and C (Figure 9-26). (A) Three folded polypeptides arrayed about the 3-fold rotational axis of pseudosymmetry at position A (at 12 o'clock), position B (at 4 o'clock), and position C (at 8 o'clock), respectively. (B) Two folded polypeptides arrayed about the global 5-fold rotational axis of icosahedral symmetry both in positions A. The lines drawn through the centers of the two subunits meet at an angle of 72° on the global 5-fold rotational axis of symmetry. Because this is an axis of symmetry, the two subunits are exactly superposable, and all interfaces around this axis are identical. (C) Two folded polypeptides arrayed about the global 3-fold rotational axis of icosahedral symmetry, which is a local 6-fold rotational axis of pseudosymmetry, at position C (2 o'clock) and position B (at 4 o'clock). The lines drawn through the centers of the two subunits meet at an angle of 60° on the local 6-fold rotational axis of pseudosymmetry. The conformations of the B and C subunits are noticeably different because they must flexibly adjust to different dispositions within the overall icosahedral array. In both panels B and C, the lines meeting at the rotational axes of symmetry pass through the α carbon of Threonine 198 and then through the respective subunit along the same trajectory. These drawings were produced with MolScript.⁴⁸⁵



that gave rise to two genes producing two separate polypeptides followed by a gene duplication of one of these genes producing a single polypeptide with two internally repeating domains followed by a division of the latter duplicated gene so that it then produced two smaller polypeptides, each containing the complete fold of its ancestral polypeptide. Following the triplication, each of the three genes evolved independently to produce three polypeptides, each of different sequence and each presumably incorporating changes rendering it more successful at occupying its respective quasi-equivalent position in the viral protein coat.

There are crystallographic molecular models for icosahedral protein coats of four viruses that have accomplished the complete evolutionary transition: rhinovirus,²²⁶ poliovirus,²²⁷ Mengo virus,²²⁸ and foot-and-mouth disease virus.²²⁹ The protein coats of these viruses are spherical shells (Figure 9–28A,B),²²⁷ the surfaces of which are paved with **heterotrimeric protomers** in icosahedral array (Figure 9–28C).²²⁸ That the three different folded polypeptides forming these four viral protein coats have arisen from a gene triplication follows from the fact that, in each case, the three different polypeptides folded in their native conformations are superpos-

able.^{226–229} If this is a definitive correspondence, it demonstrates that the ancestors of each of these protein coats were constructed from 180 folded polypeptides of identical sequence.

It has been shown that the single folded polypeptides composing the protomers of the protein coats of satellite panicum mosaic virus and satellite tobacco necrosis virus, another small virus the protein coat of which has only 60 subunits,^{215,230} are superposable on any one of the folded polypeptides forming the trimeric protomer of the protein coat of either tomato bushy stunt virus or southern bean mosaic virus.^{201,231} It has also been pointed out that, even though the former have 60 subunits and the latter have 180 subunits, the packing of the subunits of the protein coat of satellite panicum mosaic virus and satellite tobacco necrosis virus is similar to the packing of the subunits of the protein coats of both southern bean mosaic virus and tomato bushy stunt virus. When the global 3-fold rotational axis of symmetry in the protein coat of satellite tobacco necrosis virus is aligned with the local 3-fold rotational axis of pseudosymmetry within one of the trimeric protomers of the protein coat from one of the other viruses (Figure 9–29),²³⁰ the three global 5-fold rotational axes of symmetry in the protein coat of satellite tobacco necrosis virus coincide with one of the global 5-fold rotational axes of symmetry and two of the local 6-fold rotational axes of pseudosymmetry in the protein coat of the other virus.

It has also been observed²³² that when the first 61 amino acids of the protein coat of southern bean mosaic virus are removed, the remainder of the folded polypeptide can assemble to produce a hollow icosahedral shell

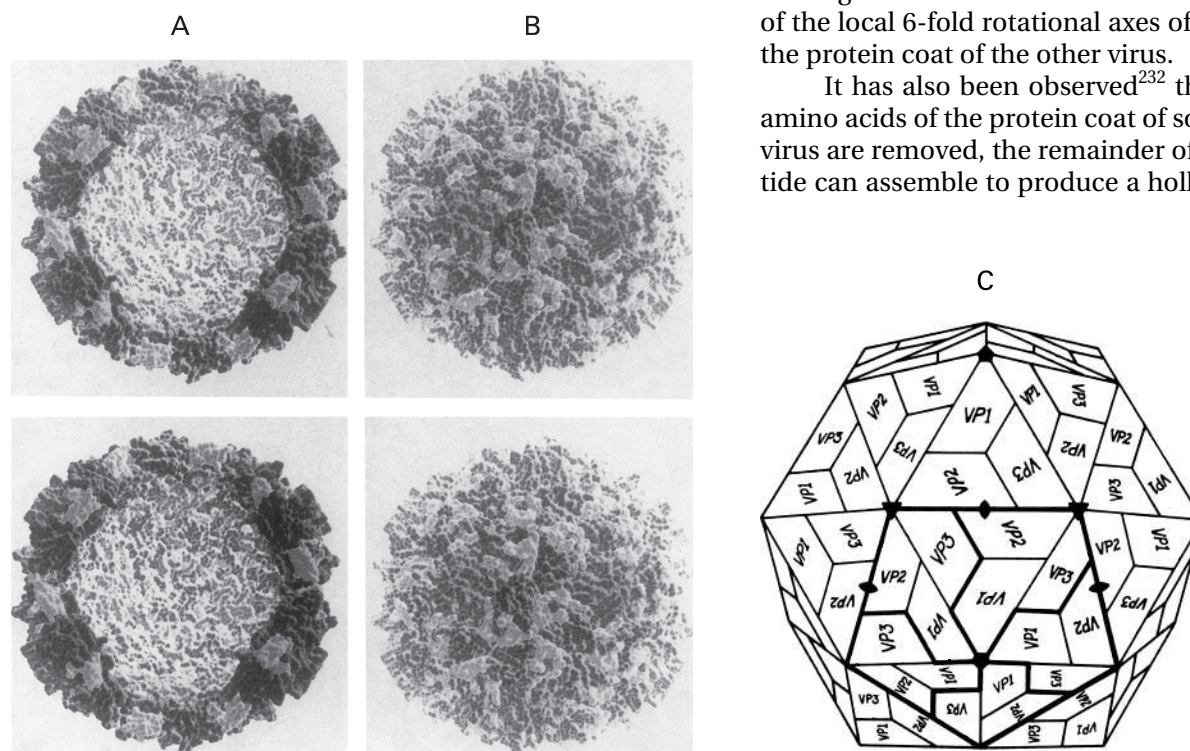


Figure 9–28: Space-filling representations of the folded polypeptides assembled into the icosahedral protein coat of poliovirus as drawn from the crystallographic molecular model of this oligomeric protein.²²⁷ (A) View into the central cavity of the viral protein coat into which the viral RNA is packed. (B) View of the surface of the viral protein coat in which the atoms contributed by each of the three different types of folded polypeptides, VP1, VP2, and VP3, have been given different shades of gray. Panels A and B reprinted with permission from ref 227. Copyright 1985 American Association for the Advancement of Science. (C) Diagrammatic representation of the surface of an icosahedral viral protein coat²²⁸ in the same orientation as panel B to illustrate the distribution of the various folded polypeptides around the rotational axes of icosahedral symmetry and local pseudosymmetry. Panel C reprinted with permission from ref 228. Copyright 1987 American Association for the Advancement of Science.

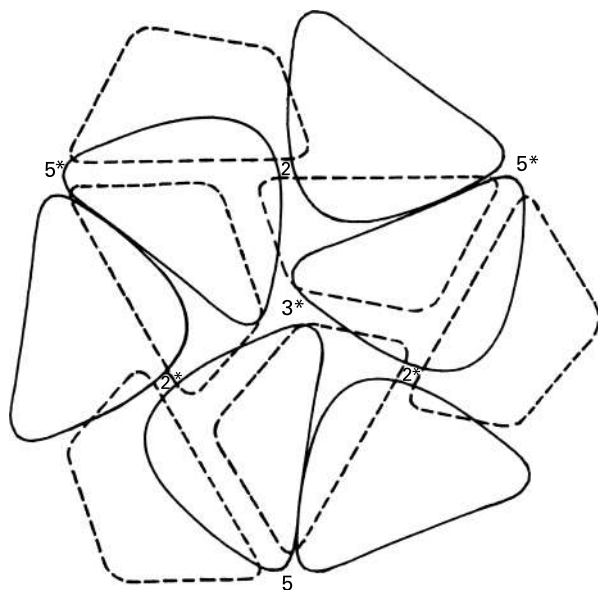


Figure 9-29: Comparison of the packing of the respective folded polypeptides in the protein coat of satellite tobacco necrosis virus (unbroken lines) and the protein coat of tomato bushy stunt virus or southern bean mosaic virus (broken lines).²³⁰ The former viral protein coat has 60 folded polypeptides in exact icosahedral symmetry; the latter have 180 folded polypeptides in $T=3$ icosahedral symmetry. The exact global 3-fold rotational axis of icosahedral symmetry for satellite tobacco necrosis virus is aligned with the local 3-fold rotational axis of pseudosymmetry (both designated as 3*) in the center of each of the protomers of the other two. This alignment causes the global 5-fold rotational axes of icosahedral symmetry at the bottom of the diagram (designated as 5) to coincide and concurrently causes the two other global 5-fold rotational axes of symmetry at the other two vertices of the three protomers from satellite tobacco necrosis virus to coincide with the two local 6-fold rotational axes of pseudosymmetry (all of these axes designated as 5*) at the other two vertices of the protomer from tomato bushy stunt virus or southern bean mosaic virus. Reprinted with permission from ref 230. Copyright 1982 Academic Press.

containing only 60 subunits instead of the usual 180. In this new oligomeric protein, the original arrangements around the global 5-fold rotational axes of symmetry have been retained and the local 3-fold rotational axes of pseudosymmetry in the centers of the protomers of the original structure have become the global 3-fold rotational axes of symmetry in the new structure (as depicted in Figure 9-29). This latter result demonstrates that these two icosahedral structures, the pseudosymmetric with 180 quasi-equivalent polypeptides and the symmetric with 60 polypeptides, are readily interconverted.

All of these results taken together indicate that the protein coats of these four viruses, tomato bushy stunt, southern bean mosaic, satellite panicum mosaic, and satellite tobacco necrosis, share a common ancestor.²³¹ The fact that satellite tobacco necrosis virus and satellite panicum mosaic virus are parasites on other viruses and the fact that viral protein coats of their size cannot carry enough nucleic acid suggests that their ancestors originally had a larger protein coat built from 60 homotrimeric protomers.

The symmetric and pseudosymmetric icosahedral protein coats of satellite tobacco necrosis, satellite panicum mosaic, tomato bushy stunt, southern bean mosaic, cowpea mosaic, beanpod mottle, Mengo, and foot-and-mouth disease viruses, black beetle nodavirus, poliovirus, and rhinovirus are all those of viruses carrying single-stranded ribonucleic acids that have positive copies of the viral messenger ribonucleic acids. These viruses infect eukaryotic cells, both animals and plants. From the descriptions that have just been given of the crystallographic molecular models of the protein coats of these viruses, it follows that the ancestors of all of these eukaryotic positive-strand RNA viruses had icosahedral protein coats constructed from 60 homotrimeric protomers, each of which was constructed from three folded polypeptides of identical sequence arranged about a local 3-fold rotational axis of pseudosymmetry. The most remarkable discovery, however, is that all of the folded polypeptides comprising the protein coats of all of these eukaryotic positive-strand RNA viruses, whether their hosts are plants or animals, are superposable.^{201,223,225-229,231} Furthermore, the single polypeptides forming the protein coats of 60 identical subunits enclosing the single-stranded DNA of the bacteriophage ϕ X174²¹⁴ and the single-stranded DNA of canine parvovirus²⁰² also have folds superposable on those of these viral protein coats that enclose single-stranded RNA. These similarities among all of these various proteins suggest that, unless convergent evolution has occurred, all of the viral protein coats they respectively compose share one **common ancestor**.*

This remarkable possibility, if it is true, would not be hard to explain. Three unique faces, each creating an independent set of repeating interfaces among the 180 identical folded polypeptides, would have had to evolve on the surface of the same monomer. The three unique interfaces produced by these three unique faces are the interface responsible for the local 3-fold rotational axis of pseudosymmetry within the protomer (Figure 9-27A), the interface responsible for the global 5-fold rotational axis of symmetry at one vertex of each subunit, and the interface responsible for the global 2-fold rotational axis of symmetry that orients the two pentagons it connects. Proteins containing molecular 5-fold rotational axes of symmetry are rare products of evolution (Table 9-1) and the 2-fold rotational axis of symmetry, though common, would be required to be located at a particular disposition relative to the 5-fold axis of symmetry. These constraints are probably not so rigid as they seem. If the

* Each of these polypeptides is folded into a 10-stranded jelly roll, antiparallel β barrel (Figure 7-19E). The same fold occurs in the subunits of the protein coat of double-stranded DNA viruses such as adenoviruses and iridoviruses. In these latter instances, however, there are substantial differences in the quaternary structures of the viral protein coats.

angles are close enough, the significant stability of such an edifice, each of whose associations strengthens all of the others, could force the interfaces to rearrange sufficiently to accommodate the icosahedral arrangement. This **cooperativity in the construction of the shell**, which should resemble the cooperativity among the supports of a building, also permits the interfaces to be weaker than interfaces that must stand alone. Nevertheless, that such a monomer, with such faces, has arisen rarely during evolution would not be a surprising fact.

There is, however, a positive-strand RNA virus, MS2, that infects bacterial hosts. It also has an icosahedral protein coat composed of 60 pseudosymmetric homotrimers, but the folded polypeptides do not appear to be related to the folded polypeptides of the coat proteins of other positive-strand RNA viruses.¹⁷³

All icosahedral viral protein coats have 60 identical protomers arrayed around the global 5-fold, 3-fold, and 2-fold rotational axes of icosahedral symmetry of point group 532(*I*). It is the identity and the number of subunits within each protomer that differ among them. The smallest protein coats have only one subunit in each protomer, those with three subunits in each protomer are somewhat larger, and those with more than three are larger still. The **number of subunits within a protomer** is designated with a capital *T*. For example, the expanded viral protein coats that have been discussed so far, in which there are three subunits in each protomer (Figures 9–26 and 9–28), have $T=3$ icosahedral symmetry. The numbers of subunits found in the protomers of the larger protein coats are those numbers that permit the subunits to assume quasi-equivalent positions around local rotational axes of pseudosymmetry and that at the same time are compatible with the global rotational axes of icosahedral symmetry. For example, in viral protein coats with $T=3$ icosahedral symmetry, the fact that a global 5-fold rotational axis of symmetry and two global 3-fold rotational axes of symmetry are arranged equilaterally (Figure 9–21C) produces the local 3-fold rotational axis of pseudosymmetry around which three subunits can be arranged within a protomer while still being compatible with those global axes of symmetry.

The viral protein coats with expanded $T=3$ icosahedral symmetry are the simplest cases of a common strategy^{218,219} used to expand the number of subunits in a protomer. Consider a **hexagonal array of cyclic hexamers** with their respective 6-fold rotational axes of symmetry normal to the plane of the array (Figure 9–30). Such a hexagonal array of hexamers automatically creates an array of global 6-, 3-, and 2-fold rotational axes of symmetry, also all normal to the plane, that operate on the entire array. In certain combinations, these global axes of symmetry have the same spacing and almost the same fold relative to each other that the global axes of icosahedral symmetry have around one of its protomers. For example, in each of the four nested quadrilaterals in

Figure 9–30A, the global 6-, 2-, 3-, and 2-fold rotational axes of hexagonal symmetry at its four vertices have the same relative spacing as the global 5-, 2-, 3-, and 2-fold rotational axes of symmetry around a quadrilateral protomer in an icosahedral array (Figure 9–30, right panel). Consequently, the array of subunits within any one of these boundaries is able to be one of the 60 protomers in an icosahedral array if the subunit at the global 6-fold rotational axis at the top vertex is able to adapt quasi-equivalently to the global 5-fold rotational axis of icosahedral symmetry. Similar compatibilities between the global axes of symmetry at the vertices and on the edges of the nested equilateral triangles of Figure 9–30B and the quadrilaterals of Figure 9–30C–E also allow the arrays of subunits within their boundaries to be one of the protomers in an icosahedral array. The nested sets in Figure 9–30A produce protomers with 4, 9, 16, and 25 subunits ($T=4$, $T=9$, $T=16$, and $T=25$); those in Figure 9–30B produce protomers with 3, 12, and 27 subunits ($T=3$, $T=12$, and $T=27$); and the skewed quadrilaterals in Figure 9–30 panels C, D, and E produce protomers with 7, 13, and 19 subunits ($T=7$, $T=13$, and $T=19$), respectively.

In each of these protomers, the three designated global rotational axes of symmetry of the hexagonal array at the vertices and the edges become global axes of icosahedral symmetry with the same fold. The global 6-fold rotational axis of symmetry of the hexagonal array at the undesignated vertex becomes a quasi-equivalent global 5-fold rotational axis of symmetry when the protomer is in the icosahedral array. The other global rotational axes of symmetry of the hexagonal array that fall within and on the edges of the boundaries become local rotational axes of pseudosymmetry in the icosahedral array, or if the protomer is large enough some of them become local rotational axes of symmetry. It was the realization of the fact that such a hexagonal array is compatible with icosahedral symmetry that allowed Fuller to design the **geodesic domes**,²¹⁹ which preceded the realization that viral protein coats are geodesic domes.²¹⁸

The protein coats of Sindbis virus²³³ and Nudaurelia ω Capensis virus²³⁴ each have 240 identical subunits arranged with **$T=4$ icosahedral symmetry**. In the protomer of $T=4$ icosahedral symmetry (Figure 9–30A), there is a local 3-fold rotational axis of pseudosymmetry (gray symbol in Figure 9–30A) equidistant from the two 2-fold rotational axes of symmetry and the upper vertex, in a position equivalent to the local 3-fold rotational axis of symmetry in the center of a protomer of $T=3$ icosahedral symmetry (gray symbol in Figure 9–30B). This local 3-fold rotational axis of pseudosymmetry is retained in the $T=4$ icosahedral shell in each protomer and is the local rotational axis of pseudosymmetry for a trimer of quasi-equivalent subunits, just as is the equivalent local 3-fold rotational axis of pseudosymmetry in a protein coat of $T=3$ icosahedral symmetry. Another copy of the same trimer necessarily occupies each position in the shell at which one of the 10 global 3-fold rotational axes

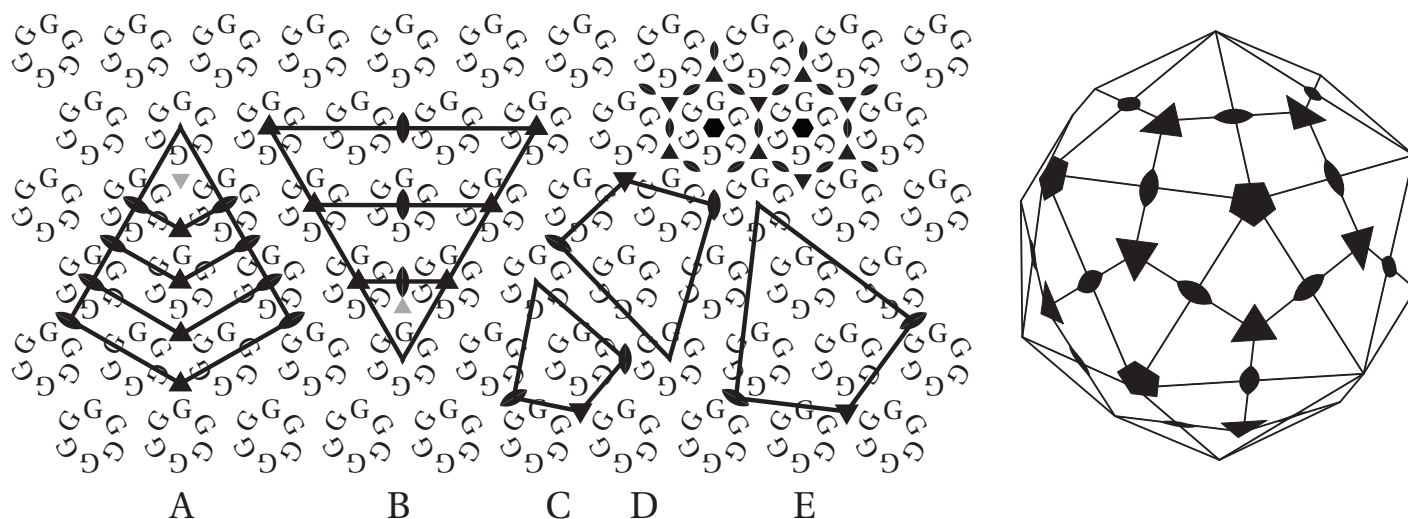


Figure 9-30: Compatibility of a hexagonal array of hexamers with icosahedral symmetry. The positions of the global axes of symmetry normal to an infinite hexagonal array of homohexamers are indicated in the distributions surrounding the two hexamers in the upper right of the array. A solid hexagon indicates a global 6-fold rotational axis of symmetry. These rotational axes of symmetry are located in the respective positions throughout the array. (A) Segments of the hexagonal array containing 4, 9, 16, and 25 subunits, respectively, that can be protomers of icosahedral symmetry. The quadrilaterals enclosing these four segments and those in C, D, and E are the protomers of an alternative icosahedral hexacontahedron (drawn to the right of the hexagonal array). In each of the protomers of this hexacontahedron, a 5-fold, a 2-fold, a 3-fold, and a 2-fold rotational axis of global symmetry are consecutively joined by line segments. (B) Segments of the hexagonal array containing 27, 12, and 3 subunits, respectively, that can be protomers of icosahedral symmetry that are each enclosed by the boundaries of the equilateral triangle connecting a global 5-fold and two global 3-fold rotational axes of symmetry as in the icosahedral hexacontahedron in Figure 9-21C. The gray global 3-fold rotational axes of symmetry of the hexagonal array in A and B become local 3-fold rotational axes of pseudosymmetry when the protomers for $T=4$ and $T=3$, respectively, are inserted into an icosahedral array (the hexacontahedron to the right of the hexagonal array and the hexacontahedron in Figure 9-21C, respectively). (C–E) Skewed segments of the hexagonal array containing 7 (C), 13 (D), and 19 (E) subunits that also can be protomers of icosahedral symmetry. The 2-fold and 3-fold rotational axes of symmetry noted at the boundary of each of the potential protomers coincide with global 2-fold and 3-fold rotational axes of symmetry of the symmetry, and the unmarked 6-fold rotational axis of symmetry becomes a quasi-equivalent 5-fold rotational axis of symmetry, when the protomer is inserted into the icosahedral array.

of symmetry emerges because that position is itself quasi-equivalent to that occupied by a trimer on one of the local 3-fold rotational axes of pseudosymmetry. Consequently, in a coat protein with $T=4$ icosahedral symmetry, there are 60 trimers located at the 60 local 3-fold rotational axes of symmetry, one in each protomer, and 20 of the same trimers located on the respective ends of the 10 global 3-fold rotational axes of symmetry.

Because the requirements placed upon the 80 trimers in $T=4$ icosahedral symmetry are similar to those placed upon the 60 trimers in $T=3$ icosahedral symmetry, there are proteins that can adopt either $T=3$ or $T=4$ icosahedral symmetry depending on the conditions,²³⁵ and within the family of alphaviruses there are protein coats with either 180 subunits ($T=3$) or 240 subunits ($T=4$).²³³

The protein coat of bacteriophage P22^{236,237} has seven identical subunits (429 aa) arranged with $T=7$ icosahedral symmetry (Figure 9-30C) in each of its 60 protomers. The subunits adapt quasi-equivalently to form the pentamers sitting on the global 5-fold rotational axes of symmetry and the hexamers sitting on the local 6-fold rotational axes of pseudosymmetry within each protomer. The mirror image of the arrangement of the

rotational axes of symmetry in each of the various skewed quadrilaterals in Figure 9-30C–E does not superpose on itself, so there are **right-handed and left-handed versions** of each of these quadrilaterals. The coat protein of bacteriophage P22 is a left-handed $T=7$ array.

The family of papilloma viruses, simian virus 40, and polyoma viruses have protein coats that are a peculiar variation on $T=7$ icosahedral symmetry.^{238–240} Each of these viral protein coats is assembled from 72 identical pentamers with cyclic symmetry the subunits of which are held together by extensive interfaces. Rather than using the same subunit to form pentamers and hexamers, both the global 5-fold rotational axes of icosahedral symmetry and the local 6-fold rotational axes within the protomer are occupied by copies of the pentamer. Consequently, dramatically different interfaces, on the same outer surfaces of identical folded polypeptides in each pentamer, must be made around the global 2-fold rotational axis of icosahedral symmetry and the local 2- and 3-fold rotational axes of hexagonal symmetry in the spaces between the homopentameric subunits. This is accomplished by forming the interfaces with flexible, structurally swapped strands of polypeptide rather than the usual rigid interfaces of interdigitated secondary structure.

Bluetongue virus²¹⁶ and reovirus²⁴¹ have protein coats with **$T = 13$ icosahedral symmetry** (Figure 9–30D). In the former, the most extensive interfaces are those holding trimers of subunits together around the global and local 3-fold rotational axes, a fact suggesting that the fundamental unit of the structure is a trimer. The trimers, however, must adapt quasi-equivalently to being arranged about both the global 5-fold rotational axes of symmetry and the local 6-fold rotational axes of symmetry.

Herpes simplex virus^{242–244} has a protein coat with **$T = 16$ icosahedral symmetry** (Figure 9–30A). Pentamers called pentons occupy the global 5-fold rotational axes of symmetry, and hexamers called hexons occupy the local 6-fold rotational axes of symmetry; but both pentons and hexons are formed from five copies and six copies, respectively, of the same folded polypeptide (1374 aa). There are 16 copies of this folded polypeptide in each protomer so each protein coat contains 1,319,040 aa.

Adenovirus²⁴⁵ has a protein coat with **$T = 25$ icosahedral symmetry** (Figure 9–30A). In this arrangement each protomer contains four hexons that occupy the local 6-fold rotational axes of pseudosymmetry, but a hexon is not a hexamer, it is a homotrimer. Each of the three identical subunits of one of these homotrimers contains two internally duplicated domains, and six domains, two from each subunit, are arranged around each local 6-fold rotational axis to produce the pseudosymmetrically displayed faces for the interfaces in which each hexon must participate with its six neighbors.²⁴⁵ Unlike those in herpes simplex virus, the pentons in adenovirus, centered on the global 5-fold rotational axes of symmetry, are formed by different folded polypeptides (571 aa) from those (967 aa) forming the hexons.

In all of these viral protein coats based on hexagonal expansion of the basic icosahedral protomer, the problem of closing the protomer arises. A hexagonal array is a potentially infinite array, yet the protomers within the boundaries of the various equilateral triangles and quadrilaterals of Figure 9–30 are finite portions of that infinite array. During the assembly of the protein coat, a mechanism for measuring out the size of that portion is required, and this role seems to be filled by proteins accessory to the subunits of the coat.²⁴⁶

The viral protein coats in which the protomer contains only a few subunits have the appearance of spheres, even though their surfaces are often quite irregular (Figure 9–28). As the number of subunits in the hexagonal array of hexamers becomes greater, there is a tendency for the structure to look polyhedral^{243,247,248} because of the tendency of the protomers to adopt the plane of the hexagonal array. In some viral protein coats, the same quasi-equivalent interface will hold its two subunits in a plane at one location, fitting into a polyhedral face, and at an angle to each other in another, forming a crease along a polyhedral edge.²²³ Usually, however, the hexagonally packed protomers are bowed out so that

their subunits end up evenly distributed over the surface of an almost spherical oligomeric protein, just as the modular units of a geodesic dome end up producing an almost spherical exploded icosahedron and for the same reason. The joints adjust to distribute uniformly the strain produced by creasing²¹⁸ the hexagonal array consequent to requiring certain of its 6-fold rotational axes of symmetry to become 5-fold rotational axes of symmetry.²¹⁹

The protein clathrin forms isometric cagelike structures that assemble around small pinocytotic vesicles as they bud inward from the plasma membrane of an animal cell. The polypeptide is 1600 aa in length, and when folded it produces a tubular protein, 45 nm in length and 2.5 nm in diameter.²⁴⁹ The protomer from which the **cages of clathrin** are formed is a trimer of these polypeptides, all three joined together at one end around a 3-fold rotational axis of symmetry to produce a triskelion with bent arms.²⁴⁹ Different numbers of these triskelia can assemble to produce intact cages of various shapes between 70 and 200 nm in diameter.²⁵⁰ The wires of the mesh forming these cages are presumed to be formed from two or more intertwined arms of the triskelia.²⁵¹ Each and every vertex in each and every cage is a junction of three wires, and each vertex must contain the local 3-fold rotational axis of symmetry at the nexus of an individual triskelion. The mesh itself is always formed of pentagons and hexagons of wire producing polyhedra with as many as 32 faces, but most of them are not based on isometric symmetries.²⁵⁰ It is the elongated and flexible nature of the subunit that permits the one protein to generate such a wide variety of oligomeric proteins.

Suggested Reading

- Caspar, D.L.D., & Klug, A. (1962) *Cold Spring Harbor Symp. Quant. Biol.* 27, 1–24.
- Rossmann, M.G., Abad-Zapatero, C., Hermodson, M.A., & Erickson, J.W. (1983) Subunit interactions in southern bean mosaic virus, *J. Mol. Biol.* 166, 37–83.
- Izard, T., Aevansson, A., Allen, M.D., Westphal, A.H., Perham, R.N., de Kok, A., & Hol, W.G. (1999) Principles of quasi-equivalence and Euclidean geometry govern the assembly of cubic and dodecahedral cores of pyruvate dehydrogenase complexes, *Proc. Natl. Acad. Sci. U.S.A.* 96, 1240–1245.

Problem 9–9: In Figure 9–22 there are four 3-fold rotational axes of symmetry, each of which superposes four different triplets of subunits. For example, one of these axes superposes A, B, and C; J, E, and H; L, D, and G; and K, F, and I.

- (A) List the four triplets for each of the other three 3-fold rotational axes of symmetry.

In Figure 9–22 there are three 2-fold rotational axes of symmetry, each of which superposes six different twins of subunits; for example, one of them superposes A and K, B and L, C and J, D and H, E and I, and F and G.

- (B) List the six twins for each of the other two 2-fold rotational axes of symmetry.

Problem 9–10: Make a xerographic copy of Figure 9–26. On that xerographic copy designate every global 2-, 3-, 5-, and 6-fold rotational axis of symmetry. Use the standard symbols for this designation. In the same figure designate some of the local 2-, 3-, and 6-fold rotational axes of pseudosymmetry by the symbols P_2 , P_3 , and P_6 , respectively.

Problem 9–11: Kepler derived the rhombic dodecahedron from the intersection of a cube and an octahedron. Draw the intersection of a cube and an octahedron, and connect its vertices to produce a rhombic dodecahedron.

Helical Polymeric Proteins

Helical fibers formed from identical subunits of protein have useful properties, and there are many examples of them. For example, to accomplish its function of hybridizing homologous single strands of DNA, the RecA protein binds to DNA in a long helical polymeric sheath that matches the helical symmetry of the DNA. The helicity of the sheath, however, is built into the protein because it spontaneously forms a sheath with almost the same helical symmetry even in the absence of the DNA.²⁵²

Every time an interface is created by evolution between two complementary faces on the surface of copies of the same globular protein, no matter how those two faces are disposed on its surface, a distinct screw axis of symmetry is defined. Most of these screw axes would be open and generate helical polymers of the monomer, so the existing helical polymers must be the few that have escaped elimination by natural selection. The surprising fact is that there are so few helical polymeric proteins.

A single **geometric helix** can be defined by three parameters: its radius, its hand, and its pitch. The pitch of a helix is the distance it rises for each complete turn. It can also be defined by four parameters: its **radius**, its **hand**, a recurring **radial angle** dividing the helix into equal segments of arc, and the **rise** for each of these equal segments of arc. In a helical polymeric protein the second definition makes more sense because the successive subunits can be considered to be the repeating segments of arc.

An interface between two identical molecules of a protein can generate several types of helical polymers. The actin helix (Figure 9–1B,C) is an example of the simplest type in which the protomers ascend one step at a time around the screw axis. In the actin polymer, the screw axis of symmetry passes through a corner of each protomer, the helix is not much wider than the protomer (Figure 9–1C), and the radial angle between successive protomers is fairly large (-166°). It has already been

noted that the actin helix can be represented as a single helix or as a double helix.

If the generating interface creates a screw axis of symmetry where the radial angle between successive subunits is fairly small, so that there are a number of subunits in one turn of the helix, and where the rise for each subunit is just enough to cause the subunits in each turn of the helix to lie upon the subunits from the turn below it, then a **singly threaded cylinder** is formed. Because of steric problems, such a generating interface usually creates a screw axis of symmetry that is separated from the subunits themselves, rather than passing through each of them, and the threaded cylinder is hollow inside. An example of such a hollow, singly threaded helical cylinder is tobacco mosaic virus (Figure 9–31).^{253–256} In tobacco mosaic virus, the angle between successive subunits is 22.03° , and the rise for each subunit is 0.14 nm.^{255,256} These dimensions bring the subunits in the next turn of the helix into contact with the subunits in the preceding turn, and the pitch of the helix is 2.3 nm, the height of a subunit. Between two successive turns there are interfaces among the protomers. Because of the screw symmetry, each protomer provides at its lower surface the upper faces for the interfaces with the two protomers below it and at its upper surface the lower faces for the interfaces with the two protomers above it. Because every protomer is the same, each of these respective interfaces is the same and repeats along the thread every 22.03° .

The generating helix emphasized in the drawing of tobacco mosaic virus (Figure 9–31B) is a right-handed helix of pitch 2.3 nm. If the structure is examined closely, however, it can be seen that there are sets of helices of steeper pitch running through it. The subunits in tobacco mosaic virus are arranged upon a **helical surface lattice**²⁵⁷ that contains all of these other sets of helices. The helical surface lattice can be displayed in two dimensions by cutting the cylinder along a line parallel to the central axis and flattening it upon the page (Figure 9–31C). When the helical surface lattice of tobacco mosaic virus is viewed in this format, it can be seen that, in addition to the single right-handed helix of pitch 2.3 nm running through the lattice, there are also sets of 17 parallel right-handed helices and sets of 16 parallel left-handed helices (lower and upper sets of arrows, respectively, in Figure 9–31C).²⁵⁸ Any one of these sets of helices can also define the structure. A helical surface lattice can be uniquely designated by the number of parallel strands of left-handed twist (a negative number) and the number of parallel strands of right-handed twist (a positive number) for any one of the sets of each respective hand.²⁵⁷ For example, the helical surface lattice of tobacco mosaic virus can be designated $(-16, 1)$ or $(-16, 17)$. The surface lattice of the helical polymer of flagellin from *Salmonella typhimurium* also contains a singly threaded right-handed helix as well as a set of five parallel left-handed helices, a set of six parallel right-handed helices, and a set of 11 parallel left-handed helices.²⁵⁹

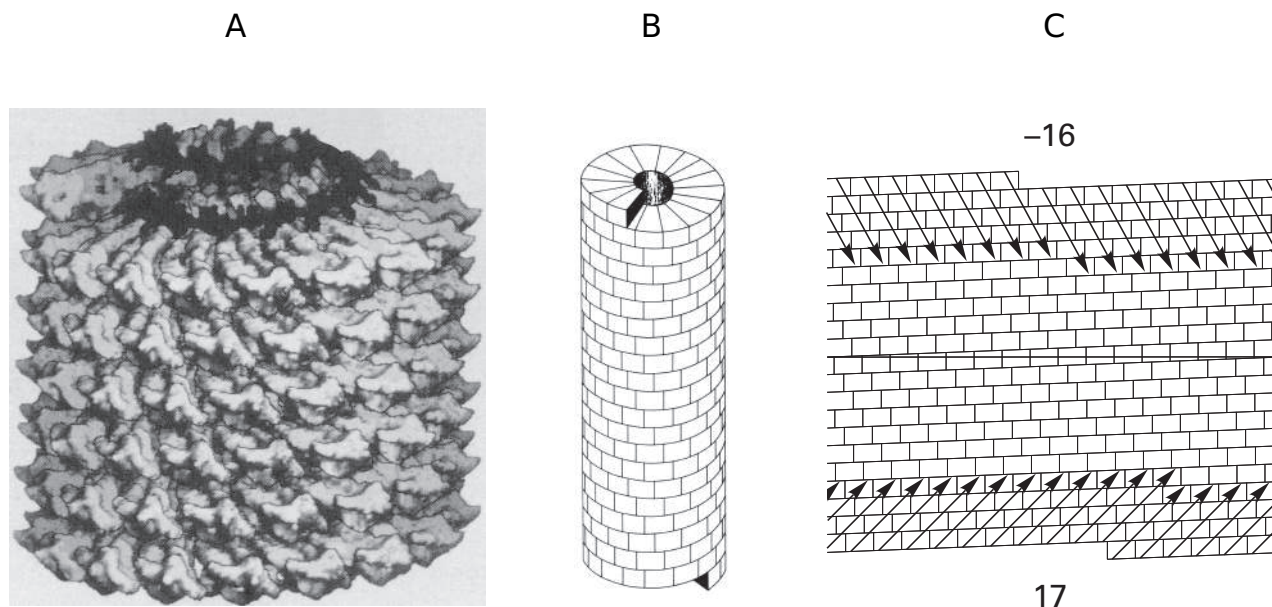


Figure 9-31: Helical surface lattice of tobacco mosaic virus. (A) Molecular model derived from a map of electron density calculated from the helical diffraction pattern of X-radiation (Bragg spacing ≥ 0.29 nm) emerging from an oriented preparation of the viruses.²⁵⁶ The diffraction from a specimen prepared in 1960 was gathered to Bragg spacing of 0.29 nm in 1982. The timing of this sequence of events illustrates how rare it is to obtain a well-aligned preparation of a helical polymeric protein. Reprinted with permission from ref 256. Copyright 1989 Elsevier B.V. (B) Diagrammatic representation of the helical surface lattice of the protein coat of tobacco mosaic virus.²⁵⁴ Individual protomers form a singly threaded helical screw that is a hollow cylinder. Each protomer has the same orientation and the successive protomers are related by a rise of 0.14 nm and a rotation of 22° along the helix. Reprinted with permission from ref 254. Copyright 1972 Federation of American Societies for Experimental Biology. (C) The surface helix of tobacco mosaic virus flattened onto the page.²⁵⁵ The hollow cylinder in panel B was split along a vertical plane passing through its wall and then flattened onto the page. The horizontal line represents a circle around the cylinder, the plane of which is normal to the central axis of the cylinder, that has been also split and flattened. It has been added to assist in counting the numbers of parallel helices in the various sets. The arrows at the upper end of the lattice indicate the set of 16 left-handed helices, and those at the lower end, the set of 17 right-handed helices that run through the lattice. These are referred to as 16 start and 17 start arrays, respectively. Reprinted with permission from ref 255. Copyright 1974 Academic Press.

A helical surface lattice does not have to have a single shallow helix of one or the other hand acting as the generating operation. The extended tail of the T4 bacteriophage is built from hexamers with cyclic symmetry. The hexamers are stacked upon each other by successive interfaces that cause them to be out of alignment in a right-handed sense by 17° .²⁶⁰ This creates a helical surface lattice that is a hexuply threaded cylinder. The $(-6, 6)$ lattice contains both a set of six parallel helices of shallow pitch with left-handed sense and a set of six parallel helices, of steeper pitch, with a right-handed sense.

The CA protein from type 1 immunodeficiency virus spontaneously forms several different sizes of helical tubes. It forms two different tubes with distinct $(-12, 11)$ and $(-11, 10)$ surface lattices, respectively, that each have a single generating helix of one strand with a left-handed pitch. The same protein, however, also forms tubes with $(-10, 13)$, and $(-8, 5)$ surface lattices.²⁶¹

The radial angle relating successive subunits to each other in a helix or each of the identical helices in a set of helices in a helical polymer is determined by the interfaces among them. It is these interfaces that produce a helical filament such as actin or a singly or multiply threaded cylinder such as tobacco mosaic virus,

flagellin, or the extended tail of the T4 bacteriophage. This **radial angle** between successive subunits that these interfaces dictate can have **any numerical value**. Technically, this means that the helix or helices never position a monomer exactly above any monomer below it in the helix. For example, tobacco mosaic virus has 49.02 subunits in three turns (22.03° subunit⁻¹), not 49.²⁵⁶ Therefore, there is no translationally repeating unit in a rigid, biological helical structure. What this means is that a helical polymeric protein has difficulty crystallizing in three dimensions, and crystals of helical polymers suitable for high-resolution crystallographic studies have not been produced. One interesting example, which is almost an exception to this uniform failure, is the protofilament running through a crystal of the protein encoded by the *mreB* gene in *E. coli*. This protein is a homologue of actin and forms filaments similar to those formed by actin (Figure 9-1B) but without the helical twist, so a filament can be incorporated into a crystal.⁶

One approach to determining the structure of a helical polymeric protein is to crystallize the monomer, construct its crystallographic molecular model at atomic resolution, and simultaneously determine the structure of the helical polymer at high enough resolution to posi-

tion crystallographic models of these monomers in the proper orientation at the locations they occupy in the polymer. Such a strategy has been applied successfully to the polymer of actin (Figure 9-1B).^{3*} The details of the structure of an intact polymer at the resolution required by this strategy can be determined by image reconstruction²⁶³ of electron micrographs.

Under appropriate circumstances, an **electron microscope** is capable of magnifying the image of a protein sufficiently that individual subunits can be distinguished and their shapes almost distinguished. A beam of collimated electrons is impinged upon the sample, and those that pass through it without being deflected sufficiently to leave the beam are then focused by electromagnetic lenses. The contrast in the image is caused by the distribution within the sample of the ability to deflect or scatter electrons from their path. The sample placed in an electron microscope usually has to reside in a chamber under high vacuum. This means that the sample has to be a solid with a low vapor pressure. In the case of molecules of protein, this requires that they be encased in a solid matrix, which also must be a **glass** to avoid the problems of the diffraction caused by crystalline solids.

To enhance contrast and provide a solid support simultaneously, the glass is often formed by drying a solution of a salt containing a heavy metal such as uranium or tungsten. Either uranyl acetate or sodium phosphotungstate, for example, will form an electron-dense glass when a film of a solution containing it is dried. This electron-dense glass will surround, encase, and support a molecule of protein that had been present in the solution from which the film was made. The glass of the salt of heavy metal encasing the molecule of protein forms a three-dimensional boundary or mold that has the shape of the molecule of protein. It is the dark image of this mold of electron-dense glass encasing the light image of the electron-translucent molecule of protein that is observed in the micrograph. This procedure is known as **negative staining**.

It is also possible to insert thin layers of **amorphous ice**, which is a glass, into a **cryoelectron microscope**, an electron microscope not with cold electrons but with a stage for the sample that is cooled to a very low temperature.^{264,265} When a molecule of protein is embedded in such a glass (Figure 9-32A),²⁶⁶ the contrast observed results from the fact that the atoms in a molecule of protein are more efficient at deflecting electrons than the molecules of water in amorphous ice. A positive image of the molecule, rather than a negative image, is observed.

The three-dimensional distribution of the ability to deflect electrons within the sample is known as the distribution of scattering density, $\theta(x,y,z)$. If a map of scat-

tering density can be produced at as high a resolution as possible, details of the structure of either the mold in which the molecule of protein is encased or the molecule of protein itself can be observed. **Image reconstruction** is any computational method that is used to calculate $\theta(x,y,z)$ from the images of molecules of proteins observed in electron micrographs.²⁶³ In all cases, the electron micrograph is the experimental data submitted to these calculations, and the electron micrograph used in a particular reconstruction must be presented to the reader so that she may appreciate the point of departure (Figure 9-32A).

A molecule of protein has a certain distribution of electron density $\rho(x,y,z)$, and when molecules of protein are arrayed in a crystal they create a periodic, three-dimensional distribution of electron density. This periodic array diffracts X-rays to produce a diffraction pattern that is also periodic. The angular dispositions of the reflections in the diffraction pattern are determined by both the angles among the axes of the fundamental unit cell and its dimensions. The dimensions and axial angles of the unit cell can be calculated from these angular dispositions. The diffraction pattern of the crystal is the Fourier transform of the periodic distribution of electron density it contains. The magnitudes and phases of the maxima in the diffraction pattern can be calculated from the distribution of electron density within the unit cell by digital Fourier transformation. Conversely, the distribution of electron density in the unit cell can be calculated from the amplitudes and phases of the diffraction maxima by digital Fourier transformation.

A helical polymer of protein in its mold of negative stain or amorphous ice has a certain **distribution of scattering density**, $\theta(x,y,z)$, which is a **periodic function** because the helix is periodic. Each protomer of the polymer is a unit cell in this helical array. The **computed Fourier transform** of this periodic array is also a periodic function. From the spatial disposition of its maxima, the angle between successive unit cells in the helix, the rise for each unit cell, and the number of helical threads in the structure can be calculated. From the amplitudes and phases of the maxima of the Fourier transform, the distribution of scattering density within the unit cell can be calculated.

The depth of focus in an electron microscope is larger than the width of a specimen containing a helical polymer, and all points in the specimen are in focus in the final micrograph. As such, the micrograph represents the projection of the three-dimensional distribution of scattering density onto a two-dimensional surface.²⁶³ The Fourier transform, $F(X,Y,Z)$, of any three-dimensional distribution of scattering density is²⁶⁷

$$F(X, Y, Z) = \iiint_{\text{object}} \theta(x, y, z) \exp[2\pi i(xX + yY + zZ)] dx dy dz \quad (9-4)$$

* A more recent crystallographic molecular model of actin²⁶² has a somewhat different structure than the one used to construct Figure 9-1B, but the differences are not significant enough to affect the choice of the orientation of the monomer within the polymer.

When $Z = 0$

$$F(X, Y, 0) = \iint \sigma(x, y) \exp[2\pi i(xX + yY)] dx dy \quad (9-5)$$

where

$$\sigma(x, y) = \int \theta(x, y, z) dz \quad (9-6)$$

The function $\sigma(x, y)$ is the projection of the three-dimensional distribution of scattering density. This set of relationships states that the two-dimensional Fourier transform of the projection of the distribution of scattering density is the **central section of the three-dimensional Fourier transform** of the unprojected distribution of scattering density. This central section of the three-dimensional Fourier transform is obtained by digitizing the optical density of the micrograph and calculating a digitized Fourier transform of the image by computer. The significant advantage of performing the Fourier transform computationally rather than by diffraction is that the computed Fourier transform comes with both **amplitudes and phases** instead of just amplitudes. The disadvantage is that there are far fewer unit cells contributing to the Fourier transform.

In the case of a helical polymeric protein embedded in negative stain or amorphous ice, the central section of its three-dimensional Fourier transform systematically intersects all of the maxima in its three-dimensional Fourier transform. In this two-dimensional central section of its Fourier transform (Figure 9-32B), the amplitudes and the associated phases of the transform (Figure 9-32C) are arrayed along **layer lines** (the horizontal lines in the figure) that arise from the repeating patterns in the helix.^{257,268} If the correct helical lattice has been assigned to the structure so that the layer lines can be properly **indexed**,²⁶⁹ the indexed amplitudes and phases distributed along the layer lines (Figure 9-32C) can be used to calculate $\theta(x, y, z)$ by a **Fourier-Bessel transform**,²⁶⁸ just as the properly indexed amplitudes and phases of the pattern of the X-rays diffracted from a crystal can be used to calculate $\rho(x, y, z)$ by a Fourier transform. Figure 9-32D presents an example of such a reconstruction from the electron micrograph of Figure 9-32A.

Image reconstruction succeeds in producing the molecular structure of a helical polymeric protein when it is able to produce an image of the helical polymer of sufficient resolution to position and orient unambiguously a crystallographic molecular model of the monomer within the polymer. In addition to the success this approach has achieved with the helical polymer of actin (Figure 9-1B),³ it has been possible to insert a crystallographic molecular model of the $\alpha\beta$ heterodimer of tubulin²⁷⁰ into image reconstructions of the micro-

tubule^{271,272} and to insert a crystallographic molecular model of flagellin into an image reconstruction of a bacterial flagellar filament.²⁷³ In the latter case, as with the protein encoded by the *mreB* gene of *E. coli*, the flagellin crystallized within a protofilament of the overall flagellar filament. Consequently, its position and orientation within the image reconstruction of the complete filament could be assigned with greater certainty.

Usually the Fourier transform of a digitized electron-microscopic image of a helical polymeric protein embedded in amorphous ice has measurable amplitudes that arise from helical spacings of 2 nm or greater,^{261,274,275} and as in crystallography, this determines the **lower limit of the resolution**. Images of tobacco mosaic virus, however, have produced Fourier transforms with terms arising from spacings of 1 nm;²⁷⁶ and images of helical tubes of acetylcholine receptor, terms arising from spacings as little as 0.5 nm.²⁷⁷

Difference maps of scattering density can also be useful. A helical polymeric protein, for example, helical fibers of tubulin, often associates with another protein, for example, kinesin, at sites on its surface, one of which is located on each of the asymmetric units in the helical lattice, for example, on each subunit of tubulin.²⁷⁸ When the helical polymer is decorated with the other protein, for example, when a helical fiber of tubulin is decorated with kinesin, the distribution of that other protein will assume the helical distribution of the underlying filament. The Fourier transform of an image of the undecorated filament is subtracted from that of a decorated filament, and the Fourier-Bessel transform of this difference is a map of scattering density arising just from the bound protein.^{266,278}

A population of **oriented helical polymeric proteins** produces **X-ray diffraction**. As a diffraction pattern is the Fourier transform of the distribution of electron density in the helical specimen, these X-ray diffraction patterns also display the layer lines seen in Fourier transforms of digitized images from electron micrographs of those same specimens. Because the reflections are produced by diffraction rather than computationally, they must be phased by multiple isomorphous replacement.²⁵⁶ This has been accomplished for oriented filaments of the protein coat of tobacco mosaic virus incorporating reflections arising from spacings to 0.3 nm to obtain a map of electron density into which a model of the polypeptide could be inserted (Figure 9-31A).^{256,279} It is also possible to use phases from electron micrographs and amplitudes from X-ray diffraction to produce a map of electron density.²⁵⁹

There are helical polymeric proteins that are constructed from long, flexible strands of polypeptide rather than globular protomers. These proteins resemble the **helical cables** that are used in the construction of suspension bridges. The smallest structural element in a cable is a **strand**. Two or more strands are twisted around each other to make a **rope**. Two or more ropes are then

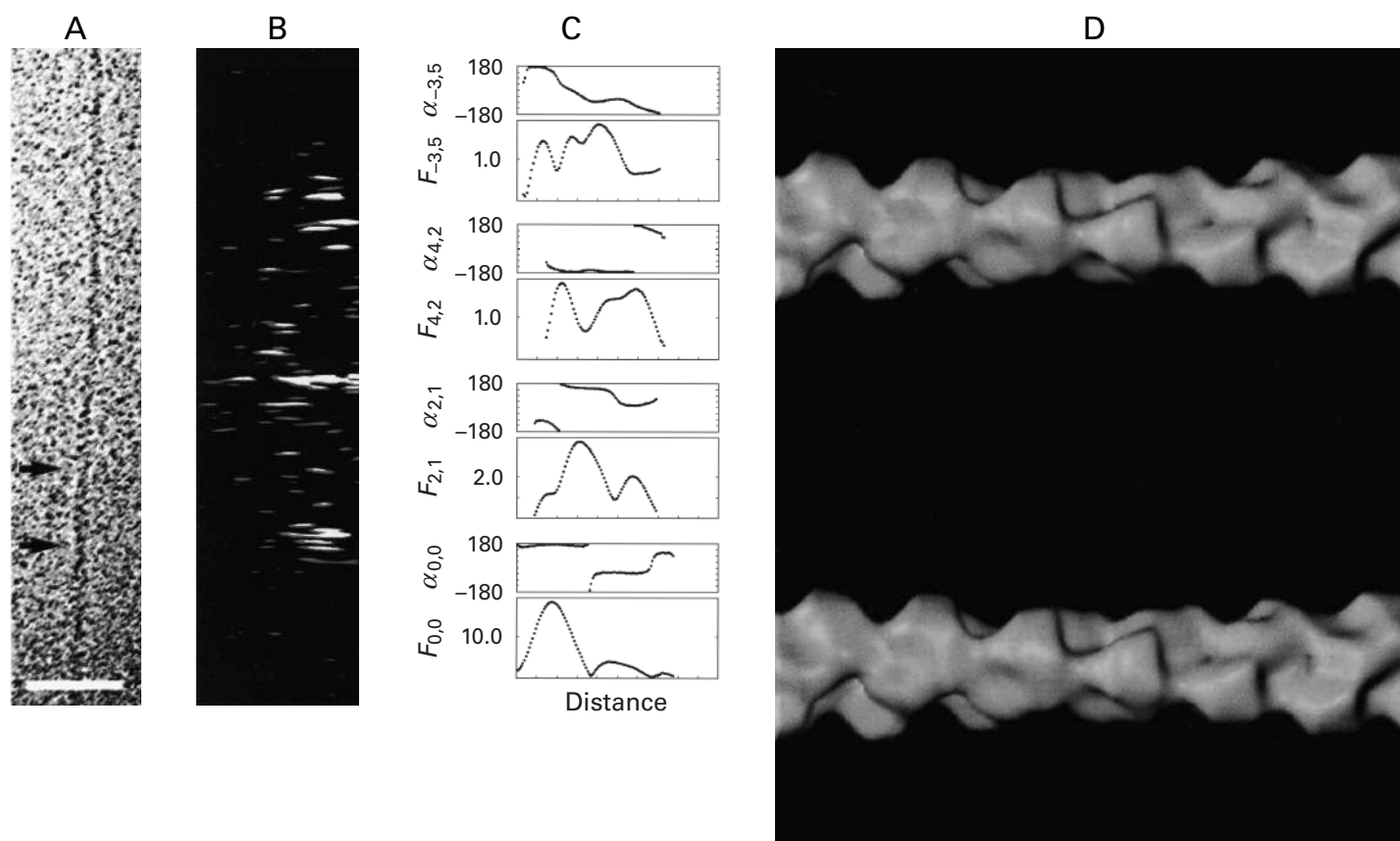


Figure 9-32: Image reconstruction of the helical array of subunits in a filament of actin.²⁶⁶ Filaments of human cytoskeletal β actin were suspended in a buffered aqueous solution. A small sample of the suspension ($4 \mu\text{L}$) was spread over a grid for electron microscopy and rapidly frozen to obtain a thin sheet of amorphous ice in which the filaments were embedded. (A) An electron micrograph ($36000\times$) of an actin filament embedded in the ice. The arrows mark points of helical crossovers. The electron micrograph was scanned and a Fourier transform of the digitized image was performed. (B) The Fourier transform (Equation 9-5) of the image in panel A is presented graphically on a two-dimensional field such that the brightness of the image is directly proportional to the amplitude of the function at that point. Because the array is helical, the maxima in the Fourier transform are found along layer lines. (C) Distribution of the amplitude and phase of the Fourier transform along several of the layer lines in panel A. Phases ($\alpha_{n,l}$; degrees) are presented in the upper plot of each pair and amplitude ($F_{n,l}$; relative units) in the lower plot. The indexing of the layer lines (n, l) designates the Bessel order (n) and the number of the layer line (l). (D) Reconstructed image in stereo resulting from a Fourier-Bessel transform of the amplitudes and phases along the properly indexed layer lines. Reprinted with permission from ref 206. Copyright 2000 Elsevier B.V.

twisted around each other to make a cable. The strands from which ropes of protein are made are polypeptides read from messenger RNA, and for this reason, each strand must have a discrete length. This in turn means that the cable is built from strands of uniform length that are overlapped to provide the necessary tensile strength.

The arrangement of the strands in the molecular rope and the ropes in the cable is elucidated by permitting a macroscopic fiber to diffract X-radiation. A macroscopic fiber, built from billions of the molecular cables, is placed in a beam of X-rays. The cables are all more or less aligned with the axis of the macroscopic fiber. The helical arrays of the strands in the ropes and the helical arrays of the segments of rope in the cable have certain regularly recurring dimensions and angles associated with them that give rise to diffraction. The dimensions and helical parameters of these arrays can be established from the angles at which the reflections emerge from the

fiber. For example, a serial set of reflections is produced by a tendon from the tail of a rat when the tendon is placed in a beam of X-rays. This serial set of reflections arises from a helical array that repeats every 67 nm ,^{280,281} and this dimension, along with others, must be incorporated into the model for the complete cable of the collagen from which the tendon is formed.

Collagen is the helical polymeric protein from which is formed the tough, flexible material composing tendons, intercellular matrix, the matrix of bone, and many strong, plastic sheets of various shapes and sizes found in animals. The basic structural element in these macroscopic structures is the fibril of collagen, which is a cylindrical thread of indefinite length, $200\text{--}800 \text{ nm}$ wide. This thread in turn is formed from molecular cables of collagen, each probably as long as the fibril, packed side by side in register.

The strand from which a molecular cable of colla-

gen is formed is a polypeptide that contains significant amounts of **4-hydroxyproline** and that has a characteristic repeating sequence in which **glycine** is every third amino acid. In humans, there are 30 different polypeptides containing segments with such a repeat. These polypeptides vary in length from 666 to 3039 aa; most of them have lengths of 1000–2000 aa. The segment with the repeating sequence can be as long as 1530 aa or as short as 15 aa. Polypeptides with segments longer than 500 aa in which every third amino acid is a glycine usually have only one continuous segment of this repeating pattern; polypeptides with segments shorter than 300 aa usually have multiple segments of different lengths. The segment or segments with the repeating sequence are usually found in the middle of the polypeptide. It is such a segment that forms the strand of the rope; the remainder of the polypeptide forms appendages to the rope usually at its two ends. If the rope forms a cable, these amino-terminal and carboxy-terminal appendages jut out from the cable at regular intervals.

Three strands of polypeptide with a repeating sequence in which every third amino acid is a glycine can wrap around each other in parallel to form a triple-helical rope (Figure 9–33).²⁸² The collagen **triple helix** is a coiled coil of three helices, each formed by one of the strands. The structure is that proposed by Rich and Crick.²⁸³ The three helices are all left-handed with the same helical parameters. Each has on average a rise of 0.29 nm aa^{-1} over an angle of $-107^\circ \text{ aa}^{-1}$, and this conformation produces a complete turn every 3.36 aa.^{282,284} Suppose each of the three helices had an angle of $-120^\circ \text{ aa}^{-1}$ instead of $-107^\circ \text{ aa}^{-1}$ so that every third amino acid in each of them would point in the same direction. They could then be brought together in a triangular bundle in which every glycine in each helix was directed into the center. As with a coiled coil of α helices, the deficit between -107° and -120° is accommodated by coiling the three individual left-handed helices around each other, but with a right-handed twist of $+13^\circ$ rather than the left-handed twist of -4° . As a result, the core of the triple helix contains only glycines along its entire length.

The three strands fit together so that they are in register. Only one of the three side chains at each level in the rope is a glycine, and the *pro-S* hydrogen of this glycine is pointed towards the center. The side chains that necessarily occupy this *pro-S* position in the other two amino acids at that level from the other two strands point away from the center of the rope. The α carbon of each glycine is snug against the acyl carbon–oxygen of the glycine in the level above. Within each level there are three amino acids arrayed cylindrically about the central axis: the glycine, an amino acid to the carboxy-terminal side of the glycine in the level above, and an amino acid to the amino-terminal side of the glycine in the level below. Traditionally, these three amino acids are designated glycine, amino acid X, and amino acid Y, respectively. The small size of the *pro-S* hydrogen of each

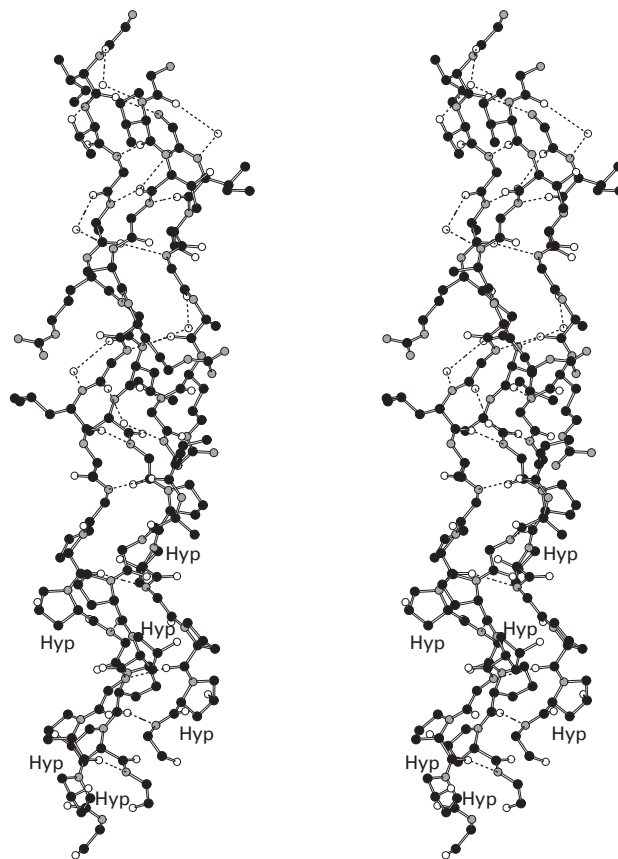


Figure 9–33: Triple helix of collagen. A portion of the crystallographic molecular model²⁸² of the synthetic triacontapeptide with the sequence $(\text{Pro-Hyp-Gly})_3\text{ITGARGLAG}(\text{Pro-Hyp-Gly})_4$, where Hyp is 4-hydroxyproline, is presented. The carboxy-terminal segment $(\text{Pro-Hyp-Gly})_2$ of the portion of the model that has been drawn represents the carboxy-terminal region of nucleation for a triple helix; the amino-terminal segment, $-\text{ITGARGLAG}-$, represents the bulk of the interior of the triple helix. Molecules of water are represented by white spheres attached to the structure only by hydrogen bonds. This drawing was produced with MolScript.⁴⁸⁵

glycine permits the three strands to approach each other so closely that an **interstrand hydrogen bond** forms at each level in the triple helix between the amido oxygen of amino acid X and the amido nitrogen–hydrogen of the glycine in the level below. In all of the regions of the triple helix in which the amino acids X are not prolines or 4-hydroxyprolines, there is a molecule of water hydrogen-bonded to the amido oxygen of the glycine at one level and the amido nitrogen–hydrogen of amino acid X in the level two steps below. These molecules of **water** are integral components of the structure.²⁸⁵ If amino acid Y in the same level as the amino acid X providing a donor to the molecule of water is a threonine, a serine, a 4-hydroxyproline, or an asparagine, the donor in its side chain can form a third hydrogen bond to one of these integral molecules of water.

The **carboxy-terminal regions** of each of the three polypeptides in such a triple helix are often composed of repeating sequences of prolyl-4-hydroxyprolylglycine or prolylprolylglycine.²⁸² These segments are thought to

nucleate the structure.²⁸⁶ Beyond this region of nucleation in the triple helix, the only requirement is for a glycine at every third position. In the short region of nucleation, the helix governing the conformation of a strand rises 0.28 nm aa^{-1} over an angle of $-104^\circ \text{ aa}^{-1}$ and the structure has angles ϕ and ψ of around -70° and $+160^\circ$, which are in the range of those for the polyproline helix. It is thought that these regions spontaneously form a triple polyproline helix, which then propagates into the rest of the structure.^{282,287} Synthetic peptides of the sequence $(\text{Pro-Pro-Gly})_n$ or $(\text{Pro-Hyp-Gly})_n$ spontaneously form such triple helices.^{288,289}

Many of the polypeptides containing sequences in which glycine is every third amino acid form triple helices that do not associate further to form fibrils. In some of these polypeptides the triple-helical regions are frequently interrupted by segments incompatible with a triple-helical structure.²⁹⁰ In others, the two or three triple-helical segments are of insufficient length to effect fibrillar formation.²⁹¹ In yet others, the globular domains flanking the triple-helical segments are too large to permit further association.²⁹² The polypeptides of collagen that associate further beyond the triple-helical state usually contain continuous segments of a thousand or more amino acids with uninterrupted sequences in which glycine is at every third position flanked by carboxy-terminal and amino-terminal portions of several hundred amino acids or less.²⁹⁰

The paradigm of such a **fibrillar collagen** is type I. The triple helix of type I collagen is a heterotrimer of two $\alpha 1$ polypeptides (1057 aa in the human) and one $\alpha 2$ polypeptide (1024 aa in the human) and is formed from a sequence $(\text{Gly-X-Y})_{338}$ of 1014 aa from each chain. The three chains in the triple helix are in register so the rope produced by these three strands is a finite segment, and each segment of rope has the same length. The carboxy-terminal and amino-terminal flanking portions, the telopeptides, are short ($<30 \text{ aa}$), so there is little to interfere with formation of the fibril. This type of collagen is the major component of tendon, and it is fiber diffraction from tendons that has provided the angle of 107° and rise of 0.29 nm characterizing the collagen triple helix²⁸⁴ as well as evidence for repeats of 67 and 30 nm.²⁸¹ These latter repeats arise from the cable formed from the segments of triple-helical rope.

The **segments of rope** (represented by vertical lines in Figure 9-34A)^{281,293} forming this cable are 298 nm in length ($1014 \times 0.29 \text{ nm}$) and are placed side by side in a **right-handed helical array**. In this helical array, the segments of rope are not butted up against each other, and gaps of 37 nm in which the cable has only four ropes alternate with segments of 30 nm in which it has five ropes. The radial angle between each successive protomer, each of which is an individual segment of rope, would be 72° and the helix would repeat every five protomers if the ropes were held perfectly vertical (Figure 9-34A), but they are not.

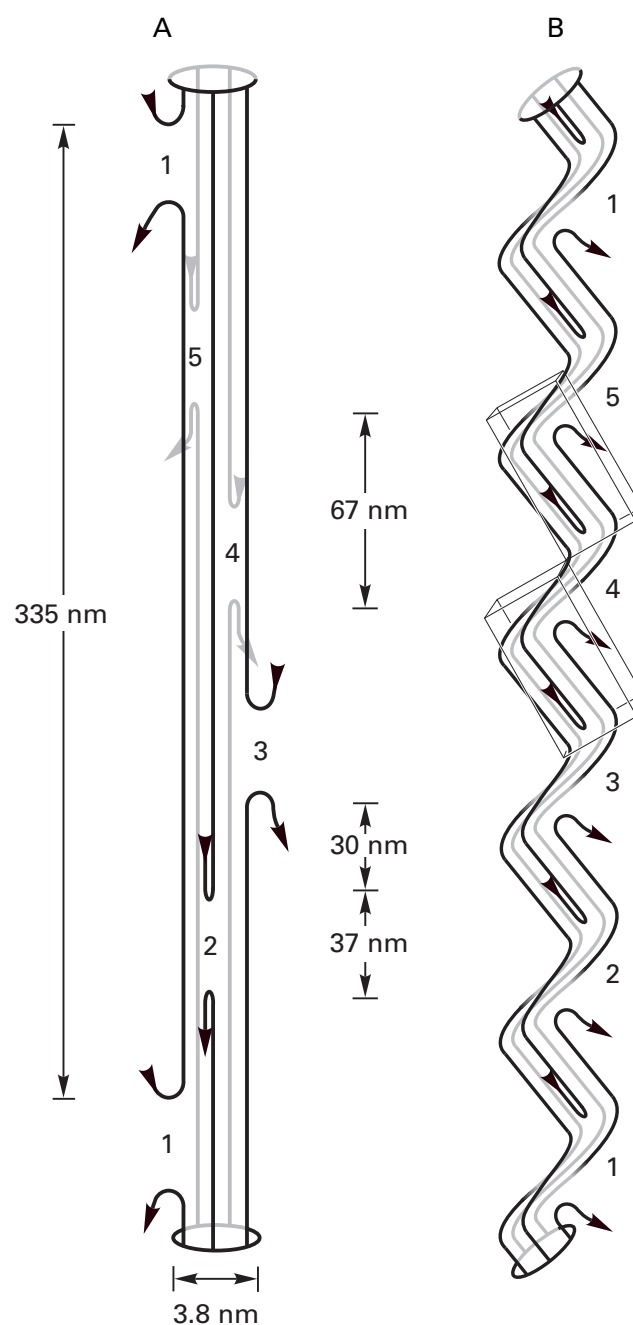


Figure 9-34: Arrangement of the triple-helical ropes of collagen in the cable-forming tendon.²⁹³ The individual segments of rope, each a triple helix formed from three segments of 1014 aa from the central regions of three collagen polypeptides, are represented by vertical lines with splayed amino- and carboxy-terminal ends. The individual segments of rope are arrayed in a right-handed helical distribution (left panel).²⁸¹ In tendon,²⁹³ the cable is twisted in a left-handed sense by one turn over its repeat of 335 nm (right panel). Because the most rigid segments of the cable are where five strands overlap, the twist occurs in the regions where only four strands overlap. The cable is twisted by -72° in each of these gap regions to give each of the five strands a left-handed helical path in these spaces. The cable is kinked as well as twisted. Two of the unit cells for the crystalline array in tendon are drawn. The cable shifts from one column of unit cells into the column of unit cells diagonally adjacent to it at each level.

The collagen in a tendon is arranged in a crystalline array the unit cell of which is triclinic (Figure 4–2)²⁹⁴ with a length of 67 nm, equal to only one-fifth of the 335 nm repeat of the cable. Consequently, the cable must be **twisted** so that its structure repeats translationally every 67 nm. This is accomplished by twisting it by -72° over each 67 nm segment of its length to produce repeating conformations that are translationally superposable on each other (Figure 9–34B).²⁹³ Because the regions with only four ropes are the most flexible, the twist is confined to them. The cable is kinked so that as it ascends at each step it shifts from one column of unit cells to the column diagonally adjacent to it. The pentagonal array of the five ropes is compressed in one dimension into a layer of two and a layer of three ropes to permit the pentamer of ropes to pack in a hexagonal array. The helix of each strand is left-handed, the superhelix of the triple helix of the three strands forming the rope is right-handed, and the twist of the cable of five ropes is left-handed. Alternating the hand of the elements in a cable increases its strength.²⁹³

As the cables enter these rather contorted arrays, they are aligned by the crystallization in register, and the crystallographic asymmetric units, each containing the equivalent of 67 nm of cable, create a **repeating pattern on the surface of the fibril** itself. This pattern can be seen in the electron microscope. It appears as alternating thickenings and thinnings along a desiccated fibril that repeat every 67 nm. These thickenings and thinnings are believed to represent the alternation in register of regions within the molecular cables five ropes in thickness and those of four ropes in thickness, respectively (Figure 9–34). Fibrils of collagen also stain positively by chelating heavy metals at specific locations on the surfaces of the ropes where there are high, unbalanced constellations of negative charge. Because the segments of rope are placed in register by the side to side crystallization of the cables, these positions to which heavy metals bind form bands across the fibrils or across sheets of collagen. The pattern of bands is quite reproducible,²⁹⁵ and it repeats every 67 nm.²⁹⁶ From an examination of the patterns in which charged amino acids occur in the amino acid sequences of the polypeptides, it can be shown that the pattern in which these bands occur on the ropes is entirely consistent²⁹⁶ with the triple-helical array of strand and pentahelical array of ropes in the hypothetical model (Figure 9–34B), the Fourier transform of which mimics the reflections in the X-ray diffraction pattern from an oriented tendon.²⁹³ All of these correlations provide independent support for this model of the structure of the cable.

Three classes of filaments can be observed within animal cells. **Thin filaments**, or microfilaments, are filaments of actin the basic structural element of which is the actin helix (Figure 9–1C). Tropomyosin, a fibrous protein that is one continuous coiled coil of two parallel α helices,²⁹⁷ lies in the grooves of the actin helix. Each of

these coiled coils of tropomyosin spans seven actins. The globular protein troponin decorates the microfilament at regular intervals. The structure of the thin filament has been elucidated by image reconstruction.^{298–300} In transmission electron micrographs, thin filaments appear to be about 8 nm wide. **Microtubules** are hollow cylinders,³⁰¹ constructed from the globular protein tubulin. They are about 20 nm in width. **Intermediate filaments** are the third class of filament. In transmission electron micrographs, they appear to be about 10 nm wide, intermediate between thin filaments and microtubules.

Intermediate filaments were originally considered to be a heterogeneous class of polymeric proteins grouped together only because they were similar in width. Within this class are tonofilaments, neurofilaments, cellular keratin filaments, desmin filaments, glial filaments, and vimentin filaments. Each of these subclasses occurs in a different set of tissues, and they form intermediate filaments that often seem quite different in their appearance and their distribution through a cell (Figure 9–35).³⁰² It is now known, however, that all of these filaments are constructed from polypeptides that are homologous in sequence (Figure 9–36)³⁰³ and thus necessarily share a common, superposable structure. That an intermediate filament can be constructed from one of these polypeptides all by itself has been demonstrated by reassembling filaments from a pure homogeneous preparation of a given polypeptide.³⁰⁴ These polypeptides form helical cables of indefinite length.

One of these intermediate filaments, keratin, composes the fibers in the composite material that forms skin, hair, and horn. For example, the quill of a porcupine represents a large array of more or less aligned keratin cables. Diffraction of X-radiation from such specimens³⁰⁵ provides dimensions of the helical arrays in these cables.³⁰⁶ Meridional reflections representing a repeat of 0.51 nm are strong features of such diffraction patterns, and this demonstrates that these cables are built from **coiled coils of α helices**.

The strand from which an intermediate filament is constructed is a polypeptide folded into an α helix. Two α helices twist around each other in a left-handed coiled coil (Figure 6–29) to produce the rope.^{307,308} The **heptad repeat** permitting the formation of this rope can be noticed in those regions of the sequences that are involved in its formation when sequences from several of these polypeptides are aligned (Figure 9–36).³⁰³ The amino acids in the heptad positions at which the side chains are directed into the core of the coiled coil are not always nonpolar, but at four of the positions where they are not, remarkable conservation is displayed.

As in collagen, the sequences producing the rope are found in the central portions of the polypeptides. In each of the amino acid sequences from this central region, there are three consecutive segments of heptad repeat, about 30, 100, and 130 aa in length,³⁰⁹ separated by short segments (about 20 aa in length) lacking the

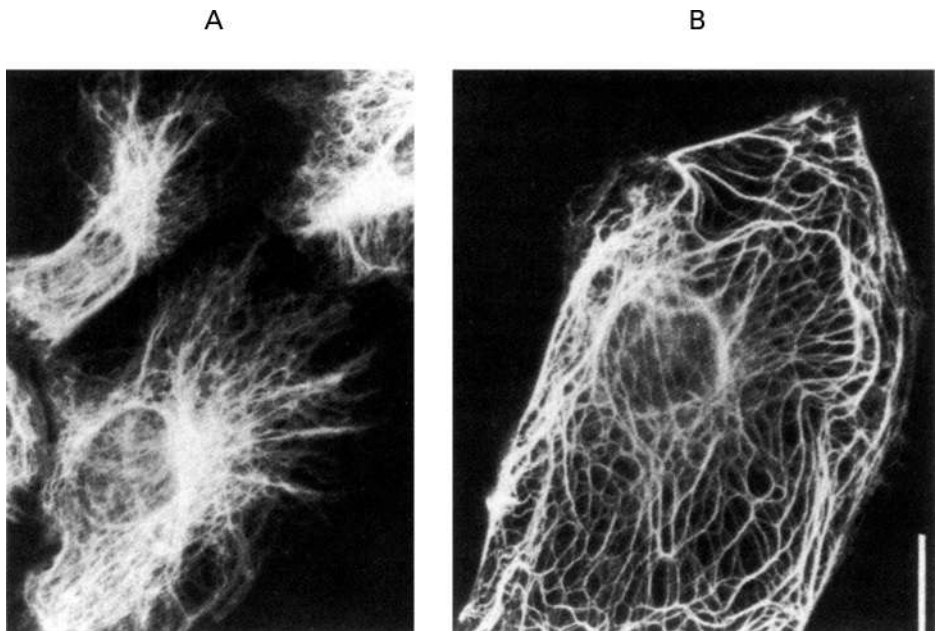


Figure 9-35: Distribution of intermediate filaments of vimentin (A) and keratin (B) in animal cells.³⁰² Hamster Nil-8 cells (A) or kangaroo rat PtK-2 cells (B) were grown on glass cover slips. The cells were fixed with methanol, rinsed with acetone, and dried in the air. Antiserum raised in guinea pigs to either vimentin (A) or epidermal prekeratin (B) was then applied to the respective cells. After the antiserum was rinsed away, those antibodies bound to the intermediate filaments could be visualized by adding fluorescein-labeled immunoglobulins specific for guinea pig immunoglobulin. The covalently attached fluorescein causes the intermediate filaments to which it is bound to fluoresce. Reprinted with permission from ref 302. Copyright 1978 National Academy of Sciences.

human keratin II 5	219	FEQY INNLRQ LDS IVGE RGR LDSE LRNMQDL VED FKNK YEDE EINK RTTAE
ovine keratin II 7c	158	FEGY IETLRRE AECVEAD SGR LSSE LNHVQEV LEGYKKK YEQEVAL RATAE
ovine keratin I 8c1	106	YFRT IEE LQOK ILC AKSE NAR LVVQ IDNAKLA ADD FRTK YET ELGL RQLVE
human keratin I 14	164	YFKT IEDLRNK ILT ATVD NANVLLQ IDNARLA ADD FRTK YET ELNL RMSVE
desmin from <i>G. gallus</i>	146	YEEE LRE LRRQ VDA LTGQ RAR VEVE RDN LLDN LQKLKQK LQE EIQL KQEA E
vimentin from <i>C. griseus</i>	148	YEEE MRE LRRQ VDQ LTND KAR VEVE RDN LAED I IRLREK LQE EMLQ REEA E
murine glial fibril	113	YQAE LRE LRLR LDQ LTAN SAR LEVE RDN FAQD LGT LRQK LQDETNL RLEAE
porcine neurofilament-M	150	YDQE IRE LRAT LEL VNHE KAQVQLD SDH LEED I HRLKER FEE EARL RDDTE
human neurofilament-H	146	YERE VRE MRGA VLR LGAA RGQLRLE QEH LLED IAH VRQR LDD EARQ REEA E
human keratin II 5	270	NE FVMLKKD VDA AYMN KVE LEAK VDA LMDE INF MKMF FDAE LSQM QTH VSD
ovine keratin II 7c	209	NE FVALKKD VDC AYVR KSD LEAN SEAL IQE IDFLRRL YQEE IRVLQAN ISD
ovine keratin I 8c1	157	SD INGLRRI LDE LTLC KSD LEAQ VES LKEE LIC LKSN HEEE VNT LRSQ LGD
human keratin I 14	215	AD INGLRRV LDE LTLA RAD LEMQ IES LKEE LAY LKKN HEEE MNA LRGQ VGG
desmin from <i>G. gallus</i>	197	NN LAA FRAD VDA ATLA RID LERR IES LQEE IAF LKKV HEEE IRE LQAQ LQE
vimentin from <i>C. griseus</i>	199	ST LQS FRQD VDN ASLA RLD LERK VES LQEE IAF LKKL HDEE IQE LQAQ IQE
murine glial fibril	164	NN LAA YRQE AHE ATLA RVD LERK VES LEEE IQFLRKI YEEE VRD LREQ LAQ
porcine neurofilament-M	201	AA I RALRKD IEE ASLV KVE LDKK VQS LQDE VAF LRSN HEEE VAD LLAQ IQA
human neurofilament-H	197	AAARA LARF AQE AEAARVD LQKK AQALQEE CGY LRRH HQEE VGE LLGQ IQG

Figure 9-36: Alignment of portions of the amino acid sequences of the polypeptides composing various intermediate filaments.³⁰³ The aligned segments come from different locations in the amino acid sequences of the various polypeptides. The numbers indicate the sequence positions in the various polypeptides at which the alignment on that line commences. The proteins are isoform 5 of human type II cytoskeletal keratin, isoform 7c of ovine type II microfibrillar keratin, isoform 8c1 of type I keratin from intermediate filaments of ovine wool, isoform 14 of human type I cytoskeletal keratin, desmin from *Gallus gallus*, vimentin from *Cricetulus griseus*, glial fibrillary acidic protein from murine astrocytes, triplet M protein from porcine neurofilaments, and triplet H protein from human neurofilaments. The pattern of the heptad repeat is highlighted in boldface type. The highlighted amino acids are those the side chains of which are directed into the cores of the coiled coils.

repeat. The approximately 300 aa strand⁻¹ should produce an interrupted rope 35–45 nm long. On either side of this rope are amino-terminal domains (100–200 aa) and carboxy-terminal domains (100–1600 aa) that must either be incorporated into the body of the cable or pro-

trude from its sides. These protrusions presumably cause each type of intermediate filament to have a different width and a different tissue-specific function. An intermediate filament is a cable formed from these ropes. The filament is a helical polymeric protein

that has a $(-1, 3)$ surface lattice. In the cables of keratin in a fiber of wool, the rise for each step of the left-handed single-stranded helix generating the structure is 6.7 nm,³¹⁰ and each step is -111° from the preceding one.³¹¹ The measured mass of protein in each nanometer of an intermediate filament indicates that its cross section contains about 30 strands of α -helical polypeptide³¹² and that each rope is a coiled coil of two parallel α helices. Adjacent ropes are antiparallel to each other and staggered, and in the resulting staggered pattern there are two different alignments, one in which the ropes are staggered by half their length and the other in which they are side by side, unstaggered.^{307,308} The precise arrangement of the ropes within the helical lattice of the cable, however, is as yet unknown.

In all of the cables that have been discussed, the faces and interfaces can be divided into three groups. There is a continuous interface between or among the strands, containing within itself the central axis of the rope and twisting around that axis with the twist of the rope. Between the ropes in the cable there are interfaces, but they are formed from faces on each rope composed of small regions of surface on each strand alternating between or among the strands as the rope is ascended (Figure 9–33). The cables themselves may have faces on them to promote side to side associations, and these faces are composed of small regions of surface on strands from the same or different ropes that are encountered in turn as the cable is ascended (Figure 9–34).

There is a class of polymers that form extracellularly during the progress of systemic amyloidosis, maturity-onset diabetes, Alzheimer's disease, and spongiform encephalopathy, which are diseases affecting mammals. The diffraction of X-radiation by these fibers shows reflections arising from a repeat of 0.48 nm along the axis of the fiber.^{313–315} This dimension is the spacing between the strands of a continuous β sheet (Figure 6–9) with individual β strands perpendicular to the axis of the fiber. The β sheets in these polymers can be either parallel β structure³¹⁵ or antiparallel β structure.³¹³ The length of each strand in one of the continuous β sheets forming these polymers varies from 2.5 nm (7 aa) to 3.5 nm (10 aa) depending on the protein forming the fiber. This length determines the width of the ribbon formed by each of the continuous β sheets. Several of these **ribbons of continuous β sheet** are packed against each other about 1 nm apart, a spacing determined by the interdigitations of the side chains (Figure 6–32) to create a fibril about 3–4 nm in width. The β sheets can twist as they usually would (Figure 6–9) to give a helical twist to the ribbons that are packed together,³¹⁵ and hence to the fibril, or they can be untwisted.³¹³

Suggested Reading

Kramer, R.Z., Bella, J., Mayville, P., Brodsky, B., & Berman, H.M. (1999) Sequence dependent conformational variations of collagen triple-helical structure, *Nat. Struct. Biol.* 6, 454–457.

Amos, L.A., & Klug, A. (1975) Three-Dimensional Image Reconstructions of the Contractile Tail of T4 Bacteriophage, *J. Mol. Biol.* 99, 51–73.

Heterologous Oligomeric Proteins

The oligomeric proteins described so far, with a few exceptions, have been homooligomers of identical subunits. There are many proteins in which the subunits are not identical to each other but are homologous, such as the α and β subunits of hemoglobin or the VP1, VP2, and VP3 subunits of the protein coat of poliovirus (Figure 9–28). In such cases, the **homologous subunits** are arrayed around rotational axes of pseudosymmetry that are the descendants of the rotational axes of symmetry of their homooligomeric ancestral proteins. For example, there are seven distinct β subunits, each with a unique sequence and each present in two copies, in the 20S multicatalytic endopeptidase complex from *S. cerevisiae*. Although they have distinct sequences, their tertiary structures are all homologous and the 14 β subunits in the oligomer are arranged at the center of the protein with dihedral pseudosymmetry of point group $722(D_7)$.³¹⁶ In the homologous complex from *Thermoplasma acidophilum*, representing the ancestor of the complex from *S. cerevisiae*, the 14 β subunits are all identical to each other, homologous to those from *S. cerevisiae*, and also arranged with the same dihedral symmetry.¹¹⁶ Often the homologous subunits in a heterooligomer are interchangeable with each other, as are those in fructose-bisphosphate aldolase (Figure 8–18) or those in the dimer of creatine kinase.³¹⁷ In all of these uncertain heterooligomers, there is no significant difference between the **pseudosymmetric arrangement** of their subunits and the symmetric arrangement of the identical subunits in their homooligomeric siblings or homooligomeric ancestors.

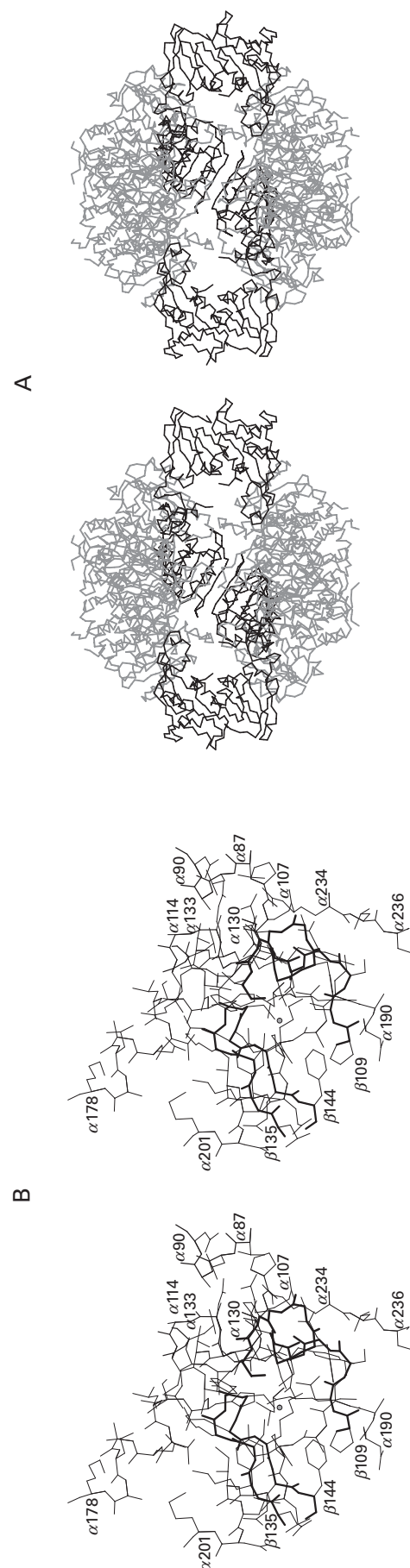
There are also **heterooligomeric proteins** formed from two or three distinct, unrelated subunits each present in equal numbers of copies and held together by heterologous interfaces. A **heterologous association** is the association between two folded polypeptides of unrelated sequence and unrelated tertiary structure. A **heterologous interface** is the interface between two unrelated subunits in heterologous association. The heterologous association between two unrelated subunits produces the protomer of aspartate carbamoyltransferase.

Aspartate carbamoyltransferase (Figure 9–37A)³¹⁸ is a hexamer with dihedral symmetry of point group $322(D_3)$ in which each protomer contains two different subunits, a catalytic α subunit and a regulatory β subunit. There are three types of interfaces holding the protein together, six interfaces among the six α subunits of the two trimers, three interfaces between each of the three pairs of β subunits, and six heterologous interfaces,

Figure 9-37: Heterologous interface in aspartate carbamoyltransferase.^{318,319} (A) α -Carbon skeletal drawing of the protomers of aspartate carbamoyltransferase arranged around the molecular axes of symmetry. The drawing was made from the crystallographic molecular model of this heterododecameric protein. The six folded α polypeptides, each with 310 aa, are drawn with gray line segments; the six folded β polypeptides, each with 152 aa, are drawn with black line segments. The view is down the 2-fold rotational axis of dihedral symmetry that passes through the center of the β_2 dimer in the rear. The two other 2-fold rotational axes of dihedral symmetry pass through the centers of the other two β_2 dimers, one to the left and one to the right. The 3-fold rotational axis of symmetry is vertical, in the center of the molecule, and in the plane of the page. The front α subunit of the upper α_3 trimer forms its heterologous interface with the upper subunit of the β_2 dimer to the right to produce one of the protomers of the overall molecule. The participants in each of the other heterologous interfaces, and hence each of the other protomers, can be related to this one by rotations around the axes of symmetry. (B) Detailed view of the heterologous interface between a β subunit and an α subunit. The participants are drawn in exactly the same orientation as those in the interface between the upper β subunit in the β_2 dimer on the left and the α subunit to the back left of the upper α_3 trimer in panel A. Two segments from the β subunit, from Proline $\beta 109$ to Valine $\beta 120$ and from Leucine $\beta 135$ to Phenylalanine $\beta 144$, are drawn with thick line segments; and six segments from the α subunit, Threonine $\alpha 87$ to Aspartate $\alpha 90$, Proline $\alpha 107$ to Leucine $\alpha 114$, Glycine $\alpha 130$ to Glutamine $\alpha 133$, Leucine $\alpha 163$ to Lysine $\alpha 178$, Aspartate $\alpha 190$ to Methionine $\alpha 201$, and Arginine $\alpha 234$ to Aspartate $\alpha 236$, are drawn with thinner line segments. Side chains in all of the segments are drawn with the thinnest line segments. A Zn^{2+} cation is represented by the gray sphere shown covalently bound to three of its four surrounding cysteines. It stabilizes the long, otherwise conformationally uncertain loops from the β subunit. These drawings were produced with MolScript.⁴⁸⁵

one between each α subunit and each β subunit within each of the six protomers. The trimers of α subunits have only limited contacts with each other.

In a molecule of aspartate carbamoyltransferase, each α subunit is associated with a β subunit at a paradigmatic heterologous interface (Figure 9-37B).³¹⁹ One face of the interface is formed from two long loops of random meander from the β subunit stitched in place by a Zn^{2+} ion (gray sphere in Figure 9-37B). These two loops are surrounded by a complementary face created from six segments of the folded α polypeptide: the amino-terminal end of an α helix ($\alpha 87$ to $\alpha 89$); three loops of random meander and β turn ($\alpha 108$ to $\alpha 114$, $\alpha 130$ to $\alpha 133$, and $\alpha 163$ to $\alpha 174$), each between the carboxy-terminal end of a β strand and the amino-terminal end of an α helix; and two loops of random meander ($\alpha 190$ to $\alpha 197$ and $\alpha 234$ to $\alpha 236$), each between two α helices. Consequently, this interface serves as an example of the fact that most intersubunit interfaces, both homologous and heterologous, are formed from loops of random meander and β turns on the surfaces of the respective subunits, while interfaces within a subunit are usually between α helices, β sheets, or an α helix and a β sheet



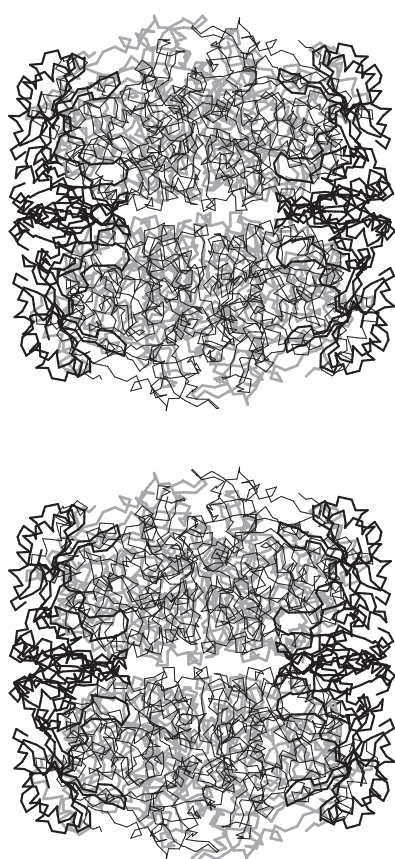


Figure 9-38: Heterologous associations in the $(\alpha_2)_4\beta_8$ heterohexamer of ribulose-bisphosphate carboxylase from *S. oleracea*.^{41,321} The drawing was made from the crystallographic molecular model of the protein. Four α_2 dimers of 2×473 aa (drawn with thin line segments in front and gray line segments behind) are held together at top and bottom by monomeric β subunits of 123 aa (drawn with thick line segments) to produce an oligomer with dihedral symmetry of point group $422(D_4)$. Each β subunit inserts two long loops of random meander, most readily observed to the right and to the left above and below, into the space between two α subunits. This drawing was produced with MolScript.⁴⁸⁵

(Figures 6-23 through 6-35) in its interior. Because it incorporates a lysine, a histidine, and four glutamates, this heterologous interface in aspartate carbamoyltransferase also reemphasizes the fact that interfaces between subunits incorporate twice as many charged side chains as do interfaces between secondary structures within a subunit.

The heterologous interface between an α subunit and a β subunit of aspartate carbamoyltransferase is formed from two **continuous faces**, one on the surface of each subunit, that fit together as a casting in a mold. The

heterologous interface between elongation factor Tu and elongation factor Ts from *E. coli*, however, incorporates 27 amino acids from the latter protein and 22 amino acids from the former, but they are situated in four **separate clusters**, one of which is formed by the carboxy-terminal α helix of elongation factor Ts that reaches over to sit in a groove on the surface of elongation factor Tu.³²⁰

In ribulose-bisphosphate carboxylase from *Spinacia oleracea* (Figure 9-38),^{41,321} heterologous interfaces between its constituent monomeric β subunits and α_2 dimers hold together the $(\alpha_2)_4\beta_8$ complex around the rotational axes of its dihedral symmetry of point group $422(D_4)$. Unlike the situations in aspartate carbamoyltransferase and the complex of elongation factors Tu and Ts, each β subunit connects two α subunits from two different α_2 dimers around the 4-fold rotational axis of symmetry. Consequently, each β subunit has **two distinct faces** on its surface, one that forms a heterologous interface with one α subunit from one dimer and one that forms a different heterologous interface with another α subunit from another dimer. The majority of each of these two distinct faces on each β subunit is formed by two long loops of random meander sandwiched between the two respective α subunits (Figure 9-38). The loop that is most detached from the β subunit and closest to the center of the heterooligomer has a conserved sequence and is required for proper assembly of the complete protein.³²² One side of each of these two loops associates with one of the α subunits, and the other side of each of these two loops, which necessarily is completely different, associates with the other of the α subunits. The four respective copies of the two different interfaces alternate with each other around the 4-fold rotational axis of symmetry. The protein from *Rhodospirillum rubrum*, which lacks the gene for the β subunit, is an α_2 dimer.³²³

Steric exclusion and mismatched symmetry are complications that can arise whenever an oligomer contains both heterologous associations between different, unrelated subunits and molecular rotational axes of symmetry relating subunits homooligomerically. **Steric exclusion** is the blocking of a face for heterologous association on one subunit of a homooligomer by the association of one of the heterologous subunits with the copy of that face on another subunit of the homooligomer. For example, the binding of a face on the extracellular domain of the immunoglobulin ϵ receptor to one of the two identical, symmetrically displayed, complementary faces on the homodimeric Fc domains of immunoglobulin E sterically blocks the other face on that molecule of immunoglobulin E from associating with another molecule of the receptor.³²⁴

Transthyretin is an $(\alpha_2)_2$ tetramer with dihedral symmetry. Its sole function is to bind retinol-binding protein. The four symmetrically arrayed faces for association with retinol-binding protein are in two adjacent pairs situated on opposite sides of transthyretin. The two

identical faces in each adjacent pair, however, are too close to the 2-fold rotational axis of symmetry around which their two respective subunits are arrayed for each to bind to a retinol-binding protein simultaneously. When a retinol-binding protein binds to one face of a pair on one side of transthyretin, the other face of that pair cannot bind another retinol-binding protein because there is not enough room for it.³²⁵ Because of this steric exclusion, only two of the four identical faces for associating with retinol-binding protein, one on each side of transthyretin, can be occupied at a time, and the complex is a closed $\beta(\alpha_2)_2\beta$ heterohexamer. During the evolution of this heterologous association, the face for binding retinol-binding protein just happened to arise on the surface of a subunit of transthyretin at this site. The realization of the resulting heterologous interface accomplished the function of the protein even though to an intelligent designer it has been accomplished inelegantly.

Mismatched symmetry occurs when the rotational axis of symmetry relating two or more identical faces on one oligomer is of a different fold from or does not coincide with the rotational axis of symmetry relating the two or more identical complementary faces on a different oligomer during the formation of the heterologous association between the two oligomers.

The 2-oxoglutarate dehydrogenase complex provides an example of mismatched symmetry. It is the multienzymatic complex responsible for the oxidative decarboxylation of 2-oxoglutarate. Three different proteins combine together in the complex to accomplish this oxidative decarboxylation: dihydrolipoyllysine-residue succinyltransferase, oxoglutarate dehydrogenase (succinyl-transferring), and dihydrolipoyl dehydrogenase. The octahedral core of the complex is formed from the 24 identical subunits of dihydrolipoyllysine-residue succinyltransferase arranged in octahedral symmetry. Oxoglutarate dehydrogenase (succinyl-transferring) is a symmetric homodimer, and dimers of this protein adorn the outer surface of the central octahedral core. They must each be attached to the core through heterologous interfaces formed between a face on the core and a face on the respective dimer. Because each of the dimers is built around a 2-fold rotational axis of symmetry, it necessarily has two identical faces, each complementary to a face on the core. Because the core is octahedral, it necessarily has 24 identical faces, each complementary to a face on a dimer of oxoglutarate dehydrogenase (succinyl-transferring). The two identical faces on each dimer are arrayed about its 2-fold rotational axis of symmetry; the 24 faces on the core are each arrayed about its octahedral rotational axes of symmetry.

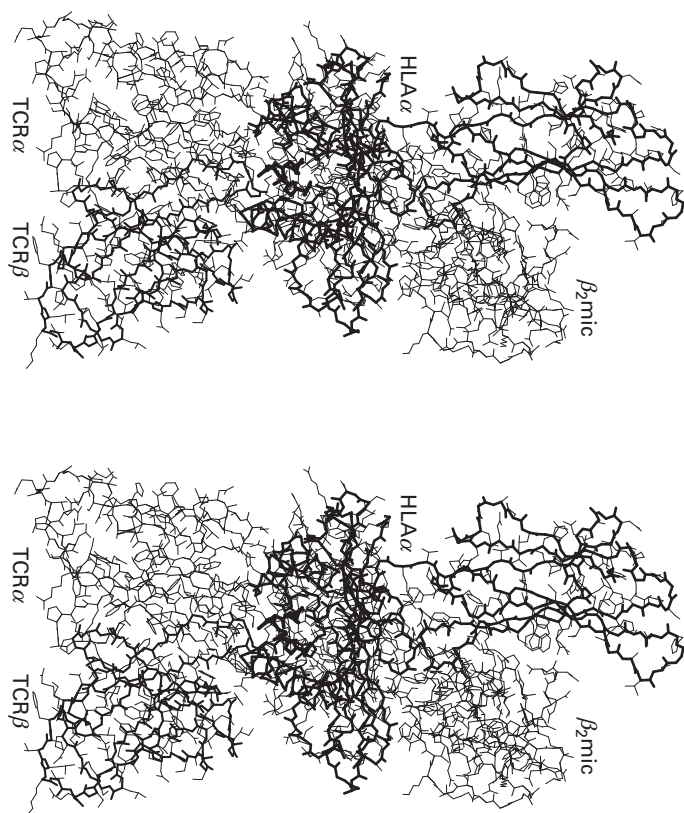
Careful examination of electron micrographs of complexes containing one tetracosamer of dihydrolipoyllysine-residue succinyltransferase but only two dimers of oxoglutarate dehydrogenase (succinyl-transferring)³²⁶ showed that a dimer of the dehydrogenase is

bound to a site on the succinyltransferase midway between one of its 2-fold rotational axes of symmetry and one of its 4-fold rotational axes of symmetry (Figure 9–21B). Therefore only one interface, formed from one face on the dimer and one face on the core, can attach each molecule of oxoglutarate dehydrogenase (succinyl-transferring) to the core because if the two identical faces on the dimer of oxoglutarate dehydrogenase (succinyl-transferring) were both occupied with complementary faces on the dihydrolipoyllysine-residue succinyltransferase, the one 2-fold rotational axis of symmetry of the dehydrogenase would have to coincide with one of the 2-fold rotational axes of symmetry of the succinyltransferase rather than being displaced from it. Only in this arrangement would the two identical faces on the oxoglutarate dehydrogenase (succinyl-transferring) be able to associate with two complementary faces on the dihydrolipoyllysine-residue succinyltransferase. The symmetry-related face on an attached dimer of the dehydrogenase finds itself empty because it is sterically inaccessible to a complementary face on another subunit of the succinyltransferase. This arrangement is a result of the fact that, as might be expected, the complementary faces on the heterologous partners, the octahedral core and the dimer, happened to evolve with no concern for their positions relative to the axes of symmetry. Consequently, their symmetries were mismatched.

In addition to this mismatch of symmetry, the complex between dihydrolipoyllysine-residue succinyltransferase and oxoglutarate dehydrogenase (succinyl-transferring) also displays steric exclusion. For each of the 24 equivalent faces on the octahedral tetracosamer of dihydrolipoyllysine-residue succinyltransferase that is occupied by a dimer of oxoglutarate dehydrogenase (succinyl-transferring), the three other identical faces arrayed around the respective 4-fold rotational axis of symmetry (Figure 9–21B) remain empty³²⁶ because the complementary face on the octahedral succinyltransferase to which a face on the dimer of the dehydrogenase attaches is too close to a 4-fold rotational axis of symmetry and the dimer is too large to permit more than one dimer to bind to the four identical faces around this axis. In the saturated complex between oxoglutarate dehydrogenase (succinyl-transferring) and dihydrolipoyllysine-residue succinyltransferase, there are six heterologous interfaces joining 12 polypeptides of the former and 24 polypeptides of the latter. These **nonstoichiometric ratios of subunits** are dictated by both mismatched symmetry and steric exclusion. Because there are only six heterologous interfaces connecting these two proteins together, only six of the 12 subunits of the dehydrogenase are directly attached to the succinyltransferase, and 18 of the faces on the succinyltransferase and six of the complementary faces on the dimers of the dehydrogenase remain unoccupied.

There are other examples within the family of oxoacid dehydrogenase complexes in which the stoichiome-

Figure 9-39: Heterologous associations among the α subunit of human HLA class I histocompatibility antigen A-2 (HCA α), β_2 microglobulin (β_{2mic}), the α subunit of the B7 isoform of human T-cell receptor (TCR α), and the β subunit of the B7 isoform of human T-cell receptor (TCR β). The skeletal drawing is from the crystallographic molecular model of this complex.^{332,333} The three extracellular amino-terminal domains of the α subunit of human HLA class I histocompatibility antigen A-2 (Glycine 1 to Alanine 90, Glycine 91 to Threonine 182, and Aspartate 183 to Glutamate 275) are drawn with thick line segments. The first two domains of this protein form the oblate disk in the center of the complex, and the carboxy-terminal domain of the three domains is the immunoglobulin modular domain (Table 7-7) at the top left of the drawing. An entire molecule (99 aa) of β_2 microglobulin, which is the β subunit of the histocompatibility antigen, is drawn with thin line segments. This single immunoglobulin modular domain is in the upper right of the drawing. The amino-terminal immunoglobulin modular domain (Leucine 1 to Arginine 111) of the α subunit of isoform B7 of the human T-cell receptor is drawn with thin line segments and is situated in the lower left of the drawing. The amino-terminal immunoglobulin modular domain (Glycine 3 to Leucine 114) of the β subunit of isoform B7 of the human T-cell receptor is drawn with line segments of intermediate thickness and is situated in the lower right of the drawing. At the center of the complex is the nonapeptide LLFEYVYV, drawn with the thickest line segments. This peptide is essential for the formation of the complex. This drawing was produced with MolScript.⁴⁸⁵



try between two different oligomers and the number of interfaces formed in the heterologous complex between the two are dictated by steric exclusion. In the pyruvate dehydrogenase complex from *Bacillus stearother-*

mophilus, only one of the two symmetrically displayed faces on dimeric dihydrolipoyl dehydrogenase is able to associate with the icosahedral dihydrolipoyllysine-residue acetyltransferase because these faces are located too close to the 2-fold rotational axis of symmetry on the dimer to accommodate two of the complementary faces on dihydrolipoyllysine-residue acetyltransferase simultaneously.³²⁷ In the pyruvate dehydrogenase complex from *S. cerevisiae*, only one molecule of E₃-binding protein can bind to each pentagonal face of the icosahedral dihydrolipoyllysine-residue acetyltransferase even though each pentagonal face must have five identical sites for associating with E₃-binding protein.³²⁸ That this low stoichiometry results from steric exclusion follows from the fact that when the pentagonal face is made less crowded by truncating the dihydrolipoyllysine-residue acetyltransferase, more molecules of E₃-binding protein can associate with it.

The stoichiometry of the heterologous association of one homooligomer with another can also be affected by **conformational changes** resulting from the association itself. Only one low-affinity immunoglobulin γ Fc region receptor can associate with the symmetrically dimeric Fc fragment of immunoglobulin G. Its association induces an asymmetric conformational change in the Fc fragment. This conformational change distorts the other face on the Fc fragment so that it cannot assume the conformation required to associate with a receptor even though it is not sterically blocked from doing so by the already bound molecule of the receptor.³²⁹

Most heterooligomeric proteins display no discernible symmetry, but often there are **vestiges of symmetry**, such as the rough 2-fold rotational axis of pseudosymmetry in the complex between growth hormone and its receptor^{330,331} and the local 2-fold rotational axis of pseudosymmetry in the complex between human HLA class I histocompatibility antigen A-2 and human T-cell receptor B7 (Figure 9-39).^{332,333} This latter complex contains four different proteins held together with several asymmetric and pseudosymmetric heterologous interfaces. There are the central heterologous interfaces between the first two domains of the histocompatibility antigen and the α and β subunits of the T-cell receptor, a heterologous interface between β_2 microglobulin and the α subunit of the histocompatibility antigen, and heterologous interfaces between the α and β subunits of the T-cell receptor. The α and β subunits of the T-cell receptor are homologous, and the interface between the two respective domains presented in the figure contains a 2-fold rotational axis of pseudosymmetry. The two amino-terminal domains of the histocompatibility antigen arose from an internal duplication, and they are related by a 2-fold rotational axis of pseudosymmetry. Both of these 2-fold rotational axes of pseudosymmetry coincide so the lower half of the complex in Figure 9-39 has two halves, right and left, that are related by an overall 2-fold rotational axis of pseudosymmetry.

Many heterooligomers are complexes between two proteins held together by **transitory heterologous associations** rather than permanent heterologous associations. These transitory complexes dissociate and associate during the lifetimes of their constituent proteins as required by their function. For example, when 3',5'-cyclic AMP is bound by the regulatory β subunits of cyclic AMP-dependent protein kinase, the catalytic α subunits dissociate from them.^{334,335} The complex between the HLA class I histocompatibility antigen A-2 and the B7 isoform of the T-cell receptor forms only after the histocompatibility antigen has bound a short peptide of a sequence recognized by the T-cell receptor (Figure 9-39).

Many asymmetrical heterooligomers, both those that are permanent and those that are transitory, contain folded polypeptides composed of multiple copies of **modular domains** (Tables 7-7 and 7-8). For example, both the nidogen and the laminin in the permanent equimolar complex between these two proteins³³⁶ contain multiple copies of EGF modular domains among others. Both the HLA class I histocompatibility antigen A-2 and the T-cell receptor are composed of internally repeating domains, most of which are immunoglobulin modular domains (Figure 9-39).

One of the main functions of modular domains is to participate in heterologous associations.³³⁷ For example, the carboxy-terminal EF hand of α actinin forms a complex with the seventh Z-repeat of titin in which the former (see Figure 7-17) wraps around the single α helix that comprises the latter.³³⁸ The WW modular domain of dystrophin forms a complex with the proline-rich motif on β -dystroglycan in which a segment of the latter 5 aa long lies in extended conformation within a complementary groove on the former.³³⁹ The fifth and sixth EF hands of thrombomodulin form a complex with the globular protein α -thrombin through an interface that is formed from two complementary faces, one including surfaces from both of the EF hands and the other a flat continuous surface on the α -thrombin.³⁴⁰

It is their lack of exact symmetry, their transitory nature, and their modularity that distinguishes asymmetric heterooligomers held together entirely by heterologous associations from homooligomers.

The **interfaces producing heterologous associations**, either permanent ones between two unrelated subunits or transitory ones between two unrelated proteins, are as variable in their structure as the interfaces between two identical subunits. They can be almost planar interfaces between two flat faces, as is the interface between T-cell receptor and HLA class I histocompatibility antigen (Figure 9-39),³³³ they can involve terminal segments from one subunit that embrace the body of the other subunit, as in the interface between the α subunit and β subunit in nitrile hydratase from *Rhodococcus*,³⁴¹ or they can be a coiled coil of α helices, one from each of the participants, as in the complex

between syntaxin-1A, synaptobrevin-II, and SNAP-25B.³⁴² In the interface between protein G from *Streptococcus* and murine immunoglobulin G, three β strands from the immunoglobulin and four β strands from the protein G (each six amino acids long) form a continuous, paradigmatic antiparallel β sheet.³⁴³ In the heterologous association between the interleukin-1 receptor and interleukin-1 β , the three consecutive immunoglobulin modular domains of the receptor wrap around the interleukin, surrounding it on three sides.³⁴⁴ The leucine-rich repeats of ribonuclease inhibitor wrap around ribonuclease, also surrounding it on three sides,³⁴⁵ as do 13 of the 18 HEAT repeats of karyopherin β 2 that wrap around GTP-binding nuclear protein RAN.³⁴⁶

An examination³⁴⁷ has been made of the composition of heterologous interfaces that form transitorily between two proteins during the performance of their normal function. The fraction of the accessible surface area of such a transitory interface formed by nonpolar atoms (0.56) is somewhat lower than that in the interfaces holding together permanent oligomers (0.65) and is indistinguishable from that on the solvent-accessible surface of a small globular protein (0.57). The fractions formed by polar but uncharged atoms (0.24) and polar and charged atoms (0.19) are consequently somewhat higher than those in the interfaces of permanent oligomers (0.22 and 0.13, respectively). The same elevation in the fraction of arginine that is observed in the interfaces of permanent oligomers is also observed in the interfaces of transitory oligomers (0.10 of the amino acids in the interfaces) and the same decreases in aspartate and glutamate, but the transient interfaces also show a significant elevation in tyrosine (0.09), a hydrophilic hydrophobe, and significant decreases in valine (0.04) and leucine (0.05), both hydrophobic hydrophobes, relative to the composition of permanent interfaces (0.05, 0.07, and 0.11, respectively). An extreme example of the **elevation in the fraction of charged side chains** in transitory interfaces is the interface between human HLA class I histocompatibility antigen Cw3 and killer cell immunoglobulin-like receptor 2DL2 in which there are five arginines, three aspartates, three glutamates, and two lysines.³⁴⁸

The elevation in the polarity of transitory interfaces is in keeping with the fact that these interfaces are not permanent features; and consequently, the complementary faces on the two proteins must be exposed to the aqueous phase through much of their lives. Within the transitory, heterologous interface between α -thrombin and thrombomodulin, however, most of the side chains are hydrocarbon. In this instance, the problem of the solubility of the separated proteins is solved by surrounding the hydrophobic patch forming the face on α -thrombin with a high density of positively charged side chains and the hydrophobic patch on thrombomodulin with a high density of negatively charged side chains.³⁴⁰ Only five of

these charged side chains actually participate in the interface itself.

A common strategy for heterologous association is the specific **binding of a disordered, structureless segment of polypeptide** within one protein by a structured binding site on another. Upon the binding of the structureless partner to the structured partner, the structureless segment of polypeptide assumes a structure complementary to the structured site.³⁴⁹ In this type of heterologous association, it is only the amino acid sequence of the disordered segment and not its conformation that is recognized by the structured binding site. Consequently, the structureless partner can be mimicked by a synthetic peptide of the proper sequence. For example, the association between troponin I and troponin C can be **blocked by a synthetic peptide** with a sequence only 12 aa in length from the center of troponin I.³⁵⁰ In such a situation, each of the members of a set of overlapping synthetic peptides comprising the complete sequence of the protein providing the structureless partner for the interface can be assayed for its ability to inhibit the heterologous association of the intact protein in order to identify the segment that participates in the association.³⁵¹

Importin- β associates with importin- α by binding the structureless, highly charged (50% arginine, lysine, aspartate, and glutamate) amino-terminal segment (40 aa) of importin- α . The amino-terminal 876 aa of importin- β form 19 internally repeating, modular HEAT domains (46 aa), each consisting of a hairpin of two α helices. In the crystallographic molecular model, the 38 α helices of these 19 internally repeating, modular domains wrap in a spiral around a synthetic peptide of 44 aa with the amino acid sequence of the amino-terminal segment of importin- α . The formation of the complex induces the carboxy-terminal 28 amino acids of the otherwise structureless synthetic peptide to form an α helix aligned with the 38 α helices from importin- β ,³⁵² in a structure resembling a thick section from the trunk of a palm. In the natural heterologous association between importin- α and importin- β , the amino-terminal segment of importin- α , represented by the synthetic peptide in the crystallographic molecular model, is inserted into the middle of the spiral formed by importin- β to form a strong complex between the two proteins.

There are a large number of examples of such heterologous associations mediated by structureless sequences of amino acids. The structureless segment can be located anywhere in the amino acid sequence of a protein. Annexin II associates with protein p11 by binding its amino-terminal 12 amino acids.³⁵³ Tumor necrosis factor receptor-associated factor 2 is a globular trimer, each subunit of which has a binding site for a sequence of six amino acids in the structureless carboxy-terminal portion of the CD40 tumor necrosis factor receptor.³⁵⁴ PDZ Domains can bind to a structureless sequence located either at the carboxy terminus of

another protein or in a portion of polypeptide in the interior of its sequence of amino acids that loops out from its surface.³⁵⁵ Nuclear localization signals are short structureless sequences (5–20 aa) containing several lysines and arginines³⁵⁶ on the surfaces of proteins destined for the nucleus of a cell. Such structureless segments are recognized by being bound to a structured domain of the nuclear import factor karyopherin α , composed of 10 internally repeated armadillo modular domains.³⁵⁷ The Eps15 homology domains are modular domains that bind short segments of polypeptide containing the sequence -Asn-Pro-Phe-³⁵⁸ and in this way produce heterologous associations between a protein containing an Eps15 homology domain and a protein containing a structureless segment of this sequence within its folded polypeptide.

Historically, homooligomers with globular subunits, all arranged around rotational axes of symmetry, accounted for most of the proteins initially purified and studied. There are several reasons for this fact. Most of the proteins present at high concentrations in the cytoplasm and most of the proteins that have enzymatic activity, and hence for which there are obvious assays, are globular homooligomers. Globular homooligomers are also compact, sturdy, and resistant to degradation by endopeptidases. Consequently, globular, homooligomeric enzymes were the easiest proteins to purify.

Proteins the sole function of which is to participate in heterologous associations are more difficult to purify. Such proteins are often assembled by those associations into **large, heterogeneous polymeric matrices**, and until recently, identifying the partners involved in a particular heterologous association has been difficult. Proteins the enzymatic activities of which are **regulated by transient heterologous associations** with other proteins or the substrates of which are other proteins are difficult to assay because several components must be mixed together in the proper ratio. Proteins that **control the enzymatic activities** of the classical enzymes are present in much lower concentrations than those enzymes. Nevertheless, it is proteins engaging in heterologous associations that form large macromolecular structures within the cell and between cells, that control the metabolism carried out by the classical enzymes, and that regulate the expression of genes. Recently, proteins involved in such functions have been purified, identified, and expressed in amounts high enough to be studied functionally and structurally.

These **new proteins** (Table 9–4) are new only because they are present normally in low concentrations, are difficult to assay, or for some other reason are difficult to purify. In addition to their novelty, however, they all seem to share the property of participating in heterologous associations, either among their unrelated subunits or with other proteins. Most of these new proteins are peculiar to eukaryotic cells and the tissues of multicellular organisms. Now that complete genomes are

available for a number of eukaryotes, it has become clear that most of the proteins encoded by the genes in those genomes are new proteins. The examples listed in Table 9–4 illustrate various properties of these new proteins.

One of the most unfortunate features of these new proteins is the **chaos of their nomenclature**. Often the same protein from two different species of organisms will have completely unrelated names. Often the proteins are designated by a number that either derives from the initial genetic screen or from the initial, invariably inaccurate estimate of the length of their constituent polypeptide by electrophoresis on gels cast in solutions of dodecyl sulfate. Often the heterologous subunits of one of these proteins will each have its own peculiar name and the complex between them has another, unrelated name. One has the suspicion that such confusion is intended to discourage individuals outside the narrow field of investigators interested in one or the other of these proteins from learning about them or even realizing that they are normal, unremarkable proteins.

Most of the new proteins contain **internally repeating domains** or widely distributed modular domains⁴⁶⁴ or both of these types of domains. The sole function of many of the modular domains, such as the SH2 domain of proto-oncogene tyrosine-protein kinase ABL1, is to form heterologous associations with other proteins.

Many of these proteins, such as laminin $\alpha 1\beta 1\gamma 1$, integrin $\alpha 2\beta 1$, and guanine nucleotide binding protein G(s), are **heterooligomers**. The $\beta 1$ and $\gamma 1$ subunits of laminin are homologous, but the $\alpha 1$ subunit is unrelated to the others. The subunits of the integrin are unrelated to each other, as are those of the guanine nucleotide binding protein. The heterologous associations between the subunits of integrin $\alpha 2\beta 1$ are permanent, as are those between the subunits of laminin, but the heterologous associations between the α , β , and γ subunits of guanine nucleotide binding protein G(s) are transitory, and they dissociate and reassociate during its normal operation.

Almost all of these new proteins are participants in extensive, intricate **networks of heterologous associations** among many proteins.^{465,466} Each of these networks is responsible for a global function such as the production of the extracellular matrix or controlling the growth and multiplication of the cell. Proteins such as SHC transforming protein 1 form heterologous associations with many different partners and act as hubs in these networks. Proteins such as gelsolin associate with only one or two proteins at the dead ends in a network, while proteins such as α actinin 1 and guanine nucleotide binding protein G(s) link one protein to the next protein within the spokes radiating from the hubs. Some of these proteins can connect one network to another network. Integrin $\alpha 2\beta 1$ connects the network of heterologous associations forming the extracellular matrix to the networks for the cytoskeleton and for cellular regulation through protein kinases. Nucleoporin Nup214 connects the network of heterologous associations forming the

nuclear pore to networks for nuclear import and for cellular regulation.

Many of these proteins are responsible for shifts in the steady state of the cell such as changes in metabolism or the initiation of growth. Consequently, they must detect changes and respond to change by altering the heterologous associations in which they participate. Proteins such as SHC transforming protein 1 and proto-oncogene tyrosine-protein kinase ABL1 form heterologous associations with some proteins only when particular tyrosines on the surfaces of those proteins have been phosphorylated. In this way, they recognize changes produced by intracellular signalling. The heterologous associations in which cyclin participates are permanent, but their status is transitory, changing systematically as the cyclin is rapidly degraded and then more is synthesized. Protein kinases form specific, transitory heterologous associations with their substrate proteins in order to phosphorylate specific serines, threonines, or tyrosines on their surfaces.

Some of the heterologous associations, such as those between α actinin and actin or between laminin and nidogen, are exclusive. Others, such as those between ankyrin and various proteins embedded in the plasma membrane serving as sites of attachment for the cytoskeleton or those between the SH2 modular domains on proto-oncogene tyrosine-protein kinase ABL1 or SHC transforming protein 1 and an array of proteins containing phosphorylated tyrosines signalling changes in the regulatory status of the cell, are **promiscuous**. Transcription initiation factor TFIID recognizes TATA boxes that precede many different genes and then initiates the assembly of the large complex of different proteins responsible for initiating transcription.

Proteins involved in these networks of interactions responsible for particular functions are often identified and assigned a role on the basis of the heterologous associations themselves. There are several ways to detect a heterologous association between two proteins. A complex between two native proteins can be detected by their coelectrophoresis.⁴⁵⁹ A complex between two proteins can be immunoprecipitated with immunoglobulins specific for one of the two, and the fact that the other coprecipitates demonstrates the existence of the complex. Glutathione transferase can be fused to a protein during its expression, and any proteins that participate in heterologous associations with that protein can be identified after isolation of the complex by affinity adsorption with a solid phase to which glutathione has been attached.⁴⁶⁷

It is also possible to screen a library by **phage display**⁴⁶⁸ for a cDNA or a gene encoding a protein that participates in a heterologous interaction with a protein of interest. Fragments of cDNA or genomic DNA are inserted at a particular position in the gene encoding the coat protein pIII of the f1 or M13 bacteriophage. A population of *E. coli* is then infected with these bacteriophage.

Table 9–4: Examples of New Proteins

protein (length of human version)	stoichiometry of subunits	modular domains and internal repeats ^a	function	heterologous associations
E-cadherin ³⁵⁹ (728 aa)	α_2	cadherin modular (5) serine-rich modular (1)	cellular adhesion	integrin $\alpha E\beta 7$, ³⁶⁰ β -catenin, ³⁶¹ other molecules of E-cadherin ^{362,363}
integrin $\alpha 2\beta 1$ ³⁶⁴ (1152 aa, 778 aa)	$\alpha\beta$	von Willebrand factor type A modular (2), cysteine-rich repeat (4), FG-GAP repeat (7)	attaches cell to extracellular matrix	collagen, ³⁶⁵ laminin, ³⁶⁶ chondroadherin, ³⁶⁷ interstitial collagenase, ³⁶⁰ filamin, ³⁶⁸ α actinin, ³⁶⁹ skelemin, ³⁷⁰ integrin cytoplasmic associated protein I, ³⁷¹ receptor 1 for activated protein kinase C, ³⁷² paxillin, ³⁷³ focal adhesion kinase, ³⁷³ integrin-linked kinase, ³⁷⁴ calnexin ³⁷⁵
laminin $\alpha 1 \beta 1 \gamma 1$ ^{376,377} (3058 aa, 1765 aa, 1576 aa)	$\alpha\beta\gamma$	laminin G modular (5), EGF modular (41), laminin modular (2), laminin amino-terminal modular (3)	extracellular matrix	nidogen, ³⁷⁸ integrin $\alpha 2\beta 1$, ³⁶⁶ thrombospondin ³⁷⁹
vitronectin ^{380–382} (459 aa)	α	hemopexin modular (2)	protein in serum and extracellular matrix	integrin $\alpha V\beta 1$, ^{383–385} proteoglycan, ³⁸⁶ plasminogen activator inhibitor type 1 ³⁸⁷
ankyrin 1 ^{388,389} (1880 aa)	α	ankyrin repeat (23), death modular (1)	cytoskeleton	anion exchanger, ³⁹⁰ spectrin, ³⁹¹ Na ⁺ , K ⁺ -exchanging ATPase, ³⁹² Na ⁺ channel, ³⁹³ neuroglian, ³⁹⁴ CD44 antigen ³⁹⁵
α actinin 1 ³⁹⁶ (892 aa)	α_2	calponin modular (2), spectrin modular (4), calcium-binding EF-hand modular (2)	cytoskeleton	actin, ³⁹⁶ vinculin, ^{397–399} titin ^{338,400,401}
gelsolin ^{402,403} (755 aa)	α	gelsolin repeat (6)	sculpts actin	actin, ⁴⁰² caspase-3 ⁴⁰⁴
synaptotagmin 1 ^{405,406} (422 aa)	α	C2 modular (2)	controls traffic of synaptic vesicles	neurexin, ⁴⁰⁷ syntaxin, ⁴⁰⁸ clathrin assembly protein 2 ⁴⁰⁹
nucleolin ^{410,411} (706 aa)		Asp/Glu-rich repeat (3), RNA-binding modular (4), nucleolin repeat (8)	nucleolar component ⁴¹²	chromatin, ⁴¹³ preribosomal particles, ⁴¹⁴ insulin receptor substrate I, ⁴¹⁵ nucleophosmin ⁴¹⁶

nucleoporin Nup214 ⁴¹⁷ (2090 aa)		Pro/Ser/Thr-rich region (1), coiled coil (2), nucleoporin FG repeats (40)	component of nuclear pore	other nucleoporins, mitogen-activated protein kinase, ⁴¹⁸ nuclear RNA export factor 1, ⁴¹⁹ CRM1 protein, ⁴²⁰ CREB-binding protein ⁴²¹
HLA type I histocompatibility antigen A-2 ^{422,423} (341 aa, 99 aa)	$\alpha\beta$	immunoglobulin modular (4)	mediates immune response	T-cell receptor ³³³
cyclin A2 ⁴²⁴ (432 aa)			control of cell cycle	cyclin-dependent kinase 2, ⁴²⁵ cell division control protein 2 homologue, ⁴²⁶ protein CDC20, ⁴²⁷ β 3 endonexin ⁴²⁸
SHC transforming protein 1 ⁴²⁹ (583 aa)		phosphotyrosine interaction domain (1), SH2 modular (1)	intracellular signalling	activated receptors for growth factors, ^{429,430} proteins phosphorylated at tyrosine, ^{431–433} Grb2 protein, ⁴³⁴ mPAL protein, ⁴³⁵ phosphotyrosine phosphatase-PEST, ⁴³⁶ Gads protein ⁴³⁷
proto-oncogene protein-tyrosine kinase ABL1 ^{438–441} (1130 aa)		eukaryotic protein kinase modular (1), SH3 modular (1), SH2 modular (1), Pro-rich domain (1)	intracellular signalling, protein tyrosine kinase	protein substrates, proteins phosphorylated at tyrosine, EphB2 protein tyrosine kinase, ⁴⁴² Wiskott–Aldrich syndrome protein family member 1, ⁴⁴³ actin, ⁴⁴⁴ Abl interactor protein 2b, ⁴⁴⁵ proteins with proline-rich segments ^{446,447}
myosin light chain kinase ⁴⁴⁸ (1914 aa)		protein kinase modular (1), flbronectin type III modular (1), immunoglobulin C2 modular (1), myosin light chain kinase repeats, type I (5) and type II (6)	intracellular signalling, protein serine/threonine kinase	calmodulin, ⁴⁴⁹ myosin regulatory light chain 2
CD45 protein-tyrosine phosphatase ⁴⁵⁰ (1281 aa)		flbronectin type III modular (2), tyrosine-protein-phosphatase modular (2)	intracellular signalling	proteins phosphorylated at tyrosine, semaphorin 4D, ⁴⁵¹ protein CD2, ⁴⁵² protein p56lck ⁴⁵³
guanine nucleotide-binding protein G(s) ^{454,455} (394 aa, 340 aa, 75 aa)	$\alpha\beta\gamma$	G-protein β WD-40 repeat (7)	intracellular signalling	adenylate cyclase, ⁴⁵⁴ β -adrenergic receptor ⁴⁵⁶
transcription initiation factor TFIID ⁴⁵⁷ (339 aa)		polyglutamine domain (1), transcription factor TFIID repeat (2)		TATA box on DNA, ⁴⁵⁷ transcription initiation factor TFIIA, ^{457,458} transcription initiation factor TFIIB, ^{459,460} negative cofactor 1, ⁴⁶¹ negative cofactor 2, ⁴⁶² general transcription factor II-I ⁴⁶³

^aAs listed in the SwissProt data base (www.expasy.ch). Numbers in parentheses are the numbers of each type of domain in the protein.

Each of the resulting bacteriophage that carries an insert has pIII coat proteins on its surface, all of which display the segment of amino acids encoded by that insert. Because the point of insertion was chosen to be within an exposed loop in the coat protein, the segment of amino acids displayed is accessible to the solution and capable of interacting with the protein of interest. The protein of interest is covalently attached to a solid phase to produce an affinity adsorbent, and bacteriophage that carry segments of amino acids with which it associates can be purified by successive rounds of affinity adsorption.⁴⁶⁹ The bacteriophage purified by the affinity adsorbent can be plated, plaques can then be individually replicated, and the particular inserts that they carry can be sequenced. The method, however, that has been used most widely to catalogue heterologous associations between large sets of proteins has been the **yeast two-hybrid assay** (Figure 9–40).^{470–472}

At various points on the chromosomal DNA of *S. cerevisiae*, in the vicinities of the genes encoding enzymes involved in the metabolism of galactose, there are sequences of 17 base pairs that are recognized by the regulatory protein GAL4, which activates transcription of these genes when the cell is grown in the presence of galactose or an oligosaccharide containing galactose.⁴⁷¹ These upstream activating sequences for galactose metabolism can be anywhere from 100 to 400 base pairs away from the sites at which transcription of the gene is initiated, and yet binding of regulatory protein GAL4 still activates transcription.

Protein GAL4 contains a domain that binds to the upstream activating sequence in the DNA and a domain that acts at the site of initiation.⁴⁷³ These two domains are connected flexibly so that either the short or the long distances between the binding site on the DNA and the site of initiation can be spanned. The domain of regulatory protein GAL4 that binds to the upstream activating sequence in the DNA is fused to protein X or a portion of protein X. If the domain of regulatory protein GAL4 that activates transcription is fused to protein Y or a portion of protein Y that normally forms a heterologous association with protein X, the heterologous association between protein X and protein Y or the portions of these two proteins will usually be sufficient to position the activating domain effectively when the DNA-binding domain associates with an activating sequence on the DNA. The activation is effective because the requirements of the complex are so flexible that the only property required is a physical connection between the two domains of regulatory protein GAL4. Even a complex in which two other proteins form a noncovalent bridge that then connects protein X and protein Y is sufficient for activation.⁴⁷⁴

When β -galactosidase from *E. coli* is inserted into the DNA of *S. cerevisiae* under the control of a site for the initiation of transcription normally activated by regulatory protein GAL4, the normal version of regulatory

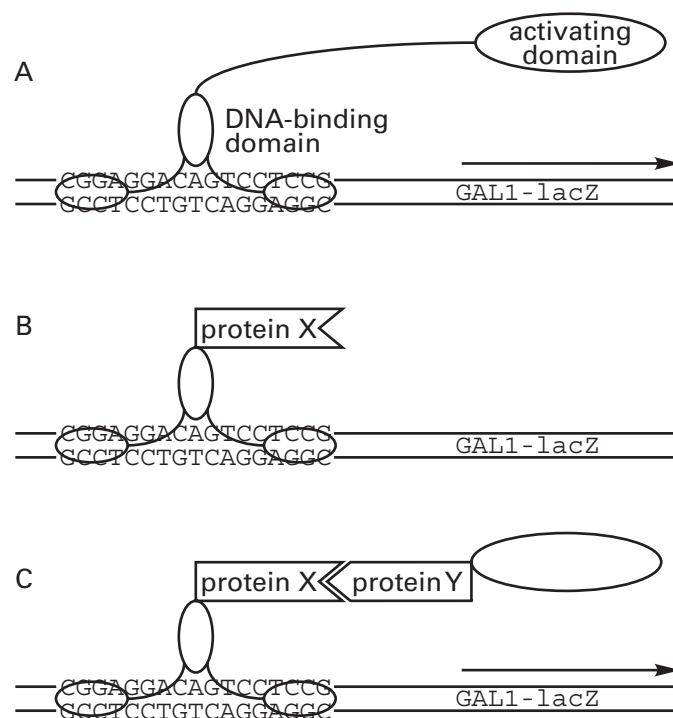


Figure 9–40: Yeast two-hybrid system for detecting heterologous associations.⁴⁷⁰ (A) Regulatory protein GAL4 from *S. cerevisiae* binds to a palindromic sequence of 17 base pairs (the consensus sequence for which is shown)⁴⁷¹ through its DNA-binding domain⁴⁷² and then activates the transcription of a gene, the initiation site for which is from 100 to 400 base pairs away, through the action of its activating domain, which is thought to be flexibly tethered to the DNA-binding domain. When a portion of the GAL1 gene, which is normally activated by regulatory protein GAL4, is fused to the *lacZ* gene from *E. coli*, which encodes β -galactosidase, the activation of the GAL1 gene by regulatory protein GAL4 causes high levels of β -galactosidase to be produced by the yeast cell (as indicated by the arrow). (B) Protein X, the partners of which in normally occurring heterologous associations are being sought, is fused to the DNA-binding domain of regulatory protein GAL4. When this fusion protein is expressed alone in a cell that is lacking regulatory protein GAL4, the GAL1 gene is not activated and β -galactosidase does not accumulate. (C) Protein Y, which participates in a normal heterologous association with protein X, is fused to the activating domain of regulatory protein GAL4. When it is coexpressed with the fusion protein between the DNA-binding domain and protein X, protein X and protein Y associate with each other, the activating domain of regulatory protein GAL4 is reassociated with its DNA-binding domain, the GAL1 gene is activated, and the cell fills with β -galactosidase.

protein GAL4 is replaced by the two pieces linked to protein X and protein Y, respectively, and the cell is grown on galactose, β -galactosidase, a protein for which there is an assay producing a bright blue color, will accumulate. Any colonies of cells containing two hybrid GAL4 domains, the respective proteins X and Y of which associate with each other, become blue when the assay is performed, but any colonies containing hybrid GAL4 domains the proteins X and Y of which do not associate do not turn blue.

The yeast two-hybrid assay is performed by choos-

ing one protein or a portion of one protein and fusing its complementary DNA in phase with the DNA encoding the domain of regulatory protein GAL4 that binds to the DNA. This acts as the bait. It is then possible to fuse DNA encoding the domain responsible for activation to fragments of genomic DNA at random. If a protein that associates heterologously with the bait happens to be encoded by the DNA in one of these fragments, it will be caught, the colony containing that fragment will turn blue, and the DNA in the fragment can be sequenced to identify the protein it encodes.⁴⁷⁵

This assay has been automated and applied to discover large numbers of heterologous associations. For example, in one of these **large screenings**, 957 heterologous associations involving 1004 different proteins were identified.⁴⁶⁶ Usually, however, the fishing is for the partners of a particular protein of interest to the investigator. For example, when cDNA for human cyclin A (Table 9–4) was used as the bait and fragments of the yeast genome as the fish, three positive colonies were identified.⁴²⁸ One of the proteins identified in this way was cyclin-dependent kinase inhibitor 1, a protein already known to form complexes containing cyclin A, but the other two, protein CDC20 and $\beta 3$ endonexin, represented novel associations. Once candidates for heterologous association have been identified with the yeast two-hybrid system, the validity of the associations must be established in more extensive studies of the two isolated proteins, as was done for the associations between cyclin A and $\beta 3$ endonexin⁴²⁸ and between cyclin A and protein CDC20.⁴²⁷

To establish the strength of the heterologous association between two proteins, its **dissociation constant** can be measured. As with any measurement of a dissociation constant, the molar concentration of the complex is followed as a function of the molar concentrations of the two unassociated proteins (Problem 5–7). The complex between the proto-oncogene protein c-fos and transcription factor AP-1 could be identified by the quenching of a fluorescent functional group on proto-oncogene protein c-fos by a fluorescent functional group on transcription factor AP-1, and a dissociation constant of 20 nM could be calculated from the changes in fluorescence as a function of the concentrations of the two proteins.⁴⁷⁶

The heterologous interfaces between two proteins are often probed by **cross-linking**^{477,478} or by **site-directed mutation**. To make sense of either of these types of experiments, a crystallographic molecular model of at least one of the participants must be available. In the case of site-directed mutation, changes are made at particular sites on the surface of the model, and the effects of these mutations on the strength of the association between the two proteins are assessed.⁴⁷⁹ It has also been possible to identify neighbors across the interface by discovering mutations in one of the proteins that compensate for mutations in the other protein that weaken the association by restrengthening it.⁴⁸⁰

Such experiments often identify a cluster of amino acids on the surface of the crystallographic molecular model of the protein, and this cluster is then assumed to represent the face participating in the interface holding the two proteins together. The heterologous interface between human somatotropin and human somatotropin receptor was probed by site-directed mutation of the somatotropin.^{481,482} Sixty-six side chains, located consecutively in three segments of the overall amino acid sequence of somatotropin, were mutated one by one to alanine, and the effect of each of these mutations on the association between somatotropin and its receptor was quantified by measuring the dissociation constant of the complex. Fourteen of these mutations produced what were judged to be significant increases in the dissociation constant, and those 14 side chains were found to form a cluster on the surface of the crystallographic molecular model of somatotropin, which was available at that time. In the crystallographic molecular model of the complex between human somatotropin and its receptor that became available subsequently, seven of those 14 side chains were found to be located within one of the interfaces.³³⁰

The difficulty in evaluating such sets of site-directed mutations is that often the change in the strength of the interaction produced by each individual mutation is not large,^{479,483} so a distinction between a mutation within the interface and one without the interface is difficult to make. When an amino acid critically involved in an interface is mutated, changes as large as 500-fold in the dissociation constant have been observed,⁴⁸⁴ so it is difficult to evaluate changes of less than 10-fold. For example, those 14 mutations judged to have a significant effect on the association of somatotropin with its receptor increased the dissociation constant by factors of only 4–20, while 14 of the mutations judged to be insignificant nevertheless increased the dissociation constant by factors of 2–3. The distinction appears to be arbitrary. Another problem is that during the formation of an interface of any kind, a conformational change may be required to occur in a portion of one or the other of the proteins in order for the proper fit between the faces to be achieved. Any mutation that hinders this conformational change will disrupt the association even if it is not in a side chain that ends up within the interface. In the case of somatotropin, four of the mutations to side chains that are not incorporated into the interface, but which nevertheless did increase the dissociation constant between somatotropin and its receptor by factors of 4–15, are to side chains that are located in a segment of random meander in free somatotropin that becomes ordered upon its association with the receptor.

Suggested Reading

Cingolani, G., Petosa, C., Weis, K., & Muller, C.W. (1999) Structure of importin- β bound to the IBB domain of importin- α , *Nature* 399, 221–229.

References

- Monod, J., Wyman, J., & Changeux, J.P. (1965) *J. Mol. Biol.* 12, 88–118.
- Kim, S.Y., Hwang, K.Y., Kim, S.H., Sung, H.C., Han, Y.S., & Cho, Y. (1999) *J. Biol. Chem.* 274, 11761–11767.
- Holmes, K.C., Popp, D., Gebhard, W., & Kabsch, W. (1990) *Nature* 347, 44–49.
- Kabsch, W., Mannherz, H.G., Suck, D., Pai, E.F., & Holmes, K.C. (1990) *Nature* 347, 37–44.
- Smith, P.R., Fowler, W.E., Pollard, T.D., & Aebi, U. (1983) *J. Mol. Biol.* 167, 641–660.
- van den Ent, F., Amos, L.A., & Lowe, J. (2001) *Nature* 413, 39–44.
- Ohlendorf, D.H., Weber, P.C., & Lipscomb, J.D. (1987) *J. Mol. Biol.* 195, 225–227.
- Ohlendorf, D.H., Orville, A.M., & Lipscomb, J.D. (1994) *J. Mol. Biol.* 244, 586–608.
- Steitz, T.A., Fletterick, R.J., Anderson, W.F., & Anderson, C.M. (1976) *J. Mol. Biol.* 104, 197–122.
- Finch, J.T., Perutz, M.F., Bertles, J.F., & Dobler, J. (1973) *Proc. Natl. Acad. Sci. U.S.A.* 70, 718–722.
- Li, H., & Abelson, J. (2000) *J. Mol. Biol.* 302, 639–648.
- Bailey, D.L., Fraser, M.E., Bridger, W.A., James, M.N., & Wolodko, W.T. (1999) *J. Mol. Biol.* 285, 1655–1666.
- Borchert, T.V., Abagyan, R., Jaenicke, R., & Wierenga, R.K. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 1515–1518.
- Oefner, C., & Suck, D. (1986) *J. Mol. Biol.* 192, 605–632.
- Hahn, T. (1983) *International Tables for Crystallography, Volume A. Space-Group Symmetry*, D. Reidel, Dordrecht, The Netherlands.
- Einspahr, H., Parks, E.H., Suguna, K., Subramanian, E., & Suddath, F.L. (1986) *J. Biol. Chem.* 261, 16518–16527.
- Bourne, Y., Abergel, C., Cambillau, C., Frey, M., Rouge, P., & Fontecilla-Camps, J.C. (1990) *J. Mol. Biol.* 214, 571–584.
- Holden, H.M., Ito, M., Hartshorne, D.J., & Rayment, I. (1992) *J. Mol. Biol.* 227, 840–851.
- Weiss, M.S., & Schulz, G.E. (1992) *J. Mol. Biol.* 227, 493–509.
- Tsukihara, T., Fukuyama, K., Mizushima, M., Harioka, T., Kusunoki, M., Katsube, Y., Hase, T., & Matsubara, H. (1990) *J. Mol. Biol.* 216, 399–410.
- Bianchet, M.A., Hullihen, J., Pedersen, P.L., & Amzel, L.M. (1998) *Proc. Natl. Acad. Sci. U.S.A.* 95, 11065–11070.
- Banner, D.W., Bloomer, A.C., Petsko, G.A., Phillips, D.C., Pogson, C.I., Wilson, I.A., Corran, P.H., Furth, A.J., Milman, J.D., Offord, R.E., Priddle, J.D., & Waley, S.G. (1975) *Nature* 255, 609–614.
- Epp, O., Ladenstein, R., & Wendel, A. (1983) *Eur. J. Biochem.* 133, 51–69.
- Sprang, S., & Fletterick, R.J. (1979) *J. Mol. Biol.* 131, 523–551.
- Eklund, H., Nordstrom, B., Zeppezauer, E., Seoderlund, G., Ohlsson, I., Boiwe, T., Seoderberg, B.O., Tapia, O., Breandaen, C.I., & Akeson, A. (1976) *J. Mol. Biol.* 102, 27–59.
- Leslie, A.G., Moody, P.C., & Shaw, W.V. (1988) *Proc. Natl. Acad. Sci. U.S.A.* 85, 4133–4137.
- Koellner, G., Luic, M., Shugar, D., Saenger, W., & Bzowska, A. (1997) *J. Mol. Biol.* 265, 202–216.
- Adams, M.J., Ford, G.C., Koekoek, R., Lentz, P.J., McPherson, A., Jr., Rossmann, M.G., Smiley, I.E., Schevitz, R.W., & Wonacott, A.J. (1970) *Nature* 227, 1098–1103.
- Lindqvist, Y., & Braenden, C.I. (1985) *Proc. Natl. Acad. Sci. U.S.A.* 82, 6855–6859.
- Mattevi, A., Obmolova, G., Kalk, K.H., Westphal, A.H., de Kok, A., & Hol, W.G. (1993) *J. Mol. Biol.* 230, 1183–1199.
- DeRosier, D.J., Oliver, R.M., & Reed, L.J. (1971) *Proc. Natl. Acad. Sci. U.S.A.* 68, 1135–1137.
- Konig, P., & Richmond, T.J. (1993) *J. Mol. Biol.* 233, 139–154.
- Steebhorn, C., Messerschmidt, A., Laber, B., Streber, W., Huber, R., & Clausen, T. (1999) *J. Mol. Biol.* 290, 983–996.
- Schuller, D.J., Wilks, A., Ortiz de Montellano, P.R., & Poulos, T.L. (1999) *Nat. Struct. Biol.* 6, 860–867.
- Buehner, M., Ford, G.C., Moras, D., Olsen, K.W., & Rossmann, M.G. (1974) *J. Mol. Biol.* 82, 563–585.
- Buehner, M., Ford, G.C., Olsen, K.W., Moras, D., & Rossmann, M.G. (1974) *J. Mol. Biol.* 90, 25–49.
- Lamzin, V.S., Dauter, Z., Popov, V.O., Harutyunyan, E.H., & Wilson, K.S. (1994) *J. Mol. Biol.* 236, 759–785.
- Lolis, E., Alber, T., Davenport, R.C., Rose, D., Hartman, F.C., & Petsko, G.A. (1990) *Biochemistry* 29, 6609–6618.
- Nikkola, M., Lindqvist, Y., & Schneider, G. (1994) *J. Mol. Biol.* 238, 387–404.
- Harutyunyan, E.H., Kuranova, I.P., Vainshtein, B.K., Hohne, W.E., Lamzin, V.S., Dauter, Z., Teplyakov, A.V., & Wilson, K.S. (1996) *Eur. J. Biochem.* 239, 220–228.
- Andersson, I. (1996) *J. Mol. Biol.* 259, 160–174.
- Waldrop, G.L., Rayment, I., & Holden, H.M. (1994) *Biochemistry* 33, 10249–10256.
- Sixma, T.K., Kalk, K.H., van Zanten, B.A., Dauter, Z., Kingma, J., Witholt, B., & Hol, W.G. (1993) *J. Mol. Biol.* 230, 890–918.
- Birktoft, J.J., Rhodes, G., & Banaszak, L.J. (1989) *Biochemistry* 28, 6065–6081.
- Niefind, K., Hecht, H.J., & Schomburg, D. (1995) *J. Mol. Biol.* 251, 256–281.
- Antson, A.A., Brzozowski, A.M., Dodson, E.J., Dauter, Z., Wilson, K.S., Kurecki, T., Otridge, J., & Gollnick, P. (1994) *J. Mol. Biol.* 244, 1–5.
- Jia, Z., Vandonselaar, M., Hengstenberg, W., Quail, J.W., & Delbaere, L.T. (1994) *J. Mol. Biol.* 236, 1341–1355.
- Darnell, D.W., & Klotz, I.M. (1975) *Arch. Biochem. Biophys.* 166, 651–682.
- Dewan, J.C., Grant, G.A., & Sacchettini, J.C. (1994) *Biochemistry* 33, 13147–13154.
- Britton, K.L., Asano, Y., & Rice, D.W. (1998) *Nat. Struct. Biol.* 5, 593–601.
- Argiriadi, M.A., Morisseau, C., Hammock, B.D., & Christianson, D.W. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 10637–10642.
- Schulz, G.E., Schirmer, R.H., Sachsenheimer, W., & Pai, E.F. (1978) *Nature* 273, 120–124.
- Leistler, B., & Perham, R.N. (1994) *Biochemistry* 33, 2773–2781.
- Aleshin, A.E., Zeng, C., Bourenkov, G.P., Bartunik, H.D., Fromm, H.J., & Honzatko, R.B. (1998) *Structure* 6, 39–50.

55. Somers, W.S., & Phillips, S.E. (1992) *Nature* 359, 387–393.
56. Raumann, B.E., Rould, M.A., Pabo, C.O., & Sauer, R.T. (1994) *Nature* 367, 754–757.
57. Hegde, R.S., Grossman, S.R., Laimins, L.A., & Sigler, P.B. (1992) *Nature* 359, 505–512.
58. Leslie, A.G. (1990) *J. Mol. Biol.* 213, 167–186.
59. Adachi, M., Takenaka, Y., Gidamis, A.B., Mikami, B., & Utsumi, S. (2001) *J. Mol. Biol.* 305, 291–305.
60. Ealick, S.E., Rule, S.A., Carter, D.C., Greenhough, T.J., Babu, Y.S., Cook, W.J., Habash, J., Helliwell, J.R., Stoeckler, J.D., Parks, R.E., Jr., et al. (1990) *J. Biol. Chem.* 265, 1812–1820.
61. Larsson, G., Svensson, L.A., & Nyman, P.O. (1996) *Nat. Struct. Biol.* 3, 532–538.
62. Benning, M.M., Taylor, K.L., Liu, R.Q., Yang, G., Xiang, H., Wesenberg, G., Dunaway-Mariano, D., & Holden, H.M. (1996) *Biochemistry* 35, 8103–8109.
63. Campobasso, N., Mathews, I.I., Begley, T.P., & Ealick, S.E. (2000) *Biochemistry* 39, 7868–7877.
64. Tebbe, J., Bzowska, A., Wielgus-Kutrowska, B., Schroder, W., Kazimierczuk, Z., Shugar, D., Saenger, W., & Koellner, G. (1999) *J. Mol. Biol.* 294, 1239–1255.
65. Jiang, J., Zhang, Y., Krainer, A.R., & Xu, R.M. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 3572–3577.
66. Matthews, B.W., Fenna, R.E., Bolognesi, M.C., Schmid, M.F., & Olson, J.M. (1979) *J. Mol. Biol.* 131, 259–285.
67. Xia, Z.X., & Mathews, F.S. (1990) *J. Mol. Biol.* 212, 837–863.
68. Dreyer, M.K., & Schulz, G.E. (1993) *J. Mol. Biol.* 231, 549–553.
69. Luo, Y., Samuel, J., Mosimann, S.C., Lee, J.E., Tanner, M.E., & Strynadka, N.C. (2001) *Biochemistry* 40, 14763–14771.
70. Sintchak, M.D., Fleming, M.A., Futer, O., Raybuck, S.A., Chambers, S.P., Caron, P.R., Murcko, M.A., & Wilson, K.P. (1996) *Cell* 85, 921–930.
71. Zhang, R., Evans, G., Rotella, F.J., Westbrook, E.M., Beno, D., Huberman, E., Joachimiak, A., & Collart, F.R. (1999) *Biochemistry* 38, 4691–4700.
72. Brejc, K., van Dijk, W.J., Klaassen, R.V., Schuurmans, M., van Der Oost, J., Smit, A.B., & Sixma, T.K. (2001) *Nature* 411, 269–276.
73. Emsley, J., White, H.E., O'Hara, B.P., Oliva, G., Srinivasan, N., Tickle, I.J., Blundell, T.L., Pepys, M.B., & Wood, S.P. (1994) *Nature* 367, 338–345.
74. Dreveny, I., Kondo, H., Uchiyama, K., Shaw, A., Zhang, X., & Freemont, P.S. (2004) *EMBO J.* 23, 1030–1039.
75. Niedenzu, T., Roleke, D., Bains, G., Scherzinger, E., & Saenger, W. (2001) *J. Mol. Biol.* 306, 479–487.
76. Lee, S.Y., De La Torre, A., Yan, D., Kustu, S., Nixon, B.T., & Wemmer, D.E. (2003) *Genes Dev.* 17, 2552–2563.
77. Mura, C., Cascio, D., Sawaya, M.R., & Eisenberg, D.S. (2001) *Proc. Natl. Acad. Sci. U.S.A.* 98, 5532–5537.
78. Chen, X., Antson, A.A., Yang, M., Li, P., Baumann, C., Dodson, E.J., Dodson, G.G., & Gollnick, P. (1999) *J. Mol. Biol.* 289, 1003–1016.
79. Toney, M.D., Hohenester, E., Keller, J.W., & Jansonius, J.N. (1995) *J. Mol. Biol.* 245, 151–179.
80. Shirakihara, Y., & Evans, P.R. (1988) *J. Mol. Biol.* 204, 973–994.
81. Le Bras, G., Auzat, I., & Garel, J.R. (1995) *Biochemistry* 34, 13203–13210.
82. Riley-Lovingshimer, M.R., Ronning, D.R., Sacchettini, J.C., & Reinhart, G.D. (2002) *Biochemistry* 41, 12967–12974.
83. Dayhoff, M.O. (1972) *Atlas of Protein Sequence and Structure*, Vol. 5, National Biomedical Research Foundation, Washington, DC.
84. Perutz, M.F., Muirhead, H., Cox, J.M., & Goaman, L.C. (1968) *Nature* 219, 131–139.
85. Ip, S.H., & Ackers, G.K. (1977) *J. Biol. Chem.* 252, 82–87.
86. Park, C.M. (1970) *J. Biol. Chem.* 245, 5390–5394.
87. Burrows, S.D., Doyle, M.L., Murphy, K.P., Franklin, S.G., White, J.R., Brooks, I., McNulty, D.E., Scott, M.O., Knutson, J.R., Porter, D., et al. (1994) *Biochemistry* 33, 12741–12745.
88. Pereira, P.J., Bergner, A., Macedo-Ribeiro, S., Huber, R., Matschiner, G., Fritz, H., Sommerhoff, C.P., & Bode, W. (1998) *Nature* 392, 306–311.
89. Kai, Y., Matsumura, H., Inoue, T., Terada, K., Nagara, Y., Yoshinaga, T., Kihara, A., Tsumura, K., & Izui, K. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 823–828.
90. Friedman, A.M., Fischmann, T.O., & Steitz, T.A. (1995) *Science* 268, 1721–1727.
91. Banerjee, R., Mande, S.C., Ganesh, V., Das, K., Dhanaraj, V., Mahanta, S.K., Suguna, K., Suroliya, A., & Vijayan, M. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 227–231.
92. Kopp, J., Kopriva, S., Suss, K.H., & Schulz, G.E. (1999) *J. Mol. Biol.* 287, 761–771.
93. Harrison, D.H., Runquist, J.A., Holub, A., & Mizioro, H.M. (1998) *Biochemistry* 37, 5074–5085.
94. Alphey, M.S., Bond, C.S., Tetaud, E., Fairlamb, A.H., & Hunter, W.N. (2000) *J. Mol. Biol.* 300, 903–916.
95. MacRae, I.J., Segel, I.H., & Fisher, A.J. (2001) *Biochemistry* 40, 6795–6804.
96. Frankenberg, N., Erskine, P.T., Cooper, J.B., Shoolingin-Jordan, P.M., Jahn, D., & Heinz, D.W. (1999) *J. Mol. Biol.* 289, 591–602.
97. Katti, S.K., Katz, B.A., & Wyckoff, H.W. (1989) *J. Mol. Biol.* 205, 557–571.
98. Cherfils, J., Morera, S., Lascu, I., Veron, M., & Janin, J. (1994) *Biochemistry* 33, 9062–9069.
99. Brandstetter, H., Kim, J.S., Groll, M., & Huber, R. (2001) *Nature* 414, 466–470.
100. Fritz-Wolf, K., Schnyder, T., Wallimann, T., & Kabsch, W. (1996) *Nature* 381, 341–345.
101. Helin, S., Kahn, P.C., Guha, B.L., Mallows, D.G., & Goldman, A. (1995) *J. Mol. Biol.* 254, 918–941.
102. Remaut, H., Bompard-Gilles, C., Goffin, C., Frere, J.M., & Van Beeumen, J. (2001) *Nat. Struct. Biol.* 8, 674–678.
103. Du, X., Choi, I.G., Kim, R., Wang, W., Jancarik, J., Yokota, H., & Kim, S.H. (2000) *Proc. Natl. Acad. Sci. U.S.A.* 97, 14079–14084.
104. Koellner, G., Luic, M., Shugar, D., Saenger, W., & Bzowska, A. (1998) *J. Mol. Biol.* 280, 153–166.
105. Murley, L.L., & MacKenzie, R.E. (1995) *Biochemistry* 34, 10358–10364.
106. Gulick, A.M., Schmidt, D.M., Gerlt, J.A., & Rayment, I. (2001) *Biochemistry* 40, 15716–15724.
107. Momany, C., Ernst, S., Ghosh, R., Chang, N.L., & Hackert, M.L. (1995) *J. Mol. Biol.* 252, 643–655.

108. Webb, P.A., Perisic, O., Mendola, C.E., Backer, J.M., & Williams, R.L. (1995) *J. Mol. Biol.* 251, 574–587.
109. Cooper, J.B., McIntyre, K., Badasso, M.O., Wood, S.P., Zhang, Y., Garbe, T.R., & Young, D. (1995) *J. Mol. Biol.* 246, 531–544.
110. Jelakovic, S., Kopriva, S., Suss, K.H., & Schulz, G.E. (2003) *J. Mol. Biol.* 326, 127–135.
111. Blickling, S., Beisel, H.G., Bozic, D., Knablein, J., Laber, B., & Huber, R. (1997) *J. Mol. Biol.* 274, 608–621.
112. Gallagher, T., Snell, E.E., & Hackert, M.L. (1989) *J. Biol. Chem.* 264, 12737–12743.
113. Yamashita, M.M., Almasy, R.J., Janson, C.A., Cascio, D., & Eisenberg, D. (1989) *J. Biol. Chem.* 264, 17681–17690.
114. Hohenester, E., Hutchinson, W.L., Pepys, M.B., & Wood, S.P. (1997) *J. Mol. Biol.* 269, 570–578.
115. Hutchinson, E.G., Tichelaar, W., Hofhaus, G., Weiss, H., & Leonard, K.R. (1989) *EMBO J.* 8, 1485–1490.
116. Lowe, J., Stock, D., Jap, B., Zwickl, P., Baumeister, W., & Huber, R. (1995) *Science* 268, 533–539.
117. Whitby, F.G., Luecke, H., Kuhn, P., Somoza, J.R., Huete-Perez, J.A., Phillips, J.D., Hill, C.P., Fletterick, R.J., & Wang, C.C. (1997) *Biochemistry* 36, 10666–10674.
118. Delbaere, L.T., Vandonselaar, M., Prasad, L., Quail, J.W., Wilson, K.S., & Dauter, Z. (1993) *J. Mol. Biol.* 230, 950–965.
119. Mondragon, A., Wolberger, C., & Harrison, S.C. (1989) *J. Mol. Biol.* 205, 179–188.
120. Spraggon, G., Kim, C., Nguyen-Huu, X., Yee, M.C., Yanofsky, C., & Mills, S.E. (2001) *Proc. Natl. Acad. Sci. U.S.A.* 98, 6021–6026.
121. Colloc'h, N., el Hajji, M., Bachet, B., L'Hermite, G., Schiltz, M., Prange, T., Castro, B., & Mornon, J.P. (1997) *Nat. Struct. Biol.* 4, 947–952.
122. Voegtli, W.C., Ge, J., Perlstein, D.L., Stubbe, J., & Rosenzweig, A.C. (2001) *Proc. Natl. Acad. Sci. U.S.A.* 98, 10073–10078.
123. Schulz, G.E., & Schirmer, R.H. (1979) *Principles of Protein Structure*, p 94, Springer-Verlag, New York.
124. Weiss, V.H., McBride, A.E., Soriano, M.A., Filman, D.J., Silver, P.A., & Hogle, J.M. (2000) *Nat. Struct. Biol.* 7, 1165–1171.
125. Bell, C.E., & Lewis, M. (2001) *J. Mol. Biol.* 314, 1127–1136.
126. Bell, C.E., Frescura, P., Hochschild, A., & Lewis, M. (2000) *Cell* 101, 801–811.
127. Ploegman, J.H., Drent, G., Kalk, K.H., & Hol, W.G. (1978) *J. Mol. Biol.* 123, 557–594.
128. Roderick, S.L., & Matthews, B.W. (1993) *Biochemistry* 32, 3907–3912.
129. Sielecki, A.R., Fedorov, A.A., Boodhoo, A., Andreeva, N.S., & James, M.N. (1990) *J. Mol. Biol.* 214, 143–170.
130. Gilliland, G.L., & Quiocho, F.A. (1981) *J. Mol. Biol.* 146, 341–362.
131. McLachlan, A.D. (1979) *J. Mol. Biol.* 128, 49–79.
132. Crane, B.R., Siegel, L.M., & Getzoff, E.D. (1995) *Science* 270, 59–67.
133. Priestle, J.P., Schar, H.P., & Grutter, M.G. (1988) *EMBO J.* 7, 339–343.
134. Eriksson, A.E., Cousens, L.S., Weaver, L.H., & Matthews, B.W. (1991) *Proc. Natl. Acad. Sci. U.S.A.* 88, 3441–3445.
135. Groft, C.M., Beckmann, R., Sali, A., & Burley, S.K. (2000) *Nat. Struct. Biol.* 7, 1156–1164.
136. Muller, Y.A., Schumacher, G., Rudolph, R., & Schulz, G.E. (1994) *J. Mol. Biol.* 237, 315–335.
137. Wright, C.S. (1992) *J. Biol. Chem.* 267, 14345–14352.
138. Huber, R., Romisch, J., & Paques, E.P. (1990) *EMBO J.* 9, 3867–3874.
139. Lawrence, M.C., Suzuki, E., Varghese, J.N., Davis, P.C., Van Donkelaar, A., Tulloch, P.A., & Colman, P.M. (1990) *EMBO J.* 9, 9–15.
140. Subramanya, H.S., Roper, D.I., Dauter, Z., Dodson, E.J., Davies, G.J., Wilson, K.S., & Wigley, D.B. (1996) *Biochemistry* 35, 792–802.
141. Janin, J., Miller, S., & Chothia, C. (1988) *J. Mol. Biol.* 204, 155–164.
142. De Simone, G., Menchise, V., Manco, G., Mandrich, L., Sorrentino, N., Lang, D., Rossi, M., & Pedone, C. (2001) *J. Mol. Biol.* 314, 507–518.
143. Roujeinikova, A., Raasch, C., Burke, J., Baker, P.J., Liebl, W., & Rice, D.W. (2001) *J. Mol. Biol.* 312, 119–131.
144. Kuhn, K., & Luisi, B.F. (2001) *J. Mol. Biol.* 313, 583–592.
145. Mondragon, A., & Harrison, S.C. (1991) *J. Mol. Biol.* 219, 321–334.
146. Kreusch, A., & Schulz, G.E. (1994) *J. Mol. Biol.* 243, 891–905.
147. Romao, M.J., Turk, D., Gomis-Ruth, F.X., Huber, R., Schumacher, G., Mollering, H., & Russmann, L. (1992) *J. Mol. Biol.* 226, 1111–1130.
148. Clausen, T., Huber, R., Laber, B., Pohlenz, H.D., & Messerschmidt, A. (1996) *J. Mol. Biol.* 262, 202–224.
149. Borgstahl, G.E., Rogers, P.H., & Arnone, A. (1994) *J. Mol. Biol.* 236, 817–830.
150. Ji, X., Zhang, P., Armstrong, R.N., & Gilliland, G.L. (1992) *Biochemistry* 31, 10169–10184.
151. Feese, M.D., Kato, Y., Tamada, T., Kato, M., Komeda, T., Miura, Y., Hirose, M., Hondo, K., Kobayashi, K., & Kuroki, R. (2000) *J. Mol. Biol.* 301, 451–464.
152. Schlunegger, M.P., & Grutter, M.G. (1992) *Nature* 358, 430–434.
153. Strater, N., Klabunde, T., Tucker, P., Witzel, H., & Krebs, B. (1995) *Science* 268, 1489–1492.
154. Fischmann, T.O., Hruza, A., Niu, X.D., Fossetta, J.D., Lunn, C.A., Dolphin, E., Prongay, A.J., Reichert, P., Lundell, D.J., Narula, S.K., & Weber, P.C. (1999) *Nat. Struct. Biol.* 6, 233–242.
155. Gonzalez, B., Pajares, M.A., Hermoso, J.A., Alvarez, L., Garrido, F., Sufrin, J.R., & Sanz-Aparicio, J. (2000) *J. Mol. Biol.* 300, 363–375.
156. Hecht, H.J., Kalisz, H.M., Hendle, J., Schmid, R.D., & Schomburg, D. (1993) *J. Mol. Biol.* 229, 153–172.
157. Stoddard, B.L., Howell, P.L., Ringe, D., & Petsko, G.A. (1990) *Biochemistry* 29, 8885–8893.
158. Freymann, D., Down, J., Carrington, M., Roditi, I., Turner, M., & Wiley, D. (1990) *J. Mol. Biol.* 216, 141–160.
159. Luo, Y., Frey, E.A., Pfuetzner, R.A., Creagh, A.L., Knoechel, D.G., Haynes, C.A., Finlay, B.B., & Strynadka, N.C. (2000) *Nature* 405, 1073–1077.
160. Sheriff, S., Chang, C.Y., & Ezekowitz, R.A. (1994) *Nat. Struct. Biol.* 1, 789–794.
161. Weaver, T.M., Levitt, D.G., Donnelly, M.I., Stevens, R.

- P.P., & Banaszak, L.J. (1995) *Nat. Struct. Biol.* 2, 654–662.
162. Schwede, T.F., Retey, J., & Schulz, G.E. (1999) *Biochemistry* 38, 5355–5361.
 163. Toth, E.A., Worby, C., Dixon, J.E., Goedken, E.R., Marqusee, S., & Yeates, T.O. (2000) *J. Mol. Biol.* 301, 433–450.
 164. Kostrewa, D., D'Arcy, A., Takacs, B., & Kamber, M. (2001) *J. Mol. Biol.* 305, 279–289.
 165. Hardman, K.D., & Ainsworth, C.F. (1972) *Biochemistry* 11, 4910–4919.
 166. Hirotsu, S., Abe, Y., Okada, K., Nagahara, N., Hori, H., Nishino, T., & Hakoshima, T. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 12333–12338.
 167. Banerjee, R., Das, K., Ravishankar, R., Suguna, K., Surolia, A., & Vijayan, M. (1996) *J. Mol. Biol.* 259, 281–296.
 168. Lott, J.S., Halbig, D., Baker, H.M., Hardman, M.J., Sprenger, G.A., & Baker, E.N. (2000) *J. Mol. Biol.* 304, 575–584.
 169. Hennig, M., D'Arcy, A., Hampele, I.C., Page, M.G., Oefner, C., & Dale, G.E. (1998) *Nat. Struct. Biol.* 5, 357–362.
 170. Arnold, E., & Rossmann, M.G. (1990) *J. Mol. Biol.* 211, 763–801.
 171. Bennett, M.J., Choe, S., & Eisenberg, D. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 3127–3131.
 172. Liu, Y., Hart, P.J., Schlunegger, M.P., & Eisenberg, D. (1998) *Proc. Natl. Acad. Sci. U.S.A.* 95, 3437–3442.
 173. Valegard, K., Liljas, L., Fridborg, K., & Unge, T. (1990) *Nature* 345, 36–41.
 174. Kim, S.J., Jeong, D.G., Chi, S.W., Lee, J.S., & Ryu, S.E. (2001) *Nat. Struct. Biol.* 8, 459–466.
 175. Hadden, J.M., Convery, M.A., Declais, A.C., Lilley, D.M., & Phillips, S.E. (2001) *Nat. Struct. Biol.* 8, 62–67.
 176. Saint-Jean, A.P., Phillips, K.R., Creighton, D.J., & Stone, M.J. (1998) *Biochemistry* 37, 10345–10353.
 177. Schymkowitz, J.W., Rousseau, F., Wilkinson, H.R., Friedler, A., & Itzhaki, L.S. (2001) *Nat. Struct. Biol.* 8, 888–892.
 178. Janowski, R., Kozak, M., Jankowska, E., Grzonka, Z., Grubb, A., Abrahamson, M., & Jaskolski, M. (2001) *Nat. Struct. Biol.* 8, 316–320.
 179. Ciglic, M.I., Jackson, P.J., Raillard, S.A., Haugg, M., Jermann, T.M., Opitz, J.G., Trabesinger-Ruf, N., & Benner, S.A. (1998) *Biochemistry* 37, 4008–4022.
 180. Green, S.M., Gittis, A.G., Meeker, A.K., & Lattman, E.E. (1995) *Nat. Struct. Biol.* 2, 746–751.
 181. Jeffery, C.J., Bahnson, B.J., Chien, W., Ringe, D., & Petsko, G.A. (2000) *Biochemistry* 39, 955–964.
 182. Milburn, M.V., Hassell, A.M., Lambert, M.H., Jordan, S.R., Proudfoot, A.E., Graber, P., & Wells, T.N. (1993) *Nature* 363, 172–176.
 183. Gabelli, S.B., Bianchet, M.A., Bessman, M.J., & Amzel, L.M. (2001) *Nat. Struct. Biol.* 8, 467–472.
 184. Melik-Adamy, W.R., Barynin, V.V., Vagin, A.A., Borisov, V.V., Vainshtein, B.K., Fita, I., Murthy, M.R., & Rossmann, M.G. (1986) *J. Mol. Biol.* 188, 63–72.
 185. Royer, W.E., Jr., Strand, K., van Heel, M., & Hendrickson, W.A. (2000) *Proc. Natl. Acad. Sci. U.S.A.* 97, 7107–7111.
 186. Royer, W.E., Jr., Heard, K.S., Harrington, D.J., & Chiancone, E. (1995) *J. Mol. Biol.* 253, 168–186.
 187. Goldberg, J.D., Yoshida, T., & Brick, P. (1994) *J. Mol. Biol.* 236, 1123–1140.
 188. Matsuda, K., Mizuguchi, K., Nishioka, T., Kato, H., Go, N., & Oda, J. (1996) *Protein Eng.* 9, 1083–1092.
 189. Xiang, S., Short, S.A., Wolfenden, R., & Carter, C.W., Jr. (1996) *Biochemistry* 35, 1335–1341.
 190. Boisset, N., & Mouche, F. (2000) *J. Mol. Biol.* 296, 459–472.
 191. Crick, F.H.C., & Watson, J.D. (1956) *Nature* 177, 473–475.
 192. Cromwell, P.R. (1997) *Polyhedra*, Cambridge University Press, Cambridge.
 193. Gourley, D.G., Shrive, A.K., Polikarpov, I., Krell, T., Coggins, J.R., Hawkins, A.R., Isaacs, N.W., & Sawyer, L. (1999) *Nat. Struct. Biol.* 6, 521–525.
 194. Hawkins, A.R., Lamb, H.K., Moore, J.D., Charles, I.G., & Roberts, C.F. (1993) *J. Gen. Microbiol.* 139, 2891–2899.
 195. Sun, S.M., McLeester, R.C., Bliss, F.A., & Hall, T.C. (1974) *J. Biol. Chem.* 249, 2118–2121.
 196. Marcq, S., Diaz-Ruano, A., Charlier, P., Dideberg, O., Tricot, C., Pierard, A., & Stalon, V. (1991) *J. Mol. Biol.* 220, 9–12.
 197. Baur, H., Stalon, V., Falmagne, P., Luethi, E., & Haas, D. (1987) *Eur. J. Biochem.* 166, 111–117.
 198. Isupov, M.N., Dalby, A.R., Brindley, A.A., Izumi, Y., Tanabe, T., Murshudov, G.N., & Littlechild, J.A. (2000) *J. Mol. Biol.* 299, 1035–1049.
 199. Kim, K.K., Kim, R., & Kim, S.H. (1998) *Nature* 394, 595–599.
 200. Ritsert, K., Huber, R., Turk, D., Ladenstein, R., Schmidt-Base, K., & Bacher, A. (1995) *J. Mol. Biol.* 253, 151–167.
 201. Ban, N., & McPherson, A. (1995) *Nat. Struct. Biol.* 2, 882–890.
 202. Tsao, J., Chapman, M.S., Agbandje, M., Keller, W., Smith, K., Wu, H., Luo, M., Smith, T.J., Rossmann, M.G., Compans, R.W., et al. (1991) *Science* 251, 1456–1464.
 203. Xie, Q., & Chapman, M.S. (1996) *J. Mol. Biol.* 264, 497–520.
 204. Braden, B.C., Velikovsky, C.A., Cauwerhff, A.A., Polikarpov, I., & Goldbaum, F.A. (2000) *J. Mol. Biol.* 297, 1031–1036.
 205. Trikha, J., Theil, E.C., & Allewell, N.M. (1995) *J. Mol. Biol.* 248, 949–967.
 206. Hempstead, P.D., Yewdall, S.J., Fernie, A.R., Lawson, D.M., Artymiuk, P.J., Rice, D.W., Ford, G.C., & Harrison, P.M. (1997) *J. Mol. Biol.* 268, 424–448.
 207. Ilari, A., Stefanini, S., Chiancone, E., & Tsernoglou, D. (2000) *Nat. Struct. Biol.* 7, 38–43.
 208. van Montfort, R.L., Basha, E., Friedrich, K.L., Slingsby, C., & Vierling, E. (2001) *Nat. Struct. Biol.* 8, 1025–1030.
 209. Wang, G.F., Kuriki, T., Roy, K.L., & Kaneda, T. (1993) *Eur. J. Biochem.* 213, 1091–1099.
 210. Izard, T., Aevansson, A., Allen, M.D., Westphal, A.H., Perham, R.N., de Kok, A., & Hol, W.G. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 1240–1245.
 211. Mattevi, A., Obmolova, G., Schulze, E., Kalk, K.H., Westphal, A.H., de Kok, A., & Hol, W.G. (1992) *Science* 255, 1544–1550.
 212. Knapp, J.E., Mitchell, D.T., Yazdi, M.A., Ernst, S.R., Reed, L.J., & Hackert, M.L. (1998) *J. Mol. Biol.* 280, 655–668.

213. Stoops, J.K., Baker, T.S., Schroeter, J.P., Kolodziej, S.J., Niu, X.D., & Reed, L.J. (1992) *J. Biol. Chem.* 267, 24769–24775.
214. McKenna, R., Xia, D., Willingmann, P., Ilag, L.L., Krishnaswamy, S., Rossmann, M.G., Olson, N.H., Baker, T.S., & Incardona, N.L. (1992) *Nature* 355, 137–143.
215. Montelius, I., Liljas, L., & Unge, T. (1988) *J. Mol. Biol.* 201, 353–363.
216. Grimes, J.M., Burroughs, J.N., Gouet, P., Diprose, J.M., Malby, R., Zientara, S., Mertens, P.P., & Stuart, D.I. (1998) *Nature* 395, 470–478.
217. Reinisch, K.M., Nibert, M.L., & Harrison, S.C. (2000) *Nature* 404, 960–967.
218. Caspar, D.L.D., & Klug, A. (1962) *Cold Spring Harbor Symp. Quant. Biol.* 27, 1–24.
219. Fuller, R.B. (1954) U.S. Patent 2682235.
220. Olson, A.J., Bricogne, G., & Harrison, S.C. (1983) *J. Mol. Biol.* 171, 61–93.
221. Rossmann, M.G., Abad-Zapatero, C., Hermodson, M.A., & Erickson, J.W. (1983) *J. Mol. Biol.* 166, 37–73.
222. Canady, M.A., Larson, S.B., Day, J., & McPherson, A. (1996) *Nat. Struct. Biol.* 3, 771–781.
223. Wery, J.P., Reddy, V.S., Hosur, M.V., & Johnson, J.E. (1994) *J. Mol. Biol.* 235, 565–586.
224. Prasad, B.V., Matson, D.O., & Smith, A.W. (1994) *J. Mol. Biol.* 240, 256–264.
225. Chen, Z.G., Stauffacher, C., Li, Y., Schmidt, T., Bomu, W., Kamer, G., Shanks, M., Lomonossoff, G., & Johnson, J.E. (1989) *Science* 245, 154–159.
226. Rossmann, M.G., Arnold, E., Erickson, J.W., Frankenberger, E.A., Griffith, J.P., Hecht, H.J., Johnson, J.E., Kamer, G., Luo, M., Mosser, A.G., et al. (1985) *Nature* 317, 145–153.
227. Hogle, J.M., Chow, M., & Filman, D.J. (1985) *Science* 229, 1358–1365.
228. Luo, M., Vriend, G., Kamer, G., Minor, I., Arnold, E., Rossmann, M.G., Boege, U., Scraba, D.G., Duke, G.M., & Palmenberg, A.C. (1987) *Science* 235, 182–191.
229. Acharya, R., Fry, E., Stuart, D., Fox, G., Rowlands, D., & Brown, F. (1989) *Nature* 337, 709–716.
230. Liljas, L., Unge, T., Jones, T.A., Fridborg, K., Leovgren, S., Skoglund, U., & Strandberg, B. (1982) *J. Mol. Biol.* 159, 93–108.
231. Rossmann, M.G., Abad-Zapatero, C., Murthy, M.R., Liljas, L., Jones, T.A., & Strandberg, B. (1983) *J. Mol. Biol.* 165, 711–736.
232. Erickson, J.W., Silva, A.M., Murthy, M.R., Fita, I., & Rossmann, M.G. (1985) *Science* 229, 625–629.
233. Choi, H.K., Tong, L., Minor, W., Dumas, P., Boege, U., Rossmann, M.G., & Wengler, G. (1991) *Nature* 354, 37–43.
234. Munshi, S., Liljas, L., Cavarelli, J., Bomu, W., McKinney, B., Reddy, V., & Johnson, J.E. (1996) *J. Mol. Biol.* 261, 1–10.
235. Al-Khayat, H.A., Bhella, D., Kenney, J.M., Roth, J.F., Kingsman, A.J., Martin-Rendon, E., & Saibil, H.R. (1999) *J. Mol. Biol.* 292, 65–73.
236. Prasad, B.V., Prevelige, P.E., Marietta, E., Chen, R.O., Thomas, D., King, J., & Chiu, W. (1993) *J. Mol. Biol.* 231, 65–74.
237. Thuman-Commike, P.A., Greene, B., Jakana, J., Prasad, B.V., King, J., Prevelige, P.E., Jr., & Chiu, W. (1996) *J. Mol. Biol.* 260, 85–98.
238. Liddington, R.C., Yan, Y., Moulai, J., Sahli, R., Benjamin, T.L., & Harrison, S.C. (1991) *Nature* 354, 278–284.
239. Baker, T.S., Newcomb, W.W., Olson, N.H., Cowser, L.M., Olson, C., & Brown, J.C. (1991) *Biophys. J.* 60, 1445–1456.
240. Belnap, D.M., Olson, N.H., Cladel, N.M., Newcomb, W.W., Brown, J.C., Kreider, J.W., Christensen, N.D., & Baker, T.S. (1996) *J. Mol. Biol.* 259, 249–263.
241. Metcalf, P., Cyrklaff, M., & Adrian, M. (1991) *EMBO J.* 10, 3129–3136.
242. Zhou, Z.H., Prasad, B.V., Jakana, J., Rixon, F.J., & Chiu, W. (1994) *J. Mol. Biol.* 242, 456–469.
243. Zhou, Z.H., Dougherty, M., Jakana, J., He, J., Rixon, F.J., & Chiu, W. (2000) *Science* 288, 877–880.
244. Trus, B.L., Booy, F.P., Newcomb, W.W., Brown, J.C., Homa, F.L., Thomsen, D.R., & Steven, A.C. (1996) *J. Mol. Biol.* 263, 447–462.
245. Athappilly, F.K., Murali, R., Rux, J.J., Cai, Z., & Burnett, R.M. (1994) *J. Mol. Biol.* 242, 430–455.
246. Furcinitti, P.S., van Oostrum, J., & Burnett, R.M. (1989) *EMBO J.* 8, 3563–3570.
247. Horne, R.W., Brenner, S., Waterson, A.P., & Wildy, P. (1959) *J. Mol. Biol.* 1, 84–86.
248. Valentine, R.C., & Pereira, H.G. (1965) *J. Mol. Biol.* 13, 13–20.
249. Ungewickell, E., & Branton, D. (1981) *Nature* 289, 420–422.
250. Pearse, B.M., & Robinson, M.S. (1984) *EMBO J.* 3, 1951–1957.
251. Vigers, G.P., Crowther, R.A., & Pearse, B.M. (1986) *EMBO J.* 5, 529–534.
252. Story, R.M., Weber, I.T., & Steitz, T.A. (1992) *Nature* 355, 318–325.
253. Watson, J.D. (1954) *Biochim. Biophys. Acta* 13, 10–19.
254. Klug, A. (1972) *Fed. Proc.* 31, 30–42.
255. Finch, J.T., & Klug, A. (1974) *J. Mol. Biol.* 87, 633–640.
256. Namba, K., Pattanayek, R., & Stubbs, G. (1989) *J. Mol. Biol.* 208, 307–325.
257. Klug, A., Crick, F.H.C., & Wyckoff, H.W. (1958) *Acta Crystallogr.* 11, 199–213.
258. Namba, K., & Stubbs, G. (1986) *Science* 231, 1401–1406.
259. Namba, K., Yamashita, I., & Vonderviszt, F. (1989) *Nature* 342, 648–654.
260. Amos, L.A., & Klug, A. (1975) *J. Mol. Biol.* 99, 51–64.
261. Li, S., Hill, C.P., Sundquist, W.L., & Finch, J.T. (2000) *Nature* 407, 409–413.
262. Otterbein, L.R., Graceffa, P., & Dominguez, R. (2001) *Science* 293, 708–711.
263. DeRosier, D.J., & Klug, A. (1968) *Nature* 217, 130–134.
264. Dubochet, J., Lepault, J., Freeman, R., Berriman, J.A., & Homo, J.C. (1982) *J. Microsc.* 128, 219–237.
265. McDowell, A.W., Chang, J.J., Freeman, R., Lepault, J., Walter, C.A., & Dubochet, J. (1983) *J. Microsc.* 131, 1–9.
266. Moores, C.A., Keep, N.H., & Kendrick-Jones, J. (2000) *J. Mol. Biol.* 297, 465–480.
267. Crowther, R.A., DeRosier, D.J., & Klug, A. (1970) *Proc. R. Soc. London, A* 317, 319–340.

268. Cochran, W., Crick, F.H.C., & Vand, V. (1952) *Acta Crystallogr.* 5, 581–586.
269. Toyoshima, C., & Unwin, N. (1990) *J. Cell Biol.* 111, 2623–2635.
270. Lowe, J., Li, H., Downing, K.H., & Nogales, E. (2001) *J. Mol. Biol.* 313, 1045–1057 and Nogales, E., Wolf, S.G., & Downing, K.H. (1998) *Nature* 391, 199–203.
271. Nogales, E., Whittaker, M., Milligan, R.A., & Downing, K.H. (1999) *Cell* 96, 79–88.
272. Meurer-Grob, P., Kasparian, J., & Wade, R.H. (2001) *Biochemistry* 40, 8000–8008.
273. Samatey, F.A., Imada, K., Nagashima, S., Vonderviszt, F., Kumasaka, T., Yamamoto, M., & Namba, K. (2001) *Nature* 410, 331–337.
274. Milligan, R.A., Whittaker, M., & Safer, D. (1990) *Nature* 348, 217–221.
275. Trachtenberg, S., & DeRosier, D.J. (1988) *J. Mol. Biol.* 202, 787–808.
276. Jeng, T.W., Crowther, R.A., Stubbs, G., & Chiu, W. (1989) *J. Mol. Biol.* 205, 251–257.
277. Miyazawa, A., Fujiyoshi, Y., Stowell, M., & Unwin, N. (1999) *J. Mol. Biol.* 288, 765–786.
278. Song, Y.H., & Mandelkow, E. (1993) *Proc. Natl. Acad. Sci. U.S.A.* 90, 1671–1675.
279. Wang, H., Culver, J.N., & Stubbs, G. (1997) *J. Mol. Biol.* 269, 769–779.
280. Miller, A., & Wray, J.S. (1971) *Nature* 230, 437–439.
281. Miller, A., & Parry, D.A. (1973) *J. Mol. Biol.* 75, 441–447.
282. Kramer, R.Z., Bella, J., Mayville, P., Brodsky, B., & Berman, H.M. (1999) *Nat. Struct. Biol.* 6, 454–457.
283. Rich, A., & Crick, F.H.C. (1961) *J. Mol. Biol.* 3, 483–506.
284. Fraser, R.D., MacRae, T.P., & Suzuki, E. (1979) *J. Mol. Biol.* 129, 463–481.
285. Kramer, R.Z., Bella, J., Brodsky, B., & Berman, H.M. (2001) *J. Mol. Biol.* 311, 131–147.
286. Dolz, R., Engel, J., & Kuhn, K. (1988) *Eur. J. Biochem.* 178, 357–366.
287. McLaughlin, S.H., & Bulleid, N.J. (1998) *Matrix Biol.* 16, 369–377.
288. Okuyama, K., Okuyama, K., Arnott, S., Takayanagi, M., & Kakudo, M. (1981) *J. Mol. Biol.* 152, 427–443.
289. Li, M.H., Fan, P., Brodsky, B., & Baum, J. (1993) *Biochemistry* 32, 7377–7387.
290. Rehn, M., & Pihlajaniemi, T. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 4234–4238.
291. Dublet, B., & van der Rest, M. (1991) *J. Biol. Chem.* 266, 6853–6858.
292. Chu, M.L., Zhang, R.Z., Pan, T.C., Stokes, D., Conway, D., Kuo, H.J., Glanville, R., Mayer, U., Mann, K., Deutzmann, R., et al. (1990) *EMBO J.* 9, 385–393.
293. Wess, T.J., Hammersley, A.P., Wess, L., & Miller, A. (1998) *J. Mol. Biol.* 275, 255–267.
294. Fraser, R.D., MacRae, T.P., Miller, A., & Suzuki, E. (1983) *J. Mol. Biol.* 167, 497–521.
295. Hodge, A.J., & Schmidt, F.O. (1960) *Proc. Natl. Acad. Sci. U.S.A.* 46, 186–206.
296. Meek, K.M., Chapman, J.A., & Hardcastle, R.A. (1979) *J. Biol. Chem.* 254, 10710–10714.
297. McLachlan, A.D., & Stewart, M. (1975) *J. Mol. Biol.* 98, 293–304.
298. Wakabayashi, T., Huxley, H.E., Amos, L.A., & Klug, A. (1975) *J. Mol. Biol.* 93, 477–497.
299. Narita, A., Yasunaga, T., Ishikawa, T., Mayanagi, K., & Wakabayashi, T. (2001) *J. Mol. Biol.* 308, 241–261.
300. Brown, J.H., Kim, K.H., Jun, G., Greenfield, N.J., Dominguez, R., Volkmann, N., Hitchcock-DeGregori, S.E., & Cohen, C. (2001) *Proc. Natl. Acad. Sci. U.S.A.* 98, 8496–8501.
301. Amos, L., & Klug, A. (1974) *J. Cell Sci.* 14, 523–549.
302. Franke, W.W., Schmid, E., Osborn, M., & Weber, K. (1978) *Proc. Natl. Acad. Sci. U.S.A.* 75, 5034–5038.
303. Geisler, N., Fischer, S., Vandekerckhove, J., Van Damme, J., Plessmann, U., & Weber, K. (1985) *EMBO J.* 4, 57–63.
304. Renner, W., Franke, W.W., Schmid, E., Geisler, N., Weber, K., & Mandelkow, E. (1981) *J. Mol. Biol.* 149, 285–306.
305. Fraser, R.D., & MacRae, T.P. (1971) *Nature* 233, 138–140.
306. Fraser, R.D., MacRae, T.P., & Suzuki, E. (1976) *J. Mol. Biol.* 108, 435–452.
307. Steinert, P.M., Marekov, L.N., & Parry, D.A. (1993) *J. Biol. Chem.* 268, 24916–24925.
308. Steinert, P.M., Marekov, L.N., & Parry, D.A. (1993) *Biochemistry* 32, 10046–10056.
309. Geisler, N., & Weber, K. (1982) *EMBO J.* 1, 1649–1656.
310. Fraser, R.D.B., & MacRae, T.P. (1983) *Biosci. Rep.* 3, 517–525.
311. Fraser, R.D., & MacRae, T.P. (1985) *Biosci. Rep.* 5, 573–579.
312. Parry, D.A., & Steinert, P.M. (1999) *Q. Rev. Biophys.* 32, 99–187.
313. Serpell, L.C., Blake, C.C., & Fraser, P.E. (2000) *Biochemistry* 39, 13269–13275.
314. Blake, C., & Serpell, L. (1996) *Structure* 4, 989–998.
315. Sunde, M., Serpell, L.C., Bartlam, M., Fraser, P.E., Pepys, M.B., & Blake, C.C. (1997) *J. Mol. Biol.* 273, 729–739.
316. Groll, M., Ditzel, L., Lowe, J., Stock, D., Bochtler, M., Bartunik, H.D., & Huber, R. (1997) *Nature* 386, 463–471.
317. Wyss, M., Schlegel, J., James, P., Eppenberger, H.M., & Wallimann, T. (1990) *J. Biol. Chem.* 265, 15900–15908.
318. Krause, K.L., Volz, K.W., & Lipscomb, W.N. (1985) *Proc. Natl. Acad. Sci. U.S.A.* 82, 1643–1647.
319. Ke, H.M., Lipscomb, W.N., Cho, Y.J., & Honzatko, R.B. (1988) *J. Mol. Biol.* 204, 725–747.
320. Kawashima, T., Berthet-Colominas, C., Wulff, M., Cusack, S., & Leberman, R. (1996) *Nature* 379, 511–518.
321. Knight, S., Andersson, I., & Branden, C.I. (1990) *J. Mol. Biol.* 215, 113–160.
322. Wasmann, C.C., Ramage, R.T., Bohnert, H.J., & Ostrem, J.A. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 1198–1202.
323. Schneider, G., Lindqvist, Y., Braenden, C.I., & Lorimer, G. (1986) *EMBO J.* 5, 3409–3415.
324. Garman, S.C., Wurzburg, B.A., Tarchevskaya, S.S., Kinet, J.P., & Jardetzky, T.S. (2000) *Nature* 406, 259–266.
325. Monaco, H.L., Rizzi, M., & Coda, A. (1995) *Science* 268, 1039–1041.
326. Wagenknecht, T., Francis, N., & DeRosier, D.J. (1983) *J. Mol. Biol.* 165, 523–539.
327. Mande, S.S., Sarfaty, S., Allen, M.D., Perham, R.N., & Hol, W.G. (1996) *Structure* 4, 277–286.
328. Maeng, C.Y., Yazdi, M.A., & Reed, L.J. (1996) *Biochemistry* 35, 5879–5882.

329. Sondermann, P., Huber, R., Oosthuizen, V., & Jacob, U. (2000) *Nature* 406, 267–273.
330. de Vos, A.M., Ultsch, M., & Kossiakoff, A.A. (1992) *Science* 255, 306–312.
331. Somers, W., Ultsch, M., De Vos, A.M., & Kossiakoff, A.A. (1994) *Nature* 372, 478–481.
332. Ding, Y.H., Smith, K.J., Garboczi, D.N., Utz, U., Biddison, W.E., & Wiley, D.C. (1998) *Immunity* 8, 403–411.
333. Garboczi, D.N., Ghosh, P., Utz, U., Fan, Q.R., Biddison, W.E., & Wiley, D.C. (1996) *Nature* 384, 134–141.
334. Tao, M., Salas, M.L., & Lipmann, F. (1970) *Proc. Natl. Acad. Sci. U.S.A.* 67, 408–414.
335. Kumon, A., Yamamura, H., & Nishizuka, Y. (1970) *Biochem. Biophys. Res. Commun.* 41, 1290–1297.
336. Aeschlimann, D., & Paulsson, M. (1991) *J. Biol. Chem.* 266, 15308–15317.
337. Jeon, H., Meng, W., Takagi, J., Eck, M.J., Springer, T.A., & Blacklow, S.C. (2001) *Nat. Struct. Biol.* 8, 499–504.
338. Atkinson, R.A., Joseph, C., Kelly, G., Muskett, F.W., Frenkiel, T.A., Nietlispach, D., & Pastore, A. (2001) *Nat. Struct. Biol.* 8, 853–857.
339. Huang, X., Poy, F., Zhang, R., Joachimiak, A., Sudol, M., & Eck, M.J. (2000) *Nat. Struct. Biol.* 7, 634–638.
340. Fuentes-Prior, P., Iwanaga, Y., Huber, R., Pagila, R., Rumennik, G., Seto, M., Morser, J., Light, D.R., & Bode, W. (2000) *Nature* 404, 518–525.
341. Nakasako, M., Odaka, M., Yohda, M., Dohmae, N., Takio, K., Kamiya, N., & Endo, I. (1999) *Biochemistry* 38, 9887–9898.
342. Sutton, R.B., Fasshauer, D., Jahn, R., & Brunger, A.T. (1998) *Nature* 395, 347–353.
343. Derrick, J.P., & Wigley, D.B. (1992) *Nature* 359, 752–754.
344. Vigers, G.P., Anderson, L.J., Caffes, P., & Brandhuber, B.J. (1997) *Nature* 386, 190–194.
345. Kobe, B., & Deisenhofer, J. (1995) *Nature* 374, 183–186.
346. Chook, Y.M., & Blobel, G. (1999) *Nature* 399, 230–237.
347. Lo Conte, L., Chothia, C., & Janin, J. (1999) *J. Mol. Biol.* 285, 2177–2198.
348. Boyington, J.C., Motyka, S.A., Schuck, P., Brooks, A.G., & Sun, P.D. (2000) *Nature* 405, 537–543.
349. Wu, G., Chen, Y.G., Ozdamar, B., Gyuricza, C.A., Chong, P.A., Wrana, J.L., Massague, J., & Shi, Y. (2000) *Science* 287, 92–97.
350. Van Eyk, J.E., Kay, C.M., & Hodges, R.S. (1991) *Biochemistry* 30, 9974–9981.
351. Nardese, V., Longhi, R., Polo, S., Sironi, F., Arcelloni, C., Paroni, R., DeSantis, C., Sarmientos, P., Rizzi, M., Bolognesi, M., Pavone, V., & Lusso, P. (2001) *Nat. Struct. Biol.* 8, 611–615.
352. Cingolani, G., Petosa, C., Weis, K., & Muller, C.W. (1999) *Nature* 399, 221–229.
353. Becker, T., Weber, K., & Johnsson, N. (1990) *EMBO J.* 9, 4207–4213.
354. McWhirter, S.M., Pullen, S.S., Holton, J.M., Crute, J.J., Kehry, M.R., & Alber, T. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 8408–8413.
355. Harris, B.Z., Hillier, B.J., & Lim, W.A. (2001) *Biochemistry* 40, 5921–5930.
356. Dingwall, C., & Laskey, R.A. (1991) *Trends Biochem. Sci.* 16, 478–481.
357. Conti, E., Uy, M., Leighton, L., Blobel, G., & Kuriyan, J. (1998) *Cell* 94, 193–204.
358. de Beer, T., Carter, R.E., Lobel-Rice, K.E., Sorkin, A., & Overduin, M. (1998) *Science* 281, 1357–1360.
359. Takeichi, M. (1990) *Annu. Rev. Biochem.* 59, 237–252.
360. van der Flier, A., & Sonnenberg, A. (2001) *Cell Tissue Res.* 305, 285–298.
361. Ozawa, M., & Kemler, R. (1992) *J. Cell Biol.* 116, 989–996.
362. Shapiro, L., Fannon, A.M., Kwong, P.D., Thompson, A., Lehmann, M.S., Grubel, G., Legrand, J.F., Als-Nielsen, J., Colman, D.R., & Hendrickson, W.A. (1995) *Nature* 374, 327–337.
363. Nagar, B., Overduin, M., Ikura, M., & Rini, J.M. (1996) *Nature* 380, 360–364.
364. Hynes, R.O. (1987) *Cell* 48, 549–554.
365. Wayner, E.A., Carter, W.G., Piotrowicz, R.S., & Kunicki, T.J. (1988) *J. Cell Biol.* 107, 1881–1891.
366. Carter, W.G., Wayner, E.A., Bouchard, T.S., & Kaur, P. (1990) *J. Cell Biol.* 110, 1387–1404.
367. Camper, L., Heinegard, D., & Lundgren-Akerlund, E. (1997) *J. Cell Biol.* 138, 1159–1167.
368. Loo, D.T., Kanner, S.B., & Aruffo, A. (1998) *J. Biol. Chem.* 273, 23304–23312.
369. Otey, C.A., Vasquez, G.B., Burrige, K., & Erickson, B.W. (1993) *J. Biol. Chem.* 268, 21193–21197.
370. Reddy, K.B., Gascard, P., Price, M.G., Negrescu, E.V., & Fox, J.E. (1998) *J. Biol. Chem.* 273, 35039–35047.
371. Chang, D.D., Wong, C., Smith, H., & Liu, J. (1997) *J. Cell Biol.* 138, 1149–1157.
372. Liliental, J., & Chang, D.D. (1998) *J. Biol. Chem.* 273, 2379–2383.
373. Schaller, M.D., Otey, C.A., Hildebrand, J.D., & Parsons, J.T. (1995) *J. Cell Biol.* 130, 1181–1187.
374. Hannigan, G.E., Leung-Hagesteijn, C., Fitz-Gibbon, L., Coppolino, M.G., Radeva, G., Filmus, J., Bell, J.C., & Dedhar, S. (1996) *Nature* 379, 91–96.
375. Lenter, M., & Vestweber, D. (1994) *J. Biol. Chem.* 269, 12263–12268.
376. Chung, A.E., Jaffe, R., Freeman, I.L., Vergnes, J.P., Braginski, J.E., & Carlin, B. (1979) *Cell* 16, 277–287.
377. Timpl, R., Rohde, H., Robey, P.G., Rennard, S.I., Foidart, J.M., & Martin, G.R. (1979) *J. Biol. Chem.* 254, 9933–9937.
378. Fox, J.W., Mayer, U., Nischt, R., Aumailley, M., Reinhardt, D., Wiedemann, H., Mann, K., Timpl, R., Krieg, T., Engel, J., & et al. (1991) *EMBO J.* 10, 3137–3146.
379. Mumby, S.M., Raugi, G.J., & Bornstein, P. (1984) *J. Cell Biol.* 98, 646–652.
380. Holmes, R. (1967) *J. Cell Biol.* 32, 297–308.
381. Barnes, D.W., Silnutzer, J., See, C., & Shaffer, M. (1983) *Proc. Natl. Acad. Sci. U.S.A.* 80, 1362–1366.
382. Hayman, E.G., Pierschbacher, M.D., Ohgren, Y., & Ruoslahti, E. (1983) *Proc. Natl. Acad. Sci. U.S.A.* 80, 4003–4007.
383. Vogel, B.E., Tarone, G., Giancotti, F.G., Gailit, J., & Ruoslahti, E. (1990) *J. Biol. Chem.* 265, 5934–5937.
384. Pytela, R., Pierschbacher, M.D., & Ruoslahti, E. (1985) *Proc. Natl. Acad. Sci. U.S.A.* 82, 5766–5770.
385. Bodary, S.C., & McLean, J.W. (1990) *J. Biol. Chem.* 265, 5938–5941.

386. Suzuki, S., Pierschbacher, M.D., Hayman, E.G., Nguyen, K., Ohgren, Y., & Ruoslahti, E. (1984) *J. Biol. Chem.* 259, 15307–15314.
387. Owensby, D.A., Morton, P.A., Wun, T.C., & Schwartz, A.L. (1991) *J. Biol. Chem.* 266, 4334–4340.
388. Bennett, V., & Stenbuck, P.J. (1979) *J. Biol. Chem.* 254, 2533–2541.
389. Lux, S.E., John, K.M., & Bennett, V. (1990) *Nature* 344, 36–42.
390. Hargreaves, W.R., Giedd, K.N., Verkleij, A., & Branton, D. (1980) *J. Biol. Chem.* 255, 11965–11972.
391. Bennett, V., & Stenbuck, P.J. (1980) *J. Biol. Chem.* 255, 2540–2548.
392. Koob, R., Zimmermann, M., Schoner, W., & Drenckhahn, D. (1988) *Eur. J. Cell Biol.* 45, 230–237.
393. Srinivasan, Y., Elmer, L., Davis, J., Bennett, V., & Angelides, K. (1988) *Nature* 333, 177–180.
394. Dubreuil, R.R., MacVicar, G., Dissanayake, S., Liu, C., Homer, D., & Hortsch, M. (1996) *J. Cell Biol.* 133, 647–655.
395. Kalomiris, E.L., & Bourguignon, L.Y. (1988) *J. Cell Biol.* 106, 319–327.
396. Ebashi, S., & Ebashi, F. (1965) *J. Biochem. (Tokyo)* 58, 7–12.
397. Belkin, A.M., & Koteliensky, V.E. (1987) *FEBS Lett.* 220, 291–294.
398. Wilkins, J.A., Chen, K.Y., & Lin, S. (1983) *Biochem. Biophys. Res. Commun.* 116, 1026–1032.
399. Wachstock, D.H., Wilkins, J.A., & Lin, S. (1987) *Biochem. Biophys. Res. Commun.* 146, 554–560.
400. Ohtsuka, H., Yajima, H., Maruyama, K., & Kimura, S. (1997) *FEBS Lett.* 401, 65–67.
401. Joseph, C., Stier, G., O'Brien, R., Politou, A.S., Atkinson, R.A., Bianco, A., Ladbury, J.E., Martin, S.R., & Pastore, A. (2001) *Biochemistry* 40, 4957–4965.
402. Yin, H.L., & Stossel, T.P. (1979) *Nature* 281, 583–586.
403. McLaughlin, P.J., Gooch, J.T., Mannherz, H.G., & Weeds, A.G. (1993) *Nature* 364, 685–692.
404. Kamada, S., Kusano, H., Fujita, H., Ohtsu, M., Koya, R.C., Kuzumaki, N., & Tsujimoto, Y. (1998) *Proc. Natl. Acad. Sci. U.S.A.* 95, 8532–8537.
405. Matthew, W.D., Tsavaler, L., & Reichardt, L.F. (1981) *J. Cell Biol.* 91, 257–269.
406. Perin, M.S., Fried, V.A., Mignery, G.A., Jahn, R., & Sudhof, T.C. (1990) *Nature* 345, 260–263.
407. Petrenko, A.G., Perin, M.S., Davletov, B.A., Ushkaryov, Y.A., Geppert, M., & Sudhof, T.C. (1991) *Nature* 353, 65–68.
408. Bennett, M.K., Calakos, N., & Scheller, R.H. (1992) *Science* 257, 255–259.
409. Zhang, J.Z., Davletov, B.A., Sudhof, T.C., & Anderson, R.G. (1994) *Cell* 78, 751–760.
410. Bugler, B., Caizergues-Ferrer, M., Bouche, G., Bourbon, H., & Amalric, F. (1982) *Eur. J. Biochem.* 128, 475–480.
411. Lapeyre, B., Bourbon, H., & Amalric, F. (1987) *Proc. Natl. Acad. Sci. U.S.A.* 84, 1472–1476.
412. Orrick, L.R., Olson, M.O., & Busch, H. (1973) *Proc. Natl. Acad. Sci. U.S.A.* 70, 1316–1320.
413. Olson, M.O., & Thompson, B.A. (1983) *Biochemistry* 22, 3187–3193.
414. Prestayko, A.W., Klomp, G.R., Schmoll, D.J., & Busch, H. (1974) *Biochemistry* 13, 1945–1951.
415. Burks, D.J., Wang, J., Towery, H., Ishibashi, O., Lowe, D., Riedel, H., & White, M.F. (1998) *J. Biol. Chem.* 273, 31061–31067.
416. Li, Y.P., Busch, R.K., Valdez, B.C., & Busch, H. (1996) *Eur. J. Biochem.* 237, 153–158.
417. Kraemer, D., Wozniak, R.W., Blobel, G., & Radu, A. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 1519–1523.
418. Matsubayashi, Y., Fukuda, M., & Nishida, E. (2001) *J. Biol. Chem.* 276, 41755–41760.
419. Katahira, J., Strasser, K., Podtelejnikov, A., Mann, M., Jung, J.U., & Hurt, E. (1999) *EMBO J.* 18, 2593–2609.
420. Kehlenbach, R.H., Dickmanns, A., Kehlenbach, A., Guan, T., & Gerace, L. (1999) *J. Cell Biol.* 145, 645–657.
421. Kasper, L.H., Brindle, P.K., Schnabel, C.A., Pritchard, C.E., Cleary, M.L., & van Deursen, J.M. (1999) *Mol. Cell Biol.* 19, 764–776.
422. Turner, M.J., Cresswell, P., Parham, P., Strominger, J.L., Mann, D.L., & Sanderson, A.R. (1975) *J. Biol. Chem.* 250, 4512–4519.
423. Saper, M.A., Bjorkman, P.J., & Wiley, D.C. (1991) *J. Mol. Biol.* 219, 277–319.
424. Evans, T., Rosenthal, E.T., Youngblom, J., Distel, D., & Hunt, T. (1983) *Cell* 33, 389–396.
425. Connell-Crowley, L., Solomon, M.J., Wei, N., & Harper, J.W. (1993) *Mol. Biol. Cell* 4, 79–92.
426. Pagano, M., Pepperkok, R., Verde, F., Ansorge, W., & Draetta, G. (1992) *EMBO J.* 11, 961–971.
427. Ohtoshi, A., Maeda, T., Higashi, H., Ashizawa, S., & Hatakeyama, M. (2000) *Biochem. Biophys. Res. Commun.* 268, 530–534.
428. Ohtoshi, A., Maeda, T., Higashi, H., Ashizawa, S., Yamada, M., & Hatakeyama, M. (2000) *Biochem. Biophys. Res. Commun.* 267, 947–952.
429. Pelicci, G., Lanfranccone, L., Grignani, F., McGlade, J., Cavallo, F., Forni, G., Nicoletti, I., Grignani, F., Pawson, T., & Pelicci, P.G. (1992) *Cell* 70, 93–104.
430. Zhou, M.M., Ravichandran, K.S., Olejniczak, E.F., Petros, A.M., Meadows, R.P., Sattler, M., Harlan, J.E., Wade, W.S., Burakoff, S.J., & Fesik, S.W. (1995) *Nature* 378, 584–592.
431. Cohen, G.B., Ren, R., & Baltimore, D. (1995) *Cell* 80, 237–248.
432. Zhou, S., Margolis, B., Chaudhuri, M., Shoelson, S.E., & Cantley, L.C. (1995) *J. Biol. Chem.* 270, 14863–14866.
433. Pawson, T. (1995) *Nature* 373, 573–580.
434. Rozakis-Adcock, M., McGlade, J., Mbamalu, G., Pelicci, G., Daly, R., Li, W., Batzer, A., Thomas, S., Brugge, J., Pelicci, P.G., et al. (1992) *Nature* 360, 689–692.
435. Schmandt, R., Liu, S.K., & McGlade, C.J. (1999) *Oncogene* 18, 1867–1879.
436. Charest, A., Wagner, J., Jacob, S., McGlade, C.J., & Tremblay, M.L. (1996) *J. Biol. Chem.* 271, 8424–8429.
437. Liu, S.K., & McGlade, C.J. (1998) *Oncogene* 17, 3073–3082.
438. Eckhart, W., Hutchinson, M.A., & Hunter, T. (1979) *Cell* 18, 925–933.
439. Reddy, E.P., Smith, M.J., & Srinivasan, A. (1983) *Proc. Natl. Acad. Sci. U.S.A.* 80, 3623–3627.
440. Goff, S.P., Gilboa, E., Witte, O.N., & Baltimore, D. (1980) *Cell* 22, 777–785.
441. Fainstein, E., Einat, M., Gokkel, E., Marcelle, C., Croce,

- C.M., Gale, R.P., & Canaani, E. (1989) *Oncogene* 4, 1477–1481.
442. Yu, H.H., Zisch, A.H., Dodelet, V.C., & Pasquale, E.B. (2001) *Oncogene* 20, 3995–4006.
 443. Westphal, R.S., Soderling, S.H., Alto, N.M., Langeberg, L.K., & Scott, J.D. (2000) *EMBO J.* 19, 4589–4600.
 444. Van Etten, R.A., Jackson, P.K., Baltimore, D., Sanders, M.C., Matsudaira, P.T., & Janmey, P.A. (1994) *J. Cell Biol.* 124, 325–340.
 445. Dai, Z., & Pendergast, A.M. (1995) *Genes Dev.* 9, 2569–2582.
 446. Ren, R., Mayer, B.J., Cicchetti, P., & Baltimore, D. (1993) *Science* 259, 1157–1161.
 447. Musacchio, A., Saraste, M., & Wilmanns, M. (1994) *Nat. Struct. Biol.* 1, 546–551.
 448. Pires, E.M., & Perry, S.V. (1977) *Biochem. J.* 167, 137–146.
 449. Yazawa, M., Kuwayama, H., & Yagi, K. (1978) *J. Biochem. (Tokyo)* 84, 1253–1258.
 450. Charbonneau, H., Tonks, N.K., Walsh, K.A., & Fischer, E.H. (1988) *Proc. Natl. Acad. Sci. U.S.A.* 85, 7182–7186.
 451. Herold, C., Elhabazi, A., Bismuth, G., Bensussan, A., & Boumsell, L. (1996) *J. Immunol.* 157, 5262–5268.
 452. Verhagen, A.M., Schraven, B., Wild, M., Wallich, R., & Meuer, S.C. (1996) *Eur. J. Immunol.* 26, 2841–2849.
 453. Marie-Cardine, A., Maridonneau-Parini, I., & Fischer, S. (1994) *Eur. J. Immunol.* 24, 1255–1261.
 454. Pfeuffer, T. (1977) *J. Biol. Chem.* 252, 7224–7234.
 455. Northup, J.K., Sternweis, P.C., Smigel, M.D., Schleifer, L.S., Ross, E.M., & Gilman, A.G. (1980) *Proc. Natl. Acad. Sci. U.S.A.* 77, 6516–6520.
 456. Brandt, D.R., Asano, T., Pedersen, S.E., & Ross, E.M. (1983) *Biochemistry* 22, 4357–4362.
 457. Davison, B.L., Egly, J.M., Mulvihill, E.R., & Chambon, P. (1983) *Nature* 301, 680–686.
 458. Geiger, J.H., Hahn, S., Lee, S., & Sigler, P.B. (1996) *Science* 272, 830–836.
 459. Buratowski, S., Hahn, S., Guarente, L., & Sharp, P.A. (1989) *Cell* 56, 549–561.
 460. Nikolov, D.B., Chen, H., Halay, E.D., Usheva, A.A., Hisatake, K., Lee, D.K., Roeder, R.G., & Burley, S.K. (1995) *Nature* 377, 119–128.
 461. Meisterernst, M., Roy, A.L., Lieu, H.M., & Roeder, R.G. (1991) *Cell* 66, 981–993.
 462. Inostroza, J.A., Mermelstein, F.H., Ha, I., Lane, W.S., & Reinberg, D. (1992) *Cell* 70, 477–489.
 463. Roy, A.L., Malik, S., Meisterernst, M., & Roeder, R.G. (1993) *Nature* 365, 355–359.
 464. Marcotte, E.M., Pellegrini, M., Yeates, T.O., & Eisenberg, D. (1999) *J. Mol. Biol.* 293, 151–160.
 465. Maslov, S., & Sneppen, K. (2002) *Science* 296, 910–913.
 466. Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., & Rothberg, J.M. (2000) *Nature* 403, 623–627.
 467. Park, J., Leong, M.L., Buse, P., Maiyar, A.C., Firestone, G.L., & Hemmings, B.A. (1999) *EMBO J.* 18, 3024–3033.
 468. Smith, G.P. (1985) *Science* 228, 1315–1317.
 469. Roberts, B.L., Markland, W., & Ladner, R.C. (1996) *Methods Enzymol.* 267, 68–82.
 470. Fields, S., & Song, O. (1989) *Nature* 340, 245–246.
 471. Johnston, M. (1987) *Microbiol. Rev.* 51, 458–476.
 472. Marmorstein, R., Carey, M., Ptashne, M., & Harrison, S.C. (1992) *Nature* 356, 408–414.
 473. Keegan, L., Gill, G., & Ptashne, M. (1986) *Science* 231, 699–704.
 474. Pause, A., Peterson, B., Schaffar, G., Stearman, R., & Klausner, R.D. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 9533–9538.
 475. Chien, C.T., Bartel, P.L., Sternglanz, R., & Fields, S. (1991) *Proc. Natl. Acad. Sci. U.S.A.* 88, 9578–9582.
 476. Patel, L.R., Curran, T., & Kerppola, T.K. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 7360–7364.
 477. Itoh, Y., Cai, K., & Khorana, H.G. (2001) *Proc. Natl. Acad. Sci. U.S.A.* 98, 4883–4887.
 478. Cai, K., Itoh, Y., & Khorana, H.G. (2001) *Proc. Natl. Acad. Sci. U.S.A.* 98, 4877–4882.
 479. Reidhaar-Olson, J.F., De Souza-Hart, J.A., & Seliak, H.E. (1996) *Biochemistry* 35, 9034–9041.
 480. Jespers, L., Lijnen, H.R., Vanwetswinkel, S., Van Hoef, B., Brepoels, K., Collen, D., & De Maeyer, M. (1999) *J. Mol. Biol.* 290, 471–479.
 481. Cunningham, B.C., Jhurani, P., Ng, P., & Wells, J.A. (1989) *Science* 243, 1330–1336.
 482. Cunningham, B.C., & Wells, J.A. (1989) *Science* 244, 1081–1085.
 483. Salter, R.D., Benjamin, R.J., Wesley, P.K., Buxton, S.E., Garrett, T.P., Clayberger, C., Krensky, A.M., Norment, A.M., Littman, D.R., & Parham, P. (1990) *Nature* 345, 41–46.
 484. Park, C.S., & Miller, C. (1992) *Biochemistry* 31, 7749–7755.
 485. Kraulis, P.J. (1991) *J. Applied Crystallogr.* 24, 946–950.

Chapter 10

Chemical Probes of Structure

Although there are systematic programs to crystallize as many of the available proteins as possible, crystallographic molecular models have been obtained so far for only a minority of the proteins that have been purified. This is not only a problem of time. There are a number of technical problems associated with the crystallographic method, and these have proved baffling in many instances. Often a purified protein has not been crystallized. Often the crystals of protein deteriorate too rapidly in the beam of X-rays. Often the space group is too complex to permit a ready solution. Often the crystals of a particular protein are microscopically disordered. Often usable isomorphous replacements have not been obtained.

There are indications that many of these problems will be solved eventually. It seems that the extensive purifications required for proteins that are naturally present in low concentrations produces unnoticed alterations leading to undetected heterogeneities in the final preparation that interfere with crystallization. When such a protein is produced at high concentrations in an appropriate expression system so that only a few steps of purification are required, attempts to crystallize that protein become significantly more successful. Crystals of membrane-spanning proteins, a class of proteins that are quite difficult to crystallize, have been obtained more and more frequently and used successfully for high-resolution crystallography. The development of charge-coupled detectors, which permit a complete data set to be gathered in a short period of time, has decreased the required length of exposure to the beam. Changing the conditions of crystallization¹ or the species from which the protein has been purified² can sometimes give crystals that have a different space group or are more ordered. As more reagents containing heavy metals become available, the odds against finding a suitable set of isomorphous replacements decrease. Yet it still seems possible that the majority of the proteins that have been or will be purified may never yield high-resolution maps of electron density.

In the absence of a map of electron density, the molecular structure of a protein is studied by a diverse collection of techniques. These approaches can be conveniently divided into three classes: the use of chemical probes, the use of immunochemical probes, and the use of physical measurements.

Covalent Modification

Several types of amino acids in a molecule of protein are susceptible to covalent modification. As most of the reactive functional groups in a protein are **nucleophiles**, the reagents used to modify proteins are usually **electrophiles**. Serine and threonine contain nucleophilic oxygens, but they are indistinguishable from those of the water and cannot be easily modified. The amino acids possessing nucleophilic sites that can be conveniently modified are cysteines, methionines, lysines, histidines, tyrosines, glutamates, aspartates, arginines, and tryptophans. The electrophilic reagents that are used to modify these amino acids couple the electrophile covalently to them. In the process, the ability of the amino acid to act as an acid, a base, or a donor or acceptor of a hydrogen bond is usually lost because an atom of the electrophile, usually a carbon, forms a covalent bond to the conjugate base of the central heteroatom. The proton that previously occupied the conjugate acid of the conjugate base, which is the smallest possible atom, is replaced with the whole molecular structure of the electrophile, and this also increases the size of the side chain dramatically. After its modification, the amino acid is no longer able to participate in any particular role it might have had in the function of the protein and is no longer able to fit into the same space. Accordingly, the function or the structure of the protein or both is usually disrupted.

Chemical modifications of amino acids in a protein are used for many different purposes. Most of the uses are designed imaginatively to answer a particular question about a particular protein, so it is impossible to give an exhaustive list of the reasons for covalently modifying amino acids in proteins. A few examples, however, will indicate why such experiments are so common.

The most common purpose for using covalent modification is to demonstrate that a particular type of amino acid is involved in the **function** of the protein. For example, fibrinogen, upon activation, polymerizes to form long helical polymers that produce a clot. The initial polymerization is noncovalent, but the polymer is then strengthened by posttranslational cross-links. When the lysines of fibrinogen were modified by amidination, the initial noncovalent polymer could form normally, but the covalent cross-links could not.³ This evidence was the basis for the prediction that the posttranslational cross-links were amides between gluta-

mates and lysines. The most common use of covalent modification to study the function of a protein is the observation of the inactivation of an enzyme by covalent modification of amino acids in its active site. For example, the chemical modification of Lysine 116 in spinach ferredoxin-NADP⁺ reductase inactivates the enzyme.⁴

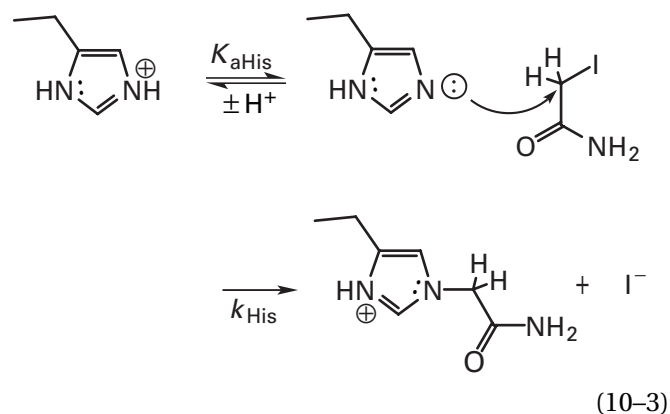
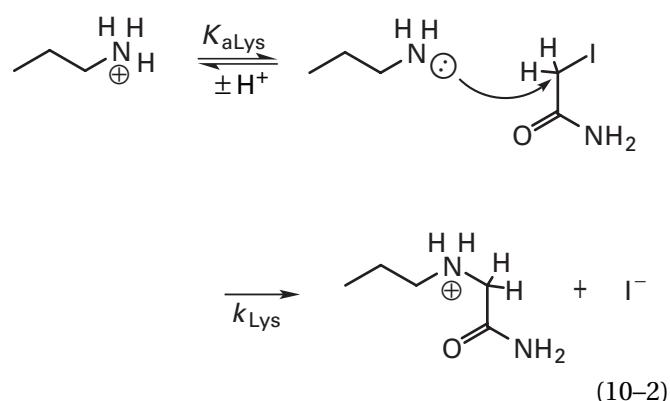
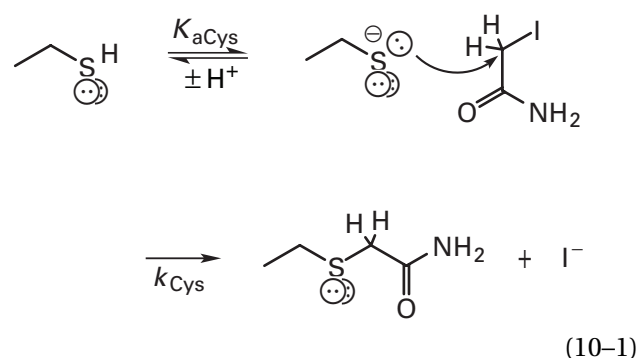
There are, however, many other purposes for covalently modifying proteins. Covalent modification can be used to **dissociate the subunits** of a protein. For example, succinylation of its lysines caused the hemerythrin of *Goldfingia gouldi* to dissociate into its eight identical subunits.⁵ Covalent modification can also be used to change the electrophoretic mobility of a protein by converting, for example, positively charged lysines into negatively charged carboxylates.⁶ When such a modification is performed reversibly, the protein will travel with a different electrophoretic mobility before the modification has been reversed than after it has been reversed. Covalent modification of a protein can be used to prevent endopeptidolytic enzymes, for example, trypsin, from digesting that protein at particular amino acids, for example, arginine.⁷ Covalent modifications are also used to **introduce foreign functional groups** into proteins. For example, functional groups that absorb visible light⁸ or have strong fluorescence⁹ may be introduced so that their spectral properties can be used in physical studies.

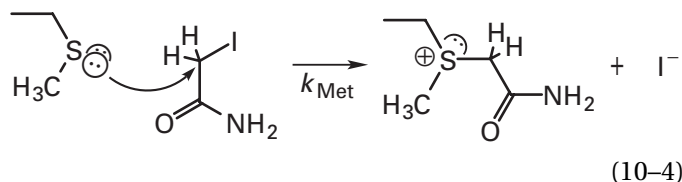
When a protein is modified covalently, either the modification is performed under conditions that produce a **high yield** of the desired product, and this is used in further experiments, or the chemical reaction itself between the protein and the reagent is monitored, and its **kinetics** are used to make arguments about the properties of a particular amino acid in the native protein. For example, the dependence of the rate of the modification of a particular amino acid on the pH of the solution can be used to estimate the pK_a of that amino acid in the native protein.^{10,11} Differences in rate constants for the reaction of amino acids of a particular type with the same electrophile can be used to assess differences in their accessibility to the solution in the native structure of the protein.¹²

In all of these experiments, the possibility that the covalent modification itself disrupts the **global conformation** of the protein must always be kept in mind. If this happens, effects of the modification on the function of the protein might be attributed to local changes around the specific amino acid that has been modified when they actually result from the disruption of the entire structure of the protein. The modified protein should always be examined to rule out this possibility. For example, it could be shown by following electrophoretic mobility, sedimentation rate in the ultracentrifuge, and optical rotation that exhaustive amidination of the lysines in either serum albumin or immunoglobulin G had no measurable effect on the structures of these proteins.¹³

In designing an experiment that involves the covalent modification of a protein, the usual desire is that the reagent chosen react with only one type of amino acid. Because cysteines, methionines, lysines, histidines, and tyrosines are similarly reactive nucleophiles, this is not a simple task. The issue of **specificity** is best addressed by examining the reaction of a simple alkylating agent, such as iodoacetamide, with the nucleophiles present in a protein.

Iodoacetamide has been shown to alkylate cysteines, lysines, histidines, and methionines.¹⁴ The four reactions are





When a protein is exposed to iodoacetamide, all four of these amino acids disappear from amino acid analyses of the reaction mixtures, at **rates that depend on the pH**,¹⁵ and the carboxymethyl products^{16,17} appear in concert. The first three reactions require that the amino acid be in the form of its **conjugate base**.

The rate of the reaction between **lysine and iodoacetamide** can be described by

$$\frac{d[\text{Lys}]_{\text{TOT}}}{dt} = -k_{\text{Lys}} [\text{iodoacetamide}] [\text{RH}_2\text{N}^\ominus] \quad (10-5)$$

where $[\text{Lys}]_{\text{TOT}}$ is the total concentration of unmodified lysine (both protonated, RNH_3^+ , and unprotonated, $\text{RH}_2\text{N}^\ominus$) at a particular time, t . Substitution of the appropriate terms into Equation 10-5 leads to

$$\frac{d[\text{Lys}]_{\text{TOT}}}{dt} = - \left(\frac{k_{\text{Lys}} K_{\text{aLys}}}{K_{\text{aLys}} + [\text{H}^+]} \right) [\text{iodoacetamide}] [\text{Lys}]_{\text{TOT}} \quad (10-6)$$

where K_{aLys} is the acid dissociation constant for lysine. If the concentration of iodoacetamide is so large that it remains constant throughout the reaction and the pH does not change, Equation 10-6 describes a pseudo-first-order reaction. The pseudo-first-order rate constant, k'_{Lys} , governing the disappearance of lysine with time is

$$k'_{\text{Lys}} = \left(\frac{k_{\text{Lys}} K_{\text{aLys}}}{K_{\text{aLys}} + [\text{H}^+]} \right) [\text{iodoacetamide}] \quad (10-7)$$

and

$$\frac{d[\text{Lys}]_{\text{TOT}}}{dt} = -k'_{\text{Lys}} [\text{Lys}]_{\text{TOT}} \quad (10-8)$$

When this is rearranged

$$\frac{d[\text{Lys}]_{\text{TOT}}}{[\text{Lys}]_{\text{TOT}}} = -k'_{\text{Lys}} dt \quad (10-9)$$

and integrated from $t = 0$ to $t = t$

$$\int_{[\text{Lys}]_{0,\text{TOT}}}^{[\text{Lys}]_{\text{TOT}}} \frac{d[\text{Lys}]_{\text{TOT}}}{[\text{Lys}]_{\text{TOT}}} = \int_0^t -k'_{\text{Lys}} dt \quad (10-10)$$

where $[\text{Lys}]_{0,\text{TOT}}$ is the initial concentration of unmodified lysine

$$\ln \frac{[\text{Lys}]_{\text{TOT}}}{[\text{Lys}]_{0,\text{TOT}}} = -k'_{\text{Lys}} t \quad (10-11)$$

It follows that

$$\ln [\text{Lys}]_{\text{TOT}} = \ln [\text{Lys}]_{0,\text{TOT}} - k'_{\text{Lys}} t \quad (10-12)$$

and

$$[\text{Lys}]_{\text{TOT}} = [\text{Lys}]_{0,\text{TOT}} \exp(-k'_{\text{Lys}} t) \quad (10-13)$$

As in all **first-order reactions**, the rate constant k'_{Lys} is estimated either by fitting the disappearance of unmodified lysine to Equation 10-13 by nonlinear least-squares analysis or from the slope of the line obtained when $\ln [\text{Lys}]_{\text{TOT}}$ is plotted against time.

The other two of the first three reactions (Reactions 10-1 through 10-3) have a formally equivalent mechanism, and the **pseudo-first-order rate constants** of each of them, k'_{Cys} and k'_{His} , are of the same form as k'_{Lys} (Equation 10-7) with the appropriate rate constants, k_{Cys} or k_{His} , and acid dissociation constants, K_{aCys} or K_{aHis} , substituted for k_{Lys} and K_{aLys} , respectively. The variation in each of these rate constants can be presented graphically (Figure 10-1).¹⁵

At values of pH greater than the $\text{p}K_{\text{a}}$ of lysine ($[\text{H}^+] < K_{\text{aLys}}$), almost all of the primary amine is the conjugate base, the rate constant k'_{Lys} is equal to k_{Lys} (Equation 10-7), and the rate of the reaction of lysine with iodoacetamide is independent of pH. At values of pH below $\text{p}K_{\text{aLys}}$ ($[\text{H}^+] > K_{\text{aLys}}$), the concentration of the unprotonated conjugate base of lysine, and hence the rate of its reaction with iodoacetamide (Equation 10-7), decreases by a factor of 10 (1 logarithmic unit) for each decrease of 1 unit in pH (Figure 2-6).

The rate of the reaction of the unprotonated conjugate base of **histidine with iodoacetamide**, which is governed solely by k_{His} (Reaction 10-3), is correlated to the rate of the reaction of the unprotonated conjugate base of lysine with iodoacetamide, k_{Lys} (Reaction 10-2), through the Brønsted relationship:

$$\log \left(\frac{k_{\text{His}}}{k_{\text{Lys}}} \right) = -\beta \log \left(\frac{K_{\text{aHis}}}{K_{\text{aLys}}} \right) \quad (10-14)$$

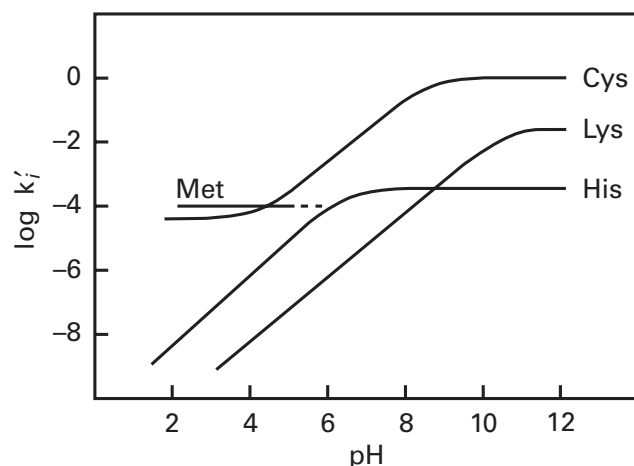


Figure 10-1: Variation with pH of the logarithm of the pseudo-first-order rate constants, k'_i , for the reaction between a fully accessible and unperturbed methionine, cysteine, histidine, or lysine in a protein with iodoacetamide. The first-order rate constants are on a relative scale with the rate constant of cysteinate anion arbitrarily assigned a value of 1.0. The individual lines were drawn according to the respective forms of Equation 10-7 with values of the pK_a of 6.6 for histidine, 8.7 for cysteine, and 10.5 for lysine. The relative vertical positions of the lines were fixed by assuming that methionine reacts 10 times more slowly than cysteine at pH 5.5, histidine reacts 30 times more slowly than cysteine at pH 5.5, methionine reacts at the same rate at pH 5.5 and 8.5, lysine reacts 2.5 times more quickly than methionine at pH 8.5, and histidine reacts 2 times more slowly than methionine at pH 8.5.¹⁵ The rate of the reaction with cysteine below pH 4 is assumed to be invariant with pH, as is the rate of the reaction with methionine, but with a rate constant less than that of methionine.

In the case illustrated, below pH 8, histidine reacts more rapidly than lysine, while above pH 8, lysine reacts more rapidly than histidine. Such an **inversion in reactivity** occurs because β is usually between 1.0 and 0. If β were 0, the unprotonated conjugate bases of both histidine and lysine would react at equal rates with iodoacetamide ($k_{\text{His}} = k_{\text{Lys}}$), and the curves for lysine and histidine would coincide at values of pH above the pK_a of lysine. If β were equal to 1.0 (Equation 10-14), the rate constants for the modification of both lysine, k'_{Lys} , and histidine, k'_{His} (Equation 10-7), would be equal at low pH and the lines for histidine and lysine in Figure 10-1 would coincide at values of pH less than $pK_{a\text{His}}$. In any case, at values of pH below the pK_a of histidine, the pseudo-first-order rate constant, k'_{His} , for its reaction with iodoacetamide decreases by a factor of 10 for each decrease of 1 unit in pH. At low pH, therefore, the rates of the reactions of both lysine and histidine with iodoacetamide decrease in concert and remain in constant ratio to each other.

The rate constant, k_{Cys} , for the reaction of the unprotonated conjugate base of **cysteine with iodoacetamide** is significantly greater than that of unprotonated lysine. Because sulfur is an element from the third row and nitrogen is an element from the second row of the periodic table, cysteine is more nucleophilic than

lysine, even though the pK_a (8.7) associated with its lone pair of electrons is less than the pK_a (10.5) associated with the lone pair of electrons on lysine.* With the appropriate substitutions, Equation 10-7 governs the behavior of the rate of the reaction of cysteine with iodoacetamide as a function of pH at values of pH above and below the pK_a of cysteine. Cysteine (Reaction 10-1), however, unlike lysine (Reaction 10-2) and histidine (Reaction 10-3), retains two nucleophilic lone pairs of electrons after its protonation and can react with iodoacetamide as the conjugate acid in a reaction analogous to that of methionine (Reaction 10-4).

The reaction of **methionine with iodoacetamide** is invariant with pH because its acid dissociation constant ($pK_{a\text{Met}} = -9$) is below accessible ranges of pH. The rate constant for the reaction of protonated cysteine with iodoacetamide should be lower than the rate constant, k_{Met} , for the reaction of methionine because of hyperconjugation. At low pH, the rate of the reaction of cysteine with iodoacetamide should level off at a value lower than k_{Met} and also should become invariant with pH because the concentration of neutral cysteine is invariant with pH.

The individual behavior of each of the reactions determines the specificity of iodoacetamide. At the lowest values of pH, methionine is the most reactive amino acid. As the pH is increased, into the range where the concentration of the thiolate anion becomes sufficiently large, cysteine becomes the most reactive amino acid at all higher values of pH. As the pH is increased, histidine becomes as reactive as methionine because the lone pair of electrons of its neutral base ($pK_{a\text{His}} = 6.6$) is so much more basic than those of methionine ($pK_{a\text{Met}} = -9$). At even higher values of pH, lysine becomes more reactive than histidine. All of these consequences determine which amino acid reacts most rapidly with iodoacetamide at a particular pH.

The reagent used to modify the amino acids of a protein may itself also be affected by alterations in pH. Methyl acetimidate is an imidoester that modifies lysines with high specificity (Figure 10-2).¹⁸ The specificity results in part from the fact that while the product of the reaction with lysine is a stable amidine, the analogous products with cysteine, methionine, glutamate, aspartate, tyrosine, and histidine are unstable derivatives of acetate and decompose as quickly as they are produced. The effect of pH on the rate of the reactions between amines and imidoesters has been explained mechanistically (Figure 10-2)¹⁸ with the assumption that the reactive form of the lysine is the free base and the reactive form of the imidoester is the cationic conjugate acid. For methyl acetimidate,¹⁹ $pK_{a\text{AI}} = 7.5$, and the concentration of cationic imidoester should be decreasing as the pH is

* The Brønsted relationship does not apply when the central atoms of the two acid-bases differ.

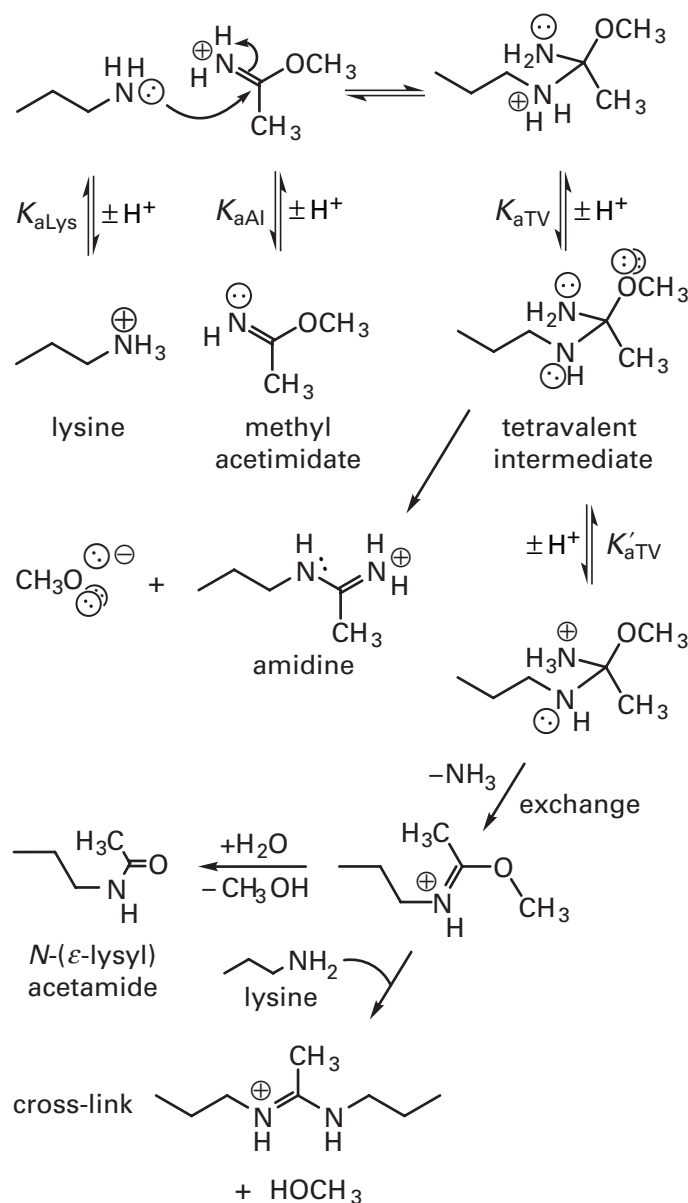


Figure 10-2: Mechanism for the reaction between the free base of lysine and the cationic conjugate acid of methyl acetimidate.¹⁸ The products of the modification can be either the amidine of one lysine, the acetamide of one lysine, or the amidine of two lysines. The latter reaction produces cross-links within the protein but rarely occurs.

raised above pH 7, while that of the free base of lysine should be increasing.

At the higher values of pH where K_{aTV} has little effect, the rate of the reaction between lysine and methyl acetimidate is governed by the equation

$$\frac{d[\text{Lys}]_{\text{TOT}}}{dt} = -k_{\text{AI}} [\text{R}'=\text{NH}_2^+] [\text{RH}_2\text{N}\ominus] \quad (10-15)$$

from which it follows that

$$\frac{d[\text{Lys}]_{\text{TOT}}}{dt} = - \left[\frac{k_{\text{AI}} K_{\text{aLys}} [\text{H}^+]}{(K_{\text{aAI}} + [\text{H}^+])(K_{\text{aLys}} + [\text{H}^+])} \right] [\text{AI}]_{\text{TOT}} [\text{Lys}]_{\text{TOT}} \quad (10-16)$$

where $[\text{AI}]_{\text{TOT}}$ is the total concentration of methyl acetimidate, both conjugate acid, $\text{R}'=\text{NH}_2^+$, and conjugate base, $\text{R}'=\text{NH}$.

If $[\text{AI}]_{\text{TOT}}$ were high and constant throughout the course of the reaction and the pH did not change, Equation 10-16 would describe a pseudo-first-order reaction. The pseudo-first-order rate constant, k'_{AI} , governing the disappearance of lysine with time would be

$$k'_{\text{AI}} = \left[\frac{k_{\text{AI}} K_{\text{aLys}} [\text{H}^+]}{(K_{\text{aAI}} + [\text{H}^+])(K_{\text{aLys}} + [\text{H}^+])} \right] [\text{AI}]_{\text{TOT}} \quad (10-17)$$

When $\text{p}K_{\text{aAI}} < \text{p}K_{\text{aLys}} < \text{pH}$

$$k'_{\text{AI}} = \frac{k_{\text{AI}} [\text{H}^+]}{K_{\text{aAI}}} [\text{AI}]_{\text{TOT}} \quad (10-18)$$

and k'_{AI} would decrease by a factor of 10 for each increase of 1 unit in pH. Between the two acid dissociation constants, when $\text{p}K_{\text{aAI}} < \text{pH} < \text{p}K_{\text{aLys}}$, Equation 10-17 predicts that

$$k'_{\text{AI}} = \frac{k_{\text{AI}} K_{\text{aLys}}}{K_{\text{aAI}}} [\text{AI}]_{\text{TOT}} \quad (10-19)$$

and the rate of the reaction should be almost invariant with pH, and when $\text{pH} < \text{p}K_{\text{aAI}} < \text{p}K_{\text{aLys}}$

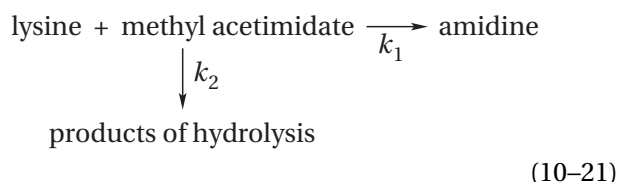
$$k'_{\text{AI}} = \frac{k_{\text{AI}} K_{\text{aLys}}}{[\text{H}^+]} [\text{AI}]_{\text{TOT}} \quad (10-20)$$

and k'_{AI} should decrease by a factor of 10 for every decrease of 1 unit in pH. While the reaction of amines with ethyl benzimidate is more complicated than this simple picture,¹⁸ in the case of the reaction of methyl acetimidate with the lysines in unfolded aldolase,¹⁹ the plateau between $\text{p}K_{\text{aAI}}$ and $\text{p}K_{\text{aLys}}$ is observed as predicted by Equation 10-19, and the decrease in the rate of amidination does not occur until below pH 8.0.

The fact that the modification of a protein is usually performed in **aqueous solution** limits the reagents that can be used. On the one hand, problems of solubility of the electrophile often arise, thereby restricting its useful

range of concentrations. On the other hand, because water is itself a nucleophile, decomposition of the reagent through hydrolysis often occurs. Methyl acetimidate, unlike iodoacetamide, reacts quite readily with water. Between pH 6.8 and 8.4, the rate constant for its hydrolysis at 20 °C is 0.02 min^{-1} .¹⁹

When the reagent chosen for a particular modification decomposes rapidly, measurements of its rate of reaction with the protein are complicated by this **decomposition**. In the case of methyl acetimidate, the situation can be represented by the kinetic mechanism



In this situation, where hydrolysis is occurring coincidentally with modification, it can be shown²⁰ that

$$f_{\text{amidine}} = 1 - \exp\left(-\frac{k_1 [\text{AI}]_{0,\text{TOT}}}{k_2}\right) \quad (10-22)$$

where $[\text{AI}]_{0,\text{TOT}}$ is the initial total molar concentration of methyl acetimidate and f_{amidine} is the fraction of the lysine that has been modified when all of the methyl acetimidate has been consumed either by reaction with lysine or by hydrolysis. From Equation 10-22 it follows that

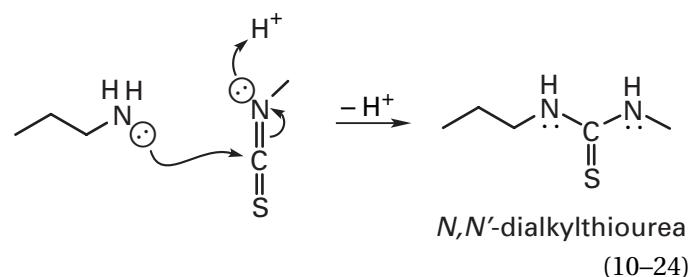
$$\ln\left(\frac{1}{1 - f_{\text{amidine}}}\right) = \frac{k_1}{k_2} [\text{AI}]_{0,\text{TOT}} \quad (10-23)$$

In effect, this relationship allows the reaction to be studied more leisurely. A series of mixtures containing the protein are prepared with increasing concentrations of methyl acetimidate at constant temperature and pH. The reaction is allowed to reach completion. The various fractions of the lysine modified are assessed. From these results and Equation 10-23, estimates of the ratio $k_1 (k_2)^{-1}$ can be determined. If the rate constant k_2 has been measured in a separate experiment under identical conditions, k_1 can be obtained directly.

Alkyl imidoesters, such as **methyl acetimidate**, react specifically and in high yield with **lysine** in proteins. In the case of myoglobin, for example, the two major products of the reaction with methyl acetimidate were proteins amidinated either at every primary amine except Lysine 77 or at every primary amine except the amino terminus.²¹ An amidine ($\text{p}K_{\text{a}} = 12.5$) is positively charged at pH 7, and as a result, no change in the charge of the protein occurs during the amidination of its lysines. The lysines, however, are no longer nucleophilic, and the size of the side chain has increased significantly.

Another way to direct the modification exclusively to lysines is to take advantage of the fact that primary amines such as lysine are the only functional groups on a protein that react with **aldehydes** to form imines (Figure 10-3). The conjugate acid of the resulting imine can then be reduced with sodium borohydride or sodium cyanoborohydride to produce the secondary amine. Both formaldehyde²² and pyridoxal 5'-phosphate²³ have been used as the aldehyde. The former is a more reactive aldehyde and produces much higher yields of alkylated lysine; the latter is more selective and under the proper conditions will modify only the most accessible and nucleophilic lysines in a protein.

Isothiocyanates are also specific for the primary amino groups of lysines, as well as the amino terminus (Figure 3-1), of a protein:



The products of the modification are *N,N'*-dialkylthioureas. Presumably, isothiocyanates are specific for lysine because the products they would form with the other nucleophilic amino acids are unstable under the

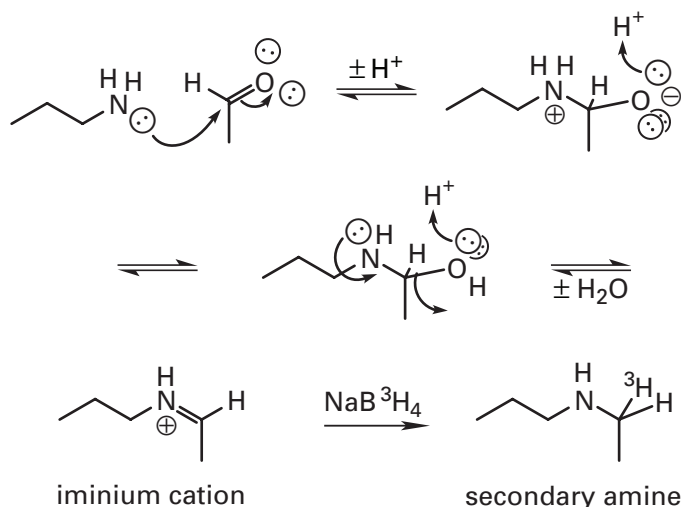
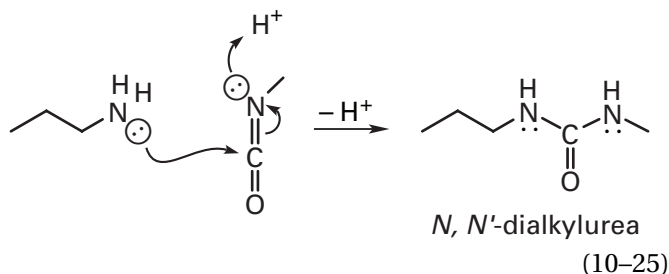


Figure 10-3: Reaction of lysine with an aldehyde to form an iminium cation, which can be reduced to the secondary amine with sodium borohydride (NaBH_4) or the less reactive sodium cyanoborohydride (NaCNBH_3). If tritiated sodium borohydride (NaB^3H_4) or tritiated sodium cyanoborohydride (NaCNB^3H_3) is used, tritium is incorporated into one of the alkyl carbons of the secondary amine.

reaction conditions. A similar reaction occurs with **isocyanates**:



The products are *N,N'*-dialkylureas. Alkyl and aryl isocyanates react with cysteine and tyrosine as well but produce products that can be hydrolyzed back to the unmodified amino acids under alkaline conditions,²⁴ to leave only the lysines modified.

A large collection of **acylating agents** react with lysine and also acylate other nucleophilic amino acids.

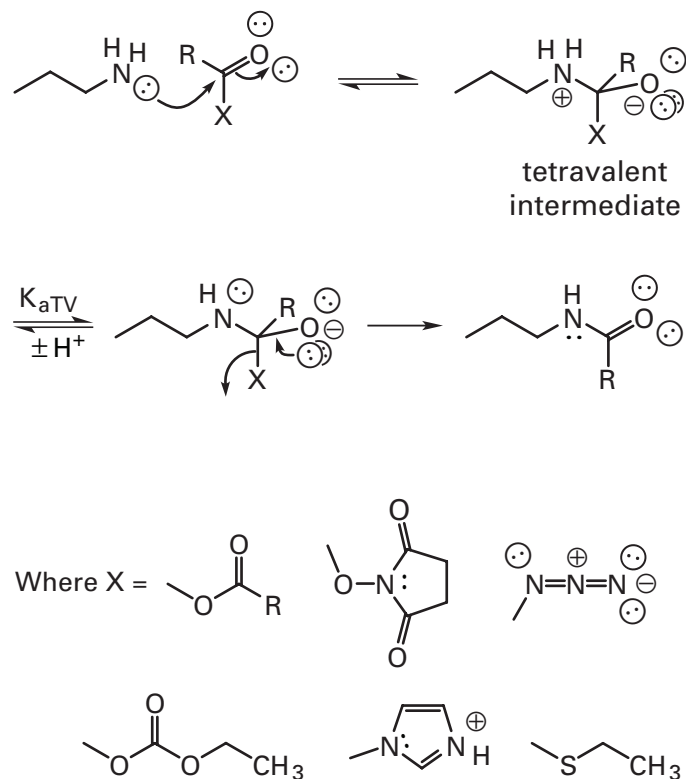
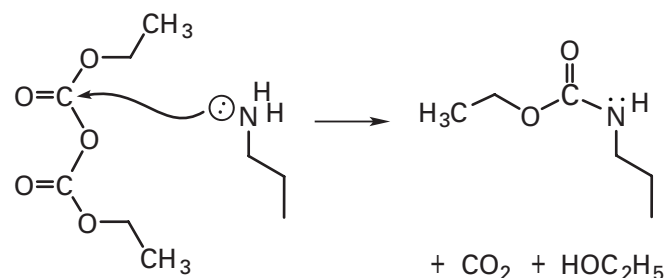


Figure 10-4: Acylation of lysine with any one of a number of acylating agents in which the acyl carbon is activated by attaching a good leaving group. In the tetra-valent intermediate, the leaving group is expelled, in preference to the nitrogen of the lysine, to produce the derivative of the carboxylic acid that is the acylating agent are the carboxylic acid itself to form the anhydride, *N*-hydroxysuccinimide to form the *N*-hydroxysuccinimide ester, azide anion to form the acyl azide, ethyl carbonic acid to form the acyl ethyl carbonate, imidazole to form the acyl imidazole, or a thiol to form the thioester.

The general reaction performed by these reagents is acyl transfer (Figure 10-4). As in synthetic organic chemistry, the appropriate reagent is chosen on the basis of its electrophilicity. The properties of the leaving group X determine both the electrophilicity and the rate at which the reagent will modify the lysines in a protein. Because the leaving group departs from a Lewis acid, the tetra-valent intermediate, the tendency for the leaving group to depart from a proton will reflect its tendency to depart from the tetra-valent intermediate (Figure 10-4). Therefore, the larger the acid dissociation constant K_{aLG} of the conjugate acid of the leaving group X, the more reactive will be the reagent. If the leaving group is the carboxylic acid itself, the reagent is an anhydride such as trifluoroacetic anhydride ($pK_{aLG} = 0.2$) or acetic anhydride²⁵ ($pK_{aLG} = 4.8$). Other leaving groups on acyl derivatives that have been used in the modification of lysine are azide²⁴ ($pK_{aLG} = 4.7$), *N*-hydroxysuccinimide²⁶ ($pK_{aLG} = 6.0$),²⁷ imidazole²⁵ ($pK_{aLG} = 7.0$), ethyl carbonate²⁸ ($pK_{aLG} = 7$), and ethanethiol²⁹ ($pK_{aLG} = 10.5$).

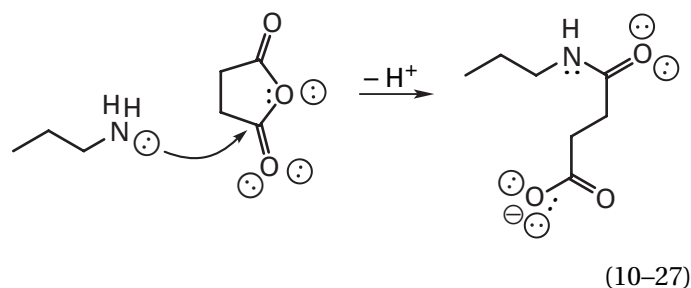
All of these acylating agents react as readily with cysteine, tyrosine, and histidine as they do with lysine to form the respective *S*-, *O*-, or *N*-acyl derivatives. Unlike the *S*-, *O*-, or *N*-alkyl derivatives formed during the reaction with an alkylating reagent such as iodoacetamide (Reactions 10-1 through 10-4), these acyl derivatives of cysteine, tyrosine, and histidine are unstable and decompose spontaneously or can be decomposed intentionally under conditions that leave the lysines in the modified form. For example, *O*-acetyltyrosine can be hydrolyzed back to tyrosine by treatment with hydroxylamine.³⁰ Often an acylating agent the structure of which causes the undesired derivatives to be particularly unstable can be chosen. For example, the ethoxycarbonyl group is added to tyrosine and histidine as well as lysine when one uses the carbonic acid anhydride, diethyl pyrocarbonate:



(10-26)

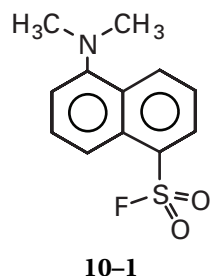
The ethoxycarbonyl group, however, can be removed from the histidine and tyrosine by treatment of the modified protein with hydroxylamine.²⁸

Cyclic anhydrides such as succinic anhydride



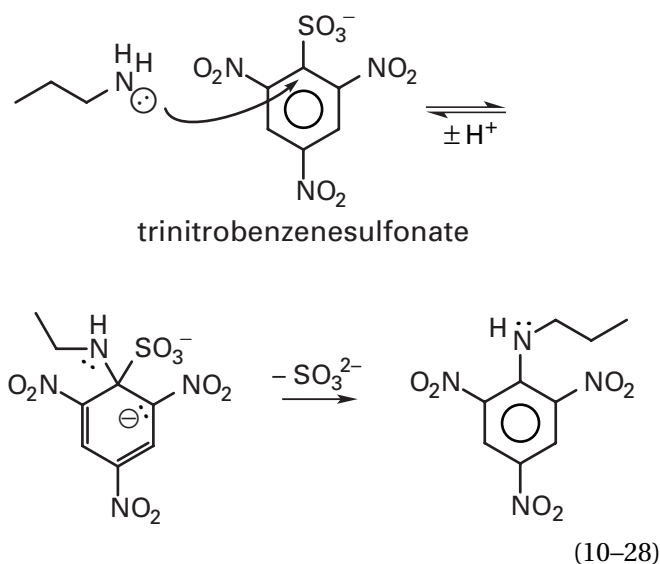
are frequently used to modify the lysines in a protein. They replace the positively charged primary ammonium cation of lysine with a negatively charged carboxylate. The acylation at lysine can be reversed to regenerate the lysine when maleic anhydride (2,3-dehydrosuccinic anhydride), citraconic anhydride,³¹ or 3,4,5,6-tetrahydrophthalic anhydride is used.⁶

Fluorosulfonic acids are general electrophilic reagents that modify lysine, by *N*-sulfonation, and tyrosine, by *O*-sulfonation. The paradigm for these reagents is 5-(dimethylamino)naphthalene-1-sulfonyl fluoride (dansyl fluoride):



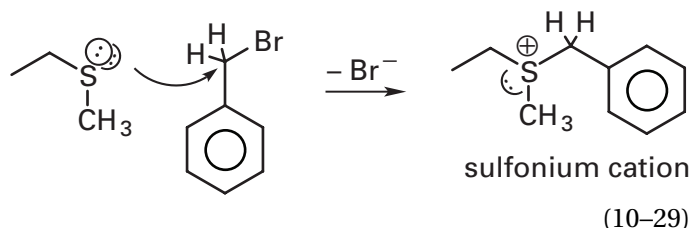
which is a fluorescent reagent for the covalent modification of proteins.³²

Both 2,4-dinitrofluorobenzene (FDNB) and **2,4,6-trinitrobenzenesulfonate** (TNBS)³³



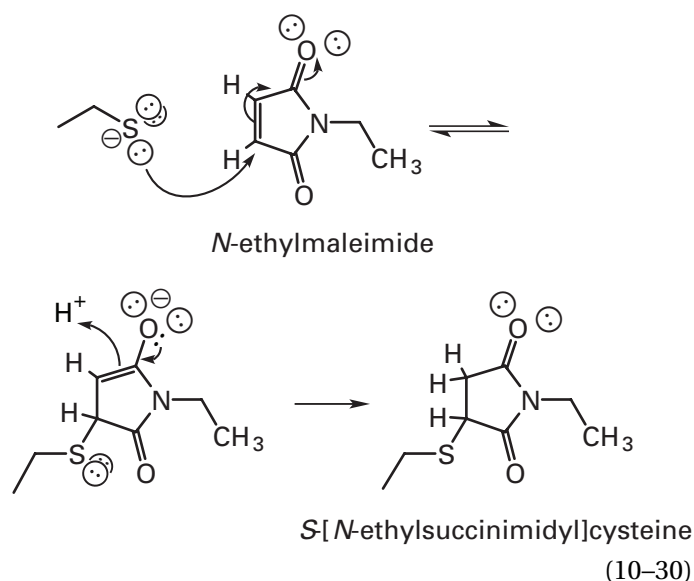
modify lysine by nucleophilic aromatic substitution. The former compound reacts with every nucleophilic amino acid.³⁴ The latter can be confined to react with only cysteine²⁴ and lysine by the proper choice of pH;³³ but the derivative formed with cysteine is unstable, so in the end only lysine is modified. In situations in which particular lysines in a protein are more nucleophilic than the others, modification of the protein even by as promiscuous a reagent as 2,4,6-trinitrobenzenesulfonate can be confined to those few sites. For example, at pH 8.0, only Lysine 332 of the α subunit of ribulose-bisphosphate carboxylase from *Spinacia oleracea* reacts significantly with this otherwise nonspecific arylating reagent.³⁵

As noted previously, alkylating agents such as iodoacetamide and other alkyl halides are electrophiles that react with every nucleophilic amino acid to yield stable products. Informed or uninformed manipulation of the pH can affect the distribution of alkylated products. At low pH, alkylating agents can be used to produce stable derivatives of **methionine** specifically.¹⁷ For example, at slightly acidic values of pH, benzyl bromide alkylates methionine in fumarase quite selectively.³⁶



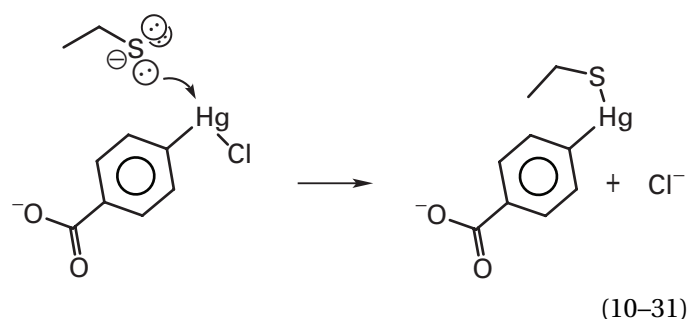
The reagent that has shown the greatest selectivity for **histidine** is **diethyl pyrocarbonate** (Equation 10-26).²⁸ Usually this selectivity is obtained by running the reaction at a pH slightly below the pK_a for histidine, where the greatest discrimination in favor of histidine, relative to lysine (Equation 10-26), should be manifested (Figure 10-1). Histidine is also susceptible to **photooxidation** in the presence of dyes such as methylene blue or rose bengal.³⁷ Under carefully controlled conditions such photooxidation can be confined to histidine,^{37,38} but usually several other amino acids are destroyed simultaneously.³⁷

One of the most readily modified amino acids in a protein is **cysteine**. At slightly alkaline pH, in the vicinity of its pK_a (Figure 10-1), cysteine is preferentially alkylated by alkyl halides such as iodoacetamide and iodoacetate (Equation 3-17), but one must remain aware of the fact that other nucleophilic amino acids can also be modified. ***N*-Ethylmaleimide** (NEM) is another reagent often used to modify cysteine.³⁹



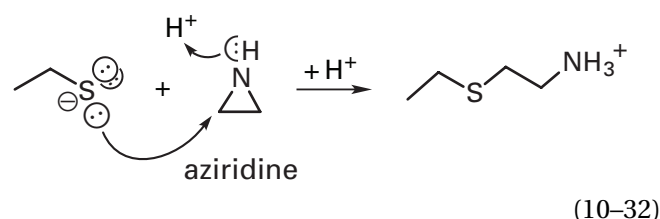
This reaction is an example of nucleophilic addition to an α,β -unsaturated acyl compound. **2-Vinylpyridine**⁴⁰ is selective for modification of cysteine in a similar reaction. The specificity of these alkylating agents for cysteine depends both upon the fact that sulfur is more nucleophilic than either nitrogen or oxygen and upon the use during the reaction of a pH just below the pK_a of cysteine (Table 2-2) so that modification of lysine is suppressed (Figure 10-1). The possibility of alkylation at other nucleophiles such as lysine⁴¹ must always be considered. For example, the inactivation of spinach ferredoxin-NADP⁺ reductase by *N*-ethylmaleimide results from alkylation of a lysine rather than a cysteine.⁴

Organic mercurials, such as ***p*-chloromercuribenzoate** (PCMB),³⁹ are usually specific for cysteine:



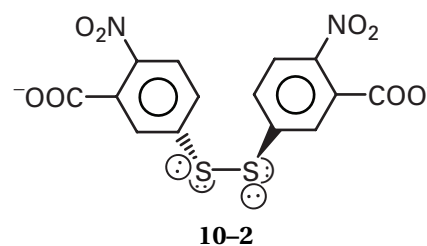
The sulfur-mercury bond, because it is a bond between a soft metal and a soft Lewis base, is significantly covalent and is particularly stable, but it is usually not stable enough to survive subsequent digestion of the protein and chromatography of the peptides.

Cysteine can be converted to *S*-(2-aminoethyl) cysteine by modification with **aziridine** (ethyleneimine):⁴²⁻⁴⁵



The modified side chain is isosteric with the side chain of a lysine and has become a site for tryptic digestion.⁴⁶

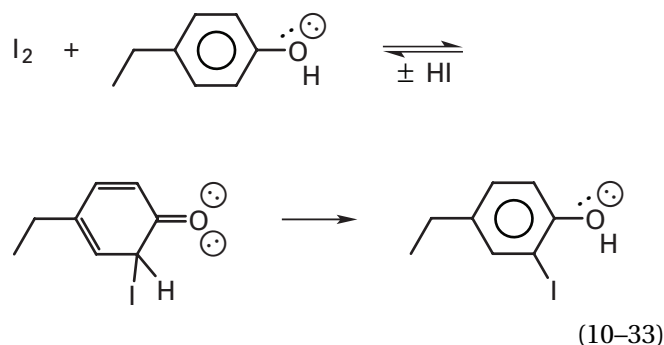
5,5'-Dithiobis(2-nitrobenzoate)



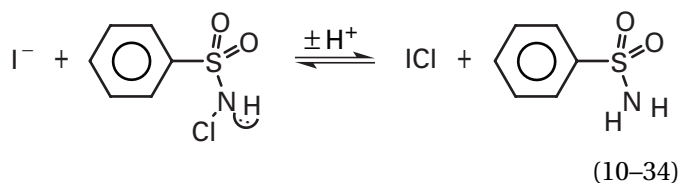
participates readily in disulfide interchange with a cysteine (Figure 3-20). The reagent contains a disulfide that is particularly electrophilic⁴⁷ because the nitrothiobenzoate dianion is such a good leaving group ($pK_a < 5$). This causes the equilibrium to lie in favor of mixed disulfides between the cysteines on the protein and 5-thio-2-nitrobenzoate. Unfortunately, these mixed disulfides are also electrophilic. Consequently, the pH of the solution should be well-buffered to prevent further reaction of the mixed disulfide with nucleophiles such as hydroxide ion. The nitrothiobenzoate dianion released during the disulfide interchange between cysteine and 5,5'-dithiobis(2-nitrobenzoate) (**10-2**) is brightly colored ($\epsilon_{412} = 13,600 \text{ M}^{-1} \text{ cm}^{-1}$), and its absorbance can be used to follow the reaction. The situation is complicated, however, by the possibility of side reactions with other nucleophiles releasing extrastochiometric nitrothiobenzoate that increases the absorbance of the solution and by the reaction of the nitrothiobenzoate itself with oxygen that decreases the absorbance of the solution. Well-buffered solutions should be used in the absence of atmospheric oxygen, and the absorbance should be measured continuously. The ease with which this reaction can be followed has led to its wide application to proteins.

Tyrosine is frequently alkylated, acylated, arylated, or sulfonylated inadvertently during modification reactions designed to be restricted to lysines. It can be modified specifically, however, by taking advantage of its elevated susceptibility to **electrophilic aromatic substitution**. As a *p*-alkylphenol, it is activated toward substitution, which is directed to its ortho positions by the

electron-releasing hydroxyl. A simple example of this susceptibility is the facile **iodination** of tyrosine:

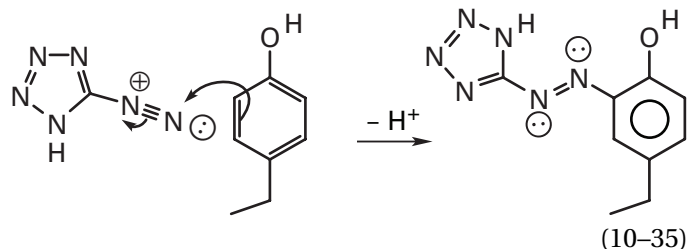


Iodide ion is used as the source of the iodine, and it is oxidized to I_2 , IOH, or ICl either chemically⁴⁸ or enzymatically.⁴⁹ Histidine is also iodinated under similar conditions but not so readily as tyrosine.⁴¹ A particularly advantageous method for activating I^- chemically uses *N*-chlorosulfamylphenyl groups covalently incorporated into a solid phase by the *N*-chlorosulfamylation of polystyrene beads.⁵⁰ The covalently attached *N*-chlorosulfamylphenyl groups produce ICl from I^- in a reaction identical to that performed by *N*-chlorobenzenesulfonamide in free solution:



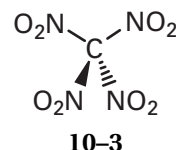
The ICl iodinate tyrosines in the protein, and the attachment of the chlorinating agent to a solid phase inhibits its ability to chlorinate the protein. The most advantageous enzymatic method uses lactoperoxidase. This enzyme converts I^- and H_2O_2 to OI^- and H_2O . Either the OI^- remains bound to the iron of the heme on the enzyme, where it can iodinate a tyrosine on the surface of another protein that has been bound by the lactoperoxidase, or it is released as HOI, which can iodinate a tyrosine on the surface of another protein free in the solution.⁵¹ Iodination is usually used to introduce the radioactive isotope of iodine, ^{125}I , into the protein to make it radioactive.

Diazonium salts also participate in the electrophilic aromatic substitution of tyrosine. 5-Diazonium-1-hydrotetrazole⁵² is a diazonium salt producing a product with tyrosine that absorbs strongly at 550 nm.⁴¹



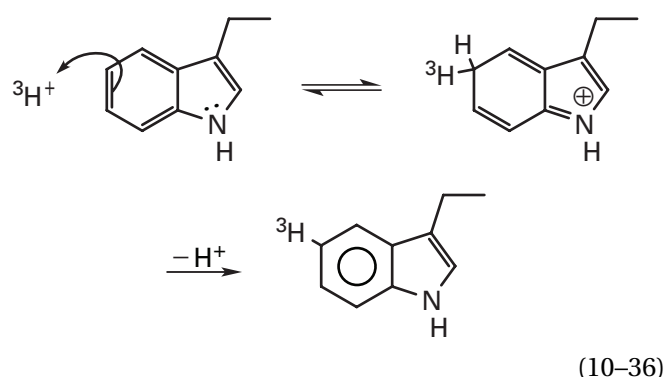
It reacts readily with histidine as well, but at low pH histidine will be mainly protonated, and the imidazolium cation is inert to electrophilic aromatic substitution.

Tetranitromethane is the reagent used most frequently to modify tyrosine:

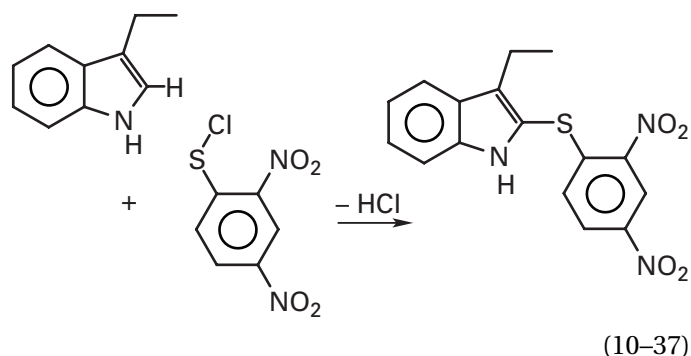


The nitration that produces the *o*-nitrotyrosine proceeds by a free radical mechanism.⁵³ The *o*-nitrotyrosine produced absorbs strongly at 428 nm as the nitrophenolate anion. It can be reduced to *o*-aminotyrosine with dithionite.⁵⁴ The hydroxyl of *o*-aminotyrosine has a uniquely low pK_a (4.8), and this fact can be exploited to direct further modification to this location in the protein.⁵⁴ Unlike *O*-acylation and *O*-alkylation, which require the tyrosine to be anionic to react as a nucleophile, the reaction with tetranitromethane proceeds with the neutral tyrosine.

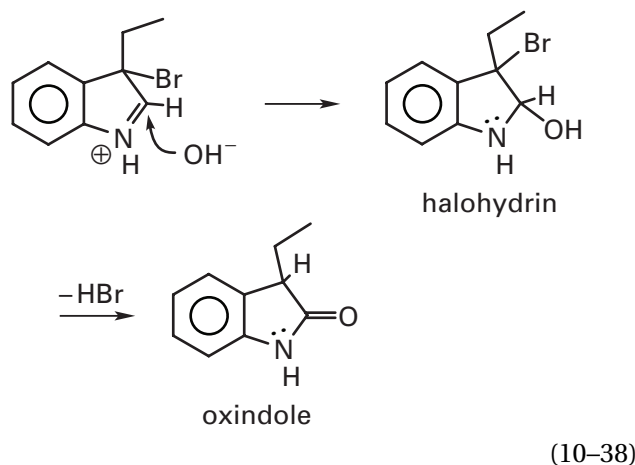
Tryptophan is susceptible to electrophilic aromatic substitution because of its similarity to aniline. For example, tritium can be incorporated specifically into tryptophan in a protein under strongly acidic conditions:⁵⁵



Sulfonyl halides such as 2,4-dinitrobenzenesulfonyl chloride participate in a formal electrophilic aromatic substitution at carbon 2 of tryptophan.⁵⁶

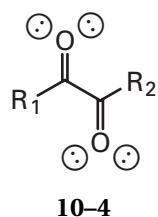


The most peculiar position in tryptophan is the π bond between carbons 2 and 3. This bond displays the properties of an olefin during **bromination** with mild brominating reagents (Equation 3-1) by participating in addition rather than substitution. Under mild conditions a relatively inert brominating agent, 2-[(2-nitrophenyl) sulfonyl]-3-methyl-3'-bromoindolenine (BNPS-skatole), oxidized the tryptophan in micrococcal nuclease to the oxindole,⁵⁷ presumably through an intermediate halo-hydrin:

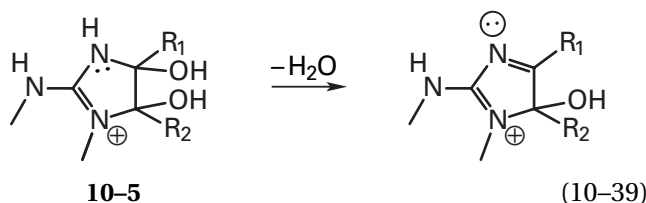


Only methionine was oxidized at the same time, and it could be regenerated readily by reduction. The use of addition reactions to the olefin in tryptophan to incorporate nucleophiles other than water might be feasible. The bromination, however, often results in cleavage of the polypeptide at the tryptophan (Equation 3-1).

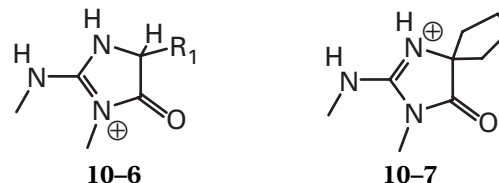
Arginine is modified specifically by **vicinal diones**:



Reagents normally used are diphenylethanedione ($R_1 = R_2 = \text{phenyl}$), *p*-nitrophenylethanedione ($R_1 = \text{C}_6\text{H}_4\text{NO}_2$; $R_2 = \text{H}$), 4-(oxoacetyl)phenoxyacetic acid ($R_1 = \text{C}_6\text{H}_4\text{OCH}_2\text{COOH}$; $R_2 = \text{H}$), and 1,2-cyclohexanedione ($R_1 = \text{CH}_2\text{CH}_2\text{CH}_2\text{CH}_2 = R_2$). In all cases the initial product is the cyclic adduct **10-5**^{7,58} that then dehydrates:



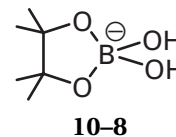
When the reagent is diphenylethanedione, this dehydrated adduct is the final product;⁵⁹ when the reagent is *p*-nitrophenylethanedione⁶⁰ or 4-(oxoacetyl)phenoxyacetic acid, the hydrogen (R_2) permits a Cannizzaro rearrangement that produces iminoimidazolidone **10-6**⁶¹



and when the reagent is cyclohexanedione, a similar rearrangement produces iminoimidazolidone **10-7**.⁶²

The modification of arginine by vicinal diones can also yield products that incorporate additional molecules of the dione. 2,3-Butanedione (**10-4**, $R_1 = R_2 = \text{CH}_3$) under appropriate conditions self-condenses to dimers and trimers that both react with arginine to yield poorly characterized, heterogeneous mixtures of products containing about 3 mol of dione for every mole of arginine.⁶³ Phenyl glyoxal (**10-4**, $R_1 = \text{phenyl}$, $R_2 = \text{H}$) reacts with arginine to produce a product containing 2 mol of dione for every mole of arginine.⁵⁸

The earliest modifications of arginine, with either diphenylethanedione⁶⁴ or 1,2-cyclohexanedione,⁶² were performed at alkaline pH (0.2 M NaOH), and the products were quite stable. The conditions, however, were too harsh to avoid destruction of the polypeptide. It was subsequently noted that the addition of **borate** during the reaction of a protein with 2,3-butanedione accelerated the rate of the reaction at neutral pH and rendered the modification irreversible as long as the borate was present.⁶⁵ The initial product of the reaction of 1,2-cyclohexanedione and arginine, presumably diol **10-5**, could also be stabilized significantly by the addition of borate.⁷ Borate is known to add to vicinal diols, such as sugars, to form cyclic borate diesters:



The addition of borate to cause the reaction with diones to be irreversible under mild conditions has permitted the isolation of modified peptides from proteins modified by 1,2-cyclohexanedione.⁶⁶

Glutamates and **aspartates** are modified with **carbodiimides** (Figure 10-5).⁶⁷ The carbodiimides used can be either hydrophobic, such as dicyclohexylcarbodiimide (DCCD; $R_1 = R_2 = \text{cyclohexyl}$), or hydrophilic and water-soluble, such as *N*-ethyl-*N'*-[3-(dimethylamino)propyl]carbodiimide [EDC; $R_1 = \text{C}_2\text{H}_5$, $R_2 = (\text{CH}_3)_2\text{NHC}_3\text{H}_6$]. The initial product of the

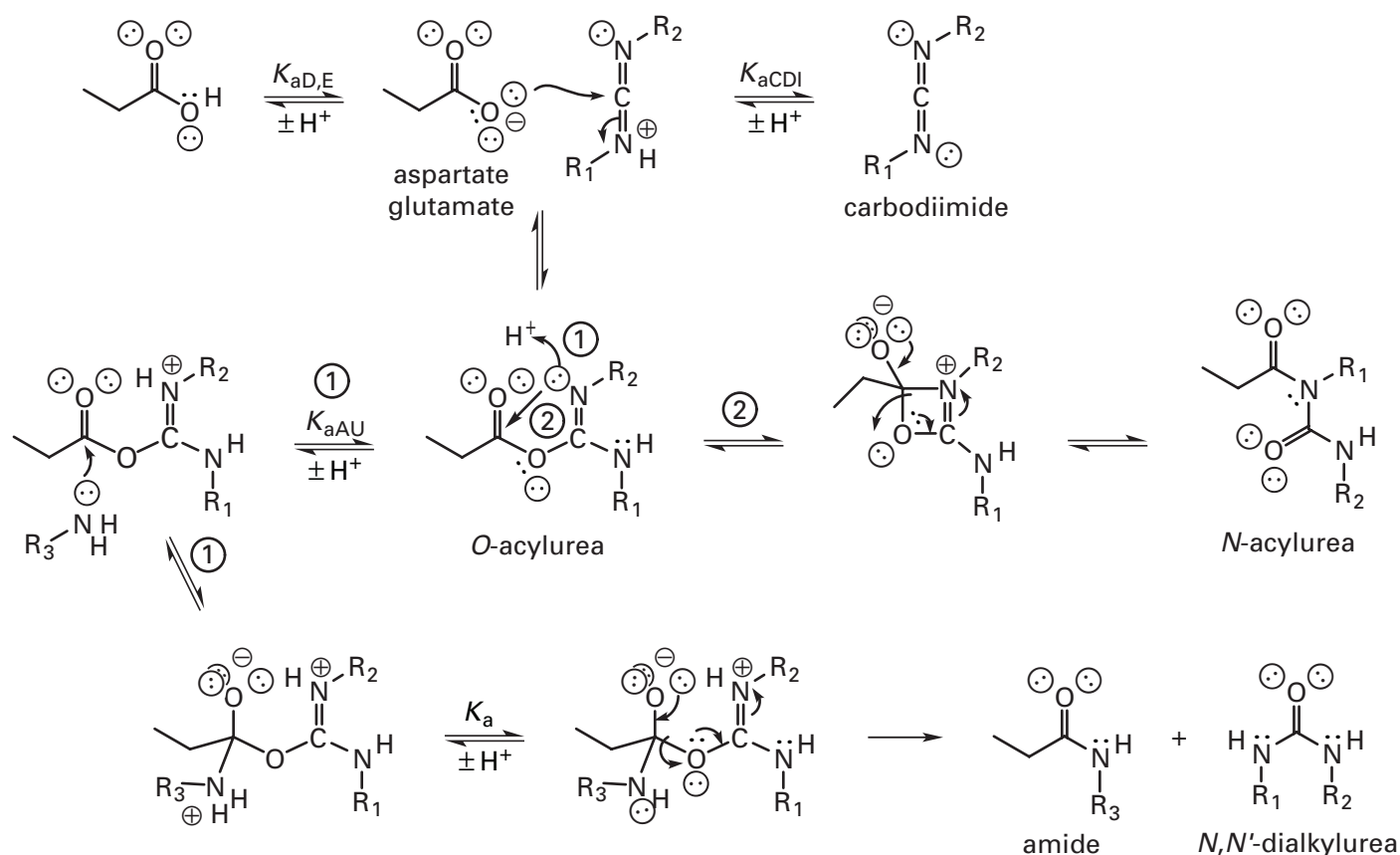


Figure 10-5: Outcomes of the reactions of carbodiimides with aspartate or glutamate. The *O*-acylurea is formed by the direct addition of the carboxylate anion to the protonated carbodiimide. In a rigid, isolated, aprotic environment, the *O*-acylurea might be the final product, but there are two other possible outcomes. If a nucleophile (usually an amine) has been added, or if there is an adjacent nucleophile in the protein (usually a lysine), the *O*-acylurea is an activated carboxylic acid derivative capable of acylating that nucleophile in an acyl exchange reaction ① to give the *N,N'*-dialkylurea as the leaving group and the acyl derivative (usually the amide) of the glutamate or aspartate with either the added nucleophile or the lysine in the protein. If there is no accessible nucleophile, the *O*-acylurea can rearrange, by intramolecular acyl exchange ②, to the *N*-acylurea. Pathway ① is initiated by the attack of the extraneous nucleophile on the activated carboxyl group, and pathway ② is initiated by intramolecular attack of the unprotonated urea nitrogen on the acyl carbon.

reaction is an *O*-acylurea⁶⁷ in which the acyl carbon of the original glutamate or aspartate has been activated by forming an acyl derivative, the leaving group of which is an excellent one because it is the oxygen of an *N,N'*-dialkylurea ($pK_a = 1$).

Four fates await this ***O*-acylurea**. First, if it is buried in a nonnucleophilic environment within the protein and sterically constrained, it will remain as the *O*-acylurea until the protein is unfolded, at which point it will usually hydrolyze back to the unmodified glutamate or aspartate. Second, if it is somewhat buried in a polar environment, but not sterically constrained, the *O*-acylurea will rearrange to the *N*-acylurea, which is stable (pathway ② in Figure 10-5). Dicyclohexylcarbodiimide is often incorporated into a protein in this way. It usually reacts with buried carboxylic acids because it is so hydrophobic, and the buried *O*-acylurea usually survives long enough to rearrange as the protein sits around after the investigator believes the reaction has finished; but the reaction rarely proceeds in high yield. Third, if there is a nucleophilic amino acid in the

protein, for example a lysine, in the vicinity of the *O*-acylurea, an intramolecular adduct, for example an amide, between that amino acid and the glutamate or aspartate will form (pathway ① in Figure 10-5).⁶⁸ Fourth, if an external amine, such as the methyl ester of glycine, has been added in high concentration to the solution, it can react with the *O*-acylurea as it is formed, if it is sterically accessible, and produce the amide between the external amine and the glutamate or aspartate.⁶⁷ In this way, a defined covalent modification of the carboxylate can be made. If the external nucleophile is ammonia, glutamates and aspartates are converted to glutamines and asparagines, respectively.⁶⁹

The practical outcome of each of these four fates is unique. In the first, the native protein is modified by the carbodiimide at glutamate or aspartate but loses the modification upon unfolding. In the second, a stable derivative between the protein and the carbodiimide is formed. In the third, the glutamate or aspartate is stably modified by being intramolecularly cross-linked,⁶⁸ but neither the carbodiimide nor an external amine is incor-

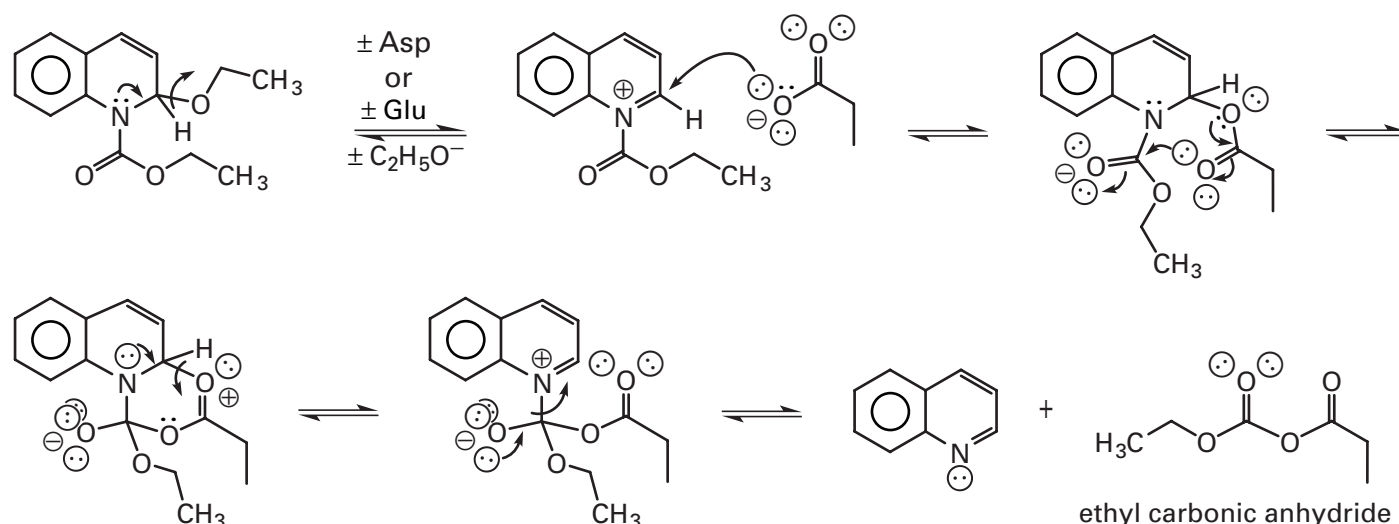
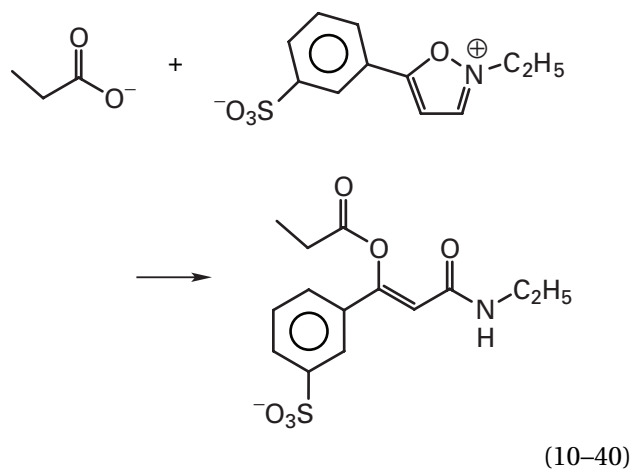


Figure 10-6: Formation of an ethyl carbonic anhydride (Figure 10-4) at a glutamate or aspartate by *N*-(ethoxycarbonyl)-2-ethoxy-1,2-dihydroquinoline. The 2-ethoxy group leaves as the alcohol when the reagent reverts in water to the aromatic *N*-(ethoxycarbonyl)iminium cation. The iminium cation reacts with the carboxylate of a glutamate or an aspartate. The adduct that is formed undergoes an intramolecular, cyclic rearrangement involving an acyl exchange at one end with quinoline as the leaving group and the expulsion of an ester from its adduct with an iminium cation at the other end.

porated into the protein. In the fourth, an external amine but not the carbodiimide is incorporated. Often, regardless of the intentions of the investigator, a combination of all of these outcomes occurs, and the complex mixture of products that results defies any attempt at quantification.

Another reagent used to activate the carboxylates of glutamates and aspartates is *N*-(ethoxycarbonyl)-2-ethoxy-1,2-dihydroquinoline (EEDQ). It activates the carboxylate (Figure 10-6) by forming a mixed ethyl carbonic anhydride ($pK_{aLG} = 7$),⁷⁰ which is an acylating agent (Figure 10-4) that is somewhat less reactive than an *O*-acylurea ($pK_{aLG} = 1$) but capable of the same types of intramolecular or intermolecular reactions with nucleophiles such as a lysine on the same protein or another protein⁷¹ or an amine that has been added to the solution.

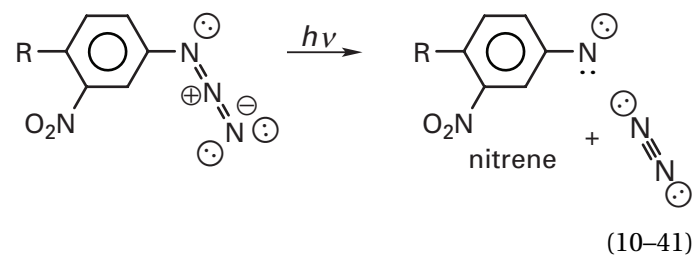
***N*-Ethyl-5-phenylisoxazolium-3'-sulfonate** (Woodward's reagent K)⁷² activates glutamates and aspartates⁷³ by forming an enol ester:



In this intermediate enol ester, the acyl carbon of the aspartyl or glutamyl side chain is activated sufficiently to react readily with nucleophiles elsewhere on the protein or with nucleophiles such as amines that have been purposely added to the solution, as does the *O*-acylurea that is the intermediate in the reactions of carbodiimides (Figure 10-5). The acyl group that has been activated with *N*-ethyl-5-phenylisoxazolium-3'-sulfonate is also reactive enough to be reduced with **borohydride ion**, (BH_4^-) to convert the side chain of the aspartate or glutamate to the respective alcohol,⁷⁴ just as an ester can be reduced to the corresponding alcohol by AlH_4^- ion.

Compounds that serve as precursors to nitrenes or carbenes through **photolytic reactions** are reagents that display even less specificity than alkylating agents in the modification of the amino acids in a protein. The fact that they are generated photolytically permits an added level of control over the reaction. The reagent can be equilibrated with the protein and then activated.

Aryl azides, such as phenyl azides or nitrophenyl azides, are the usual precursors for nitrenes. A nitroaryl azide produces a nitroaryl nitrene upon photolysis:



A convenient, widely used reagent for attaching a nitrophenyl azide to other compounds (the R in Reaction 10-41) by nucleophilic aromatic substitution is 4-azido-

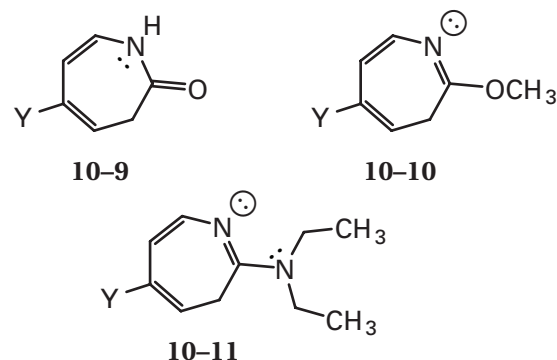
2-nitrofluorobenzene. Aliphatic azides have also been observed to insert photolytically into proteins.⁷⁵

A **nitrene** is a nitrogen the four valence orbitals of which are occupied by only six valence electrons. Therefore, it is dramatically electron-deficient and electrophilic. In a **singlet nitrene**, three of these orbitals are occupied by pairs of electrons and one orbital is vacant. In theory, a singlet nitrene, because of its vacant orbital, has a higher preference for insertion into nitrogen–hydrogen or oxygen–hydrogen bonds than carbon–hydrogen bonds because atoms of oxygen or nitrogen attached to carbon are electron-rich. In a **triplet nitrene**, two of the orbitals on nitrogen are each occupied by only one unpaired electron and the other two are occupied by two pairs of electrons. Consequently, a triplet nitrene is a diradical. Theoretically, triplet nitrenes should be able to modify proteins by hydrogen abstraction followed by rebound of the two adjacent monoradicals.⁷⁶ Because hydrogen is usually more easily abstracted from carbon than from oxygen or nitrogen, triplet nitrenes should abstract hydrogen more readily from carbon–hydrogen bonds than either nitrogen–hydrogen or oxygen–hydrogen bonds.

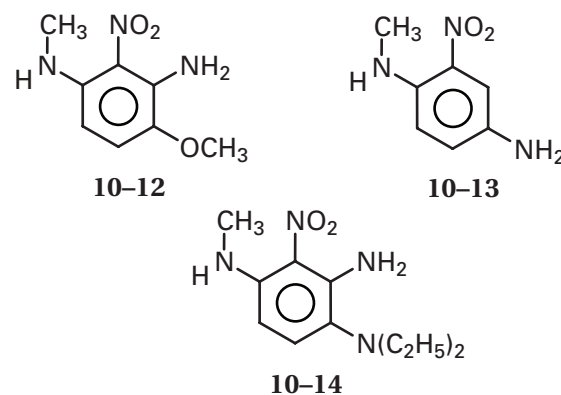
When light is absorbed by an aryl azide, which itself is a singlet, the excited state is initially a singlet excited state that must produce a singlet nitrene because N_2 is a singlet molecule. If the singlet excited state lasts long enough, it can convert to a triplet excited state by **inter-system crossing**. The triplet excited state produces a triplet nitrene and singlet N_2 . The yield of triplet excited state can be increased by adding a triplet sensitizer.⁷⁷ Singlet nitrene itself can turn into triplet nitrene if it survives long enough. In the absence of a sensitizer, only about 10% of the nitrene produced by photolysis of phenyl azide is triplet.⁷⁸

Although it is widely believed that **aryl nitrenes**, such as the phenyl nitrenes or the 3-nitro-4-(alkylamino)phenyl nitrenes usually employed in the modification of proteins, should insert into carbon–hydrogen bonds, a reaction that would require significant yields of the triplet state, the chemistry of such nitrenes belies this belief. In ideal situations, such as the intramolecular insertion in the vapor phase of an aryl nitrene into a tertiary carbon–hydrogen bond four carbons away, a reasonable yield of the *N*-alkylaniline (50%)⁷⁹ is obtained. When, however, phenyl nitrene is generated in cyclohexane by photolysis, no insertion (<30%)⁷⁸ into the solvent is observed, and most of the reaction proceeds with either dimerization of the nitrene itself or the production of aniline by two successive hydrogen abstractions by the triplet. Phenyl nitrene generated by photolysis under the same conditions in hydroxylic solvents such as methanol or propanol inserts into those solvents in high yield (80%).⁷⁸ The products of the photolytic reactions of 4-substituted phenyl nitrenes with water, methyl alcohol, or diethylamine as solvent are the lactams **10-9** (60–90% yield), the 2-methoxy-3-hydroazepines **10-10** (40–80% yield), or the 2-di-

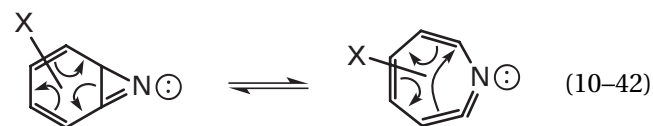
ethylamino-3-hydroazepines **10-11** (90–100% yield), respectively:⁸⁰



The products of the photolytic reaction of 4-methylamino-3-nitrophenyl nitrene with methanol as solvent or 1% diethylamine in methanol are aniline **10-12** (40% yield) and aniline **10-13** (40% yield) or aniline **10-12** (30% yield) and aniline **10-14** (70% yield), respectively:⁸⁰

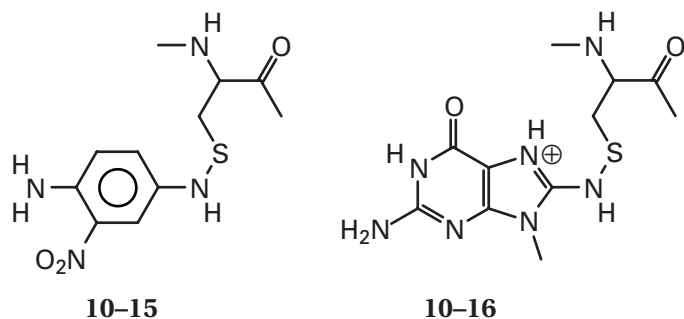


These products can be explained as the results of nucleophilic addition of solvent or solute to the two electrophilic species engaged in the following equilibrium:



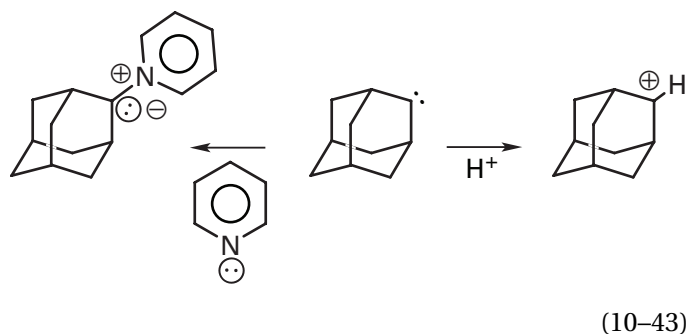
Consequently, the majority of the products from the reaction of an aryl nitrene with a protein should result from reaction with nucleophilic functional groups.

The identity of the amino acids modified by aryl nitrenes are consistent with these general considerations. The modification of rabbit glyceraldehyde-3-phosphate dehydrogenase (phosphorylating) by a *p*-amino-*m*-nitrophenyl nitrene produces the sulfenamide of Cysteine 149 (**10-15**)⁸¹



which, however, could have arisen from either the triplet or the singlet. The photolytic modification of rat phosphoenolpyruvate carboxykinase (GTP) with 8-azidoguanosine 5'-triphosphate produces an intramolecular disulfide between two cysteines in the protein,⁸² presumably arising from the nucleophilic displacement of the 8-aminoguanosine from the initially formed sulfenamide (**10-16**) by an adjacent cysteine. In addition to cysteine, tyrosine and lysine have usually been identified as the reactants with aryl nitrenes, but a leucine, two alanines, and a phenylalanine have also been reported to be modified.^{76,83,84} Singlet aryl nitrenes can insert intramolecularly by electrophilic aromatic substitution into phenyl rings,⁷⁷ and this reaction would explain the modification of phenylalanine.

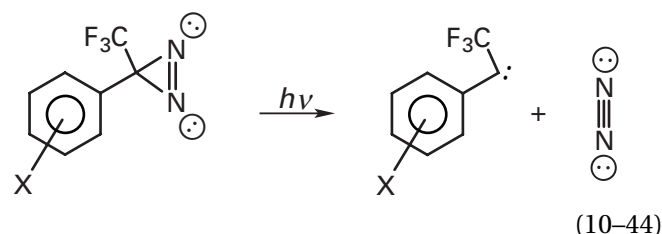
Carbenes, like nitrenes, have only six valence electrons on one atom, but they are distributed around a carbon instead of nitrogen. The carbenes generally used for the modification of proteins are on secondary carbons. They can be singlets or triplets, and relatively stable examples of both of these electronic states for carbenes have been synthesized.^{85,86} Again, the singlet is the first product and has a significant preference for **insertion into nucleophilic locations** such as nitrogen-hydrogen or oxygen-hydrogen bonds⁸⁷ and, in an aqueous solution of protein, probably reacts before much triplet is formed. The carbene of adamantane, formed by photolysis of adamantyldiazirine, is entirely (>99.9%) singlet and reacts as a strong electrophile to form ylides with the nitrogen of pyridine (Equation 10-43) and the sulfur of thiophene.⁸⁸



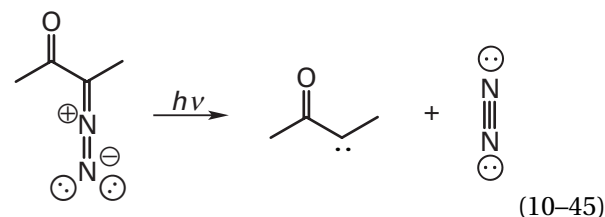
The singlet adamantyl carbene can also be protonated on its lone pair of electrons to produce the carbocation

that reacts readily with nucleophilic heteroatoms in reactions analogous to S_N1 substitutions.

Most aliphatic carbenes are prone to **intramolecular rearrangements** that are more rapid than their intermolecular reaction with other molecules in solution.⁸⁸ Reagents must be designed to avoid such rearrangement. The compounds that have proven to be the most efficient and uncomplicated precursors of carbenes unsusceptible to rearrangement are derivatives of 1-trifluoromethyl-1-phenyldiazirines:⁸⁹



The carbene is generated by photolysis. Prior to the advent of these reagents, α -diazoketones, α -diazoacetyl esters, and ethyldiazomalonyl esters were used as precursors of carbenes:⁷⁶



The first application of a carbene as a reagent for the modification of a protein used a diazoacetyl ester of Serine 195 in α chymotrypsin to increase the chances of insertion into other amino acids in the protein. Modifications of a cystine,⁹⁰ a serine other than Serine 195,⁹¹ an alanine,⁹² and a tyrosine⁹¹ were observed. Carbenes have usually been found to display a preference for insertion into oxygen-hydrogen and nitrogen-hydrogen bonds in proteins. They have been observed to modify lysines,⁷⁶ tyrosines,⁸⁷ tryptophans,⁹³ glutamic acids,⁹⁴ and aspartic acids,⁸⁷ but incorporation has also been noted into valine⁸⁷ and glycine.⁷⁶

Two remarkable illustrations of the **preference of carbenes for nitrogen-hydrogen and oxygen-hydrogen bonds** can be found in the use of these reagents to modify amino acids found in polypeptides that in the cell span membranes of phospholipid. In both instances, several precursors for the carbene were used that should have placed it at different respective locations along the polypeptide crossing the phospholipid bilayer; yet in each of the two experiments, the several different carbenes reacted respectively with the same amino acid, in one case a tryptophan⁹³ and in the other case a glutamic acid.⁹⁴ Other amino acids, with readily abstracted hydrogens at tertiary carbons, must have been more accessible

to at least some of these carbenes than the respective tryptophan or glutamic acid into which they eventually inserted. These results demonstrate that carbenes, like nitrenes, are not so promiscuous in their choices of reactants as they are often thought to be.

The products of the reactions between nitrenes and carbenes and proteins have rarely been characterized. In part, this is due to the **low yields** encountered in most of these reactions, presumably because of the tendency of the singlet carbene or singlet nitrene and its rearranged products to insert into water,⁹⁵ the inescapable solvent.

There are other compounds that can be photoactivated to form reactive species that modify proteins. For example, a 5-bromouracil was inserted in place of a thymine within the DNA sequence recognized by general control protein GCN4. When the complex between the protein and the modified DNA was irradiated with ultraviolet light ($\lambda_{\text{max}} = 254 \text{ nm}$), Alanine 238 of the protein had been covalently modified in low yield.⁹⁶ Vanadate anion binds specifically to proteins such as rabbit myosin⁹⁷ and isocitrate lyase from *Escherichia coli*.⁹⁸ Upon photolysis of these complexes, serines in the sites at which the vanadate had bound were photooxidized to products that could be converted by reduction with $\text{Na}[\text{}^3\text{H}]\text{BH}_4$ to $[\text{}^3\text{H}]\text{serine}$.⁹⁷ The incorporated tritium tagged the specific serines modified.

Oxidative cleavage is used to modify the **polyamide backbone** of a protein. In the presence of reducing agents such as ascorbate or dithiothreitol, complexes between Fe^{2+} or Cu^+ and chelators such as tetracycline, *N,N,N',N'*-tetracarboxymethyl-1,2-diaminoethane (EDTA), phenanthroline, or the protein itself convert O_2 or H_2O_2 into reactive species capable of cleaving peptide bonds.⁹⁹⁻¹⁰¹ The products of this cleavage that have been identified so far suggest that several different reactions can bring it about, so no unique mechanism seems to predominate.¹⁰² When the chelated metal is attached in some way to the protein, the cleavages that occur are confined to a few peptide bonds rather than being widely distributed along the polypeptide,^{103,104} an observation suggesting that the reactive species responsible for the cleavage either remain bound to the metallic cation or cannot diffuse very far without being discharged.

The **chelating ligand** surrounding the **metallic cation** is usually attached purposefully to a defined location on the surface of the protein to be modified. The activated products of the reaction cleave a peptide bond immediately adjacent to the place where the cation ends up. For example, the polypeptide in the vicinity of the site with which tetracycline associates on the tetracycline repressor was identified by the oxidative cleavage of the protein produced by the complex between tetracycline and Fe^{2+} .¹⁰³ In this case, the major targets for cleavage were the peptide bonds between positions 103 and 104 and positions 104 and 105. Lower yields of cleavage were observed between positions 55 and 56 and positions 135 and 136 and even lower yields between positions 143 and

144 and positions 146 and 147. If the protein itself has a site at which a structural metallic cation is bound, oxidative cleavage induced by the appropriate cation can map the polypeptide surrounding that site.¹⁰⁴ One advantage of oxidative cleavage of the polypeptide is that it can be readily detected by submitting complexes between dodecyl sulfate and the protein and its fragments to electrophoresis. Mass spectrometry of the resulting fragments can be used to locate the exact points of cleavage.¹⁰³

Site-directed mutation^{105,106} produces the covalent modification of a protein by converting one particular amino acid in its sequence into another of the 20 amino acids. It is also possible to delete amino acids from the sequence of a polypeptide or insert extra amino acids at a particular location with similar techniques. A common goal of both chemical modification and site-directed mutation is to correlate the structure of the protein with its function. An assessment of the effects of either type of modification on the normal function of the protein provides information about the role of the modified amino acid in that function. In order for either approach to place that side chain in a structural context, a crystallographic molecular model of the protein is required to define the location of the modified amino acid.

A strategic distinction, however, exists between covalent modification with chemical reagents and covalent modification by site-directed mutation. In the former procedure, the protein of interest is modified, and the effects of the modification on the function of the protein are assessed before the outcome of the reaction is defined by digesting the protein and then identifying the modified peptides either by mass spectrometry or by chromatographic separation and sequencing. In the latter procedure, any amino acid in the sequence of the protein can be chosen, and this particular modification is then performed before the results are assessed by the effect of the modification on some property of the protein. In the former procedure, the selectivity of the chemical modification is determined by the accessibility and the inherent nucleophilicity of the side chains that are potential targets, properties controlled by the protein and not by the investigator, so the results often contain unexpected information about the relationship between structure and function. In the latter, modification can be performed at any site selected by the investigator with absolute specificity, but the choice of which amino acid to modify and which of the 19 mutations to perform at that site relies on his intuition. Because the intuition of the investigator is usually fallible, informative experiments using site-directed mutation usually require that a crystallographic molecular model be already available to assist in the choice of the site to be modified.

It is also possible to decide to focus one's attention on the **nucleophilicity of a particular side chain** in the sequence of a protein on the basis of the location of that side chain in a crystallographic molecular model, its

position relative to specific features in the sequence, or its identification as an important structural or functional site from other experiments, just as one would choose a particular amino acid in the sequence of a protein for site-directed mutation. To assess its role in the function of the protein, the rate of the reaction of this one particular nucleophilic side chain with one or more appropriately chosen electrophilic reagents can then be followed to probe its accessibility,¹⁰⁷ changes in accessibility under different circumstances,¹⁰⁸ or its intrinsic nucleophilicity.¹⁰⁹ The rate or yield of the modification at this particular side chain can be monitored by repetitively purifying a short peptide containing this amino acid from digests of the protein modified under different conditions or for various intervals of time.¹¹⁰

An effective way to perform the requisite purifications of the products of these reactions rapidly is with an **immunoabsorbent**. Immunoglobulins raised against a synthetic peptide¹¹¹ with the same amino- or carboxy-terminal sequence as the peptide containing the amino acid being modified are used to produce an affinity adsorbent able to capture out of a digest of the protein only the particular peptide containing the modified target.^{112–114} Because the immunoabsorbent recognizes sequences that do not contain the target of the modification, both modified and unmodified versions of the peptide are isolated simultaneously, and the fraction of the side chain chosen for investigation that has been modified in each sample is immediately apparent.¹¹⁰

During the covalent modification of a protein, even with as nonspecific a reagent as iodoacetamide,¹⁵ the various types of amino acids do not react as homogeneous populations. This is most readily discerned when the amount of incorporation into a particular type of amino acid is plotted as a function of the duration of the reaction.¹⁵ If the concentration of reagent remains constant, the natural logarithm of the amount of unmodified amino acid should decrease as a linear function of time because the reaction is pseudo-first-order (Equation 10–12). This is usually not observed, and the disappearance of a particular type of amino acid, such as histidine, lysine, or cysteine, as the modification proceeds usually displays **inhomogeneous kinetics** (Figure 10–7).¹⁵ This can be ascribed to the fact that each histidine, each lysine, or each cysteine in the sequence of the protein is in a different environment in the folded polypeptide and reacts at a unique rate with the reagent. For example, the three histidines of α -lactalbumin can be modified by iodoacetamide, but each reacts at a significantly different rate¹¹⁵ so that the disappearance of total histidine displays inhomogeneous kinetics.

It has already been noted that the environment surrounding a particular amino acid in a protein shifts its apparent acid dissociation constant from the value it would have in an unfolded polypeptide (Table 2–2). Such shifts of the acid dissociation constants from their intrinsic values move each of the inflections of the profiles of

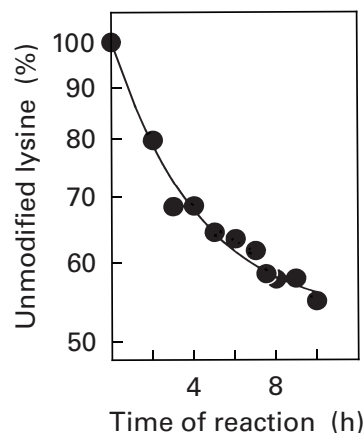


Figure 10–7: Reaction of iodoacetamide with the lysines of glucose-6-phosphate isomerase.¹⁵ Glucose-6-phosphate isomerase (0.12 mM) was mixed with iodoacetamide (49 mM) at pH 8.5 and 40 °C. At the noted times a sample was removed from the solution, the reaction was quenched with 0.3 M 2-mercaptoethanol, and the sample was subjected to total amino acid analysis. The amount of unalkylated lysine (percentage of total) is plotted on a logarithmic scale as a function of time. Reprinted with permission from ref 15. Copyright 1970 *Journal of Biological Chemistry*.

log k'_i against pH (Figure 10–1) horizontally to coincide with the altered value of the respective pK_a and simultaneously move the plateau at high pH vertically in response to the Brønsted relationship (Equation 10–14). A unique shift occurs for each amino acid in the protein. An example of such an effect of environment on the nucleophilicity of acid–bases is provided by a pair of cysteine side chains, Cysteine 3 and Cysteine 32, in seminal ribonuclease.¹¹⁶ These two adjacent cysteines react more rapidly with 5,5′-dithiobis(2-nitrobenzoate) at pH 7 than do model compounds such as cysteine itself or cysteinyl-cysteine. The synthetic peptide MCCRKM, which incorporates the sequence of seminal ribonuclease around these two cysteines, however, has the same enhanced reactivity. It was shown that this enhanced reactivity is due to the fact that the cysteines are adjacent to an arginine and a lysine. These cationic amino acids lower the values of the acid dissociation constants for the two cysteines. Although the nucleophilicities of the thiolate anions also decrease accordingly, the Brønsted coefficient β is small (<0.2). Therefore, the increase in reactivity at pH 7 results from a significant increase in the concentration of the respective thiolate anions, which each have almost the same intrinsic nucleophilicity, k_{Cys} , as a cysteine the pK_a of which has not been lowered.

Although experiments involving covalent modification are usually designed to answer a specific question about a particular protein, there are several **common themes**. Covalent modification is often performed to identify particular amino acids within a binding site for a particular ligand such as a substrate for an enzyme, or the agonist or antagonist of a receptor, or for a segment of DNA. An amino acid involved in the function of a pro-

tein can be identified when its covalent modification inhibits that function. The acid dissociation constant for a particular amino acid in a protein can be determined from the rate of its modification as a function of the pH. The accessibility of particular amino acids to the solvent can be inferred from the rates of their covalent modification. The amino acids incorporated into a transient heterologous interface can be identified by changes in the rates of their modification upon formation of the complex between the two proteins. The close proximity of two amino acids in a protein can be revealed by their covalent cross-linking.

The most common use of covalent modification is to identify amino acids in a **site on a protein for binding a specific ligand**. One way this can be done is to measure the change in the accessibility of a particular amino acid upon the binding of the ligand to the protein. For example, the greater than 4-fold decrease in the rate constant for the reaction between acetic anhydride and Lysine 501 in ovine Na^+/K^+ -exchanging ATPase when MgATP is bound to the enzyme suggested that this lysine participates in the specific interactions between the protein and MgATP when it is bound.¹¹⁷ This suggestion was later verified by a crystallographic molecular model of the complex between ATP and a closely related enzyme.¹¹⁸ Lysines involved in the interface forming the complex between DNA topoisomerase of vaccinia virus and a double helix of DNA were identified by noting significant decreases in the yield of their modification by citraconic anhydride upon formation of the complex.¹¹⁹ In this latter experiment, advantage was taken of the reversibility of the acylation of lysines by citraconic anhydride. The products of the modification were unfolded and modified at the unacylated lysines with *N*-hydroxysuccinimide acetate (Figure 10-4), the citraconyl groups were removed, and the locations of the previously citraconylated lysines were identified by digesting the protein at the deprotected lysines with lysyl endopeptidase and examining the pattern of fragments produced.

Another way that an amino acid in a binding site on a protein is identified is to incorporate a reactive functional group into the ligand itself. For example, it was demonstrated that the amino-terminal threonine produced upon the normal posttranslational cleavage of γ -glutamyltransferase from *E. coli* between Glutamine 390 and Threonine 391 is in the active site of the enzyme because the threonine was covalently modified by 2-amino-4-(fluorophosphono)butanoic acid, an electrophilic mimic of the γ -glutamyl group in glutathione, a substrate of the enzyme.¹²⁰

Covalent modification often leads to the inactivation of a protein and identifies candidates for **functionally important amino acids**. The modification of Lysine 85, Histidine 88, and Histidine 161 in bovine cytochrome *b*₅₆₁ by diethyl pyrocarbonate leads to the inactivation of the fast electron transfer performed by

this protein,¹²¹ a fact suggesting that these two histidines and the lysine participate in this function.

The **pK_a for a particular amino acid** is often estimated from the rate of its reaction with an electrophile as a function of pH. The reaction of acetic anhydride with particular lysines in a protein has been used to monitor their individual acid dissociation constants and nucleophilicities.¹⁰ Because acetic anhydride is rapidly hydrolyzed, the yield of its incorporation at a set pH into a particular lysine in the protein, relative to the yield of its incorporation into an added standard amine, after the reaction has reached completion provides a direct measurement of the relative bimolecular rate constant for its reaction with that particular lysine at that pH. The situation is formally equivalent to Equation 10-21 with k_2 being the rate constant for reaction of the acetic anhydride with the standard amine. If the absolute rate constant for the reaction between acetic anhydride and the standard amine is known, the absolute bimolecular rate constant for the reaction between the lysine of interest and acetic anhydride at that pH can be calculated from the relative rate constant, $k_1(k_2)^{-1}$ in Equation 10-23. The behavior of this absolute rate constant as a function of pH (Figure 10-1) provides an estimate of the pK_a of the lysine.

It is also possible to measure a pK_a directly. For example, the pK_a of Cysteine 25 in papain¹²² was determined to be 8.5 at 25 °C and an ionic strength of 0.5 M¹²³ by following the rate of its reaction with chloroacetamide as a function of pH, the pK_a of Cysteine 115 in UDP-*N*-acetylglucosamine 1-carboxyvinyltransferase from *Enterobacter cloacae* was determined to be 8.3 by following the rate of its reaction with iodoacetamide as a function of pH,¹¹ and the pK_a of Lysine 166 in ribulose-bisphosphate carboxylase from *Rhodospirillum rubrum* was determined to be 7.9 by following the rate of its reaction with 2,4,6-trinitrobenzenesulfonate.³⁵

The reactivity of particular amino acids in the sequence of a protein can provide an indication of their **accessibility** to the aqueous phase. For example, a comparison between the observed rate constant for the reaction of a particular lysine in a protein with acetic anhydride and the rate constant calculated from its observed pK_a provides an estimate of the accessibility of that lysine to the solvent. Knowledge of the Brønsted coefficient β (0.48) and the absolute bimolecular rate constant for the free base of an unhindered lysine ($2700 \text{ M}^{-1} \text{ s}^{-1}$) of normal pK_a (10.8) with acetic anhydride at 10 °C¹⁰ permits the bimolecular rate constant expected for the modification of the free base of a fully accessible lysine with a particular pK_a to be calculated. For example, the pK_a of Lysine 501 in ovine Na^+/K^+ -exchanging ATPase was found to be 10.4, so the rate constant of the reaction of its free base with acetic anhydride at 10 °C should have been $1700 \text{ M}^{-1} \text{ s}^{-1}$.¹⁰⁹ The fact that the rate constant was only $400 \text{ M}^{-1} \text{ s}^{-1}$ indicated that Lysine 501 was not fully exposed on the surface of the protein. In most instances, the apparent accessibility of a particular amino acid is

determined by **steric effects**, engendered by neighboring amino acids in the folded polypeptide or a decrease or increase in its nucleophilicity brought about by its participation in intramolecular **hydrogen bonds**.

Tetranitromethane, which is large and quite polar (**10-3**), reacts with the un-ionized, neutral form of tyrosine. At neutral pH, all of the tyrosines in a protein should be un-ionized, and their modification by tetranitromethane should reflect only their accessibility.¹² There are eight tyrosines in human carbonate dehydratase B, and only three of them, Tyrosine 20, Tyrosine 88, and Tyrosine 114, react with tetranitromethane.¹² Subsequent to this assessment, the protein was studied crystallographically, and in the map of electron density only Tyrosine 20, Tyrosine 88, Tyrosine 114, and Tyrosine 129 were found to be "located on the surface of the molecule."¹²⁴ Aspartate 194 in bovine chymotrypsinogen A could not be modified in the native protein with ethyl glycinate and *N*-ethyl-*N'*-[3-(dimethylamino)propyl]carbodiimide even under conditions where 13 of its 15 carboxylates were modified completely.¹²⁵ Subsequently it was observed that Aspartate 194 is "buried" in the interior of the crystallographic molecular model of chymotrypsinogen.¹²⁶ When fructose-bisphosphate aldolase was modified with methyl acetimidate at high concentrations, only 20 of its 30 lysines were modified.¹²⁷ The 10 unmodified lysines reacted readily when the protein was unfolded, and it could be shown that these were 10 unique lysines in the sequence of the protein, presumably made unreactive by their surroundings in the folded polypeptide.

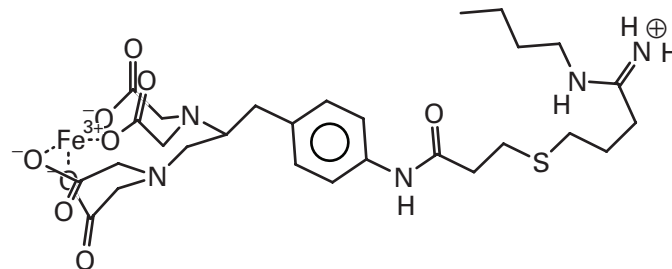
Site-directed mutation can also be used to monitor the accessibility of particular locations in the amino acid sequence of the protein. An α helix passes across the surface of the crystallographic molecular model of λ repressor.¹²⁸ In this α helix Isoleucine 84 and Methionine 87 are on the face of the α helix directed toward the interior of the protein while Tyrosine 85, Glutamate 86, Tyrosine 88, and Glutamate 89 are on the surface of the α helix that is accessible to the solution. After this observation had been made crystallographically, it was shown that only isoleucine at position 84 and either methionine or isoleucine at position 87, of all of the 20 amino acids, produces a functional protein, but 10-14 of the 20 amino acids can be substituted at the other four positions and still produce a functional protein.¹²⁹

Another covalent modification that has been used to assess the accessibility of particular amino acids in a folded polypeptide is **endopeptidolytic cleavage**. For an endopeptidase to cleave a peptide bond, the polypeptide at that location must be able to enter its active site. This has usually been assumed to require that the susceptible peptide bond be located on a somewhat flexible loop, on the outside surface of the protein, well exposed to the solvent. In the case of chymotrypsinogen A, the endopeptidolytic cleavages of the folded polypeptide that remove the amino acids between Leucine 13 and

Isoleucine 16 and between Tyrosine 146 and Alanine 149 to produce α chymotrypsin occur within two such loops.¹²⁶ In the case of deoxyribonuclease I, however, a less easily explained endopeptidolytic cleavage of the folded polypeptide has been observed. Under the proper set of conditions, chymotrypsin cleaves deoxyribonuclease I completely and exclusively at the peptide bond on the carboxy-terminal side of Tryptophan 178.¹³⁰ In the refined crystallographic molecular model of deoxyribonuclease I,¹³¹ Tryptophan 178 is found in the middle of an α helix that is a rigid feature of the structure. This α helix traverses the outer surface of the protein, but Tryptophan 178 is on the side of the α helix pointed toward the interior and itself is inaccessible. There are, however, no more accessible sites in the protein at which chymotrypsin could cleave, and it may be the case that in solution the α helix containing Tryptophan 178 is in equilibrium with a disordered loop.

Changes in the accessibility of amino acids on the surface of a protein brought about by its participation in an **association with another protein** can be monitored as changes in the yields of their covalent modification. For example, the yields of the reductive methylations of Lysines 50, 61, 68, 113, 284, and 291 with [¹⁴C]formaldehyde and NaCNBH₃ decreased 2-4-fold upon conversion of monomeric actin to helical filaments of actin.¹³² Covalent modification can also introduce bulky groups that sterically inhibit an association between two proteins. For example, when Histidine 40 of actin is modified with diethyl pyrocarbonate¹³³ or when Lysine 61 of actin is modified with fluorescein isothiocyanate,¹³⁴ the modified actin is no longer able to form helical polymers. All of these observations were used as evidence in favor of the molecular model of the helical polymer of actin (Figure 9-1B),¹³⁵ in which all of these amino acids ended up in the interfaces between the monomers. If the model is correct, the formation of the interfaces in the homopolymer should sterically hinder the reductive methylation of the lysines, and the addition of bulky functional groups to these histidines or these lysines on the actin monomers (Equations 10-24 and 10-26) should sterically hinder their polymerization.¹³⁵

The Fe³⁺ chelate **10-17**



10-17

was covalently attached at random to lysines on the surface of sigma factor rpoD isolated from DNA-directed

RNA polymerase of *E. coli*. When the normally occurring complex was formed between this modified protein and DNA-directed RNA polymerase, the tethered Fe^{3+} was able to catalyze cleavage at specific peptide bonds in the DNA-directed RNA polymerase when ascorbate and H_2O_2 were added.¹³⁶ These sites of cleavage identified locations on the surface of the RNA polymerase within or adjacent to the interface between it and sigma factor rpoD. This interface has also been probed by footprinting.

Footprinting is the identification of those peptide bonds on the surface of a protein that are protected from random nonspecific cleavage when that protein forms a complex with another protein. The peptide bonds protected are assumed to be within the footprint of the other protein upon the surface of the protein being examined. That footprint is the portion of the surface of the protein being examined that falls within the heterologous interface. DNA-Directed RNA polymerase from *E. coli* stripped of its sigma factor rpoD is cleaved at 83 different peptide bonds on its surface when it is exposed to the Fe^{3+} chelate of *N,N,N',N'*-tetracarboxymethyl-1,2-diaminoethane in the presence of ascorbate and H_2O_2 . When sigma factor rpoD is reassociated with the DNA-directed RNA polymerase, the cleavage at seven of these sites is prevented.¹³⁷ It was assumed that these seven peptide bonds are within the footprint of sigma factor rpoD upon DNA-directed RNA polymerase.

Covalent cross-linking uses covalent modification to assess the proximity of particular amino acids. The

functional groups at the two ends of a bifunctional cross-linking reagent are usually electrophiles commonly used in monofunctional reagents for the modification of proteins. They can be identical to each other (**8-1** and **8-2**), or they can be two electrophiles with different specificities (**8-3**). They can be connected by a chain of atoms stably bonded or a chain of atoms containing a bond that can be cleaved as desired (**8-3**) to permit later separation and identification of the cross-linked species.

Intramolecular cross-linking can be used to determine juxtapositions in a single folded polypeptide. The simplest example of such cross-links are naturally occurring cystines, which automatically provide evidence for the juxtaposition of two segments of polypeptide,¹³⁸ but there are unnatural, chemical methods for forming cross-links. The bifunctional reagent 2-(*p*-nitrophenyl)-3-(3-carboxy-4-nitrophenyl)thio-1-propene (**10-18**) can undergo a series of reversible addition-eliminations to form bridges between two nucleophilic amino acids (Figure 10-8), either lysines or cysteines.¹³⁹ The reaction is reversible as long as the nitrophenyl group is present to stabilize the carbanion but can be made irreversible by reducing the nitro group with dithionite. Therefore, the reagent can be permitted to step around the protein until the most stable cross-link is formed, and this cross-link can then be locked in by reduction. In this way, two pairs of intramolecular cross-links on bovine pancreatic ribonuclease, one between Lysine 7 and Lysine 37 and the other between Lysine 31 and Lysine 41, could be

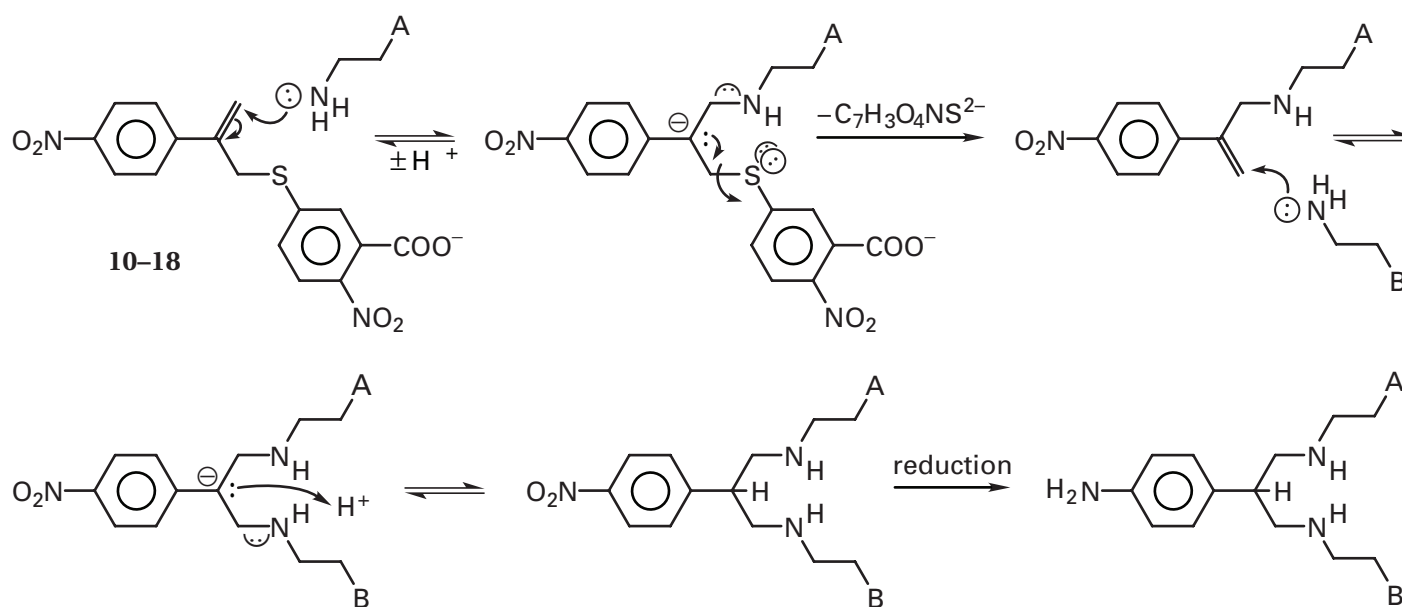


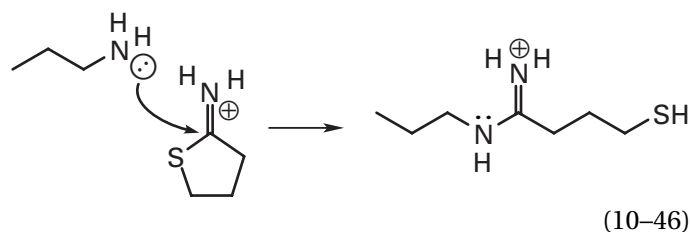
Figure 10-8: Mechanism by which 2-(*p*-nitrophenyl)-3-(3-carboxy-4-nitrophenyl)thio-1-propene (**10-18**) cross-links adjacent lysines. The olefin on the *p*-nitrostyrene is activated by the electron-withdrawing capacity of the nitro group and participates in a sequence of reversible, nucleophilic addition-eliminations. Because the adduct is symmetric, two nucleophiles are cross-linked reversibly. In the first step of the reaction, the nitrothiobenzoate (Structure **10-2**) is the preferred leaving group from the asymmetric carbanion, but when the carbanion is then formed between two lysines, either can be the leaving group and the reagent can be passed from lysine to lysine over the surface of the protein. The nitrobenzyl carbanion is not that basic, so its protonation is reversible and this allows the cycles of addition and elimination to proceed. When the reaction is quenched with acid and the nitro group is reduced to the amine, the aminobenzyl proton is no longer acidic and the reagent is fixed in place.

formed in high yield when only 2 molar equivalents of the reagent was added initially to the protein. The β carbons of these two pairs of lysines are 1.3 and 1.1 nm apart, respectively, and each partner in a pair is on the same side of the crystallographic molecular model.

The reagent bromopyruvate is bifunctional by virtue of its alkyl bromide, which is an alkylating agent, and its carbonyl, which can form an imine with a lysine reversibly that can be reduced to the permanent secondary amine with NaCNBH_3 . Bromopyruvate is able to form an imine with Lysine 144 of 2-dehydro-3-deoxy-6-phosphogluconate aldolase and then alkylate Glutamate 56.^{140,141} This observation established the proximity of these two amino acids in the folded polypeptide before the crystallographic molecular model became available.¹⁴²

The participants in **heterologous interfaces** have also been probed by cross-linking. During the contraction of muscle, a complex must form between a subunit of myosin in its helical polymer and actin in its helical polymer (Figure 9-1B). This complex between actin and myosin from rabbit muscle was cross-linked with 1-ethyl-3-[3-(dimethylamino)propyl]carbodiimide, a reagent that couples carboxylates to lysines on the surfaces of proteins (Figure 10-5). The amino-terminal peptide produced by cleavage of actin by hydroxylamine between Asparagine 12 and Glycine 13 and the carboxy-terminal peptide produced by cleavage of actin by cyanogen bromide at Methionine 354 were both found to be cross-linked to myosin in the covalently cross-linked complex.¹⁴³ It was also found that amino acids within the segment of actin between Histidine 40 and Lysine 113 were covalently attached to myosin when the complex between it and actin was cross-linked by *N*-(ethoxycarbonyl)-2-ethoxy-1,2-dihydroquinoline, a reagent that also can couple carboxylates to lysines (Figure 10-6). On the basis of these results, amino acids within these three segments of actin are thought to be within the heterologous interface it forms with myosin.¹⁴⁴

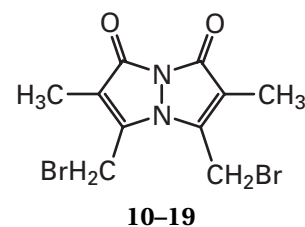
Thiols, either in the form of cysteines in the sequence of the protein or introduced as covalent modifications, can be sites for cross-linking. It is possible to insert cysteines at specific positions in the amino acid sequence of a protein by site-directed mutation and then attempt to form cystines between them.¹⁴⁵ The actual formation of a cystine demonstrates that the two cysteines participating in it were adjacent to each other in the tertiary or quaternary structure of the protein. Proteins can also be modified by 2-iminothiolane to convert lysines to thiols:¹⁴⁶



These thiols are then oxidized to mixed disulfides to cross-link various lysines on the proteins. The 21 folded polypeptides within the 30S subunit of the ribosome have been cross-linked to each other in this way. The various products of the intramolecular cross-linking were identified by two-dimensional gel electrophoresis. During the electrophoresis, the disulfides linking pairs of these polypeptides were reduced by disulfide interchange to unlink them from each other between the first and the second dimension.

With 2-iminothiolane, as well as with dimethyl suberimide,^{147,148} dimethyl adipimide,¹⁴⁹ *N,N'*-1,4-phenylenedimaleimide,^{150,151} tetranitromethane,¹⁵² tartryldi (ϵ -aminocaproyl azide),¹⁵³ and dimethyl 3,3'-dithiobis(propionimide),¹⁵⁴ 26 pairs of covalently cross-linked polypeptides could be unambiguously identified among the products from the reactions of the 30S subunit of the ribosome from *E. coli*.¹⁵⁵ After these results were reported, a crystallographic molecular model of the 30S subunit became available.^{156,157} Five of the cross-linked pairs involved polypeptides S1 and S21, which were not identified in the maps of electron density of the 30S subunit.¹⁵⁷ Of the remaining 21 pairs, nine have significant portions of their folded structure touching each other so intimately that their cross-linking would be expected and four may be near enough to each other in the crystallographic molecular model to be intramolecularly cross-linked by a long linker (Equation 10-46), but three have only a few positions in their sequences close enough to be cross-linked and five are not even near each other in the crystallographic molecular model. These last five cross-linked products must have been the result of intermolecular cross-linking, and some of the others may be as well.

It is also possible to covalently cross-link thiols that have been introduced into a protein with 2-iminothiolane. For example, such inserted thiols can be cross-linked with 4,6-di(bromomethyl)-3,7-dimethyl-1,5-diazabicyclo[3.3.0]octadiene-2,7-dione (dibromobimane):



This reagent makes the final cross-link strongly fluorescent so that peptides containing the cross-linked lysines, still joined together, can be purified to identify their positions in the sequence of the protein.¹⁵⁸

In most experiments involving covalent modification of a protein with an electrophilic reagent, the effect of the modification on the normal function of that protein is first monitored. When a reagent that has an inter-

esting or desirable effect is discovered, the position or positions in the amino acid sequence of the protein at which the modification has occurred must be identified in order to correlate this functional effect with the structure of the protein. This identification is usually made by digesting the modified protein either chemically or enzymatically, isolating the peptides that have been modified, and analyzing them by direct sequencing or by mass spectrometry. The precise **position of the modification in the amino acid sequence** is defined by the appearance of the modified amino acid itself in one of the cycles of sequencing, by the disappearance of the amino acid normally found at that cycle, or from the masses of the fragments produced by bombardment of the vaporized peptide with helium in a tandem **mass spectrometer** (Figure 3–8).

Two inactivated products from the alkylation of bovine pancreatic ribonuclease with iodoacetate could be separated from each other in their native state before they were digested. Each had incorporated one carboxymethyl group. From one of the products, a tryptic peptide (Histidine 105 to Valine 124) containing 1-carboxymethylhistidine at position 119 was isolated; from the other product, a tryptic peptide (Glutamine 11 to Lysine 31) containing 3-carboxymethylhistidine at position 12 was isolated.¹⁵⁹

In the chromatogram of the tryptic digest of ferredoxin–NADP⁺ reductase inactivated with *N*-ethyl[2,3-¹⁴C₂] maleimide (Equation 10–30), there was one major radioactive peak that had the amino acid sequence SVSLCVXR, comprising positions 110–117 in the sequence of the protein. In the sequence of the unmodified protein, the amino acid at position X is a lysine. Because lysine was not observed in that cycle of Edman degradation, because trypsin failed to cleave between the lysine and the arginine as it usually does, and because amino acid analysis of the peptide following its hydrolysis in acid produced a peak at the position of *N*-succinyllysine on the chromatogram (Figure 1–3), the modification could be assigned to Lysine 116.⁴

When γ -glutamyltransferase from *E. coli* that had been covalently modified with 2-amino-4-(fluorophosphono)butanoic acid was digested with lysyl endopeptidase (Figure 3–2) and the resulting peptides were separated by reverse-phase adsorption chromatography, only one previously unobserved peak of absorbance appeared on the chromatogram. A mass spectrum of the peptide responsible for that peak identified it as the peptide TTHYSVDDK (positions 391–399 in the sequence of the protein) into which 1 mole of the phosphorylating agent had been incorporated. A mass spectroscopic determination of the sequence of that peptide (Figure 3–8) identified Threonine 391 as the site of phosphorylation.¹²⁰

When bovine cytochrome *b*₅₆₁ that had been modified with diethyl pyrocarbonate was submitted to matrix-assisted-laser-desorption ionization in a time-of-flight

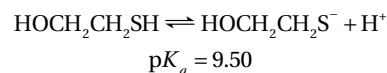
mass spectrometer, it was observed that the modification had increased the mass of the protein by the equivalent of three ethylpyrocarbonyl groups ($3 \times 72 \text{ Da} \approx 28,253 \text{ Da} - 28,033 \text{ Da}$). When a tryptic digest of the protein was submitted to mass spectrometry, 33 new peptides appeared in the spectrum,* the masses of all but two of which could be explained as the result of modification only at Lysine 85, Histidine 88, and Histidine 161.¹²¹

The advantage of identifying a site of modification chemically is that several different methods of analysis can be applied to the isolated product; the advantages of mass spectrometry are its rapidity and its sensitivity.

Suggested Reading

- Aliverti, A., Gadda, G., Ronchi, S., & Zanetti, G. (1991) Identification of Lys116 as the target of *N*-ethylmaleimide inactivation of ferredoxin:NADP⁺ oxidoreductase, *Eur. J. Biochem.* 198, 21–24.
- Inoue, M., Hiratake, J., Suzuki, H., Kumagai, H., & Sakata, K. (2000) Identification of the catalytic nucleophile of *Escherichia coli* γ -glutamyltranspeptidase by a γ -monofluorophosphono derivative of glutamic acid: *N*-terminal Thr-391 in small subunit is the nucleophile, *Biochemistry* 39, 7764–7771.
- Buechler, J.A., & Taylor, S.S. (1989) Dicyclohexylcarbodiimide cross-links the side chains of two conserved amino acids, Asp-184 and Lys-72, at the active site of the catalytic subunit of c-AMP-dependent protein kinase, *Biochemistry* 28, 2065–2070.

Problem 10–1: Mercaptoethanol undergoes the following dissociation:



Suppose the total amount of mercaptoethanol in solution is equal to $[\text{SH}]_{\text{TOT}}$, a quantity you know since you added that much. Suppose, also, that there is a chemical reaction that occurs between only the anion, $\text{HOCH}_2\text{CH}_2\text{S}^-$, and an electrophile, X, and the rate of this reaction is

$$\text{rate} = k[\text{HOCH}_2\text{CH}_2\text{S}^-][\text{X}]$$

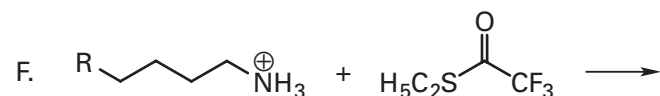
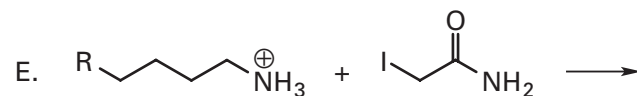
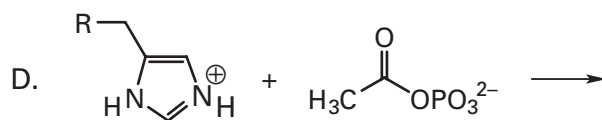
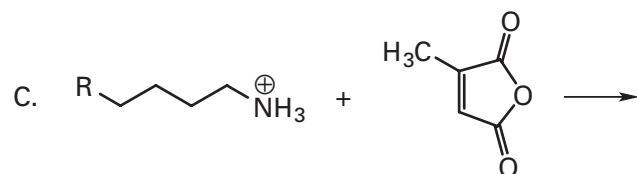
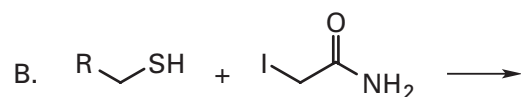
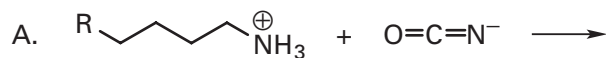
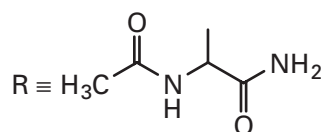
As the pH changes, $[\text{HOCH}_2\text{CH}_2\text{S}^-]$ changes although $[\text{SH}]_{\text{TOT}}$ is always the same.

- (A) Show that $\text{rate} = k\{f([\text{H}^+])\}[\text{SH}]_{\text{TOT}}[\text{X}]$
- (B) Give an explicit equation for $f([\text{H}^+])$.

* Of the 31 tryptic peptides, 24 were derivatives of the tryptic peptide from Threonine 83 to Arginine 111 containing Lysine 85 and Histidine 88, and seven were derivatives of the tryptic peptide from Tyrosine 157 to Lysine 191 containing Histidine 161. The various derivatives resulted from incomplete yields of modification, low yields of modification at other histidines in the peptides, and incomplete tryptic digestion.

- (C) Plot $\log [k_{\text{obs}}(k)^{-1}]$ against pH, where $k_{\text{obs}} \equiv k\{f([H^+])\}$. Indicate on the plot where $\text{pH} = \text{p}K_a$.
- (D) At what pH does the rate of the reaction become zero?
- (E) By what factor does the rate decrease for each decrease in pH of 1.00 when $\text{pH} < \text{p}K_a$?

Problem 10-2: Give all of the products and a mechanism for each of the following reactions.



Problem 10-3: Diisopropyl fluorophosphate inhibits serine endopeptidases by specific phosphorylation of the serine in the active site. Papain is an endopeptidase that does not have a serine in its active site. Nevertheless, it reacts with diisopropyl fluorophosphate with the result that 1 mol of phosphate is bound for every mole of enzyme but without loss of enzymatic activity.¹⁶⁰ The reaction between papain and the reagent was carried out with radioactive diisopropyl [³²P]fluorophosphate, the modified protein was digested with chymotrypsin, and the segment of enzyme containing the radioactive label

was isolated. It corresponded to amino acids 112–123 in the sequence of papain: –QVQPYNQGALLY–. What is the most likely site of alkylphosphorylation of papain?

Problem 10-4:

- (A) Draw the structure of the peptide RDVLMKE in the ionization state in which it would exist at pH 1.4. Indicate all lone pairs.

The peptide was modified with iodoacetamide at pH 1.4, 40 °C, for 20 h. Digestion with carboxypeptidase yielded the full complement of glutamic acid from the resulting peptide. Edman degradation yielded the full complement of arginine. It is possible to estimate the number of charges a peptide bears at a certain pH from its behavior on electrophoresis. This was done for the initial peptide and the product from its reaction with iodoacetamide.

pH	charge	
	original peptide	alkylated product
6.5	0	+1
2.1	+3	+4

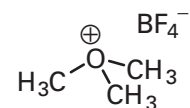
- (B) Write a stoichiometric mechanism for the reaction that occurred between iodoacetamide and one of the side chains on this peptide.
- (C) How would the product of the reaction with iodoacetamide move on chromatography by cation exchange relative to the unreacted peptide?

Problem 10-5: Write the complete reaction that occurs between a cysteine in a protein and 5,5'-dithiobis (2-nitrobenzoate). Write out the resonance forms of the nitrothiobenzoate dianion. What properties of this dianion do the resonance forms explain?

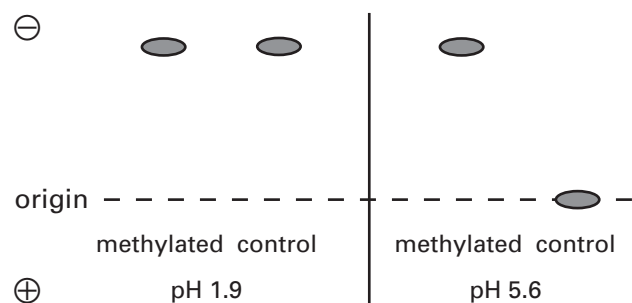
Problem 10-6: A peptide has the sequence SVEKCYEKP.

- (A) How many charges does the peptide bear at pH 1.9? At pH 5.6?

The peptide was reacted with trimethyloxonium tetrafluoroborate



in aqueous solution at pH 6.0. Three methyl groups were covalently attached to the peptide. When the methylated and unmethylated peptides were examined by electrophoresis, the following result was observed.



- (B) What nucleophiles on the peptide have reacted with the trimethyloxonium cation? Write a mechanism for this reaction. Why is the trimethyloxonium cation so reactive?

References

- Remington, S., Wiegand, G., & Huber, R. (1982) *J. Mol. Biol.* 158, 111–152.
- Stammers, D.K., & Muirhead, H. (1975) *J. Mol. Biol.* 95, 213–225.
- Fuller, G.M., & Doolittle, R.F. (1966) *Biochim. Biophys. Acta* 25, 694–700.
- Aliverti, A., Gadda, G., Ronchi, S., & Zanetti, G. (1991) *Eur. J. Biochem.* 198, 21–24.
- Klotz, I.M., & Keresztes-Nagy, S. (1962) *Nature* 195, 900–901.
- Gibbons, I., & Schachman, H.K. (1976) *Biochemistry* 15, 52–60.
- Patthy, L., & Smith, E.L. (1975) *J. Biol. Chem.* 250, 557–564.
- Koshland, D.E., Karkhanis, Y.D., & Latham, H.G. (1964) *J. Am. Chem. Soc.* 86, 1448–1450.
- Wu, C.W., & Stryer, L. (1972) *Proc. Natl. Acad. Sci. U.S.A.* 69, 1104–1108.
- Kaplan, H., Stevenson, K.J., & Hartley, B.S. (1971) *Biochem. J.* 124, 289–299.
- Krekel, F., Samland, A.K., Macheroux, P., Amrhein, N., & Evans, J.N. (2000) *Biochemistry* 39, 12671–12677.
- Dorner, F. (1971) *J. Biol. Chem.* 246, 5896–5902.
- Wofsy, L., & Singer, S.J. (1963) *Biochemistry* 2, 104–116.
- Gurd, F.R.N. (1967) *Methods Enzymol.* 11, 532–541.
- Schnackerz, K.D., & Noltmann, E.A. (1970) *J. Biol. Chem.* 245, 6417–6423.
- Gundlach, H.G., Stein, W.H., & Moore, S. (1959) *J. Biol. Chem.* 234, 1754–1761.
- Gundlach, H.G., Moore, S., & Stein, W.H. (1959) *J. Biol. Chem.* 234, 1761–1764.
- Hand, E.S., & Jencks, W.P. (1962) *J. Am. Chem. Soc.* 84, 3505–3514.
- Makoff, A.J., & Malcolm, A.D. (1981) *Biochem. J.* 193, 245–249.
- Makoff, A.J., & Malcolm, A.D. (1980) *Eur. J. Biochem.* 106, 313–320.
- DiMarchi, R.D., Garner, W.H., Wang, C.C., Hanania, G.I., & Gurd, F.R. (1978) *Biochemistry* 17, 2822–2828.
- Means, G.E., & Feeney, R.E. (1968) *Biochemistry* 7, 2192–2201.
- Rippa, M., Spanio, L., & Pontremoli, S. (1967) *Arch. Biochem. Biophys.* 118, 48–57.
- Stark, G.R. (1970) *Adv. Protein Chem.* 24, 261–308.
- Riordan, J.F., & Vallee, B.L. (1967) *Methods Enzymol.* 11, 565–570.
- Anderson, G.W., Zimmerman, J.E., & Callahan, F.M. (1964) *J. Am. Chem. Soc.* 86, 1839–1842.
- Serjeant, E.P., & Dempsey, B. (1979) *Ionisation Constants of Organic Acids in Aqueous Solution*, Pergamon Press, Oxford, England.
- Miles, E.W. (1977) *Methods Enzymol.* 47, 431–442.
- Goldberger, R.F., & Anfinsen, C.B. (1962) *Biochemistry* 1, 401–405.
- Riordan, J.F., & Vallee, B.L. (1967) *Methods Enzymol.* 11, 570–576.
- Dixon, H.B., & Perham, R.N. (1968) *Biochem. J.* 109, 312–314.
- Weber, G. (1952) *Biochem. J.* 51, 155–167.
- Fields, R. (1972) *Methods Enzymol.* 25, 464–468.
- Henkart, P. (1971) *J. Biol. Chem.* 246, 2711–2713.
- Hartman, F.C., Milanez, S., & Lee, E.H. (1985) *J. Biol. Chem.* 260, 13968–13975.
- Rogers, G.A., Shaltiel, N., & Boyer, P.D. (1976) *J. Biol. Chem.* 251, 5711–5717.
- Westhead, E.W. (1965) *Biochemistry* 4, 2139–2144.
- Bond, J.S., Francis, S.H., & Park, J.H. (1970) *J. Biol. Chem.* 245, 1041–1053.
- Riordan, J.F., & Vallee, B.L. (1967) *Methods Enzymol.* 11, 541–548.
- Banas, T., Gontero, B., Drews, V.L., Johnson, S.L., Marcus, F., & Kemp, R.G. (1988) *Biochim. Biophys. Acta* 957, 178–184.
- Cohen, L.A. (1968) *Annu. Rev. Biochem.* 37, 695–726.
- Raftery, M.A., & Cole, R.D. (1963) *Biochem. Biophys. Res. Commun.* 10, 467–472.
- Raftery, M.A., & Cole, R.D. (1966) *J. Biol. Chem.* 241, 3457–3461.
- Cole, R.D. (1967) *Methods Enzymol.* 11, 315–317.
- Hollenbaugh, D., Aruffo, A., & Senter, P.D. (1995) *Biochemistry* 34, 5678–5684.
- Lindley, H. (1956) *Nature* 178, 647–648.
- Ellman, G.L. (1959) *Arch. Biochem. Biophys.* 82, 70–77.
- McConahey, P.J., & Dixon, F.J. (1966) *Int. Arch. Allergy Appl. Immunol.* 29, 185–189.
- Hubbard, A.L., & Cohn, Z.A. (1972) *J. Cell Biol.* 55, 390–405.
- Markwell, M.A. (1982) *Anal. Biochem.* 125, 427–432.
- Sun, W., & Dunford, H.B. (1993) *Biochemistry* 32, 1324–1331.
- Horinishi, H., Hachimori, Y., Kurihara, K., & Shibata, K. (1964) *Biochim. Biophys. Acta* 86, 477–489.
- Bruice, T.C., Gregory, M.J., & Walters, S.L. (1968) *J. Am. Chem. Soc.* 90, 1612–1619.
- Sokolovsky, M., Riordan, J.F., & Vallee, B.L. (1967) *Biochem. Biophys. Res. Commun.* 27, 20–25.
- Holt, L.A., Milligan, B., & Rivett, D.E. (1971) *Biochemistry* 10, 3559–3564.
- Scoffone, E., Fontana, A., & Rocchi, R. (1968) *Biochemistry* 7, 971–979.
- Omenn, G.S., Fontana, A., & Anfinsen, C.B. (1970) *J. Biol. Chem.* 245, 1895–1902.
- Takahashi, K. (1968) *J. Biol. Chem.* 243, 6171–6179.

59. Nishimura, T., & Kitajima, K. (1979) *J. Org. Chem.* 44, 818–824.
60. Soman, G., Hurst, M.O., & Graves, D.J. (1985) *Int. J. Pept. Protein Res.* 25, 517–525.
61. Duerksen-Hughes, P.J., Williamson, M.M., & Wilkinson, K.D. (1989) *Biochemistry* 28, 8530–8536.
62. Toi, K., Bynum, E., Norris, E., & Itano, H.A. (1967) *J. Biol. Chem.* 242, 1036–1043.
63. Yankeelov, J.A. (1970) *Biochemistry* 9, 2433–2439.
64. Itano, H.A., & Gottlieb, A.J. (1963) *Biochem. Biophys. Res. Commun.* 12, 405–408.
65. Riordan, J.F. (1973) *Biochemistry* 12, 3915–3923.
66. Patthy, L., Vaaradi, A., Thaes, J., & Kovaacs, K. (1979) *Eur. J. Biochem.* 99, 309–313.
67. Hoare, D.G., & Koshland, D.E., Jr. (1967) *J. Biol. Chem.* 242, 2447–2453.
68. Buechler, J.A., & Taylor, S.S. (1989) *Biochemistry* 28, 2065–2070.
69. Lewis, S.D., & Shafer, J.A. (1973) *Biochim. Biophys. Acta* 303, 284–291.
70. Belleau, B., & Malek, G. (1968) *J. Am. Chem. Soc.* 90, 1651–1652.
71. Bertrand, R., Chaussepied, P., Kassab, R., Boyer, M., Roustan, C., & Benyamin, Y. (1988) *Biochemistry* 27, 5728–5736.
72. Woodward, R.B., Olofson, R.A., & Mayer, H. (1961) *J. Am. Chem. Soc.* 83, 1010–1012.
73. Llamas, K., Owens, M., Blakeley, R.L., & Zerner, B. (1986) *J. Am. Chem. Soc.* 108, 5543–5548.
74. Jennings, M.L., & Anderson, M.P. (1987) *J. Biol. Chem.* 262, 1691–1697.
75. Rajasekharan, R., Mariani, R.C., Shockey, J.M., & Kemp, J.D. (1993) *Biochemistry* 32, 12386–12391.
76. Bayley, H., & Knowles, J.R. (1977) *Methods Enzymol.* 46, 69–114.
77. Iddon, B., Meth-Cohn, O., Scriven, E.F.V., Suschitzky, H.J., & Gallagher, P.T. (1979) *Angew. Chem., Int. Ed. Engl.* 18, 900–917.
78. Reiser, A., & Leyshon, L.J. (1971) *J. Am. Chem. Soc.* 93, 4051–4052.
79. Abramovitch, R.A., & Davis, B.A. (1964) *Chem. Rev.* 64, 149–185.
80. Nielsen, P.E., & Buchardt, O. (1982) *Photochem. Photobiol.* 35, 317–323.
81. Chen, S., Lee, T.D., Legesse, K., & Shively, J.E. (1986) *Biochemistry* 25, 5391–5395.
82. Lewis, C.T., Haley, B.E., & Carlson, G.M. (1989) *Biochemistry* 28, 9248–9255.
83. Garin, J., Boulay, F., Issartel, J.P., Lunardi, J., & Vignais, P.V. (1986) *Biochemistry* 25, 4431–4437.
84. Richards, F.F., Lifter, J., Hew, C.L., Yoshioka, M., & Konigsberg, W.H. (1974) *Biochemistry* 13, 3572–3575.
85. Buron, C., Gornitzka, H., Romanenko, V.V., & Bertrand, G. (2000) *Science* 288, 834–836.
86. Hirai, K., & Tomioka, H. (1999) *J. Am. Chem. Soc.* 121, 10213–10214.
87. Westerman, J., Wirtz, K.W., Berkhout, T., van Deenen, L.L., Radhakrishnan, R., & Khorana, H.G. (1983) *Eur. J. Biochem.* 132, 441–449.
88. Morgan, S., Jackson, J.E., & Platz, M.S. (1991) *J. Am. Chem. Soc.* 113, 2782–2783.
89. Brunner, J., Senn, H., & Richards, F.M. (1980) *J. Biol. Chem.* 255, 3313–3318.
90. Hexter, C.S., & Westheimer, F.H. (1971) *J. Biol. Chem.* 246, 3934–3938.
91. Hexter, C.S., & Westheimer, F.H. (1971) *J. Biol. Chem.* 246, 3928–3933.
92. Vaughan, R.J., & Westheimer, F.H. (1979) *J. Am. Chem. Soc.* 101, 217–218.
93. Brunner, J., & Richards, F.M. (1980) *J. Biol. Chem.* 255, 3319–3329.
94. Ross, A.H., Radhakrishnan, R., Robson, R.J., & Khorana, H.G. (1982) *J. Biol. Chem.* 257, 4152–4161.
95. Shafer, J., Baronowsky, P., Laursen, R., Finn, F., & Westheimer, F.H. (1966) *J. Biol. Chem.* 241, 421–427.
96. Blatter, E.E., Ebright, Y.W., & Ebright, R.H. (1992) *Nature* 359, 650–652.
97. Cremo, C.R., Grammer, J.C., & Yount, R.G. (1988) *Biochemistry* 27, 8415–8420.
98. Ko, Y.H., Cremo, C.R., & McFadden, B.A. (1992) *J. Biol. Chem.* 267, 91–95.
99. Hoyer, D., Cho, H., & Schultz, P.G. (1990) *J. Am. Chem. Soc.* 112, 3249–3250.
100. Schepartz, A., & Cuenoud, B. (1990) *J. Am. Chem. Soc.* 112, 3247–3249.
101. Rana, T.M., & Meares, C.F. (1990) *J. Am. Chem. Soc.* 112, 2457–2458.
102. Gallagher, J., Zelenko, O., Walts, A.D., & Sigman, D.S. (1998) *Biochemistry* 37, 2096–2104.
103. Ettner, N., Metzger, J.W., Lederer, T., Hulmes, J.D., Kisker, C., Hinrichs, W., Ellestad, G.A., & Hillen, W. (1995) *Biochemistry* 34, 22–31.
104. Ue, K., Muhrad, A., Edmonds, C.G., Bivin, D., Clark, A., Piechowski, W.V., & Morales, M.F. (1992) *Eur. J. Biochem.* 203, 493–498.
105. Hutchison, C.A., III, Phillips, S., Edgell, M.H., Gillam, S., Jahnke, P., & Smith, M. (1978) *J. Biol. Chem.* 253, 6551–6560.
106. Zoller, M.J., & Smith, M. (1982) *Nucleic Acids Res.* 10, 6487–6500.
107. Dwyer, B.P. (1988) *Biochemistry* 27, 5586–5592.
108. Dwyer, B.P. (1991) *Biochemistry* 30, 4105–4112.
109. Xu, K.Y. (1989) *Biochemistry* 28, 6894–6899.
110. Erickson, H.K. (2000) *Biochemistry* 39, 9241–9250.
111. Walter, G., Scheidtmann, K.H., Carbone, A., Laudano, A.P., & Doolittle, R.F. (1980) *Proc. Natl. Acad. Sci. U.S.A.* 77, 5197–5200.
112. Wilchek, M., Bocchini, V., Becker, M., & Givol, D. (1971) *Biochemistry* 10, 2828–2834.
113. Kyte, J., Xu, K.Y., & Bayer, R. (1987) *Biochemistry* 26, 8350–8360.
114. Thibault, D. (1993) *Biochemistry* 32, 2813–2821.
115. Castellino, F.J., & Hill, R.L. (1970) *J. Biol. Chem.* 245, 417–424.
116. Parente, A., Merrifield, B., Geraci, G., & D'Alessio, G. (1985) *Biochemistry* 24, 1098–1104.
117. Xu, K.Y., & Kyte, J. (1989) *Biochemistry* 28, 3009–3017.
118. Toyoshima, C., Nakasako, M., Nomura, H., & Ogawa, H. (2000) *Nature* 405, 647–655.
119. Hanai, R., & Wang, J.C. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 11904–11908.
120. Inoue, M., Hiratake, J., Suzuki, H., Kumagai, H., & Sakata, K. (2000) *Biochemistry* 39, 7764–7771.

121. Tsubaki, M., Kobayashi, K., Ichise, T., Takeuchi, F., & Tagawa, S. (2000) *Biochemistry* 39, 3276–3284.
122. Light, A., Frater, R., Kimmel, J.R., & Smith, E.L. (1964) *Proc. Natl. Acad. Sci. U.S.A.* 52, 1276–1283.
123. Chaiken, I.M., & Smith, E.L. (1969) *J. Biol. Chem.* 244, 5087–5094.
124. Kannan, K.K., Notstrand, B., Frideborg, K., Leovgren, S., Ohlsson, A., & Petef, M. (1975) *Proc. Natl. Acad. Sci. U.S.A.* 72, 51–55.
125. Abita, J.P., Maroux, S., Delaage, M., & Lazdunski, M. (1969) *FEBS Lett.* 4, 203–206.
126. Freer, S.T., Kraut, J., Robertus, J.D., Wright, H.T., & Xuong, N.H. (1970) *Biochemistry* 9, 1997–2009.
127. Lambert, J.M., Perham, R.N., & Coggins, J.R. (1977) *Biochem. J.* 161, 63–71.
128. Pabo, C.O., & Lewis, M. (1982) *Nature* 298, 443–447.
129. Reidhaar-Olson, J.F., & Sauer, R.T. (1988) *Science* 241, 53–57.
130. Hugli, T.E. (1973) *J. Biol. Chem.* 248, 1712–1718.
131. Oefner, C., & Suck, D. (1986) *J. Mol. Biol.* 192, 605–632.
132. Lu, R.C., & Szilagyi, L. (1981) *Biochemistry* 20, 5914–5919.
133. Hegyi, G., Premecz, G., Sain, B., & Muhrad, A. (1974) *Eur. J. Biochem.* 44, 7–12.
134. Burtnick, L.D. (1984) *Biochim. Biophys. Acta* 791, 57–62.
135. Holmes, K.C., Popp, D., Gebhard, W., & Kabsch, W. (1990) *Nature* 347, 44–49.
136. Traviglia, S.L., Datwyler, S.A., Yan, D., Ishihama, A., & Meares, C.F. (1999) *Biochemistry* 38, 15774–15778.
137. Greiner, D.P., Hughes, K.A., Gunasekera, A.H., & Meares, C.F. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 71–75.
138. Spackman, D.H., Stein, W.H., & Moore, S. (1960) *J. Biol. Chem.* 235, 648–659.
139. Mitra, S., & Lawton, R.G. (1979) *J. Am. Chem. Soc.* 101, 3097–3110.
140. Meloche, H.P. (1973) *J. Biol. Chem.* 248, 6945–6951.
141. Suzuki, N., & Wood, W.A. (1980) *J. Biol. Chem.* 255, 3427–3435.
142. Allard, J., Grochulski, P., & Sygusch, J. (2001) *Proc. Natl. Acad. Sci. U.S.A.* 98, 3679–3684.
143. Sutoh, K. (1982) *Biochemistry* 21, 3654–3661.
144. Kabsch, W., Mannherz, H.G., Suck, D., Pai, E.F., & Holmes, K.C. (1990) *Nature* 347, 37–44.
145. Cai, K., Klein-Seetharaman, J., Altenbach, C., Hubbell, W.L., & Khorana, H.G. (2001) *Biochemistry* 40, 12479–12485.
146. Jue, R., Lambert, J.M., Pierce, L.R., & Traut, R.R. (1978) *Biochemistry* 17, 5399–5406.
147. Clegg, C., & Hayes, D. (1974) *Eur. J. Biochem.* 42, 21–28.
148. Expert-Bezancon, A., Barritault, D., Milet, M., Guerin, M.F., & Hayes, D.H. (1977) *J. Mol. Biol.* 112, 603–629.
149. Lutter, L.C., Bode, U., Kurland, C.G., & Stoffler, G. (1974) *Mol. Gen. Genet.* 129, 167–176.
150. Chang, F.N., & Flaks, J.G. (1972) *J. Mol. Biol.* 68, 177–180.
151. Lutter, L.C., Zeichhardt, H., & Kurland, C.G. (1972) *Mol. Gen. Genet.* 119, 357–366.
152. Shih, C.Y., & Craven, G.R. (1973) *J. Mol. Biol.* 78, 651–663.
153. Lutter, L.C., Kurland, C.G., & Stoffler, G. (1975) *FEBS Lett.* 54, 144–150.
154. Peretz, H., Towbin, H., & Elson, D. (1976) *Eur. J. Biochem.* 63, 83–92.
155. Sommer, A., & Traut, R.R. (1976) *J. Mol. Biol.* 106, 995–1015.
156. Wimberly, B.T., Brodersen, D.E., Clemons, W.M., Jr., Morgan-Warren, R.J., Carter, A.P., Vonnrhein, C., Hartsch, T., & Ramakrishnan, V. (2000) *Nature* 407, 327–339.
157. Pioletti, M., Schlunzen, F., Harms, J., Zarivach, R., Gluhmann, M., Avila, H., Bashan, A., Bartels, H., Auerbach, T., Jacobi, C., Hartsch, T., Yonath, A., & Franceschi, F. (2001) *EMBO J.* 20, 1829–1839.
158. Sinz, A., & Wang, K. (2001) *Biochemistry* 40, 7903–7913.
159. Crestfield, A.M., Stein, W.H., & Moore, S. (1963) *J. Biol. Chem.* 238, 2413–2419.
160. Chaiken, I.M., & Smith, E.L. (1969) *J. Biol. Chem.* 244, 4247–4250.

Chapter 11

Immunochemical Probes of Structure

Immunoglobulins are proteins found, among other locations, in the blood serum of birds and mammals. Immunoglobulins are also called **antibodies**. In an animal, the function of an immunoglobulin is to recognize a foreign macromolecule, the antigen, by binding to it tightly. An **antigen** is any foreign macromolecule that elicits, upon its introduction into an animal, the production of immunoglobulins capable of binding to it with high affinity. Within the animal, when an antigen has been recognized by being bound to the immunoglobulin, it is usually destroyed. An important point which should be kept in mind is that the primary biological purpose for a particular immunoglobulin is to distinguish one particular foreign, undesirable macromolecule from the myriad of necessary macromolecules indigenous to the animal. Whenever an immunoglobulin makes a mistake by recognizing and binding not only to its antigen but also to one or more of the macromolecules normally present in the animal, these indigenous macromolecules are also destroyed in autoimmune processes detrimental to the animal. Therefore, the immune system has evolved to produce immunoglobulins that are highly specific in their recognition of molecular structure. Almost any macromolecule can serve as an antigen, but proteins are the most common antigens.

Because no predictions can be made as to what foreign antigens will have to be recognized and destroyed during the life of the animal, the immune system must be prepared to make immunoglobulins capable of **binding with high specificity** to any foreign molecule when it is presented to the animal in an antigenic form. An extreme example of the ability of the immune system to produce immunoglobulins able to bind any foreign molecule is the production of immunoglobulins that bind C_{60} fullerene,¹ a form of elemental carbon that is not encountered in any natural setting and that does not resemble any natural antigen.

The **serum** from any mammal or bird contains a wide variety of immunoglobulins, each with its own distinct amino acid sequence and each present in its own distinct concentration. They are the immunoglobulins that have been produced in response to all of the foreign antigens encountered by that particular individual during its peculiar lifetime. This mixture of immunoglobulins is present in the serum at a total concentration of 10–20 mg mL⁻¹.

The immune system of an animal can be stimulated

to produce new immunoglobulins recognizing a particular protein of interest by injecting that protein into the animal. Because the repertoire of immunoglobulins that an animal is capable of producing contains none that would recognize any of its own proteins, or they would be destroyed, the protein injected has to be from another species and different enough from any related indigenous protein to be recognized as foreign. Because the oligosaccharides found on the proteins of animals are so similar and because most species contain every possible sequence of these common oligosaccharides as a result of microheterogeneity, an immunoglobulin specific for an oligosaccharide on the protein from an animal will rarely be produced. There are no such problems, however, in producing immunoglobulins in animals that recognize bacterial oligosaccharides.² Because the immune system has evolved to recognize and destroy foreign organisms such as viruses and bacteria, the surfaces of which are large aggregates of many subunits or many different proteins, small proteins sometimes have to be covalently cross-linked to make them antigenic.³ If the **immunization** is successful, immunoglobulins that bind tightly to the protein that was injected appear at high concentration in the serum of the animal within two months.

The paradigm of the various types of immunoglobulins found in serum is **immunoglobulin G** (Figures 7–13 and 11–1).^{4–6} An immunoglobulin G is composed of two identical heavy α polypeptides ($n_{aa} = 440\text{--}450$) and two identical light β polypeptides ($n_{aa} = 210\text{--}220$).⁵ Each heavy α polypeptide is folded into four **internally repeating, superposable domains** designated V_H , C_{H1} , C_{H2} , and C_{H3} in the order in which they occur in the sequence of the protein. Each light β polypeptide is folded into two internally repeating, superposable domains, V_L and C_L . Each of these six different domains, each approximately 110 amino acids in length and each present in two copies in the intact immunoglobulin G, is superposable in its folded form on each of the other five (Figure 7–13). The V_H and V_L domains associate with each other, the C_L and C_{H1} domains associate with each other, and these associations produce an $\alpha\beta$ heterodimer of one heavy α subunit and one light β subunit. Two of these $\alpha\beta$ heterodimers are associated through their C_{H2} and C_{H3} domains to form the entire immunoglobulin.

An immunoglobulin G can be cut into three pieces. When the intact native protein is treated with papain,^{7,8}

cleavage occurs to the amino-terminal side of the cystines connecting the two heavy α polypeptides within the open, structureless segments (Figure 11-1), and two identical Fab fragments and one **Fc fragment** are produced. The designation Fab arises from the fact that this fragment contains the site that binds the antigen. The designation Fc originally referred to the fact that this fragment could be crystallized. It is now more informative and consistent to consider this the constant fragment. It is because it is constant that it can crystallize. Each of these fragments is a well-behaved, independent, soluble, globular protein. Each contains four of the original 12 internally repeating, superposable domains.

The advantage of using **Fab fragments** in experiments is that they are **univalent**. Each Fab fragment contains only one binding site for the antigen. An intact molecule of immunoglobulin G is necessarily bivalent because, as an $(\alpha\beta)_2$ homodimer, it must have two identical binding sites for antigen (Figure 11-1). The fact that two antigens can be bound by intact immunoglobulin G

complicates some experiments. A **bivalent analogue** of the Fab fragment can be produced by digesting intact immunoglobulin G with pepsin.^{9,10} The pepsin cleaves to the carboxy-terminal side of the cystines between the two heavy α subunits and produces a fragment, $(\text{Fab}')_2$, containing two Fab fragments joined by two or more cystines. The advantage of an $(\text{Fab}')_2$ fragment is that when it is reduced with dithiothreitol or 2-mercaptoethanol, it dissociates into two monovalent Fab' fragments the $\text{C}_{\text{H}}1$ domains of which are slightly longer than those of an Fab fragment. A permanently bivalent fragment that is the same size as an Fab fragment can be made by fusing the cDNAs encoding the V_{L} and V_{H} domains of a particular immunoglobulin. If the two cDNAs are connected to each other by a segment of DNA encoding a segment of flexible polypeptide too short to permit the intramolecular association of the domains in the expressed protein, they associate intermolecularly to form an antiparallel $(\text{V}_{\text{L}}-\text{V}_{\text{H}})_2$ homodimer with two identical sites for binding antigen.¹¹

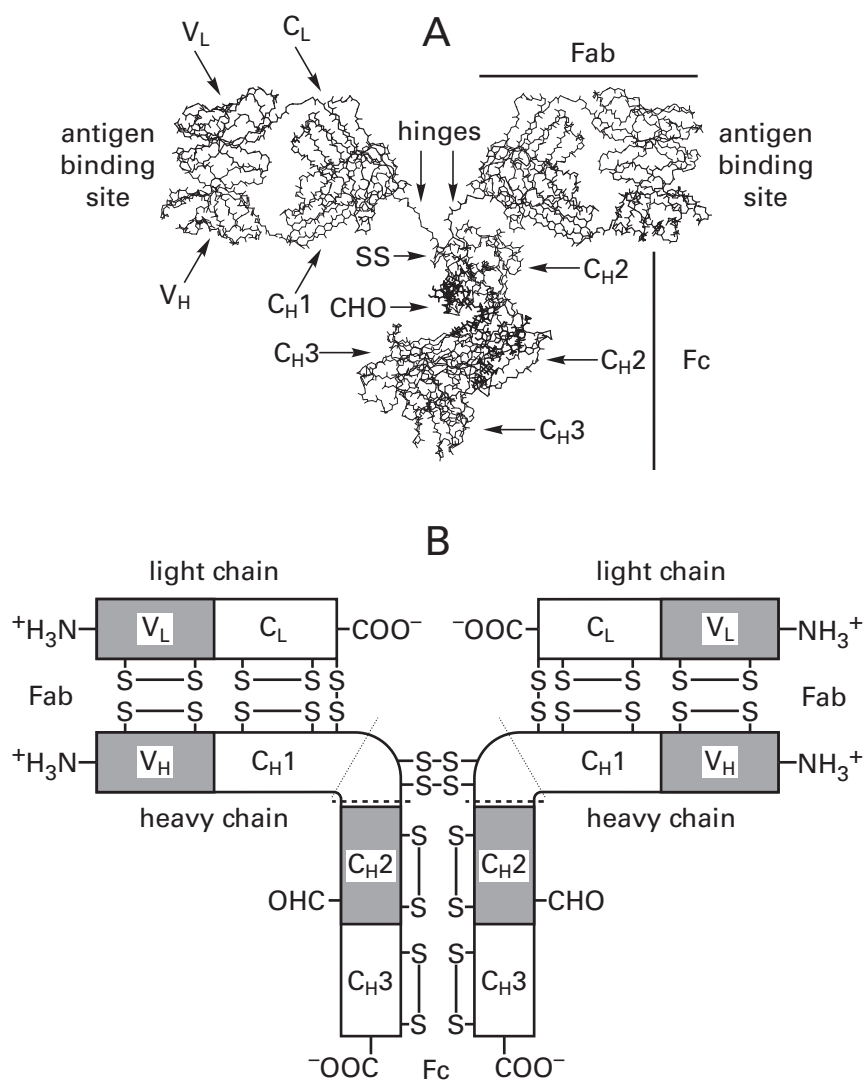


Figure 11-1: Structure of a molecule of immunoglobulin G. (A) Skeletal drawing of the polypeptide backbone of the crystallographic molecular model of immunoglobulin G⁶ that is presented in stereo in Figure 7-13. This drawing was produced with MolScript.¹⁰³ (B) Diagrammatic representation of the molecule based on the internally repeating domains observed in its amino acid sequences.⁵ (Adapted with permission from ref 5. Copyright 1969 National Academy of Sciences.) The complete molecule is composed of 12 superposable domains, all homologous in amino acid sequence, each about 110 amino acids in length. In the center of each domain is a cystine (S—S) formed between two structurally adjacent cysteines, 60 amino acids apart in the amino acid sequence. A light β polypeptide and a heavy α polypeptide are linked by a cystine at the carboxy terminus of the light β polypeptide, and heavy α polypeptides are linked together by two or more cystines between the hinges that join the three arms. The 12 domains are referred to as variable domain, heavy α polypeptide (V_{H}); constant domain 1, heavy α polypeptide ($\text{C}_{\text{H}}1$); constant domain 2, heavy α polypeptide ($\text{C}_{\text{H}}2$); constant domain 3, heavy α polypeptide ($\text{C}_{\text{H}}3$); variable domain, light β polypeptide (V_{L}); and constant domain, light β polypeptide (C_{L}). Oligosaccharides (CHO) are attached to the two $\text{C}_{\text{H}}2$ domains. The three arms are the two antigen-binding fragments (Fab) and the constant fragment (Fc). The binding sites for the antigens are at the tips of the Fab arms and are formed by the variable domains from heavy and light subunits, respectively. The light dotted lines indicate where papain cleaves the molecule to produce the two Fab fragments and the Fc fragment; the heavy dashed lines indicate where pepsin cleaves the molecule to produce the $(\text{Fab}')_2$ fragment. The Fab fragments are missing the cystines holding the fragments of the heavy α subunits together; the $(\text{Fab}')_2$ fragments include these cystines. The hinges at which the cleavages by endopeptidases occur are indicated by the arrows in panel A.

The flexible, unsupported segments of polypeptide that connect the Fab portions to the Fc portion of an intact immunoglobulin G (Figure 11-1) are usually about 20 aa long but can be as long as 70 aa.¹² These are its **hinges**. It is the open structure of these hinges that permits the papain or the pepsin to cleave the immunoglobulin G into its fragments. Because these hinges are so long, the two Fab portions in an intact immunoglobulin G are constantly moving relative to each other and relative to the Fc portion. When these segments are shortened sufficiently by site-directed mutation, the immunoglobulin becomes rigid.¹³

The **major immunoglobulins** in the serum of a mammal are immunoglobulins G (10–20 mg mL⁻¹), immunoglobulins M (1 mg mL⁻¹), and immunoglobulins A (1 mg mL⁻¹). Each of these immunoglobulins contains light β subunits that are indistinguishable from one type to the next. It is the heavy α subunits, always present in equimolar ratio to the light β subunits, that distinguish one type of immunoglobulin¹⁴ from the other. The heavy α subunits of all of these immunoglobulins are homologous to each other over the first three domains. It is in the peripheral portions of their Fc segments that they differ.

Immunoglobulin M has a longer heavy α polypeptide ($n_{aa} = 570$ –580) than the one in immunoglobulin G by one extra domain, $C_{\mu}4$, which would be the analogue of $C_{H}4$, if $C_{H}4$ existed. Immunoglobulin M is a pentameric complex of five $(\alpha\beta)_2$ heterotetramers held together by cystines among themselves. The cystines cross-link pairs of $C_{\mu}3$ domains to form a pentameric ring of the heterotetramers.

Immunoglobulin A has a heavy α polypeptide only about 30 amino acids ($n_{aa} = 470$ –480)¹⁵ longer than that of immunoglobulin G. Immunoglobulin A is a mixture of monomeric $(\alpha\beta)_2$ heterotetramers similar to those of immunoglobulin G and higher oligomers of $(\alpha\beta)_2$ heterotetramers held together by cystines between their $C_{\alpha}3$ domains.¹⁶

Both immunoglobulins M and A have a short polypeptide J associated with them that may promote their initial oligomerization even before the intertetrameric cystines are formed.¹⁶ Immunoglobulins M and A have their binding sites for antigens in a similar location to those on immunoglobulins G and distant from the regions ($C_{\mu}3$, $C_{\mu}4$, and $C_{\alpha}3$) that account for their distinct oligomeric structures. The only significant differences between these types of immunoglobulins and immunoglobulins G is their size and, hence, their valence. Monovalent Fab fragments can be produced from each.^{14,15,17} Although the injection of an antigen usually stimulates the production of immunoglobulins G, it is not unusual for it to stimulate production of the other types, either instead of or along with immunoglobulins G.

An immunoglobulin of a particular sequence is produced by a **colony of lymphocytes**, all derived from one single cell that was initially stimulated to divide and

manufacture. All members of the colony secrete immunoglobulins with identical α polypeptides and identical β polypeptides. The colony assumes its identity from its pedigree and not from its situation. All of the lymphocytes in the colony are descendants of the same cell, but each member of the colony, like all other lymphocytes, is dispersed by the bloodstream and lymphatic system and wanders independently and at random through the animal as it continuously manufactures its particular immunoglobulin. The sole product of the members of a particular colony is this one immunoglobulin continuously released into the serum and extracellular fluid. Each time a lymphocyte is **stimulated to divide and manufacture** its particular immunoglobulin against a particular antigen, a new colony is established. The particular amino acid sequences of the two subunits of the immunoglobulin produced by a particular colony confer the ability to bind a particular antigen.

Because many (10–100) different lymphocytes are stimulated to divide and manufacture by molecules of the same antigen, many different colonies continuously produce immunoglobulins after exposure of an animal to a particular antigen. Each of these immunoglobulins has a different amino acid sequence, but all are specific for that one antigen. They differ, however, in the location on the surface of the antigen that they recognize and to which they bind and in the strength with which they bind to that location, as reflected in their individual dissociation constants for the binding of antigen. Such a set of immunoglobulins, each capable of recognizing the same antigen but each different from the others, is referred to as a **polyclonal** set. The product of the reaction of an intact animal to an antigen is always a polyclonal set of immunoglobulins, which are present as a complex mixture in the serum of the animal.

In a normal animal, the various colonies are stable contributors to the mixture of immunoglobulins in the serum necessary to deal with antigens in the environment. Occasionally, the controls maintaining the stable population of the colony fail, and one lymphocyte begins to multiply malignantly. This uncontrolled cancerous growth causes an enormous increase in the number of lymphocytes producing an immunoglobulin of just one unique sequence and structure. Such a cancer is referred to as a myeloma. The serum of such individuals contains high concentrations of only one type of immunoglobulin. Such **myeloma proteins** are present in sufficient quantities to be purified, sequenced,⁵ and crystallized.⁴ Myeloma proteins appear by chance as the products of the random malignant transformation of normal lymphocytes, and the antigens to which most myelomas are directed are unknown.

The disadvantage of a myeloma protein is that the investigator cannot choose the antigen against which it is directed. Its advantage is that it can be purified to homogeneity, and the purified protein is necessarily composed of identical copies of the same molecule, each with an

identical ability to recognize the antigen. The advantage of the ability to select the antigen and the advantage of the homogeneity of the product have been combined in the production of **monoclonal immunoglobulins**.¹⁸ In this procedure, lymphocytes from the spleen of a mouse that has been immunized with the antigen of interest are fused with cultured murine myeloma cells that normally secrete a particular myeloma protein or, even better, myeloma cells that have lost their incitement to secrete an immunoglobulin. These cultured myeloma cells are immortal cell lines that were originally derived from a myeloma in a mouse and that continuously grow and divide either in flasks in an incubator or as solid tumors in mice. Hybrids, each produced by the fusion of one lymphocyte and one myeloma cell, are selected on the basis of their ability to grow on a particular medium. The hybrids are then reproduced in dishes as single colonies of cells.

Because each colony in the dish arose from one single cell, the cells in a particular colony produce only the immunoglobulin originally secreted by the parental lymphocyte that fused to the myeloma cell and the myeloma protein originally secreted by that myeloma cell. Therefore, each colony is the offspring of only one lymphocyte in the mouse from which the spleen was taken. If that lymphocyte happened to be producing one of the immunoglobulins directed against the antigen originally injected into the mouse, its offspring can be cultivated for the production of a homogeneous monoclonal immunoglobulin recognizing that antigen. To identify the colonies secreting monoclonal immunoglobulins against the antigen of interest, each colony is individually screened. When a colony producing a monoclonal immunoglobulin that has the desired specificity has been identified by the screen, it is expanded either in culture or in an animal so that significant amounts of that monoclonal immunoglobulin can be produced and purified.

Although a few intact myeloma proteins⁴ and intact monoclonal immunoglobulins (Figure 11-1)¹⁹ have been crystallized and submitted to crystallographic analysis, most of the crystallographic molecular models containing complexes with antigens are those of Fab fragments.^{20,21} The sites on the surface of an intact immunoglobulin that bind tightly to the two respective copies of the antigen are at the far ends of the Fab arms (Figure 11-1), at the tip of the portion of the intact molecule formed from the association of a V_H domain and a V_L domain. The Fab fragment retains this site in its entirety, and it is on the opposite end of the fragment from the carboxy-terminal point of cleavage that produced the Fab fragment. An example of a complex between an Fab fragment and its antigen is that between lysozyme from *Gallus gallus* and the Fab fragment of a murine monoclonal immunoglobulin (Figure 11-2).^{20,22}

In the **complex between an immunoglobulin and its antigen**, the single site on the Fab fragment for binding the antigen is formed from six loops of random

meander, three from each polypeptide, heavy and light. These loops are the **complementarity-determining regions** of the structure. Each of these six loops is one of the connections between two of the strands of the antiparallel β structure that form the superstructure of the core of the respective domains (Figure 11-2).

The amino acid sequences in the loops of these six complementarity-determining regions show remarkable variation among the different immunoglobulins, and they are referred to as the **hypervariable regions** of the sequences.²³ It is this variety of amino acid sequence that gives the immunoglobulins as a class their ability to provide individual proteins each tailored to bind a particular antigen. The specific sequences in these loops define the specificity of the particular immunoglobulin. Both the specific amino acid sequence and the lengths of these loops differ among the various immunoglobulins. The variations in sequence, as well as the variations in length, cause both the structure of the polypeptide forming these loops and the distribution of functional groups over the surface formed by these loops to differ dramatically from one immunoglobulin to the next.²⁴ As an illustration of the opportunism of biological processes, it is also possible for a sequence that dictates the glycosylation of an asparagine to be found in a complementarity-determining region and for the oligosaccharide attached to that asparagine to contribute favorably to the binding of the antigen.²⁵ It is these variations in structure and chemical character that produce the array of potential specificities at the site formed by the six loops. In contrast to this wild variation, the structures of the cores of the V_H domain and the V_L domain remain constant because the amino acid sequences forming these cores are well conserved.²⁶

Each immunoglobulin that appears in response to an antigen possesses a binding site that is formed by the complementarity-determining regions and that binds an epitope on the antigen. An **epitope** is the region on the antigenic protein that interacts directly with the binding site on the immunoglobulin. An antigen can have one or many epitopes, and each epitope elicits many different colonies each producing a different immunoglobulin recognizing that epitope and binding to it with its own particular dissociation constant. Each epitope is one of the regions on the antigen that induced the reproduction of the members of the colony of lymphocytes and their production of that particular immunoglobulin. Usually, an epitope is one or two short sequences of amino acids plus two or three other side chains in the antigen that are all adjacent to each other on its surface and that associate specifically with the surface formed by the six complementarity-determining regions. The epitope and the binding site on the immunoglobulin combine noncovalently as if they were two faces forming a heterologous interface in a heterooligomeric protein.

In the crystallographic molecular model of the complex between lysozyme and the Fab fragment

Because of their almost limitless variety and because they are adventitious, **interfaces** between immunoglobulins and their antigens are paradigms for all heterologous interfaces, and the details of the interactions within these interfaces²⁹ are typical of those found

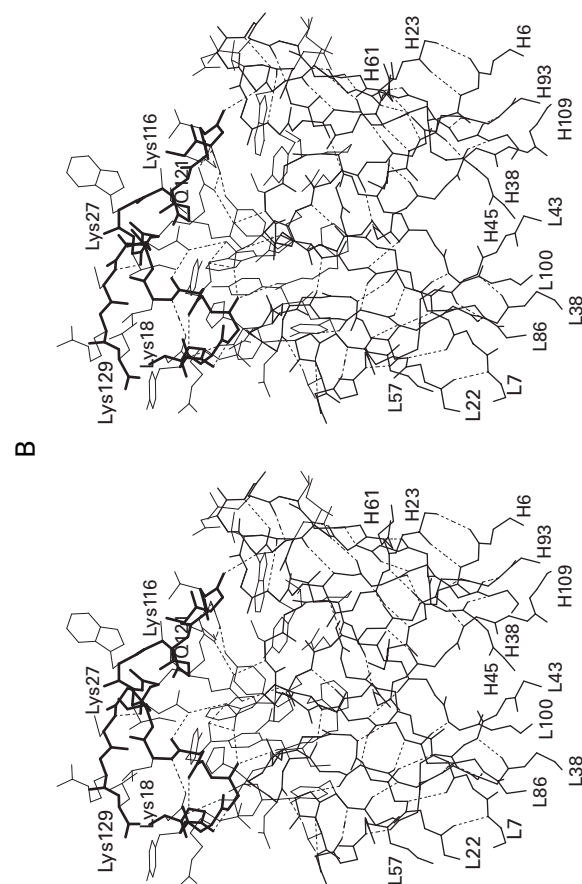


Figure 11-2: Crystallographic molecular model of lysozyme from *G. gallus* bound to the Fab fragment of a murine monoclonal immunoglobulin G.^{20,22} A complex was formed between lysozyme and the Fab fragment of murine monoclonal immunoglobulin D1.3, and the complex was crystallized from 20% poly(ethylene glycol). A crystallographic molecular model was prepared from the map of electron density. (A) Skeletal drawing of the polypeptide backbones of the complete crystallographic molecular model. The Fab fragment (lower two-thirds of the drawing) is drawn so that the heavy α subunit (thick bonds) and light β subunit (thin bonds) include both domains from each subunit (V_H and C_H1 and V_L and C_L , respectively). The lysozyme (bonds of intermediate thickness) is at the top of the structure. The two epitopic loops of random meander in lysozyme that are in contact with the complementarity-determining loops of the immunoglobulin are Aspartate 18 to Asparagine 27 and Lysine 116 to Leucine 129. (B) Detailed view of the interface between these two loops from the lysozyme (thick bonds at the top) and the six loops of the complementarity-determining regions, three from each subunit of the Fab fragment (thinner bonds). Each complementarity-determining region and the segments from each of its ends that anchor it in the two β sheets of the respective domain of the immunoglobulin are included in the drawing. The two ends of each of the eight segments of polypeptide represented in the figure, two from the lysozyme (Lys), three from the heavy α subunit (H), and three from the light β subunit (L), are labeled by their position in the respective amino acid sequence. Side chains (thinnest bonds) within or adjacent to the interface are included in the drawing. Hydrogen bonds are dashed lines. Glutamine 121 at the center of the interface between the lysozyme and the Fab fragment is identified. These drawings were produced with MolScript.¹⁰³

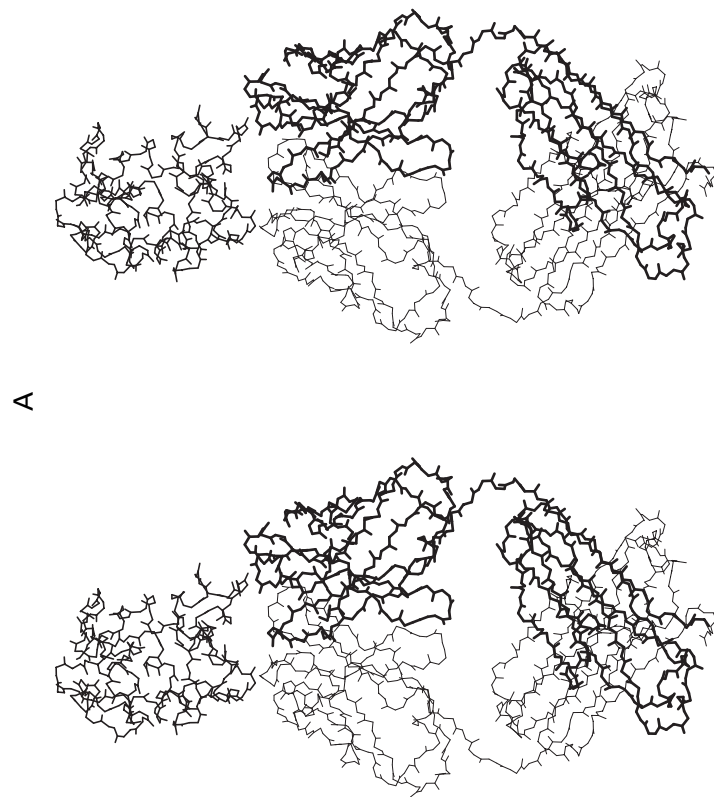


Table 11-1 Amino Acids within the Interface between a Murine Monoclonal Immunoglobulin G and Its Antigen, Micrococcal Nuclease^a

micrococcal nuclease		immunoglobulin		contacts		
secondary structure	amino acid	CDR ^b	amino acids ^c	VDW ^d	HB ^e	IHB ^f
amino terminus	Lys 9	H2	Ser 54, Thr 56	3		
α helix	Glu 57	L1	Ser 28	3		
	Ala 60	L1	Ser 28, Thr 27	7		
	Phe 61	L1	Phe 30, Ser 29	5		
	Lys 64	L1, L3	Thr 27, Trp 92	17		
	Asn 68	L3	Glu 93, ^g Trp 92	8	1	
bend	Lys 70	L3	Glu 93, ^g Ile 94	4		1
β strand	Asp 95	H2, L3	Tyr H50, Ile L94	5	1	
β turn	Gly 96	H2	Thr 52, Ser 54	3		
	Lys 97	H2, L3	Tyr H50, ^g Tyr L96	11	1	
α helix	Met 98	H2	Tyr 53	2		
	Arg 105	H3	Asn 96	7	2	
	Gln 106	L2	Tyr 50	1		
α helix	His 121	H2	Tyr 53	1		
	His 124	H1, H2	Ser 31, Tyr 53	20		
	Lys 127	H1	Tyr 27, ^g Ser 31 ^g	5	2	

^aThe crystallographic molecular model was that for the complex between the Fab fragment of murine monoclonal immunoglobulin N10 and micrococcal nuclease from *S. aureus*.²⁷ ^bComplementarity-determining region. Those from the heavy α polypeptide of the immunoglobulin are designated H and those from the light β polypeptide are designated L. Numbering is from the amino terminus. ^cAmino acids from the indicated loops of the complementarity-determining regions that contact the particular amino acid on the surface of micrococcal nuclease. ^dvan der Waals contacts. ^eHydrogen bond between two neutral atoms or between one charged atom and one neutral atom. ^fHydrogen bond between oppositely charged atoms. ^gDonor or acceptor where charge is ambiguous.

in other examples of heterologous associations.³⁰ Usually 5–10 nm² of **accessible surface area** from both antigen and immunoglobulin is buried and 5–20 hydrogen bonds form across the interface (Table 11-1), with donors and acceptors from both side chains and backbone. Normally, there are few ionized **hydrogen bonds** (Table 11-1),²⁹ but in the interface between cytochrome *c* and murine monoclonal immunoglobulin E8 there are five.³¹

Waters are also incorporated as structural elements into these interfaces. For example, in the interface within the crystallographic molecular model of equine cytochrome *c* and the Fab fragment of murine monoclonal immunoglobulin E8,³¹ there are 38 positions occupied by molecules of water,³² 16 of which are present at the same locations in crystallographic molecular models of either the uncomplexed antigen or uncomplexed Fab fragment or in both, and these locations are simply incorporated into the interface within their respective faces. Eight of these waters bridge the two proteins. In the interface between lysozyme from *G. gallus* and the murine Fab fragment HyHEL-63,³³ there are also 38 positions occupied by molecules of water,³² 14 of which are present at the same locations in uncomplexed antigen and uncomplexed Fab fragment, and eight of these

bridge the two proteins; and in the interface between human tissue factor and the murine Fab fragment D3h44 there are 46 positions occupied by molecules of water,³² 23 of which are incorporated as structural elements of the uncomplexed antigen and uncomplexed Fab fragment, and 19 of these bridge the two structures.

The central and most critical amino acid in the epitope on lysozyme recognized by murine monoclonal immunoglobulin D1.3 (Figure 11-2B) is Glutamine 121, the side chain of which occupies a distinct hole among the six loops of the complementarity-determining regions on the surface of the Fab fragment (Figure 11-2B). Each of the three amino acids lining the hole for Glutamine 121 is from a different complementarity-determining loop, one from the heavy α subunit and two from the light β subunit, and this places the hole in the very center of the binding site on the Fab fragment. If Glutamine 121 is replaced by either a histidine or an asparagine by site-directed mutation, the antigen is no longer bound by the immunoglobulin. In the normal structure of lysozyme, Glutamine 121 is fully exposed to the solvent.

It is often the case that an epitope seems to be **focused on a particular amino acid** on the surface of a

protein. For example, about 30–40% of the polyclonal immunoglobulins raised to human cytochrome *c* fail to bind to the cytochrome *c* from *Macaca mulatta*, which differs from the human protein only by the replacement of Isoleucine 58 by a threonine.³⁴ These immunoglobulins that fail to recognize cytochrome *c* from *M. mulatta* do, however, recognize cytochrome *c* from *Macropus canguru* that differs from the human at several other locations but does contain Isoleucine 58. No immunoglobulins raised to the cytochrome *c* from *M. mulatta* failed to recognize the cytochrome *c* from the human, and this result suggests that when a cytochrome *c* contains a threonine at position 58, as does the protein from *M. mulatta*, this region on the external surface³⁵ is not antigenic.³⁴ The impression left by these observations is that Isoleucine 58 is the key amino acid in this epitope, as is Glutamine 121 in the epitope of lysozyme.

In the crystallographic molecular model of the complex between lysozyme and murine monoclonal immunoglobulin D1.3, the structure of the lysozyme is identical, within the error of the models, with its structure in the absence of the immunoglobulin,²⁰ and the structures of both the V_L and the V_H domains of the Fab fragment are also identical, within the errors of the models with their structures in the uncomplexed Fab fragment,³⁶ even though the two domains have shifted slightly relative to each other by 0.1 nm. In this instance, formation of the complex between antigen and immunoglobulin is simply the docking of two **complementary faces**. In the crystallographic molecular model of the complex between human tissue factor and the Fab fragment of murine monoclonal immunoglobulin D3h44, “conformational changes upon formation of the complex are very small and almost exclusively limited to the reorientation of side-chains”.³²

Usually, however, the **conformations of both antigen and immunoglobulin change** noticeably as the complex is formed.³¹ The relative orientations of the V_H domain and V_L domain can shift by 5–10°. ²¹ One or more of the loops of the complementarity-determining regions can be reconfigured³⁷ or can pivot so that their tips move as much as 1.0 nm.³⁸ Side chains on both antigen and immunoglobulin often reorient by rotating around their carbon–carbon bonds.^{27,31} Flexible strands of polypeptide on the surface of the antigen readily rearrange upon formation of the complex.²¹ Most of these changes, however, are small ones, and the surface formed by the six complementarity-determining regions in the uncomplexed immunoglobulin has a shape that is already roughly complementary to the surface of the uncomplexed epitope,³¹ so that only a few readjustments are required upon formation of the complex.

The binding site on an immunoglobulin is usually a flat surface or a depression on the surface of a globular protein formed from the V_H and V_L domains (Figure 11–2). **Viruses** appear to take advantage of this feature of the site. The surface of picornaviruses such as rhinoviruses and

polioviruses are highly irregular. They are furnished with bosses at the 5-fold rotational axes of symmetry of the icosahedral shell. These bosses are separated from each other by deep depressions on the surface of the virus. The epitopes on polioviruses are located mainly on the bosses themselves and a few small protruding segments of polypeptide.³⁹ It is believed that the crucial regions of the surface of rhinoviruses that allow them to produce an upper respiratory infection are located in the depressions between the bosses.⁴⁰ These locations would be inaccessible to immunoglobulins. Each time the epitopes on the bosses or the smaller protrusions of rhinoviruses sustain a sufficient number of mutations to escape recognition, an antigenically novel but still infectious rhinovirus arises. The actual machinery of infection, lying as it does within the depressions, would be protected from being recognized by any immunoglobulin.

This strategy depends on the fact that most antigen binding sites are themselves flat or concave. On the murine monoclonal immunoglobulin HyHEL-5, however, Tyrosine 33 and Tyrosine 53 from complementarity-determining regions 1 and 2, respectively, of the V_H domain form a protrusion that juts out of the almost flat surface that constitutes the binding site for lysozyme, its antigen. In the complex between this immunoglobulin and its antigen, this protrusion fits into a deep groove on the surface of the protein that forms the active site of the enzyme.⁴¹ Consequently, immunoglobulins do not do so often but they can recognize their antigens by protruding into their structure instead of surrounding a protrusion on their surface.

The interface between antigen and immunoglobulin seen in the crystallographic molecular model²¹ of the neuraminidase from influenza virus and the Fab fragment from murine monoclonal immunoglobulin NC41 is formed from at least four juxtaposed strands of polypeptide from distant segments of the amino acid sequence of the neuraminidase. When amino acids in any one of three of these four strands are mutated, the neuraminidase can no longer bind to the immunoglobulin. When amino acids in the fourth strand are mutated, the affinity of the binding is noticeably diminished. All of these results indicate that the epitope on this protein for this immunoglobulin is a large region on the surface comprising all four of these strands. It seems an inescapable conclusion that this epitope would cease to exist if the protein were to unfold in this region. Such an epitope is a **conformationally specific epitope**.

Many immunoglobulins, however, will recognize their antigens even when the antigenic protein is no longer in its native structure. These are **sequence-specific immunoglobulins**. The paradigm of this class would be an immunoglobulin that, when covalently coupled to a stationary phase, can be used to isolate by affinity adsorption a peptide comprising its epitope from a digest of its antigen;⁴² such an immunoglobulin can recognize its epitope even when it is a formless peptide.

The ability of an immunoglobulin to recognize short peptides from an endopeptolytic digest of its antigen⁴³ or short synthetic peptides with sequences of amino acids from its antigen^{44,45} is used routinely, in the absence of a crystallographic molecular model of the complex between intact antigen and immunoglobulin, to **identify the epitope** on the antigen recognized by the immunoglobulin. In fact, when crystals of a complex between Fab fragment and intact antigen cannot be made, a crystallographic molecular model of a peptide from the antigen bound within the binding site on the immunoglobulin is assumed to depict accurately at least a portion if not all of the structure of the interface in the complex with the intact antigen.⁴⁶

Usually the class of sequence-specific immunoglobulins is distinguished from the class of conformationally specific immunoglobulins. Although the monoclonal immunoglobulin NC41 specific for viral neuraminidase that was used in the crystallographic studies seems to be an obvious member of the latter class, the evidence for such conformationally specific immunoglobulins is often anecdotal. A protein is irreversibly unfolded and loses its antigenic properties. To unfold a polypeptide irreversibly, however, it is usually either covalently modified⁴⁷ or noncovalently and intermolecularly polymerized by aggregation. Either the epitope could be covalently modified or it could be sterically sequestered within an aggregate of the protein during such uncontrolled reactions. If either of these events occurs, it appears as if the immunoglobulin were conformationally specific and recognized only the native structure of the protein when actually it was sequence-specific and recognized a single linear sequence of amino acids that was simply covalently modified or inaccessible. The technical difficulty is to unfold the polypeptide of the antigen to a monodisperse random polymer without doing the same thing to the immunoglobulin, which is also a folded polypeptide. Digestion of the antigen accomplishes this goal by producing structureless peptides, but if the immunoglobulin fails to recognize any of these peptides it could simply be the case that cleavage has occurred within the epitope.

It is probably the case that the distinction between sequence-specific and conformationally specific immunoglobulins is one of degree. For example, a polyclonal set of immunoglobulins was raised against native micrococcal nuclease from *S. aureus* ($n_{aa} = 149$) and purified on the basis of its ability to recognize a fragment of the intact polypeptide comprising amino acids 99–149. These immunoglobulins could bind the intact, folded polypeptide of micrococcal nuclease 2×10^4 times more tightly, as judged from the dissociation constants, than they could the fragment.⁴⁸ The fragment is a monodisperse random polymer, and it may be that the binding site on the immunoglobulin recognizes only the small portion of the random polymer that by chance has assumed the proper conformation. This would explain

the much greater dissociation constant of the fragment relative to the native protein. It is also possible, however, that the immunoglobulins still recognize the epitope or portions of the epitope after it is unfolded, but with much smaller free energy of dissociation. A difference in dissociation constant of even the magnitude observed for the complexes between the immunoglobulins and the nuclease, if the concentrations of immunoglobulin and antigen and the individual dissociation constants are in the appropriate ranges, would be observed as a complete elimination of the ability of antigen to bind to immunoglobulin when it is in fact only a finite attenuation of the ability of antigen to bind to immunoglobulin.

Because the immune system was developed to recognize and destroy foreign organisms such as viruses and bacteria and because other systems are used by animals to eliminate small toxic molecules, antigens are always large macromolecules, usually proteins, and never small molecules. This fastidiousness of the immune system can be circumvented by covalently attaching a small molecule to a carrier protein as a **hapten**. The attachment is accomplished with chemical couplings analogous to those used for covalent modification or cross-linking of proteins. In fact, many substances that are able to covalently modify proteins produce undesirable immune reactions by attaching to proteins in an animal or the incautious investigator and turning those immunologically benign proteins into malignant antigens.

A hapten, when covalently attached to the carrier protein, protrudes from its surface even more dramatically than Glutamine 121 does from the surface of lysozyme. Because of its peculiar chemical structure, the hapten is usually the focus of the immune system; and because it protrudes from the surface, the entire hapten usually ends up occupying a deep hole or deep crevice among the loops of two or three complementarity-determining regions.^{17,49,50} These facts explain why the unattached hapten can usually be bound efficiently by the immunoglobulin. For example, polyclonal immunoglobulins raised against a protein the lysines of which had been modified by 2,4,6-trinitrobenzenesulfonate (Reaction 10–28) bind N^{ϵ} -(2,4,6-trinitrophenyl)lysine with dissociation constants of 10^{-7} to 10^{-9} M.⁵¹ It is this ability of an immunoglobulin raised against a hapten to **bind tightly the small molecule** from which the hapten was derived that permits immunoglobulins to be used in highly specific assays for small molecules.⁵²

Although immunoglobulins for use in protein chemistry are usually raised by injecting an intact protein into an animal, it is also possible to raise immunoglobulins directed against **synthetic peptides** with the same amino acid sequence as a segment from a particular protein.⁵³ The synthetic peptides are attached as haptens to another protein. For example, the amino-terminal amino acid sequence of the large tumor antigen from simian virus is AcMDKVLNR-, where Ac is a post-

translationally added acetyl group. Immunoglobulins raised against a synthetic peptide with this sequence that had been covalently attached to bovine serum albumin as a hapten were able to recognize and bind exclusively to the large tumor antigen protein in crude homogenates from animal cells infected with the simian virus.⁵³ Immunoglobulins directed against a particular peptide can be purified by using a stationary phase to which the peptide is attached as an affinity adsorbent.^{54,55}

The difficulty inherent in the use of immunoglobulins raised against a particular amino acid sequence in a protein to study that protein in its native conformation is that the investigator, a fallible judge, rather than the immune system, which is less fallible, has chosen the epitope. A native protein does not expose many sequences sufficiently for immunoglobulins to recognize them. If the investigator has chosen the epitope, there is only a small chance that it will be **accessible** on the surface of the antigen and recognized by an immunoglobulin in the native protein, unless the choice is the safe one of the amino terminus or the carboxy terminus, which are usually well exposed in a native protein. The segment of sequence against which the immunoglobulin was raised, however, will usually be accessible in the unfolded polypeptide.

Any solution of a protein, even pristine cytoplasm, freshly drawn plasma, or a solution of redissolved crystals, contains some of the irreversibly **unfolded polypeptide** of any particular protein. Immunoglobulins raised against synthetic peptides necessarily bind preferentially, and often exclusively, to any unfolded protein exposing the sequence of amino acids against which they were raised because the original antigen itself was a structureless peptide of that sequence. Yet most experiments are designed with the requirement that the immunoglobulins recognize and bind to the native protein for the conclusion to be valid.

Crystallographic molecular models of complexes between immunoglobulins raised against synthetic peptides and the peptides themselves heighten these concerns.⁵⁶ In the binding site on the immunoglobulin, the peptide usually binds rigidly in a conformation that differs dramatically from the conformation that the same sequence of amino acids assumes in the native protein. Furthermore, it is, as might be expected, usually buried in a crevice among the loops of the complementarity-determining regions. Consequently, it is difficult to understand how such an immunoglobulin can ever recognize that sequence in the native protein.

The solution to this problem is analytical. When either monoclonal immunoglobulins or polyclonal immunoglobulins raised against synthetic peptides are used, one can assume that each folded polypeptide possesses only one epitope, either exposed or buried. It is also possible to purify by affinity adsorption a subset of polyclonal immunoglobulins recognizing only one epitope from a set of polyclonal immunoglobulins raised

against an intact protein.^{34,55} If every molecule of protein in a solution can bind one immunoglobulin at a particular epitope, then the immunoglobulin is recognizing the native protein. If only a small percentage of the molecules of protein in a solution can bind an immunoglobulin at that epitope, it is only the unfolded polypeptides that are presenting that epitope. Consequently, if every protomer of an antigen binds one molecule of an immunoglobulin, when only immunoglobulins directed against one epitope are present, then the immunoglobulins must be recognizing the epitope when it is in the native protein. In any experiment relying on the assumption that the immunoglobulins recognize the native protein, it must be demonstrated both that the immunoglobulins bind to only one unique epitope and that every protomer of the antigen is capable of binding one molecule of immunoglobulin. In the absence of such a demonstration, the conclusions reached can be disregarded.

In contrast to these difficulties encountered when they are used to examine a native protein, immunoglobulins raised against a synthetic peptide can be used to **purify a peptide** containing a particular amino acid in the sequence of a protein from an endopeptolytic digest of that protein.⁵⁵ The amino acid sequence surrounding Lysine 380 of the α subunit of acetylcholine receptor is -SAIEGVKYIAEHM-. The synthetic peptide KYIAE was coupled covalently as a hapten to bovine serum albumin, and polyclonal immunoglobulins were raised against this antigen in rabbits. Because the peptide had been coupled to the serum albumin through the amino groups of its lysine and of its amino terminus, the carboxy-terminal sequences -YIAE protruded as haptens from the surface of the serum albumin. The antiserum was passed over a stationary phase to which the peptide KYIAE had been covalently attached, and immunoglobulins specific for the carboxy-terminal sequence -YIAE were adsorbed by the affinity adsorbent. After all of the other proteins in the serum had been washed away, the adsorbed immunoglobulins were eluted. These purified immunoglobulins in turn were covalently attached to a stationary phase to produce an immunoabsorbent specific for the carboxy-terminal sequence -YIAE. When acetylcholine receptor was digested with glutamyl endopeptidase (Figure 3-2) and the digest was passed over the immunoabsorbent, the peptide GVKYIAE was adsorbed and eluted in high yield and high purity.⁵⁷ The covalent modification of Lysine 380 in intact, native acetylcholine receptor could be readily monitored by using this immunoabsorbent to purify rapidly the peptide containing it.

Such an immunoabsorbent can purify a peptide from the digest of a large protein, which contains so many peptides that a direct purification by chromatography would be difficult if not impossible. If a chemical modification of an amino acid in the protein within the sequence included in the peptide occurs in low yield so

that the modified peptide is only a minor component of the digest, the immunoadsorbent will still purify it.⁵⁷ In fact, the targeted amino acid can be destroyed by the modification and the protein cleaved at the point of destruction, and the immunoadsorbent will still purify the peptide containing the remaining fragment of that amino acid.⁵⁸ By use of such an immunoadsorbent, for example, the modification of a particular amino acid in a protein can be followed kinetically⁵⁹ or its accessibility to a particular electrophile under different circumstances can be monitored.⁶⁰

Immunoglobulins directed against an antigenic protein can effect immunoprecipitation. An **immuno-precipitate** is a visible, white precipitate that forms when an antigen and its immunoglobulin are present in solution at the proper concentrations. Each immunoglobulin has at least two binding sites for antigen (Figure 11-1). Each antigen, if it is a protein, is usually polyvalent. A **polyvalent antigen** is one that has more than one epitope. If the polyclonal mixture of immunoglobulins in the serum contains several monoclonal immunoglobulins directed against different epitopes on the antigen and if this mixture of immunoglobulins is mixed in the proper ratio with that antigen, an immunoprecipitate forms containing antigen and immunoglobulin cross-linked among themselves. The ratio of concentrations at which maximum precipitation occurs is known as the **equivalence point**. If there is only one epitope on the antigen that has elicited the immune response, no precipitate will form. An Fab fragment, because it is univalent, cannot produce a precipitate. If a large excess of antigen is present, each immunoglobulin has its binding sites filled with antigens that are not bound to other immunoglobulins and no precipitate forms. If a large excess of immunoglobulin is present, each antigen is surrounded by immunoglobulins, each with its other binding site vacant, and no precipitate forms. Such soluble complexes of excess antigen and excess immunoglobulin are present under all circumstances, and it is never possible to precipitate directly all of the antigen or all of the immunoglobulins even at equivalence. Finally, the molar ratio between antigenic protein and immunoglobulin in a precipitate gathered at equivalence is a complicated function of the number of epitopes, their relative affinities, and the distribution of the different monoclonal immunoglobulins in the polyclonal mixture.

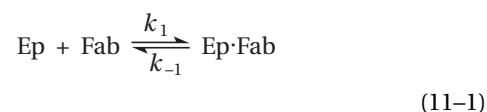
After a protein of interest has been injected as a potential antigen into an animal, the appearance of the desired immunoglobulins in the serum is often detected by the ability of those immunoglobulins to form an immunoprecipitate. The simplest way to produce the proper ratio of concentrations in order to observe a precipitate is to layer a solution of the antigen onto a sample of the serum in a narrow tube. As antigen diffuses into the serum and immunoglobulin diffuses into the solution of antigen, there will be a point along the two gradients at which the concentrations are those necessary for

precipitation to occur and a visible **disc of precipitate** will form at this location.

Immunodiffusion is a more sophisticated procedure for permitting antigen and immunoglobulin to diffuse into each other.⁶¹ Antigen and serum are placed in separate wells cut in a block of agar, and as they diffuse outward into the agar and towards each other, there will be a line between the two wells along which the ratio of concentrations is appropriate for precipitation. Along this line a white immunoprecipitate will form. When the original antigen and the same protein from a different species or a mutant variety of the antigen are placed in two adjacent wells both the same distance from the well containing immunoglobulin, the pattern of the lines of immunoprecipitate demonstrates whether or not the related protein shares all of the epitopes present on the original antigen.⁶²

Complement fixation is a procedure for detecting the relative concentrations of immunoprecipitates in a series of samples.⁶³ It is sensitive to concentrations of immunoprecipitate much smaller than can be observed visually. The procedure can be performed on a series of mixtures of antigen and immunoglobulin at different ratios of concentration to obtain a direct measurement of the equivalence point, which is the ratio at which complement fixation reaches its maximum level.

An immunoprecipitate is held together by a complex collection of interfaces formed between the binding sites on the tips of the Fab arms of the various immunoglobulins present in antiserum and their respective epitopes on the molecules of antigen. Each of the individual reactions between an epitope and the binding site on the Fab arm of an intact immunoglobulin is a simple dissociation:



where Fab is the binding site, Ep is the epitope, and Ep·Fab is the immune complex (Figure 11-2). The immune complexes between epitopes on antigens and immunoglobulins are quite strong, a fact that permits an immunoprecipitate to form even when antigen and immunoglobulins are present at the low concentrations used for procedures such as complement fixation. More useful than the strength of the association, however, is the fact that because k_{-1} is almost always a small rate constant, dissociation of the complex is slow. Most of the procedures that use immunoglobulins depend on this **slow dissociation** of the complexes between them and their antigens. The slow dissociation permits the complex between antigen and its immunoglobulin to be separated either from an excess of the specific immunoglobulin and other immunoglobulins and proteins in an antiserum or from all of the other proteins with which the antigen is mixed. For example, an

immunoprecipitate can be extensively washed without dissociating. Immunoblotting, immunostaining, and immunoadsorption also rely on this advantage of the slow dissociation.

One of the most important uses of immunochemistry is to identify the particular protein to which the antibody is directed. For example, a specific immunoglobulin can be used to stain only its antigen among all of the proteins in a heterogeneous mixture separated by electrophoresis.⁶⁴⁻⁶⁷ First the proteins that have been separated by electrophoresis on a slab of poly-

acrylamide are transferred laterally by electrophoresis onto a membrane of nitrocellulose or poly(vinylidene difluoride) placed against the slab of polyacrylamide. This **electrotransfer** produces a **blot** on which the bands of protein in the polyacrylamide have become bands of protein arrayed in the same pattern but now plastered onto the membrane of polymer (Figure 11-3A).⁶⁸ Then the blot is soaked in a solution of the specific immunoglobulin, and excess immunoglobulins are rinsed away. The immunoglobulins that were bound by the antigen are **immunostained** with a second

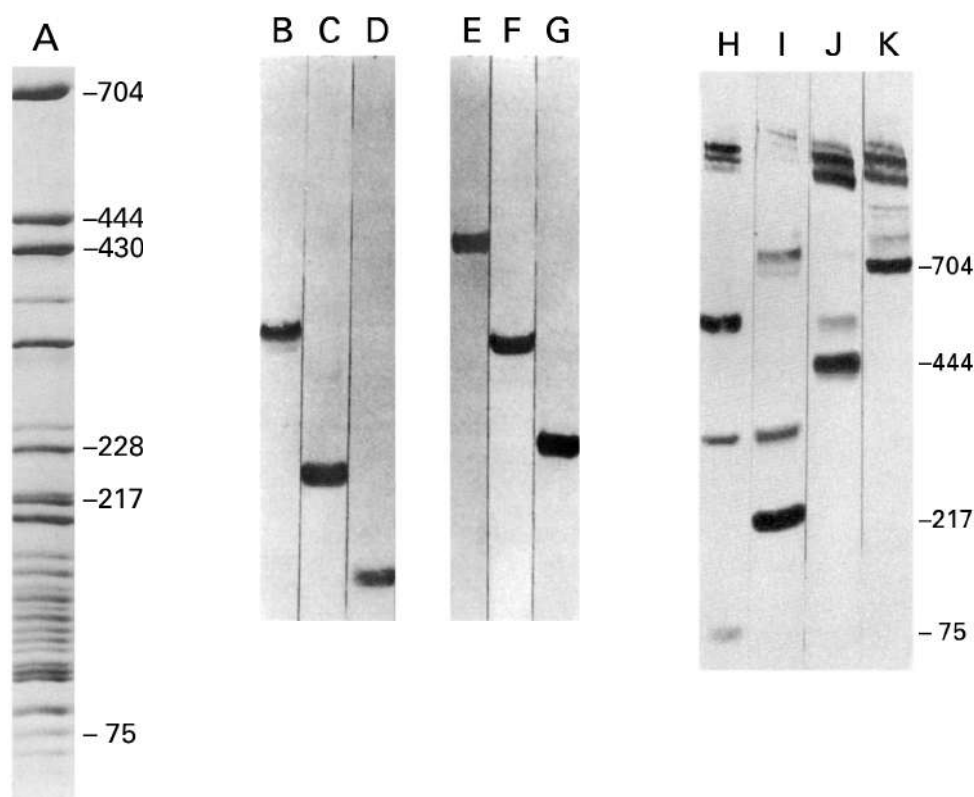


Figure 11-3: Immunostaining of immunoblots of NADH dehydrogenase (ubiquinone) from bovine heart mitochondria. (A) NADH Dehydrogenase (ubiquinone) was dissolved in a solution of dodecyl sulfate and submitted to electrophoresis on a slab of polyacrylamide. Following the electrophoresis, the separated polypeptides on the gel were electrotransferred laterally onto a membrane of poly(vinylidene difluoride) placed against the slab by applying an electric field perpendicular to the plane of the slab. The protein in each band remained at the same position on the field and adhered tightly to the poly(vinylidene difluoride). The membrane was then stained with Coomassie brilliant blue.⁶⁸ The 25 bands that appeared represent only a fraction of the more than 40 polypeptides in NADH dehydrogenase (ubiquinone). Reprinted with permission from ref 68. Copyright 1992 Elsevier B.V. (B-G) The complexes between dodecyl sulfate and the polypeptides of NADH dehydrogenase (ubiquinone) of 704, 444, 430, 228, 217, and 75 aa were each purified from subcomplexes of the enzyme.^{73,74} Following purification, each of them was injected into rabbits, and polyclonal immunoglobulins specific for each of the polypeptides were purified from the respective antiserum^{71,72} by binding them to the respective polypeptide immobilized on a membrane of nitrocellulose, washing away the other proteins, and eluting the immunoglobulins.⁷⁵ Membranes of poly(vinylidene difluoride) to which electrophoretically separated polypeptides of intact NADH dehydrogenase (ubiquinone) had been electrotransferred as in part A were then immunostained⁷⁶ with purified rabbit polyclonal immunoglobulins against the polypeptides of (B) 444 aa, (C) 217 aa, (D) 75 aa, (E) 704 aa, (F) 430 aa, and (G) 228 aa, respectively, followed by goat polyclonal immunoglobulins that were raised against rabbit immunoglobulin G and to which peroxidase had been covalently coupled. The electrophoretic separation in lane A and those in lanes B-G were performed in different laboratories on different polyacrylamide gels that produced different mobilities. (H-K) Intact NADH dehydrogenase (ubiquinone) with its more than 40 subunits was cross-linked with ethylene glycol bis(succinimidyl succinate). The product of the reaction was dissolved in a solution of dodecyl sulfate. Four identical samples of this solution were submitted to electrophoresis in separate lanes, the separated polypeptides electrotransferred to a sheet of poly(vinylidene difluoride), and the respective lanes were immunostained as in part B-H with immunoglobulins specific for the polypeptides of (H) 75 aa, (I) 217 aa, (J) 444 aa, and (K) 704 aa, respectively.⁷⁶ The respective un-cross-linked polypeptide is the lowest band in each lane, and covalent complexes in which the polypeptide participates are represented by the higher bands. Reprinted with permission from ref 76. Copyright 1993 American Chemical Society.

immunoglobulin directed against the first immunoglobulin, for example, ovine anti-murine immunoglobulin, to which either peroxidase⁶⁹ or alkaline phosphatase⁶⁶ has been attached. The peroxidase or alkaline phosphatase produces a colored precipitate at the location of the antigen (Figure 11-3, panels B-H). In this way, one protein can be picked out of a complex mixture because it is the only protein that is stained. For example, a murine monoclonal immunoglobulin that had been selected for its ability to inhibit the mitochondrial dicarboxylate transporter from *Pisum sativum* immunostained only one polypeptide on an immunoblot of a polyacrylamide gel on which the hundreds of polypeptides from intact mitochondria dissolved in a solution of dodecyl sulfate had been separated electrophoretically.⁷⁰ That polypeptide was assumed to be the transporter.

Immunostaining of immunoblots is also used to identify which polypeptides are components of a particular covalent complex produced by **cross-linking**. NADH Dehydrogenase (ubiquinone) is a large protein composed of more than 40 different polypeptides (Figure 11-3A).⁶⁸ Immunoglobulins that specifically recognize the polypeptides of 75, 217, 228, 430, 444, and 704 aa were produced in rabbits⁷¹⁻⁷⁴ and purified by binding them to their antigens and then eluting them.⁷⁵ On immunoblots of polyacrylamide gels on which the complete protein with its more than 40 polypeptides had been separated (Figure 11-3A), each of the immunoglobulins directed immunostaining only to the polypeptide against which it was raised (Figure 11-3, lanes B-G).⁷⁶

The complete protein was then cross-linked with ethylene glycol bis(succinimidyl succinate), and the polypeptides that were separated by electrophoresis were again immunoblotted and immunostained with the immunoglobulins specific for the polypeptides of 75, 217, 444, and 704 aa (Figure 11-3, lanes H-K). Covalent complexes between the polypeptides of 217 and 75 aa; 444 and 75 aa; 444 and 217 aa; 704 and 444 aa; 704, 444, and 75 aa; and 704, 444, and 217 aa could be positively identified on the basis of their mobility and, more importantly, the fact that they were immunostained by the appropriate immunoglobulins. The most interesting result was the fact that none of these four polypeptides had been cross-linked to any of the other polypeptides in the complex. Because there are many polypeptides that have mobilities similar to others in the protein, only the immunostaining could sort out the products of the cross-linking reaction.

One of the most effective ways to raise immunoglobulins that recognize a particular protein with high specificity is to use as a hapten a synthetic peptide with the sequence of the **amino terminus** or the **carboxy terminus** of that protein. Polyclonal immunoglobulins were raised against the synthetic peptide SEFIGA, the carboxy-terminal sequence of the human receptor for epidermal growth factor. The peptide had been attached as a hapten through its amino terminus to

bovine serum albumin so the immunoglobulins recognized the carboxy-terminal sequence -EFIGA. These polyclonal immunoglobulins were purified by passing the antiserum over an affinity adsorbent composed of a solid phase to which the peptide had been attached covalently. When a homogenate of cultured human cells was dissolved in a solution of dodecyl sulfate and submitted to electrophoresis and the separated polypeptides were immunoblotted, the immunoglobulin raised against the peptide SEFIGA directed immunostaining only to the polypeptide of the receptor for epidermal growth factor.⁷⁷ The solution submitted to electrophoresis contained all of the hundreds of polypeptides in the cells, and yet only the receptor for epidermal growth factor was recognized by the polyclonal immunoglobulins.

Immunoglobulins can also be used to isolate a particular protein by **immunoabsorption**. An **immunoabsorbent** is a stationary phase to which an immunoglobulin specific for a particular antigen has been covalently bound and with which that antigen can be purified by affinity adsorption.^{78,79} For example, although the protein had not been purified, the amino acid sequence of Shaker S4 K⁺ channel from *Drosophila melanogaster* had been determined genetically. The protein was expressed in Sf9 insect cells following their infection with baculovirus into the DNA of which complementary DNA encoding the protein had been inserted. The expressed protein could then be purified⁸⁰ on an immunoabsorbent to which had been covalently attached immunoglobulins raised against the synthetic peptide EEEDTLNLPKAPVSPQDKS, an amino acid sequence from a region of the protein (amino acids 333-351) thought to be an exposed loop connecting two α helices.

The gene encoding dystrophin, the protein that is missing in muscles of patients suffering from Duchenne/Becker muscular dystrophy, had been identified before dystrophin itself was known to exist.⁸¹ An immunoabsorbent, containing immunoglobulins raised against an expressed fusion protein containing a fragment of 556 aa from the amino acid sequence of dystrophin, was used to purify it.⁸²

The voltage-gated chloride channel from *Torpedo californica* had also been identified and sequenced genetically before it had been purified. It was then purified in one step of affinity adsorption by use of a stationary phase to which immunoglobulins recognizing the hydrophilic sequence EGQREGLEAVKVQTEDP from the protein had been coupled covalently.⁸³

Immunoabsorption can also be accomplished by adding an excess of specific immunoglobulins to a solution to saturate all of the antigen, passing the solution over agarose to which protein A from *S. aureus*, a protein with a high affinity for immunoglobulin G, has been covalently attached,⁸⁴ and eluting the adsorbed antigen from the agarose after the other proteins have been rinsed away.

A protein can be identified by **tagging it with an epitope**.⁸⁵ A short segment of DNA encoding the amino acid sequence of an epitope to which an immunoglobulin has already been raised is inserted in phase at one end of the reading frame for the protein of interest. The protein is then expressed from the complementary DNA into which the insertion has been made, and the expressed protein contains the amino acid sequence of the tag at one of its ends. The protein thus tagged can be identified by immunostaining and isolated by immunoadsorption with the immunoglobulins directed against the epitope. In this way, the protein encoded by an unidentified reading frame in a segment of genomic DNA or complementary DNA can be identified and purified. If a protein has been tagged by an epitope at one of its ends, an immunoblot of a digest of that protein separated by electrophoresis in a solution of dodecyl sulfate will provide a map of the positions in the amino acid sequence at which cleavage occurred during the digestion. The lengths of the successive **end-labeled fragments** will correspond to the positions of cleavage in the sequence of the protein.⁸⁶

Immunoglobulins are also used to **screen libraries** of cDNA.⁸⁷ If immunoglobulins specific for a protein of completely unknown sequence have been made, they can be used in such a screen to detect the complementary DNA for that protein. The complementary DNAs to be screened are inserted into plasmids that cause the proteins they encode to be expressed when they are transfected into bacteria. The transfected bacteria are spread onto a field and grown into individual colonies, each of which consequently contains protein expressed from the cDNA on its respective plasmid. The bacteria are lysed and their proteins are immobilized on a surface in such a way that the immobilized proteins remain in the same location on the field and produce a replica of the original pattern of the colonies. The replica is soaked with the immunoglobulins that recognized the protein of interest and then washed, and those colonies that contained antigen are identified by immunostaining or by binding radioactive protein A from *S. aureus* to the bound immunoglobulins. The bacteria in a colony identified in this way can then be replicated and the cDNA sequenced.

Immuno-electron microscopy is used to identify the region on the surface of a protein containing the epitope against which particular immunoglobulins are directed. Most proteins, including immunoglobulins, are large enough to be observed in the electron microscope when embedded in a glass of negative stain. When the embedded complex between a protein and an immunoglobulin or the Fab fragment of an immunoglobulin is observed on an electron micrograph, the immunoglobulin or Fab fragment appears as a protrusion on the surface of the protein that it recognizes. Often the Fab arms and the Fc trunk of an immunoglobulin can be distinguished; usually, however, an

immunoglobulin appears only as a vague elongated structure. If the protein has a characteristic shape, the location of the epitope recognized by the immunoglobulin on the surface of that shape can be identified. α_2 -Macroglobulin is a molecule that in an electron micrograph has the shape of a letter H; a murine monoclonal immunoglobulin specific for the domain of about 200 aa responsible for the binding of the α_2 -macroglobulin to its receptor binds to the ends of the arms on the H.⁸⁸ Fibrinogen is a molecule that, in an electron micrograph, has three globular domains arranged in a row; a murine monoclonal immunoglobulin specific for the carboxy-terminal 150 amino acids of its α polypeptide binds near the central domain of the structure.⁸⁹

The multicatalytic endopeptidase complex is a cylinder composed of 14 different subunits, each present in two copies. A murine monoclonal immunoglobulin specific for one of these subunits binds at both ends of the cylinder, consistent with the existence of a 2-fold rotational axis of symmetry at the center of the cylinder and locating the positions in the cylinder of the two copies of that subunit.⁹⁰ Murine monoclonal immunoglobulins against several of the subunits in the multicatalytic endopeptidase complex were always bound at two symmetrically displayed locations on the surface of the cylinder, and the respective angles between those positions when the cylinder was viewed along its axis could be used to position those subunits relative to the 2-fold rotational axis of symmetry that is normal to the cylindrical axis and at the middle of the cylinder.⁹¹

The most extensive application of immuno-electron microscopy has been an examination of the distribution of the constituent polypeptides over the surface of the two subunits of the ribosome from *Escherichia coli*. The application of these procedures to the 30S subunit serves as an example. Although it was unknown at the time these experiments were performed, the core of the **30S ribosomal subunit** is formed from ribosomal RNA, and luckily almost all of its constituent polypeptides are distributed over its external surface and are accessible to immunoglobulins.

The 21 unique polypeptides found in the 30S subunit of the ribosome can be separated and catalogued by two-dimensional gel electrophoresis (Figure 11-4).^{92,93} They have been separated and individually purified,⁹³ and their amino acid sequences have been determined.⁹⁴ Polyclonal sets of immunoglobulins have been raised against most of these polypeptides. When immunoglobulins specific for one of them were mixed with intact 30S ribosomal subunits and the immune complexes were then prepared for electron microscopy, individual immunoglobulins bound to individual 30S ribosomal subunits or cross-linking two 30S ribosomal subunits could be observed (Figure 11-5).⁹⁵⁻⁹⁸ The 30S ribosomal subunit has a characteristic, asymmetric shape and the epitopes recognized by these immunoglobulins could be assigned to certain regions on the surface of that shape.



Figure 11-4: Separation of the polypeptides composing the 30S subunit of ribosomes from *E. coli*.⁹³ Intact ribosomes were isolated from a homogenate of bacteria by centrifugation. They were dissociated into subunits by treatment with MgCl_2 , and the 30S subunit was separated from the 50S subunit by centrifugation through a gradient of sucrose. The protein was extracted from the 30S ribosomal subunit and dissolved in 6 M urea. The individual polypeptides were separated in the first dimension at pH 9.6 and in the second dimension at pH 4.6. Separation in each dimension was performed in 6 M urea. The proteins were stained with amido black. Although only 17 components are observed, three polypeptides coelectrophorese at one spot and two pairs of polypeptides coelectrophorese at two other spots. The total number of polypeptides is 21. Reprinted with permission from ref 93. Copyright 1973 *Journal of Biological Chemistry*.

Two different laboratories have determined the distributions of the various polypeptides over the surface of the 30S ribosomal subunit, based on the relative distributions of the antigenic sites.^{99,100} In each of these two sets of observations, the location at which each individual immunoglobulin was bound on the 30S ribosomal subunit was ascertained by deciding visually which projection and which orientation of a crude three-dimensional structure of the 30S ribosomal subunit each image in the micrograph represented. Both of these two crude structures, although significantly different from each other, incorporated a small globular domain, a large globular domain, and a significant protrusion on one side. These three features served to orient the particles, and the positions of the various polypeptides could be assigned relative to them. Luckily, these features did appear in the crystallographic molecular model of the 30S subunit (Figure 11-5).^{95,101}

Although there was initially significant disagreement about the distribution of these antigens over the surface of the 30S subunit,^{96,102} the differences were resolved, and in the final maps from the two laboratories,^{99,100} the agreement is quite close. Furthermore, these distributions determined by immunoelectron microscopy are in remarkable agreement with the distribution of the folded polypeptides over the crystallographic molecular model that was obtained 10 years later.^{95,101}

Suggested Reading

- Amit, A.G., Mariuzza, R.A., Phillips, S.E.V., & Poljak, R.J. (1986) Three-dimensional structure of an antigen-antibody complex at 2.8-Å resolution, *Science* 233, 747-753.
- Pons, F., Augier, N., Heilig, R., Leger, J., Mornet, D., & Leger, J.J. (1990) Isolated dystrophin molecules as seen by electron microscopy, *Proc. Natl. Acad. Sci. U.S.A.* 87, 7851-7855.
- Yamaguchi, M., & Hatefi, Y. (1993) Mitochondrial NADH:ubiquinone oxidoreductase (complex I): proximity of the subunits of the flavoprotein and the iron-sulfur protein subcomplexes, *Biochemistry* 32, 1935-1939.

Problem 11-1: A polyclonal set of immunoglobulins was produced against the synthetic peptide -ETYY, the carboxy-terminal sequence of Na^+/K^+ -exchanging ATPase, a protein embedded in the membranes of animal cells. Immunoglobulins recognizing this peptide were purified by affinity adsorption on a solid phase to which the peptide had been attached. The immunoglobulins were made radioactive by reductive methylation (Figure 10-3) with formaldehyde and sodium [^3H]borohydride to a final specific radioactivity of $10,760 \text{ cpm nmol}^{-1}$. Equal amounts of this immunoglobulin were mixed with increasing amounts of homogeneous Na^+/K^+ -exchanging ATPase in its membrane-bound form so that bound and unbound immunoglobulin could be separated by centrifugation after equilibrium had been reached. The amount of bound immunoglobulin increased linearly with the amount of membrane-bound protein added. It was found that each milligram of protein of purified Na^+/K^+ -exchanging ATPase could bind 670 cpm of radioactive immunoglobulin regardless of the final concentration of immunoglobulin. The asymmetric unit of Na^+/K^+ -exchanging ATPase is composed of one α polypeptide ($n_{\text{aa}} = 1020$) and one β polypeptide ($n_{\text{aa}} = 300$). What fraction of the molecules of the ATPase displays epitopes?

A monoclonal immunoglobulin was also produced against Na^+/K^+ -exchanging ATPase. This monoclonal immunoglobulin could bind the synthetic peptide HLLVMKGAPER, which has a sequence identical to a segment of the sequence from the α polypeptide of the enzyme. The relative concentration of binding sites of this immunoglobulin in a solution could be determined by an indirect immunoassay. When samples of a solution of this monoclonal immunoglobulin, at a final concentration of 11 nM in binding sites for antigen, were mixed and brought to equilibrium with increasing concentrations of Na^+/K^+ -exchanging ATPase, the concentration of unoccupied binding sites for antigen decreased. A final concentration of Na^+/K^+ -exchanging ATPase of $300 \mu\text{g mL}^{-1}$ was required to decrease the concentration of active immunoglobulin by greater than 90% (from 11 nM to less than 1 nM). What fraction of the molecules of Na^+/K^+ -exchanging ATPase displays epitopes accessible to the immunoglobulin?

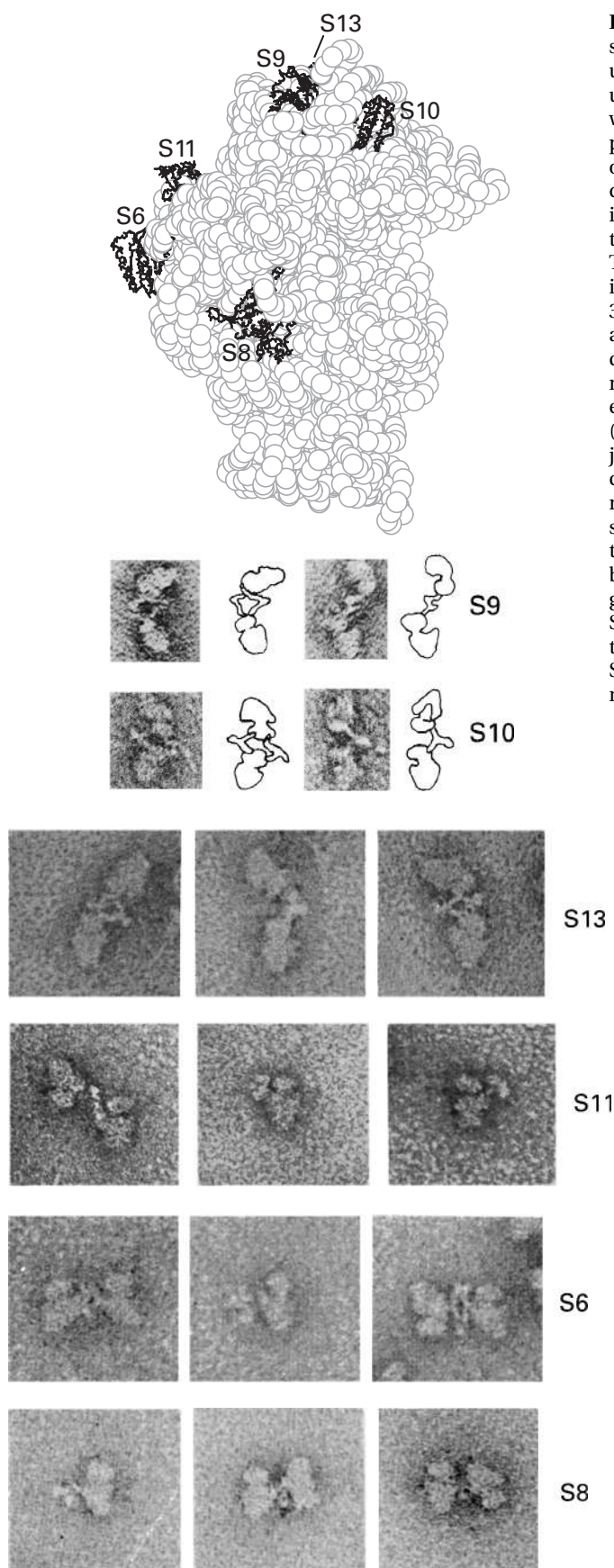


Figure 11-5: Immuno-electron microscopy of the 30S subunit of the ribosome. A drawing of the crystallographic molecular model of the 30S subunit from *Thermus thermophilus* is at the top.⁹⁵ The ribosomal RNA and unhighlighted polypeptides are drawn in a space-filling representation in which a sphere was placed at each α carbon of the proteins and at each phosphorus atom of the nucleic acid. The backbones of the polypeptides of those subunits that were used as antigens in the various experiments are drawn in skeletal representation and identified by the standard numbering. Subunit S13 is located on the surface of the 30S subunit just over the top of the view presented. This drawing was produced with MolScript.¹⁰³ The electron micrographs are of immune complexes between polyclonal immunoglobulins G and 30S ribosomal subunits from *E. coli*. Purified 30S subunits were mixed with the polyclonal immunoglobulins raised against the particular purified polypeptide: S9, S10, S13, S11, S6, or S8. The complexes that formed were adsorbed to a layer of carbon on a grid for microscopy, negatively stained with uranyl acetate, and observed in the electron microscope. The immunoglobulins G are Y-shaped proteins (Figure 11-1) that connect two globular 30S ribosomal subunits or bind to just one. The 30S subunits in the micrographs can be recognized by their characteristic shapes as illustrated by the crystallographic molecular model. At the top in the view of the crystallographic molecular model presented is a smaller globular domain, to the left a significant protrusion, at the bottom the larger globular domain, and to the right a deep cleft between the upper and lower domains. The top two panels of micrographs, for polypeptides S9 and S10, are results from the laboratory of Stöffler.⁹⁶ Reprinted with permission from ref 96. Copyright 1975 held by the authors. The lower rows of micrographs, for polypeptides S13,⁹⁷ S6,⁹⁸ S11,⁹⁷ and S8,⁹⁸ are results from the laboratory of Lake. Reprinted with permission from refs 97 and 98. Copyright 1975 and 1981 Elsevier B.V.

References

1. Chen, B.X., Wilson, S.R., Das, M., Coughlin, D.J., & Erlanger, B.F. (1998) *Proc. Natl. Acad. Sci. U.S.A.* 95, 10809–10813.
2. Villeneuve, S., Souchon, H., Riottot, M.M., Mazie, J.C., Lei, P., Glaudemans, C.P., Kovac, P., Fournier, J.M., & Alzari, P.M. (2000) *Proc. Natl. Acad. Sci. U.S.A.* 97, 8433–8438.
3. Margoliash, E., Nisonoff, A., & Reichlin, M. (1970) *J. Biol. Chem.* 245, 931–939.
4. Silverton, E.W., Navia, M.A., & Davies, D.R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* 74, 5140–5144.
5. Edelman, G.M., Cunningham, B.A., Gall, W.E., Gottlieb, P.D., Rutishauser, U., & Waxdal, M.J. (1969) *Proc. Natl. Acad. Sci. U.S.A.* 63, 78–85.
6. Harris, L.J., Larson, S.B., Hasel, K.W., & McPherson, A. (1997) *Biochemistry* 36, 1581–1597.
7. Mage, M.G. (1980) *Methods Enzymol.* 70, 142–150.
8. Porter, R.R. (1959) *Biochem. J.* 73, 119–126.
9. Nisonoff, A., Wissler, F.C., Lipman, L.N., & Woernley, D.L. (1960) *Arch. Biochem. Biophys.* 89, 230–244.
10. Masson, P.L., Cambiaso, C.L., Collet-Cassart, D., Magnusson, C.G., Richards, C.B., & Sindic, C.J. (1981) *Methods Enzymol.* 74, (Part C), 106–139.
11. Holliger, P., Prospero, T., & Winter, G. (1993) *Proc. Natl. Acad. Sci. U.S.A.* 90, 6444–6448.
12. Gregory, L., Davis, K.G., Sheth, B., Boyd, J., Jefferis, R., Nave, C., & Burton, D.R. (1987) *Mol. Immunol.* 24, 821–829.
13. Guddat, L.W., Herron, J.N., & Edmundson, A.B. (1993) *Proc. Natl. Acad. Sci. U.S.A.* 90, 4271–4275.
14. Putnam, F.W., Florent, G., Paul, C., Shinoda, T., & Shimizu, A. (1973) *Science* 182, 287–291.

15. Toraano, A., & Putnam, F.W. (1978) *Proc. Natl. Acad. Sci. U.S.A.* 75, 966–969.
16. Chapuis, R.M., & Koshland, M.E. (1975) *Biochemistry* 14, 1320–1326.
17. Segal, D.M., Padlan, E.A., Cohen, G.H., Rudikoff, S., Potter, M., & Davies, D.R. (1974) *Proc. Natl. Acad. Sci. U.S.A.* 71, 4298–4302.
18. Keohler, G., & Milstein, C. (1976) *Eur. J. Immunol.* 6, 511–519.
19. Harris, L.J., Skaletsky, E., & McPherson, A. (1998) *J. Mol. Biol.* 275, 861–872.
20. Amit, A.G., Mariuzza, R.A., Phillips, S.E., & Poljak, R.J. (1986) *Science* 233, 747–753.
21. Colman, P.M., Laver, W.G., Varghese, J.N., Baker, A.T., Tulloch, P.A., Air, G.M., & Webster, R.G. (1987) *Nature* 326, 358–363.
22. Fischmann, T.O., Bentley, G.A., Bhat, T.N., Boulot, G., Mariuzza, R.A., Phillips, S.E., Tello, D., & Poljak, R.J. (1991) *J. Biol. Chem.* 266, 12915–12920.
23. Wu, T.T., & Kabat, E.A. (1970) *J. Exp. Med.* 132, 211–250.
24. Chothia, C., Lesk, A.M., Tramontano, A., Levitt, M., Smith-Gill, S.J., Air, G., Sheriff, S., Padlan, E.A., Davies, D., Tulip, W.R., et al. (1989) *Nature* 342, 877–883.
25. Wright, A., Tao, M.H., Kabat, E.A., & Morrison, S.L. (1991) *EMBO J.* 10, 2717–2723.
26. Chothia, C., Boswell, D.R., & Lesk, A.M. (1988) *EMBO J.* 7, 3745–3755.
27. Bossart-Whitaker, P., Chang, C.Y., Novotny, J., Benjamin, D.C., & Sheriff, S. (1995) *J. Mol. Biol.* 253, 559–575.
28. Desmyter, A., Transue, T.R., Ghahroudi, M.A., Thi, M.H., Poortmans, F., Hamers, R., Muyldermans, S., & Wyns, L. (1996) *Nat. Struct. Biol.* 3, 803–811.
29. Davies, D.R., & Cohen, G.H. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 7–12.
30. Lo Conte, L., Chothia, C., & Janin, J. (1999) *J. Mol. Biol.* 285, 2177–2198.
31. Mylvaganam, S.E., Paterson, Y., & Getzoff, E.D. (1998) *J. Mol. Biol.* 281, 301–322.
32. Faelber, K., Kirchhofer, D., Presta, L., Kelley, R.F., & Muller, Y.A. (2001) *J. Mol. Biol.* 313, 83–97.
33. Li, Y., Li, H., Smith-Gill, S.J., & Mariuzza, R.A. (2000) *Biochemistry* 39, 6296–6309.
34. Nisonoff, A., Reichlin, M., & Margoliash, E. (1970) *J. Biol. Chem.* 245, 940–946.
35. Takano, T., Kallai, O.B., Swanson, R., & Dickerson, R.E. (1973) *J. Biol. Chem.* 248, 5234–5255.
36. Bhat, T.N., Bentley, G.A., Fischmann, T.O., Boulot, G., & Poljak, R.J. (1990) *Nature* 347, 483–485.
37. Braden, B.C., Souchon, H., Eisele, J.L., Bentley, G.A., Bhat, T.N., Navaza, J., & Poljak, R.J. (1994) *J. Mol. Biol.* 243, 767–781.
38. Sheriff, S., Chang, C.Y., Jeffrey, P.D., & Bajorath, J. (1996) *J. Mol. Biol.* 259, 938–946.
39. Hogle, J.M., Chow, M., & Filman, D.J. (1985) *Science* 229, 1358–1365.
40. Rossmann, M.G., Arnold, E., Erickson, J.W., Frankenberger, E.A., Griffith, J.P., Hecht, H.J., Johnson, J.E., Kamer, G., Luo, M., Mosser, A.G., et al. (1985) *Nature* 317, 145–153.
41. Padlan, E.A., Silverton, E.W., Sheriff, S., Cohen, G.H., Smith-Gill, S.J., & Davies, D.R. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 5938–5942.
42. Atassi, M.Z., Habeeb, A.F., & Ando, K. (1973) *Biochim. Biophys. Acta* 303, 203–209.
43. Macht, M., Fiedler, W., Kurzinger, K., & Przybylski, M. (1996) *Biochemistry* 35, 15633–15639.
44. Tzartos, S.J., Kokla, A., Walgrave, S.L., & Conti-Tronconi, B.M. (1988) *Proc. Natl. Acad. Sci. U.S.A.* 85, 2899–2903.
45. Evin, G., Galen, F.X., Carlson, W.D., Handschumacher, M., Novotny, J., Bouhnik, J., Menard, J., Corvol, P., & Haber, E. (1988) *Biochemistry* 27, 156–164.
46. Dokurno, P., Bates, P.A., Band, H.A., Stewart, L.M., Lally, J.M., Burchell, J.M., Taylor-Papadimitriou, J., Snary, D., Sternberg, M.J., & Freemont, P.S. (1998) *J. Mol. Biol.* 284, 713–728.
47. Ahern, T.J., & Klibanov, A.M. (1985) *Science* 228, 1280–1284.
48. Sachs, D.H., Schechter, A.N., Eastlake, A., & Anfinsen, C.B. (1972) *Proc. Natl. Acad. Sci. U.S.A.* 69, 3790–3794.
49. Arevalo, J.H., Hassig, C.A., Stura, E.A., Sims, M.J., Taussig, M.J., & Wilson, I.A. (1994) *J. Mol. Biol.* 241, 663–690.
50. Mizutani, R., Miura, K., Nakayama, T., Shimada, I., Arata, Y., & Satow, Y. (1995) *J. Mol. Biol.* 254, 208–222.
51. Barisas, B.G., Singer, S.J., & Sturtevant, J.M. (1972) *Biochemistry* 11, 2741–2744.
52. Smith, T.W., Butler, V.P., Jr., & Haber, E. (1970) *Biochemistry* 9, 331–337.
53. Walter, G., Scheidtmann, K.H., Carbone, A., Laudano, A.P., & Doolittle, R.F. (1980) *Proc. Natl. Acad. Sci. U.S.A.* 77, 5197–5200.
54. Kyte, J., Xu, K.Y., & Bayer, R. (1987) *Biochemistry* 26, 8350–8360.
55. Wilchek, M., Bocchini, V., Becker, M., & Givol, D. (1971) *Biochemistry* 10, 2828–2834.
56. Shoham, M. (1993) *J. Mol. Biol.* 232, 1169–1175.
57. Dwyer, B.P. (1988) *Biochemistry* 27, 5586–5592.
58. van der Donk, W.A., Zeng, C., Biemann, K., Stubbe, J., Hanlon, A., & Kyte, J. (1996) *Biochemistry* 35, 10058–10067.
59. Erickson, H.K. (2001) *Biochemistry* 40, 9631–9637.
60. Dwyer, B.P. (1991) *Biochemistry* 30, 4105–4112.
61. Ouchterlony, O. (1958) *Prog. Allergy* 5, 1–78.
62. Izumi, Y., Kanzaki, H., Morita, S., Futazuka, H., & Yamada, H. (1989) *Eur. J. Biochem.* 182, 333–341.
63. Levine, L., & VanVunakis, H. (1967) *Methods Enzymol.* 11, 928–936.
64. Renart, J., Reiser, J., & Stark, G.R. (1979) *Proc. Natl. Acad. Sci. U.S.A.* 76, 3116–3120.
65. Towbin, H., Staehelin, T., & Gordon, J. (1979) *Proc. Natl. Acad. Sci. U.S.A.* 76, 4350–4354.
66. Blake, M.S., Johnston, K.H., Russell-Jones, G.J., & Gotschlich, E.C. (1984) *Anal. Biochem.* 136, 175–179.
67. Pluskal, M.G., Przekop, M.B., Kavonian, M.R., Vecoli, C., & Hicks, D.A. (1986) *BioTechniques* 4, 272–283.
68. Walker, J.E., Arizmendi, J.M., Dupuis, A., Fearnley, I.M., Finel, M., Medd, S.M., Pilkington, S.J., Runswick, M.J., & Skehel, J.M. (1992) *J. Mol. Biol.* 226, 1051–1072.
69. Domingo, A., & Marco, R. (1989) *Anal. Biochem.* 182, 176–181.

70. Vivekananda, J., Beck, C.F., & Oliver, D.J. (1988) *J. Biol. Chem.* 263, 4782–4788.
71. Han, A.L., Yagi, T., & Hatefi, Y. (1989) *Arch. Biochem. Biophys.* 275, 166–173.
72. Han, A.L., Yagi, T., & Hatefi, Y. (1988) *Arch. Biochem. Biophys.* 267, 490–496.
73. Galante, Y.M., & Hatefi, Y. (1979) *Arch. Biochem. Biophys.* 192, 559–568.
74. Ragan, C.I., Galante, Y.M., & Hatefi, Y. (1982) *Biochemistry* 21, 2518–2524.
75. Bisson, R., & Schiavo, G. (1986) *J. Biol. Chem.* 261, 4373–4376.
76. Yamaguchi, M., & Hatefi, Y. (1993) *Biochemistry* 32, 1935–1939.
77. Canals, F. (1992) *Biochemistry* 31, 4493–4501.
78. Wofsy, L., & Burr, B. (1969) *J. Immunol.* 103, 380–382.
79. Schneider, C., Newman, R.A., Sutherland, D.R., Asser, U., & Greaves, M.F. (1982) *J. Biol. Chem.* 257, 10766–10769.
80. Santacruz-Toloz, L., Perozo, E., & Papazian, D.M. (1994) *Biochemistry* 33, 1295–1299.
81. Hoffman, E.P., Brown, R.H., Jr., & Kunkel, L.M. (1987) *Cell* 51, 919–928.
82. Pons, F., Augier, N., Heilig, R., Leger, J., Mornet, D., & Leger, J.J. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87, 7851–7855.
83. Middleton, R.E., Pheasant, D.J., & Miller, C. (1994) *Biochemistry* 33, 13189–13198.
84. Langone, J.J. (1982) *J. Immunol. Methods* 55, 277–296.
85. Munro, S., & Pelham, H.R. (1984) *EMBO J.* 3, 3087–3093.
86. Hanai, R., & Wang, J.C. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 11904–11908.
87. Kemp, D.J., & Cowman, A.F. (1981) *Proc. Natl. Acad. Sci. U.S.A.* 78, 4520–4524.
88. Delain, E., Barray, M., Tapon-Brethaudiere, J., Pochon, F., Marynen, P., Cassiman, J.J., Van den Berghe, H., & Van Leuven, F. (1988) *J. Biol. Chem.* 263, 2981–2989.
89. Veklich, Y.I., Gorkun, O.V., Medved, L.V., Nieuwenhuizen, W., & Weisel, J.W. (1993) *J. Biol. Chem.* 268, 13577–13585.
90. Kopp, F., Dahlmann, B., & Hendil, K.B. (1993) *J. Mol. Biol.* 229, 14–19.
91. Kopp, F., Hendil, K.B., Dahlmann, B., Kristensen, P., Sobek, A., & Uerkvitz, W. (1997) *Proc. Natl. Acad. Sci. U.S.A.* 94, 2939–2944.
92. Kaltschmidt, E., & Wittmann, H.G. (1970) *Anal. Biochem.* 36, 401–412.
93. Held, W.A., Mizushima, S., & Nomura, M. (1973) *J. Biol. Chem.* 248, 5720–5730.
94. Wittmann, H.G., Littlechild, J.A., & Wittman-Liebold, B. (1980) in *Ribosomes, Structure, Function, and Genetics* (Chambliss, G., Craven, G.R., Davies, J., Davis, K., Kahan, L., & Nomura, M., Eds.) pp 51–88, University Park Press, Baltimore, MD.
95. Pioletti, M., Schlunzen, F., Harms, J., Zarivach, R., Gluhmann, M., Avila, H., Bashan, A., Bartels, H., Auerbach, T., Jacobi, C., Hartsch, T., Yonath, A., & Franceschi, F. (2001) *EMBO J.* 20, 1829–1839.
96. Tischendorf, G.W., Zeichhardt, H., & Stöffler, G. (1975) *Proc. Natl. Acad. Sci. U.S.A.* 72, 4820–4824.
97. Lake, J.A., & Kahan, L. (1975) *J. Mol. Biol.* 99, 631–644.
98. Kahan, L., Winkelmann, D.A., & Lake, J.A. (1981) *J. Mol. Biol.* 145, 193–214.
99. Scheinman, A., Atha, T., Aguinaldo, A.M., Kahan, L., Shankweiler, G., & Lake, J.A. (1992) *Biochimie* 74, 307–317.
100. Stoffler-Meilicke, M., & Stoffler, G. (1987) *Biochimie* 69, 1049–1064.
101. Brodersen, D.E., Clemons, W.M., Jr., Carter, A.P., Wimberly, B.T., & Ramakrishnan, V. (2002) *J. Mol. Biol.* 316, 725–768.
102. Winkelmann, D.A., Kahan, L., & Lake, J.A. (1982) *Proc. Natl. Acad. Sci. U.S.A.* 79, 5184–5188.
103. Kraulis, P.J. (1991) *J. Applied Crystallogr.* 24, 946–950.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Chapter 12

Physical Measurements of Structure

Physical properties used to assess the structure of a protein are its standard diffusion constant, its standard sedimentation coefficient, its intrinsic viscosity, and the angular dependence of its ability to scatter light, X-radiation, and neutrons, all of which respond to the shape of the macromolecule; its absorption of light, which responds to the environments around particular chromophores, in particular the peptide bonds; its fluorescence or the fluorescence of chromophores covalently attached to it, which can be used to measure molecular shape and intramolecular dimensions; and its nuclear magnetic resonance spectrum, which can be used to map spatial relationships among the amino acids in the native structure. These physical properties are derived from measurements made of solutions of the protein. When a crystallographic molecular model is not available, such physical measurements provide the only structural information about a particular protein. As more and more crystallographic molecular models become available, however, physical measurements have become valuable complements to crystallography. Because they are structural measurements of the protein in solution, they can be used to validate a crystallographic molecular model, or in some situations to adjust the crystallographic molecular model to correct for differences between the structure of a molecule of protein in a crystal and its structure in solution.

Shape

A molecule of protein dissolved in an aqueous solution of moderate ionic strength is a compact solid of peculiar shape coated with a **layer of water** that is more or less fixed upon its surface and that has the effect of smoothing its roughness. The available crystallographic molecular models of various proteins are similar enough to each other to provide an accurate mental picture of the boundary between the molecule of protein proper and the liquid water of the bulk phase. Crevices on the surface of a molecule of protein are filled with molecules of water. Although these molecules of water are rapidly exchanging with their neighbors more peripherally located, the locations where they sit are always occupied and can be considered to be permanent features of the molecule of protein. Their net effect is to fill the crevices on the surface of the folded protein. Between these

crevices, over the open surface of a molecule of protein, a large number of molecules of water are situated in locations that are also permanently occupied, even though constantly exchanging. The relative positions of these locations becomes less and less fixed the farther they are situated from the atoms of the molecule of protein until a region is reached where the water is no different from the water in an otherwise identical solution lacking the protein. This continuous transition between the molecules of water fixed to two or three donors and acceptors of hydrogen bonds on the surface of a molecule of protein and the molecules of water in the bulk solvent is characterized by a gradual, rather than an abrupt, decrease of attachment. Therefore, no distinct boundary exists between the macromolecule and the solvent. Nevertheless, the concept of the hydrodynamic particle¹ is necessary if specific dimensions are to be extracted from physical measurements of the shapes of molecules of protein dissolved in free solution.

The **hydrodynamic particle** is the covalent molecule of protein and any molecules of water and any solutes that behave during the measurement as if they were affixed to the molecule of protein. An **affixed molecule of water** would be a specific location upon the surface of the protein continuously occupied by one or another molecule of water over a period of time long enough that the measurement registers it as a permanent feature.

If it is assumed that a hydrodynamic particle exists, its **mass** m_h (in grams) will be

$$m_h = \frac{M_{\text{prot}}(1 + \delta_{\text{H}_2\text{O}})}{N_A} \quad (12-1)$$

where M_{prot} is the molar mass (grams mole⁻¹) of the protein, $\delta_{\text{H}_2\text{O}}$ denotes the grams of water bound for every gram of protein (Table 6-4), and N_A is Avogadro's number (6.022×10^{23} mol⁻¹). It should be recalled that M_{prot} , the molar mass of the covalent structure of the protein, is almost always calculated directly from the amino acid sequences and stoichiometries of its constituent polypeptides and the amount of any posttranslational modifications.

The **volume** of the hydrodynamic particle (centimeters³) should be

$$V_h = \frac{M_{\text{prot}}}{N_A} (\bar{v}_{\text{prot}} + \delta_{\text{H}_2\text{O}} \nu_{\text{H}_2\text{O}}^0) \quad (12-2)$$

where \bar{v}_{prot} is the partial specific volume of the protein in $\text{cm}^3 \text{ gram}^{-1}$ and $\nu_{\text{H}_2\text{O}}^0$ is the specific volume of pure water in $\text{cm}^3 \text{ gram}^{-1}$.

If **other solutes** j are attached to the hydrodynamic particle, Equation 12-1 is expanded by adding a set of terms δ_j , each of which is the grams of each solute j for every gram of protein, and Equation 12-2 is expanded by adding a set of terms $\delta_j \bar{v}_j$, where the \bar{v}_j are the partial specific volumes ($\text{cm}^3 \text{ gram}^{-1}$) of the solutes j . An example of bound solutes for which these additional terms are major features of these equations is a case in which the protein has bound detergents or bound lipids.²

The **standard diffusion coefficient** of a protein ($\text{cm}^2 \text{ second}^{-1}$) is designated as $D_{20,w}^0$, where the superscript indicates extrapolation to a zero concentration of protein and the subscripts indicate a correction to a temperature of 20 °C and to a solvent with the viscosity of pure water. The standard diffusion coefficient is a measure of f , the **frictional coefficient** (grams second^{-1}) of the hydrodynamic particle in water at 20 °C at infinite dilution:

$$f = \frac{k_B T}{D_{20,w}^0} \quad (12-3)$$

where k_B is Boltzmann's constant ($1.381 \times 10^{-16} \text{ erg K}^{-1}$) and T is the temperature (293.15 K). A standard diffusion coefficient and a frictional coefficient are particular and **intrinsic properties** of a given protein in a given solution.

The concept of the hydrodynamic particle qualifies the meaning of the diffusion coefficient presented in Chapter 1 and the frictional ratio presented in Chapter 6. By use of the equation for the frictional coefficient of a sphere, a **minimum frictional coefficient** for the hydrodynamic particle at infinite dilution can be defined as

$$f_{0,h} \equiv 6\pi\eta R_{0,h} \quad (12-4)$$

where the subscript zero refers to the minimization and η is the viscosity (pascal seconds). The viscosity of pure water at 20 °C is 1.002 mPa s. The **hydrodynamic radius**, $R_{0,h}$, is defined as the radius (centimeters) of a sphere with the same volume as the hydrodynamic particle:

$$R_{0,h} \equiv \left(\frac{3V_h}{4\pi} \right)^{1/3} \quad (12-5)$$

Consequently

$$f_{0,h} = 6\pi\eta \left[\frac{3M_{\text{prot}}}{4\pi N_A} (\bar{v}_{\text{prot}} + \delta_{\text{H}_2\text{O}} \nu_{\text{H}_2\text{O}}^0) \right]^{1/3} \quad (12-6)$$

The hydrodynamic radius, $R_{0,h}$, the radius of a sphere with the same volume as the hydrodynamic particle, must not be confused with the apparent radius, or Stokes radius, of the particle, a , the radius of a sphere with the same standard diffusion coefficient as the particle.

The definition of the minimum frictional coefficient for the molecule of protein, that expected of a hydrated sphere of the same volume as the hydrated molecule of protein, incorporates the water bound to the protein rather than treating the protein as if it were unhydrated as was done earlier (Equation 8-39). If $\delta_{\text{H}_2\text{O}}$ is 0.3 g (g of protein)⁻¹, consistent with the values in Table 6-4, the **hydrated effective sphere** should have a volume 1.4 times as large as the unhydrated effective sphere (if \bar{v}_{prot} is taken as 0.74 $\text{cm}^3 \text{ g}^{-1}$ and $\bar{v}_{\text{H}_2\text{O}}$ as 1.00 $\text{cm}^3 \text{ g}^{-1}$), and the frictional coefficient of the hydrated effective sphere of protein, $f_{0,h}$, should be 1.12 times larger than the frictional coefficient of the unhydrated effective sphere of protein, $f_{0,unh}$. This is consistent with the fact that the smallest frictional ratios, $f/f_{0,unh}$, observed for globular proteins are always greater than or equal to 1.1 when no correction is made for hydration.

The relationship between the frictional ratio (f/f_0) and the shape of a particle has been derived for **ellipsoids of revolution**, either prolate or oblate.³ The relationships can be presented graphically (Figure 12-1A).¹ After the frictional ratio ($f/f_{0,h}$) for the hydrodynamic particle has been calculated from the observed value of the frictional coefficient f (Equation 12-3) and the calculated value of $f_{0,h}$ (Equation 12-6), the apparent **axial ratio**, a/b , of the hydrodynamic particle can be read from the graph.

Molecules of protein are neither prolate nor oblate ellipsoids of revolution, but exact solutions to the hydrodynamic equations are available only for these shapes. Such an approximation, however, may provide some insight into the shape of a particular molecule of protein, especially when the frictional ratio differs greatly from 1. Such a result cannot be explained on the basis of an unexpectedly high degree of hydration and states that the protein of interest is peculiar in its shape. In the particular case where the molecule is thought to resemble a **cylindrical rod** of length L and diameter d , it has been concluded that the dimensions of that cylindrical rod can be calculated by using the frictional ratio to determine the axial ratio of an equivalent prolate ellipsoid, a/b , and then applying the formula

$$\frac{L}{d} = \left(\frac{3}{2} \right)^{1/2} \frac{a}{b} \quad (12-7)$$

The segment of rope formed by triple-helical collagen type I (Figure 9-33), usually referred to as a protofil-

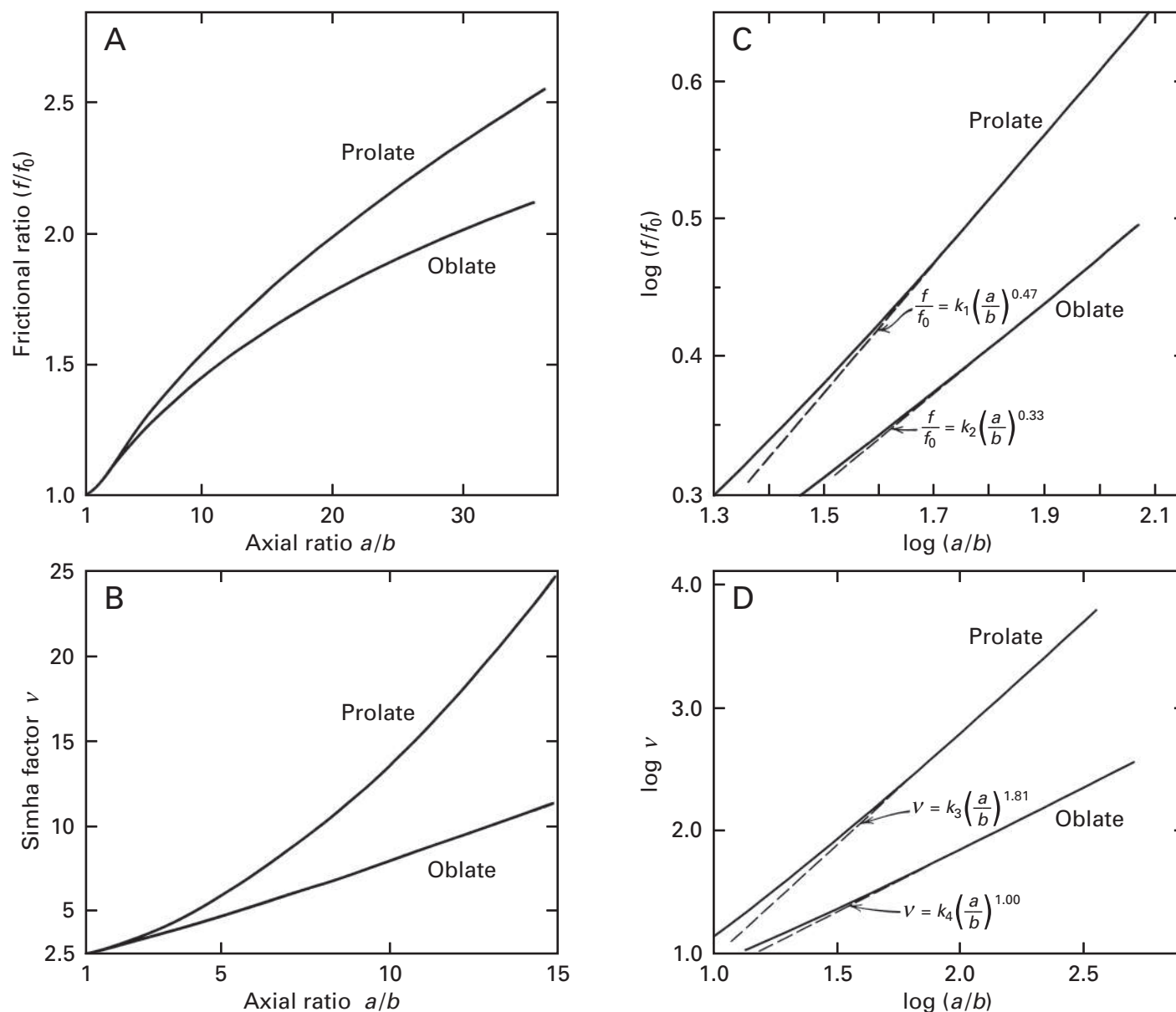


Figure 12-1: Graphic relationships¹ between the axial ratio (a/b) of an oblate ellipsoid of revolution or a prolate ellipsoid of revolution and either (A) the frictional ratio (f/f_0) or (B) the Simha factor (ν). A prolate ellipsoid of revolution is generated by rotation around the major axis of an ellipse; an oblate ellipsoid of revolution, by rotation around the minor axis. The relationships for smaller values of the axial ratio are given directly. For large values of the axial ratio (>10) the logarithm of the frictional ratio (C) or the logarithm of the Simha factor (D) is given as a function of the logarithm of the axial ratio. The frictional ratio or the Simha factor, determined experimentally, can be converted into an axial ratio with the appropriate graph. When the axial ratios are greater than 100, each of the four curves in panels C and D becomes a straight line to infinity. As a result, values of the frictional ratios or the Simha factors that are greater than those on the graphs can still be converted to values for axial ratios with the use of the slopes of these lines for extrapolation. The slope of the line in panel C for logarithms of the frictional ratios of prolate ellipsoids is 0.47; that for oblate ellipsoids, 0.33; the slope of the line in panel D for logarithms of the Simha factors for prolate ellipsoids is 1.81; and that for oblate ellipsoids, 1.00. Reprinted with permission from ref 1. Copyright 1961 John Wiley.

ament, is known to be a rod. The molar mass of a triple-helical rope of collagen type I is $281,000 \text{ g mol}^{-1}$, its partial specific volume¹ is $0.695 \text{ cm}^3 \text{ g}^{-1}$, and its standard diffusion coefficient¹ is $0.85 \times 10^{-7} \text{ cm}^2 \text{ s}^{-1}$. If it is assumed that $\delta_{\text{H}_2\text{O}} = 0.3 \text{ g g}^{-1}$, then the frictional ratio for the hydrodynamic particle, $f/f_{0,h}$, would be 5.3, for which the ratio of L/d would be 210 (Figure 12-1C and Equation 12-7). The volume of the hydrodynamic particle containing a

segment of rope of collagen type I, based on the assumption that $\delta_{\text{H}_2\text{O}} = 0.3$, would be 460 nm^3 (Equation 12-2), which would fill a cylinder 285 nm long with a diameter of 1.35 nm. From the dimensions of the triple helix (Figure 9-33) and the length of the polypeptide, a molecule of unpolymerized collagen type I is thought to be 300 nm long.

It is also possible to calculate the frictional coeffi-

cient for a **string of spherical beads**⁴ in various geometric arrangements. Fibronectin in electron micrographs appears as a flexible segment of rope 130 nm in length. A string of spherical beads of that length and with a total volume equal to that of a molecule of fibronectin has a frictional coefficient equal to that calculated from its standard diffusion coefficient.⁵ Before the length of a molecule of caldesmon had been established by electron microscopy, it was calculated that a string of spherical beads 74 nm in length with a total volume equal to that of a molecule of the protein would have a frictional coefficient equal to that observed for a molecule of the protein.⁶ It was later shown that the length of the rope-like molecule of caldesmon seen in electron micrographs of the protein is about 70 nm.⁷ A string of 12 small spherical beads ($r = 0.8$ nm) attached to a single larger spherical bead ($r = 3$ nm) produced a structure that resembles electron micrographs of vinculin and has a frictional coefficient equal to that calculated from the standard diffusion coefficient observed for the protein.⁸

The frictional coefficient of a molecule of protein can also be determined by **sedimentation velocity**. Consider a hydrodynamic particle dissolved in aqueous solution that is submitted to a high centrifugal force in the rotor of an ultracentrifuge. The centrifugal force on the particle is equal to $m_h \omega^2 r$, where ω is the angular velocity (radians second⁻¹) of the rotor, m_h is the mass (grams) of the hydrodynamic particle, and r is the distance (centimeters) the particle is from the axis of the rotor. This centrifugal force is countered by the buoyant force, which is equal to $V_h \rho_{\text{sol}} \omega^2 r$, where ρ_{sol} is the density (grams centimeter⁻³) of the solution displaced by V_h , the volume (centimeters³) of the hydrodynamic particle. The net force on the particle is

$$F = \omega^2 r (m_h - V_h \rho_{\text{sol}}) \quad (12-8)$$

Because the measurements are extrapolated to a solution of protein at a concentration of zero in only water, it can be assumed that ρ_{sol} is insignificantly different from ρ_0 , the density of water at the appropriate temperature. Because ρ_0 is the reciprocal of $V_{\text{H}_2\text{O}}^0$, when expressions for m_h (Equation 12-1) and V_h (Equation 12-2) are substituted, the net force is

$$F = \frac{M_{\text{prot}}}{N_A} \omega^2 r (1 - \bar{v}_{\text{prot}} \rho_0) \quad (12-9)$$

The term $(M_{\text{prot}}/N_A)(1 - \bar{v}_{\text{prot}} \rho_0)$ is the **buoyant mass** of the hydrodynamic particle of the protein.

The net force (Equation 12-9) causes the hydrodynamic particle to accelerate. As it accelerates, the frictional force, which is equal to fu , where f is the frictional coefficient (grams second⁻¹) and u is the velocity (centimeters second⁻¹) of the hydrodynamic particle,

increases in direct proportion to the velocity of the particle until it just balances the net centrifugal force. At that point, a steady state is achieved, the forces on the hydrodynamic particle are equal and opposite, the particle travels in the direction of the centrifugal force at its **terminal velocity**, and

$$fu = \frac{M_{\text{prot}}}{N_A} \omega^2 r (1 - \bar{v}_{\text{prot}} \rho_0) \quad (12-10)$$

This equation can be rearranged to give

$$s \equiv \frac{u}{\omega^2 r} = \frac{M_{\text{prot}} (1 - \bar{v}_{\text{prot}} \rho_0)}{f N_A} \quad (12-11)$$

The term on the left, $u\omega^{-2}r^{-1}$, which is the observed velocity normalized for all of the parameters of the instrument, can be directly measured, and it is referred to as the **sedimentation coefficient**, s . The standard sedimentation coefficient (seconds) for the hydrodynamic particle is designated as $s_{20,w}^0$, where superscript and subscript have the same meaning as before. Because sedimentation coefficients of proteins are between 10^{-13} and 10^{-11} s, the unit 10^{-13} s is designated as S, the Svedberg.

Because it is only a function of universal constants and the properties of the molecule of protein itself, the standard sedimentation coefficient is also an intrinsic property of the protein. In particular it is, as is the standard diffusion coefficient, a direct measurement of the frictional coefficient

$$f = \frac{M_{\text{prot}} (1 - \bar{v}_{\text{prot}} \rho_0)}{s_{20,w}^0 N_A} \quad (12-12)$$

To use Equation 12-12, the molar mass of the protein must be a fixed and known quantity. If the protein is normally engaged in a reaction that changes its molar mass, such as the equilibrium between the dimers and the tetramer of hemoglobin, a molar mass cannot be assigned. If the protein is participating in such a reaction, abnormally large decreases in the sedimentation coefficient will occur as the concentration of the protein is decreased upon extrapolation to the zero concentration of protein.

Aspartate carbamoyltransferase from *Escherichia coli* (Figure 9-37) serves as an example of the application of analysis by sedimentation to the study of a hydrodynamic particle of known shape. It is a protein of molar mass 308,100 g mol⁻¹, and its standard sedimentation coefficient $s_{20,w}^0$ is 11.6 S,^{9,10} from which a frictional coefficient f of 11.6×10^{-8} g s⁻¹ can be calculated. The minimum frictional coefficient, $f_{0,\text{unh}}$, for the unhydrated protein ($\delta_{\text{H}_2\text{O}} = 0$), folded as a sphere, would be 8.5×10^{-8} g s⁻¹ (Equation 12-6). If hydration of

0.3 g g^{-1} is assumed, this would give a frictional ratio $f/f_{0,h}$ of 1.22. Although the protein (Figure 9–37) does not have the axial ratio suggested by the frictional ratio (a/b would be 5 were the protein an oblate ellipsoid), the value of 1.22 probably results from both the irregular shape of the molecule and an abnormally large amount of bound water within the central cavity between the two C subunits.

It has been demonstrated crystallographically¹¹ that when aspartate carbamoyltransferase binds the enzymatic inhibitor *N*-(phosphonacetyl)-L-aspartate, the protein undergoes a conformational change that alters the disposition of its subunits significantly. A **conformational change** in a protein is any change in its structure brought about by a change in the solution, for example, the addition of an inhibitor. The net effect of this particular conformational change in aspartate carbamoyltransferase is to move the two trimers of catalytic α subunits (Figure 9–37) 1.2 nm farther apart. In the process, the water-filled space between the two C subunits widens by the same amount. This change in structure caused by the binding of the inhibitor can be detected as a change in the sedimentation coefficient of aspartate carbamoyltransferase.^{9,10} This change is accurately quantified by difference sedimentation analysis in which the two samples, with and without the inhibitor, are simultaneously monitored in separate cells in the same rotor.¹² The sedimentation coefficient^{9,10} decreases by 3.4% upon the change in structure. The diameter of the space between two trimers of catalytic α subunits of aspartate carbamoyltransferase is about 8 nm,¹¹ so the increase in the amount of water between them resulting from a movement apart of 1.2 nm should be about $40,000 \text{ g mol}^{-1}$. This change alone should increase the frictional coefficient (Equation 12–6) by 3–4%, which would account completely for the observed change of 3.4%.

Both the high frictional ratio for unexpanded aspartate carbamoyltransferase and the increase in hydration experienced upon its expansion illustrate the fact that **oligomeric proteins**, because of the spaces among the subunits, display greater hydration than monomeric proteins. From results of measurements of the scattering of X-radiation at small angles by solutions of proteins, it has been calculated that while monomeric proteins have values of $\delta_{\text{H}_2\text{O}}$ of 0.25–0.35 g g^{-1} , oligomeric proteins have significantly higher values for $\delta_{\text{H}_2\text{O}}$ of 0.35–0.7 g g^{-1} .^{13,14} If the hydration of aspartate carbamoyltransferase in its unexpanded state were 0.6 g g^{-1} , its frictional ratio would be only 1.13, a value that is easily accounted for by its irregular shape.

Desmin is one of the proteins that forms intermediate filaments (Figures 9–35 and 9–36). The monomeric unit of the polymer is an α_2 dimer of two identical polypeptides; those from chicken are 463 aa long. The core of the dimer is a coiled coil of two α helices, one from each polypeptide. This coiled coil is contained within a fragment of the dimer containing the polypep-

tides from Glycine 70 to Phenylalanine 415 that can be produced by digestion with chymotrypsin. The standard sedimentation coefficient of this dimeric fragment is 2.85 S, while the standard sedimentation coefficient of an $(\alpha_2)_2$ dimer of this dimer is 3.85 S.¹⁵ If it is assumed that both of these oligomers, the dimer and the dimer of dimers, can be represented as cylindrical rods, L/d for the α_2 dimer is 28 and that for the $(\alpha_2)_2$ dimer of dimers is 44. Consequently, the $(\alpha_2)_2$ dimer of dimers is 1.7 times longer than one α_2 dimer. If this approximation is realistic, the coiled coils of the two dimers must be staggered by about 0.7 of their length in the dimer of dimers.

The standard diffusion coefficient and the standard sedimentation coefficient provide **independent determinations** of the frictional coefficient of a molecule of protein. The force producing net flux of protein when diffusion is measured is chemical potential, which is unrelated to, as well as being somewhat less concrete of a concept than, centrifugal force. Furthermore, the theoretical derivations of the relationship between the diffusion coefficient and the frictional coefficient and that between the sedimentation coefficient and the frictional coefficient are entirely different. It is of interest to compare (Table 12–1) the two frictional coefficients, that calculated from diffusion (f_{diff}) and that calculated from sedimentation (f_{sed}). Depending upon one's prejudice, the agreement between the numbers is either as one expected or quite gratifying. The lack of any systematic deviation verifies the assumption, first made by Einstein, that the same frictional coefficient applies to both diffusion and sedimentation.

The frictional ratios ($f_{\text{av}}/f_{0,h}$), where f_{av} is the average of the two measurements, are close to 1 (1.1–1.2) for most globular proteins (Table 12–1), but even in these instances the frictional ratios of the hydrated particles predict (Figure 12–1A) an axial ratio of greater than 3, which is unrealistic. These values of 1.1–1.2 do not indicate elongation but reflect the fact that molecules of protein, even when they are almost spherical, are not smooth spheres but globular macromolecules with **irregular, rough surfaces**. Proteins such as fibrinogen, apolipoprotein(a), and plasminogen, however, which are known from other observations to be highly asymmetric, have much higher frictional ratios.

The specific examples of **elongated or excessively hydrated proteins** described in detail so far, collagen, fibronectin, caldesmon, vinculin, aspartate carbamoyltransferase, and desmin, are macromolecules about which enough is known that the hydrodynamic measurements can be evaluated comprehensively. When no other structural information is available about a protein, a frictional ratio around 1.1 is strong evidence that it is globular, but frictional ratios greater than 1.1 are difficult to interpret. For example, the fact that human bifunctional polynucleotide 3'-phosphatase/5'-kinase has a frictional ratio ($f/f_{0,h}$; $\delta_{\text{H}_2\text{O}} = 0.3$) of 1.30¹⁹ could result from an unusually irregular shape, an elongated shape,

Table 12-1: Frictional Coefficients from Sedimentation and Diffusion

protein	species	\bar{v}^a (cm ³ g ⁻¹)	$s_{20,w}^0$ ^a (s × 10 ¹³)	$D_{20,w}^0$ ^a (cm ² s ⁻¹ × 10 ⁷)	M_p ^b (g mol ⁻¹)	f_{sed}^c (g s ⁻¹ × 10 ⁸)	f_{diff}^d (g s ⁻¹ × 10 ⁸)	$f_{0,unh}^e$ (g s ⁻¹ × 10 ⁸)	$f_{0,h}^f$ (g s ⁻¹ × 10 ⁸)	$f_{av}/f_{0,h}^g$
lysozyme	chicken	0.703	1.91	11.20	14,310	3.7	3.6	2.99	3.4	1.09
alcohol dehydrogenase	horse	0.750	5.0	6.2	79,600 ^h	6.6	6.5	5.41	6.1	1.09
catalase	cow	0.730	11.30	4.10	232,800 ⁱ	9.3	9.9	7.67	8.6	1.11
β-galactosidase	<i>E. coli</i>	0.76	15.93	3.12	465,400	11.7	13.0	9.79	10.9	1.13
serum albumin	human	0.735	4.64	6.0	66,470	6.3	6.7	5.06	5.7	1.15
fructose-bisphosphate aldolase	rabbit	0.742	7.35	4.63	156,840	9.2	8.7	6.76	7.6	1.18
prothrombin	cow	0.70	4.85	6.24	72,600 ^j	7.5	6.5	5.13	5.8	1.21
manganese-stabilizing protein ^k	spinach	0.732	2.26	7.6	26,530	5.3	5.3	3.72	4.2	1.27
plasminogen	human	0.71	4.30	4.31	103,000 ^l	11.6	9.4	5.8	6.5	1.61
apolipoprotein(a) ^m	human	0.69	9.30	2.29	323,000 ⁿ	18.0	17.7	8.4	9.5	1.88
fibrinogen	human	0.725	7.63	1.98	344,000 ^o	20.7	20.4	8.7	9.8	2.10

^aUnless otherwise noted, these values are from tables in ref 16. The entries are arranged in order of asymmetry. ^bFrom sequence. ^c f_{sed} from Equation 12-12. ^d f_{diff} from Equation 12-3. ^e $f_{0,unh}$ from Equation 12-6 with δ_{H_2O} equal to 0. ^f $f_{0,h}$ from Equation 12-6 with δ_{H_2O} equal to 0.3. ^gAverage of f_{sed} and f_{diff} divided by $f_{0,h}$. ^h2 Zn²⁺ subunit⁻¹. ⁱOne heme subunit⁻¹. ^j10.4 g of oligosaccharide (100 g of protein)⁻¹. ^kReference 17. ^l17 g of oligosaccharide (100 g of protein)⁻¹. ^mReference 18. ⁿ30 g of oligosaccharide (100 g of protein)⁻¹. ^o2 g of oligosaccharide (100 g of protein)⁻¹.

or an abnormally large amount of bound water or, most likely, some combination of all of these factors.

Measurement of the **viscosity** of a solution of a protein also provides an evaluation of the shape of the hydrodynamic particle. When a fluid flows through a cylindrical capillary under the appropriate circumstances, laminar flow occurs. The fluid immediately adjacent to the walls of the capillary is stationary, and the fluid at the center of the capillary has the highest rate of flow. Each cylindrical lamina between the center and the wall moves with an intermediate velocity that, as the distance from the center increases, monotonically decreases to a value of zero at the wall. **Laminar flow** requires that each cylindrical lamina move more slowly than its neighbor toward the center. As such, shear occurs between adjacent lamina throughout the capillary. The surfaces at which shear occurs are all parallel to the axis of the capillary. The more viscous the fluid, the more difficult it will be for these surfaces of shear to move across one another, and the more slowly the fluid will flow through the capillary. The time required for a given volume of a fluid to move through a given capillary at a given hydrostatic pressure is directly proportional to η , the viscosity (pascal seconds) of the fluid.

The addition of macromolecules such as proteins to the fluid in the capillary interrupts the shear that otherwise would occur in the solution in their vicinity and increases the viscosity of the solution. This increase can be expressed in terms of the **specific viscosity**, η_{sp} , which is defined by

$$\eta_{sp} \equiv \frac{\eta' - \eta}{\eta} = \frac{\eta'}{\eta} - 1 \quad (12-13)$$

where η' is the viscosity of the solution containing a particular concentration of the protein and η is the viscosity of an otherwise identical solution lacking the protein. The specific viscosity is a positive number because η' is always greater than η . The specific viscosity is the normalized incremental increase in the viscosity caused by the protein.

If the flow through the capillary is driven only by the weight of the fluid, the specific viscosity is readily measured because

$$\frac{\eta'}{\eta} = \frac{t'\rho'}{t\rho} \quad (12-14)$$

where t is the time for a given volume of a solution to flow through the capillary, ρ is the density of the solution, and the primed and unprimed terms refer to the solution of protein and an identical solution lacking the protein, respectively.

The specific viscosity, η_{sp} , is the fractional increase in the viscosity of the solution due to addition of the protein, and it increases monotonically as the concentration of protein is increased. To render this increase an intrinsic property of the protein, regardless of its concentration, the **intrinsic viscosity**, $[\eta]$ (centimeters³ gram⁻¹), is defined as

$$[\eta] \equiv \lim_{\gamma_{prot} \rightarrow 0} \frac{\eta_{sp}}{\gamma_{prot}} \quad (12-15)$$

where γ_{prot} is the concentration of protein in grams centimeter⁻³. At low concentrations of protein, η_{sp} should be directly proportional to γ_{prot} , and $[\eta]$ is simply the slope of

the line of η_{sp} plotted against γ_{prot} . Neither the specific viscosity nor the intrinsic viscosity is itself a viscosity. The intrinsic viscosity is sometimes called the **limiting viscosity number** to avoid this confusion.

For macromolecules such as proteins, it can be shown¹ that

$$[\eta] = v \frac{V_h N_A}{M_{prot}} \quad (12-16)$$

$$[\eta] = v \left(\bar{v}_{prot} + \delta_{H_2O} v_{H_2O}^0 \right) \quad (12-17)$$

where v is a dimensionless coefficient of proportionality referred to as the **Simha factor**. On the basis of Einstein's calculations, the value of v for a spherical hydrodynamic particle is 2.5. As with the frictional ratio, $f/f_{0,h}$, the relationship between the Simha factor v and shape has been derived for ellipsoids of revolution.²⁰ The relationships can be presented graphically (Figure 12-1B). If a value for δ_{H_2O} is assumed, v can be calculated from

$$v = \frac{[\eta]}{\bar{v}_{prot} + \delta_{H_2O} v_{H_2O}^0} \quad (12-18)$$

and the apparent value of the axial ratio can be read from the graph.

From Equation 12-17, if $\delta_{H_2O} = 0.3 \text{ g g}^{-1}$, $\bar{v}_{prot} = 0.74 \text{ cm}^3 \text{ g}^{-1}$, and $v_{H_2O}^0 = 1 \text{ cm}^3 \text{ g}^{-1}$, $[\eta]$ would be $2.6 \text{ cm}^3 \text{ g}^{-1}$ if the hydrodynamic particle were a sphere regardless of its molar mass. What this means is that as long as δ_{H_2O} and \bar{v}_{prot} do not vary significantly, the viscosities observed for a set of solutions, each of a different spherical molecule of protein and each at the same concentration in grams centimeter⁻³, will be the same regardless of whether the mass is distributed among only a few large spheres because the protein has a large molar mass or is distributed among many small spheres because the protein has a small molar mass. Most globular proteins do have intrinsic viscosities between 3.0 and $4.0 \text{ cm}^3 \text{ g}^{-1}$ regardless of their molar masses. An observation of an intrinsic viscosity in this range demonstrates that a protein is globular.

Intrinsic viscosity is dramatically **more sensitive to the asymmetry of a molecule** of protein than is the frictional coefficient. The frictional coefficient of a molecule of collagen type I is only 5 times larger than the frictional coefficient it would have if it were a hydrated sphere, but the intrinsic viscosity of a solution of collagen type I is 460 times larger than the intrinsic viscosity it would have if it were a hydrated sphere. The intrinsic viscosity of collagen type I¹ is $1150 \text{ cm}^3 \text{ g}^{-1}$. If it is assumed that $\delta_{H_2O} = 0.3 \text{ g g}^{-1}$, the Simha factor is 1160 (Equation 12-18), and if it is assumed that the molecule is cylindrical, the axial ratio (a/b) of the hydrodynamic

particle would be 140 (Figure 12-1D) and the ratio of cylindrical length to diameter (L/d) would be 170 (Equation 12-7). This ratio is that of a cylinder of length 260 nm with a diameter of 1.5 nm and a volume of 460 nm^3 . As noted before, collagen type I is 300 nm in length.

Another procedure that can provide information about the shape of a molecule of protein is the **scattering** of electromagnetic radiation or neutrons. In the earlier discussion of light scattering, it was mentioned that the intensity of the scattered light from a solution of protein can depend on the angle at which the measurement is made. This is due to the fact that if the molecule of protein has at least one dimension that is an appreciable fraction of the wavelength of the light, photons scattered from different points in the same molecule of protein will be out of phase, and **intramolecular interference** due to these mismatched phases will diminish the overall intensity of the scattered light. This interference increases as the angle at which the scattered light is measured, the scattering angle θ (Figure 8-4), is increased. At a scattering angle of 0, the angle of the **forward scattering** i_0 , there is no interference. It is the forward scattering that contains information about the molar concentration of particles in the solution, and hence the molar mass of those particles.

It can be shown that, when the contribution of the virial coefficients to the scattering is eliminated by extrapolating the measurements to zero concentration of protein

$$\lim_{\gamma_{prot} \rightarrow 0} \frac{K \gamma_{prot}}{R_\theta} \left(\frac{\partial \bar{n}}{\partial \gamma_{prot}} \right)_{T,P,\mu}^2 = \left(\frac{1}{M_{prot}} \right) \left[\frac{1}{P(\theta)} \right] \quad (12-19)$$

where K is the optical constant (moles centimeter⁻⁴) defined by Equation 8-28, R_θ is the Rayleigh ratio (centimeters⁻¹) calculated from the measurements by Equations 8-30 or 8-31, γ_{prot} is the concentration of protein in the units of grams centimeter⁻³, $(\partial \bar{n} / \partial \gamma_{prot})_{T,P,\mu}$ is the change (centimeters³ gram⁻¹) in the refractive index of the solution as a function of the concentration of the protein, and M_{prot} is the molar mass (grams mole⁻¹) of the protein. As noted previously, the incremental scattering i_θ is the scattering that results only from the molecules of protein and is measured as the difference in scattering between the solution containing protein and an identical solution not containing protein.

The function $P(\theta)$ is the factor by which the intensity of the light scattered only by the protein, the incremental scattered light (i_θ), is decreased as a result of the interference:²¹

$$P(\theta) = 1 - \frac{16\pi^2 R_G^2}{3\lambda^2} \sin^2 \frac{\theta}{2} + \dots \quad (12-20)$$

where R_G is the radius of gyration (centimeters) of the molecule of protein and θ is the angle relative to the incident radiation at which the scattered radiation is measured. The value for the wavelength of the light, λ , is its wavelength in the solution

$$\lambda = \frac{\lambda_0}{\tilde{n}} \quad (12-21)$$

where λ_0 is its wavelength in a vacuum. Equation 12-20 is an infinite series, but at small values of θ the higher terms become negligible and the approximation

$$\lim_{\theta \rightarrow 0} \frac{1}{P(\theta)} = \frac{1}{1 - \frac{16\pi^2 R_G^2}{3\lambda^2} \sin^2 \frac{\theta}{2}} = 1 + \frac{16\pi^2 R_G^2}{3\lambda^2} \sin^2 \frac{\theta}{2} \quad (12-22)$$

can be used. In practice

$$\lim_{\substack{\theta \rightarrow 0 \\ \gamma_{\text{prot}} \rightarrow 0}} \frac{\gamma_{\text{prot}}}{R_\theta} = \frac{\gamma_{\text{prot}}}{R_0} \left(1 + \frac{16\pi^2 R_G^2}{3\lambda^2} \sin^2 \frac{\theta}{2} \right) \quad (12-23)$$

A plot of the left-hand quotient, extrapolated to zero concentration of protein at each value of θ , against $\sin^2(\theta/2)$, for small values of θ , will be a straight line from the slope and intercept of which a value of R_G , the radius of gyration, can be calculated. Equation 12-19 emphasizes that the interference arising from the shape of the molecule of protein is independent from the colligative property of light scattering from which its molar mass can be estimated.

The radius of gyration of the molecule of protein is the molecular parameter that is obtained from the angular dependence of the intensity of the scattered radiation and that provides information about the shape of the molecule of protein. The **radius of gyration** of a solid of uniform scattering density, as is usually assumed to be the case for proteins, is defined by the relationship

$$R_G^2 = \frac{\int_{\text{vol}} r^2 dV}{\int_{\text{vol}} dV} \quad (12-24)$$

where r is the distance of a volume element dV from the center of mass. The integration is performed over the whole volume of the solid. The advantage of the radius of gyration is that it can be calculated by numerical integration for any structure, for example, a crystallographic molecular model, and compared to the value obtained

from the measurement. For example, in an elongated protein such as fibronectin, which is known to be constructed from internally repeating domains, radii of gyration can be calculated for various rigid structures built from a string of spheres each representing one of the individual domains of the molecule, and these calculated values can be compared to the observed value for the radius of gyration estimated by light scattering.²²

The radius of gyration for a single sphere of uniform density is

$$R_G = \left(\frac{3}{5} \right)^{1/2} R_{\text{sph}} \quad (12-25)$$

where R_{sph} is the radius of the sphere. The radius of gyration for a cylindrical rod is

$$R_G = \frac{L_r}{12^{1/2}} \quad (12-26)$$

where L_r is the length of the rod. The radius of gyration for a prolate ellipsoid²³ is

$$R_G = \left(\frac{2a^2 + b^2}{5} \right)^{1/2} \quad (12-27)$$

where a and b are the semi-major and semi-minor axes, respectively.

The effect of the finite size of the molecule of protein on the scattered light is that its intensity, as reflected in R_θ (Equation 8-30 or 8-31), decreases as θ increases, owing to intramolecular interference, but its intensity will decrease significantly only if the term $[16\pi^2 R_G^2 \sin^2(\theta/2)]/3\lambda^2$ in Equation 12-23 is large enough to cause a measurable effect. In practice,¹ this means that at least one dimension of the protein must be greater than $\lambda/20$. The sizes of most molecules of protein are too small for this to be the case when visible light ($\lambda = 300\text{--}500$ nm in water) is used as the radiation. For example, the decrease in light scattering from a solution of fibronectin measured at a wavelength of 436 nm (in vacuo) was only about 10% at the maximum possible $\sin^2(\theta/2)$ of 0,²⁴ even though fibronectin has a molar mass of $519,000 \text{ g mol}^{-1}$, is a string of domains with a total length of 180 nm, and has a radius of gyration of 8.6 nm.²²

For most proteins, significant decreases in angular light scattering are observed only when **X-radiation** is used ($\lambda = 0.1\text{--}0.2$ nm). Unfortunately, this is radiation of such **short wavelength** that complete intramolecular interference occurs at quite small values of θ (Equation 12-20), and the scattered radiation from a solution of protein becomes equal to that from the solution lacking protein when θ is only $1\text{--}2^\circ$. Fortunately, accurate measurements of scattered X-radiation can be made at the necessary small angles. The values for R_G obtained from

small-angle X-ray scattering, for example, 1.75 nm for myoglobin,²⁵ 2.3 nm for cyclic AMP-dependent protein kinase,²⁶ and 1.36 nm for reduced cytochrome *c*,²⁷ demonstrate the ability of this technique to provide information about small globular proteins.

When the observations of **X-ray scattering** are presented, a different convention is used to approximate $P(\theta)$. Because

$$\exp(-x) = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \dots \quad (12-28)$$

the first two terms in Equation 12-20 are identical to the first two terms in the expansion of $\exp[(16\pi^2 R_G^2 \sin^2(\theta/2)/3\lambda^2)]$, and at small angles^{21,28}

$$\lim_{\theta \rightarrow 0} \ln P(\theta) = - \frac{16\pi^2 R_G^2}{3\lambda^2} \sin^2 \frac{\theta}{2} \quad (12-29)$$

Because at such small angles none of the terms in Equation 12-19 except i_θ (see Equations 8-30 and 8-31) and $P(\theta)$ change as θ is varied, a plot of $\ln i_\theta$ as a function of $\sin^2(\theta/2)$ at the smallest angles will give a straight line with a slope of $-16\pi^2 R_G^2/(3\lambda^2)$.^{*} From this slope R_G is readily determined.

In reports of studies of X-ray scattering, there are several ways in which the observations are analyzed. First, the data can be presented directly (Figure 12-2A)²⁶ as the natural logarithm of the observed incremental scattering intensity ($\ln i_\theta$) as a function of q , where

$$q \equiv \frac{4\pi \sin(\theta/2)}{\lambda} \quad (12-30)$$

The advantage of this presentation is that the scattering calculated from a particular model of the molecule of protein as a function of q can be compared to the complete set of scattering data. Second, the data can be presented in a **Guinier plot** as $\ln i_\theta$ as a function of q^2 (Figure

12-2B). The advantage of this presentation is that a radius of gyration can be estimated from the limiting slope at the smallest values of scattering angle θ , typically those less than 1° (Figure 12-2A). Third, the **distance distribution function** $p(r)$, which is the Fourier transform of the scattering function

$$p(r) = \frac{1}{2\pi} \int i_\theta q r \sin(qr) dq \quad (12-31)$$

can be calculated. The distance distribution function $p(r)$ is the frequency with which vectors of a length r connect two volume elements within the molecule of protein (Figure 12-2C). In practice, the inverse Fourier transform of Equation 12-31

$$i_\theta = 4\pi \int_0^{d_{\max}} p(r) \frac{\sin(qr)}{qr} dr \quad (12-32)$$

where d_{\max} is the maximum dimension of the particle, is used to compute $p(r)$ in reverse.²⁹

In a plot of the distance distribution function $p(r)$ against r , the **longest dimension of the molecule** of protein is the intercept of the function with the abscissa. For example, the longest dimension of a molecule of cyclic AMP-dependent protein kinase is 7.2 nm (Figure 12-2C). From scattering curves of myoglobin in its native state (Figure 4-18), after the removal of its heme, and then in solution at pH 2, the gradual expansion of the protein as its structure was disrupted could be followed by monitoring the gradual increase in the value of this intercept.²⁵

The shape of the distance distribution function provides information about the shape of the molecule of protein.²⁹ If the molecule of protein is globular, the distance distribution function $p(r)$ has a fairly symmetric shape with a single maximum (Figure 12-2C). If the molecule of protein has an elongated structure, the distance distribution function will be skewed. If it is elongated in only one dimension so that it is prolate in shape, the maximum will be shifted to short distances because there are more short vectors in a prolate solid than there are long vectors.²⁹ There is a slight indication of such an elongation in Figure 12-2C. If the molecule of protein contains two well-separated globular domains, there will be two maxima in the distance distribution function, the one at shorter distances for vectors confined within each domain and the one at longer distances for vectors between domains.

A distinction can be made between small-angle scattering and solution scattering. Measurements of **small-angle scattering** are confined to the region of the scattering function for angles that are only large enough to define accurately the distance distribution function $p(r)$. This range of small angles also includes scattering at the smallest angles, which provides an estimate of the

* Unfortunately, investigators who study the scattering of X-radiation and neutrons use a different convention for the scattering angle (Figure 8-4) from those who study the scattering of light. The same scattering angle routinely designated as θ during measurements of light scattering is routinely designated as 2θ during measurements of X-ray scattering and neutron scattering. Consequently, the angles θ from measurements of X-ray scattering and neutron scattering must be multiplied by a factor of 2 before they are used as angles θ in the equations presented in this text. The term $\sin^2(\theta/2)$ used when results of light scattering are presented thus becomes $\sin^2 \theta$ when results of small-angle X-ray scattering and neutron scattering are presented, and values of $\sin^2 \theta$ from X-ray scattering and neutron scattering are equivalent to values of $\sin^2(\theta/2)$ in the equations used in this text. In measurements of X-ray scattering and neutron scattering, the slope of the line when $\ln i_\theta$ is plotted as a function of the $\sin^2 \theta$ used by these investigators has the value $-16\pi^2 R_G^2/3\lambda^2$.

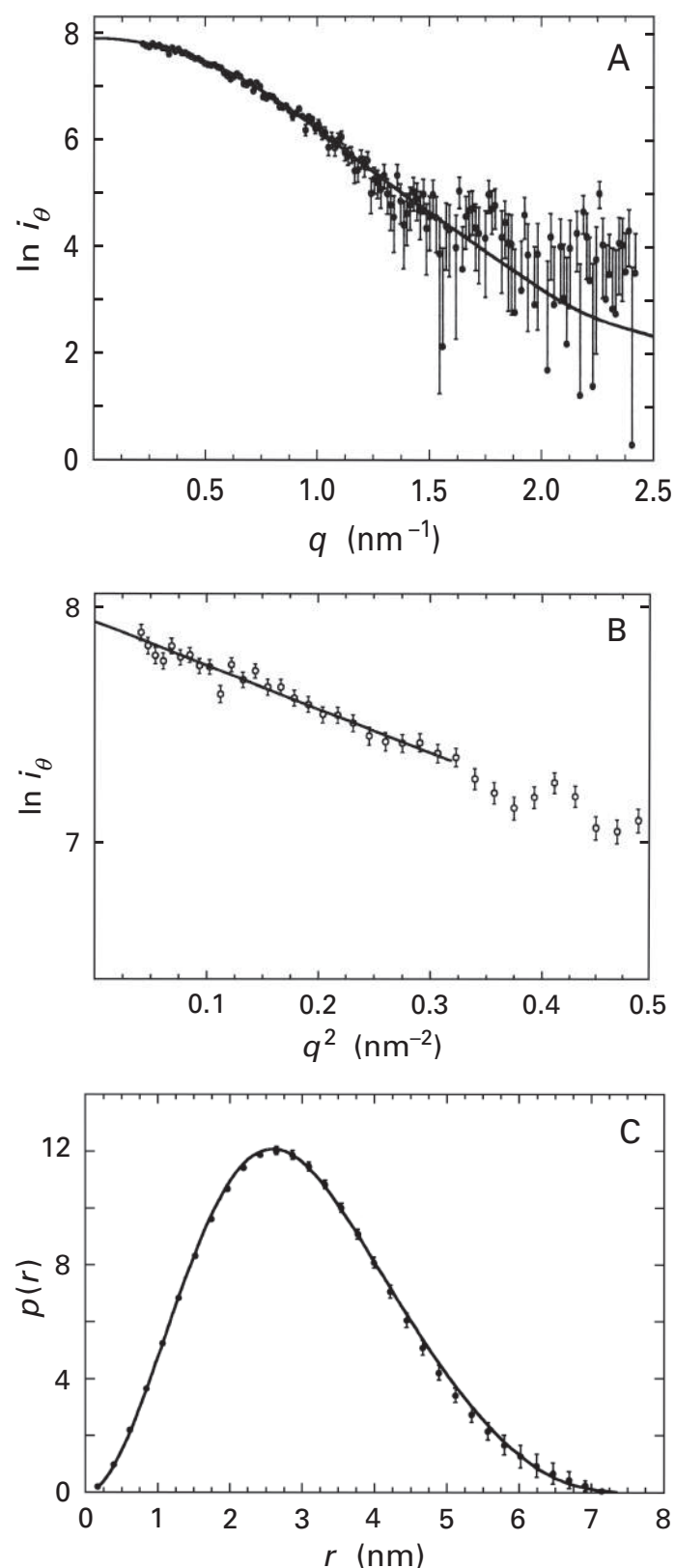


Figure 12-2: Scattering of X-radiation of wavelength 0.154 nm by a solution of cyclic AMP-dependent protein kinase at a concentration of 66 μM .²⁶ (A) Solution scattering curve. The natural logarithm of the incremental scattering intensity ($\ln i_\theta$) of the entire set of data, which includes intensities to 1% of the maximum scattering, is plotted as a function of q (Equations 8–30, 8–31, 12–19, 12–29 and 12–30). The set of data includes scattering angles θ to 3.5° , the largest angle at which incremental scattering intensity could be measured. (B) Small-angle X-ray scattering curve. In a Guinier plot, the natural logarithm of the incremental scattering intensity is plotted as a function of q^2 at scattering angles less than 1° ($q < 0.7$) for all of the data from the same set as in panel A. The limiting slope (the line drawn in the figure) provides the radius of gyration (2.31 nm) of the unliganded cyclic AMP-dependent protein kinase. (C) Distance distribution function. The Fourier transform (Equation 12–31) of the scattering profile in panel A is the distance distribution function $p(r)$, the frequency with which two volume elements within the actual structure of cyclic AMP-dependent protein kinase in solution are a distance r from each other. It provides an estimate of the longest dimension (7.2 nm) of the molecule of protein. The crystallographic molecular model of cyclic AMP-dependent protein kinase has a peptide bound in the active site. A molecular model of the empty protein was constructed in which the two structural domains enclosing the peptide were opened by 39° relative to their orientation in the crystallographic molecular model. Each carbon, oxygen, and nitrogen in this hypothetical model was converted into an equivalent sphere of scattering density. The interference expected from this arrangement of spheres as a function of scattering angle is the curve drawn through the complete set of data in panel A, and the distance distribution function $p(r)$ calculated from that theoretical scattering curve in panel A is the curve drawn through the data in panel C. Reprinted with permission from ref 26. Copyright 1993 American Chemical Society.

radius of gyration (Figure 12-2B; Equations 12–19 and 12–29) and an estimate of the forward scattering i_0 and hence the molar mass of the protein (Equation 12–19). **Solution scattering** includes the data at larger scattering angles, where more features are observed (Figure 12-2A), such as subsidiary peaks in the scattering function. In the region of small-angle scattering, the connection between measurements of scattering and hydrodynamic measurements is most apparent. In the region of solution scattering, information about the internal structure of the protein is revealed.

The **complete solution scattering curves** for different proteins are distinct (Figure 12-3),³⁰ and these distinctions indicate that each solution scattering curve contains information about the structure of the molecule of protein beyond just its radius of gyration and distance distribution function. There are several methods for extracting this information.³¹ Small uniform spheres can be arranged to produce a hypothetical structure thought to represent the structure of the molecule of protein, and a theoretical solution scattering curve for this arrangement can be calculated and compared to the observed solution scattering curve.³² The spheres can be the individual atoms in a crystallographic molecular model,^{33,34} and it is possible to calculate the **theoretical solution scattering curve** that a particular crystallographic molecular model should produce.²⁹ In order to duplicate the observed solution scattering curve with the theoretic-

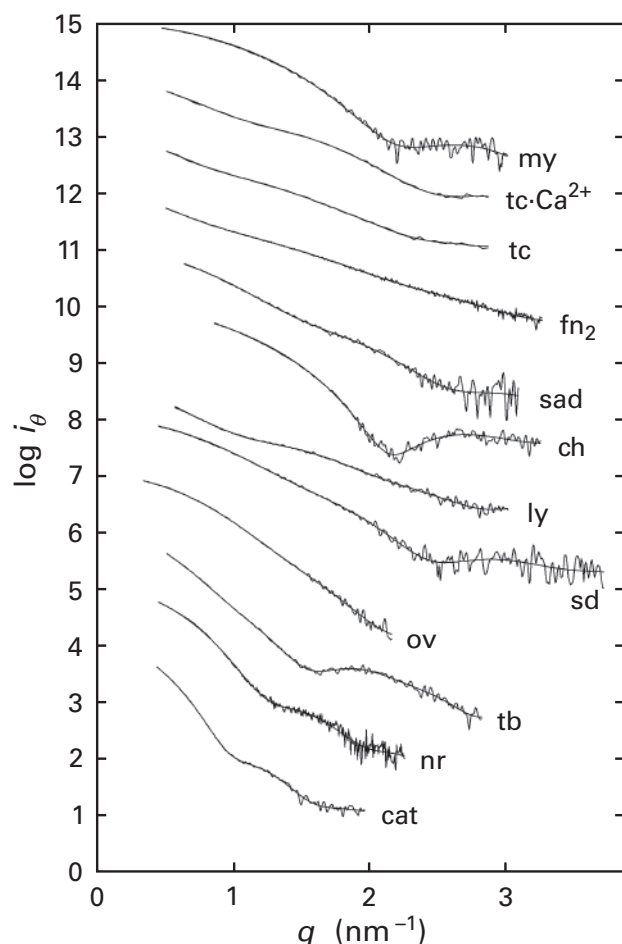


Figure 12-3: Solution scattering curves for, in descending order, myoglobin (my), troponin C in the presence of Ca^{2+} (tc- Ca^{2+}), troponin C in the absence of Ca^{2+} (tc), a tandem pair of fibronectin type 3 domains from $\beta 4$ integrin (fn_2), spermadhesin PSP-I/PSP-II (sad), chymotrypsinogen A (ch), the domain C-lytA from pneumococcal autolysin (ly), superoxide dismutase (sd), ovalbumin (ov), tubulin (tb), nitrite reductase (NO-forming) (nr), and catalase (cat).³⁰ The observed curves of scattering density were normalized by dividing each along its length by the intensity of the forward scattering i_0 , and the logarithms of these normalized profiles are presented, displaced by one logarithmic unit from each other. For example, $\log i_0$ for the scattering curve of myoglobin was arbitrarily designated as 15; that for ovalbumin, 7; and that for catalase, 4. The data are the irregular curves. The smooth curves drawn through the data are theoretical scattering curves of interference as a function of q calculated from respective arrangements of sets of spheres of 0.3 nm radius, each arrangement with the same total volume as the respective molecule of protein. The spheres were systematically rearranged until their arrangement reproduced the experimental curve as closely as possible. The final arrangement of spheres in each case ended up resembling in its shape the crystallographic molecular model of the respective protein. Reprinted with permission from ref 30. Copyright 2000 Elsevier B.V.

cal scattering curve, however, it is necessary to include a layer of hydration around the molecule of protein that is about 0.3 nm thick.³⁵ The density of the water in this layer has a density 1.05–1.20 times that of the bulk water.²⁹ It is also possible to represent the domains of a molecule of protein as ellipsoids of the appropriate

dimensions in a particular arrangement and calculate the scattering curve of this representation (Figure 12-2A, solid line).²⁶

A rough estimate of the shape of a molecule of protein of unknown structure can be derived from the complete scattering curve alone with no preconceptions. An envelope can be generated by spherical harmonics. By systematically adjusting the parameters of the spherical harmonics, the scattering curve calculated from the envelope can be made to fit to the experimental scattering curve of the protein.³⁶ The resulting envelope will have two to four ellipsoidal protrusions of size, shape, and orientation determined by the fit to the data. It is also possible to rearrange systematically a set of uniform spherical beads until the structure they form produces a scattering curve matching the one observed (solid lines in Figure 12-3).³⁰

The ideal wavelength for studying the shape of most proteins by scattering would be somewhere between 1 and 10 nm. Unfortunately, light of these wavelengths cannot be readily generated. Neutrons produced in nuclear reactors, however, have velocities high enough that a beam with a wavelength of around 1 nm can be produced.^{37,38} This wavelength permits measurements of neutron scattering of solutions of proteins to be made to an angle θ of 20–40° before interference becomes too large. Measurements of **neutron scattering** from a solution of protein as a function of scattering angle are traditionally treated just as are those from X-ray scattering (Figures 12-2A and 12-3) even though observations can be made to much greater angles. In the range of small angles, Guinier plots provide radii of gyration and distance distribution functions, and complete solution scattering curves are fit with molecular models of the protein.^{37,39,40}

Neutrons are scattered by atomic nuclei and each isotope of each element scatters neutrons with a characteristic efficiency, quantified in its **scattering length**. The most dramatic difference in scattering length is that between the nucleus of hydrogen, a proton, and the nucleus of deuterium, a deuteron. Their neutron scattering lengths are –3.74 fm and 6.67 fm, respectively. The negative sign for that of a proton indicates that it scatters neutrons 180° out of phase to those scattered by a deuteron, so that the contrast produced by interference between a neutron scattered from a proton and that scattered from a deuteron is dramatic.

This large **difference in scattering length** between proton and deuteron has been used to map the distances between the proteins in the 30S ribosomal subunit from *E. coli* (Figure 11-5) by neutron scattering.⁴¹ The 21 deuterated proteins found in a 30S subunit were produced by growing bacteria on $[\text{D}_2]\text{H}_2\text{O}$. Pairs of deuterated proteins were reassembled into the same 30S subunit with the RNA and the other proteins all undeuterated. The incremental scattering of neutrons due to just the interference between each pair of deuter-

ated proteins was measured by subtracting the scattering from a mixture of the two respective types of 30S subunits that each contained only one of the two deuterated proteins.⁴² The Fourier transform (Figure 12-2C) of this incremental neutron scattering function provides the frequency with which vectors of length r connect a volume element in one deuterated protein with a volume element in the other. The maximum in such a curve is an estimate of the distance between the centers of mass of the two proteins in the 30S subunit. Enough of these distances (92 out of a possible 210) were measured to establish the relative positions of all 21 of the proteins in the 30S subunit.⁴¹ The majority of these relative positions agreed with their relative positions in the crystallographic molecular model of the 30S subunit.⁴³

Because the molecules of protein in solution are randomly oriented, solution scattering curves for X-radiation or neutrons are rotationally averaged and by themselves can provide only **spherically symmetric structural information**. Because molecules of protein are not spherically symmetric, some information about the structure of the protein must be available to constrain the model. For example, scattering of X-radiation or neutrons can be used together with crystallographic information. Crystallographic molecular models of the domains of a protein may be available but not a crystallographic molecular model of the complete protein, and scattering curves can provide models for the dispositions of the domains in the intact protein.⁴⁴ Intact immunoglobulin M is a pentameric ring of five subunits each composed of two Fab portions and one Fc portion formed from folded polypeptides related to those forming immunoglobulin G (Figure 11-1). Ten copies of a crystallographic molecular model of an Fab fragment of immunoglobulin G and five copies of a crystallographic molecular model of an Fc fragment of immunoglobulin G could be arranged to produce a structure with a calculated solution scattering curve in agreement with that observed for immunoglobulin M, and the appropriate combinations of Fab and Fc fragments could be arranged to produce structures with calculated solution scattering curves in agreement with those observed for the respective fragments.³²

Measurements of X-ray or neutron scattering from a solution of a protein for which a complete crystallographic molecular model is available have also proven to be valuable complements to the crystallographic observation. The usual result is that the theoretical solution scattering curve calculated from the crystallographic molecular model agrees closely with the one that is observed.^{29,45} Such coincidences are further evidence that the crystallographic molecular model represents the structure of the protein when it is in solution.

There are, however, a number of instances in which measurements of scattering have been used to **adjust crystallographic molecular models**. Two independently

shifting domains in a crystallographic molecular model of a protein are often confined to a particular orientation either by the packing forces of the crystal or by the binding of a ligand. A measurement of the radius of gyration from small-angle X-ray scattering of the protein in solution unconfined by the packing forces or in the absence of that ligand can be matched with a value calculated from a molecular model of the protein in which the domains have a different orientation from those observed crystallographically.⁴⁶ From a crystallographic molecular model of the protein²⁷ or a systematically altered conformation of that molecular model, the frequency with which vectors of length r actually do connect volume elements within the model can be calculated and compared with the observed distance distribution function. In Figure 12-2C, the line through the points was calculated from an altered conformation of the crystallographic molecular model of cyclic AMP-dependent protein kinase. This altered conformation was proposed to represent the structure that the molecule assumes in solution in the absence of the peptide that was bound to the protein when it was crystallized.

The disagreement between an observed solution scattering curve and that calculated from a crystallographic molecular model is also an indication that the protein assumes a conformation different from that of the crystallographic molecular model when it is dissolved in a solution of a particular composition,³⁴ for example, in which ligands for the protein may be dissolved.³³ Theoretical solution scattering curves calculated from likely alternative conformations often indicate how the structure of the protein has changed. In the case of aspartate carbamoyltransferase, it had been possible to crystallize the protein in the conformation it assumes when its regulatory β subunits bind MgATP. Even so, the crystallographic molecular model of this conformation⁴⁷ had to be adjusted before it would produce a theoretical X-ray solution scattering curve that agreed with the one observed for it in solution.⁴⁸ Only when the distance between the two catalytic α_3 trimers in the crystallographic molecular model (Figure 9-37) was increased 0.3 nm by reasonable rotations of the regulatory β_2 dimers did the theoretical curve agree with the observed curve.

The examples of measurements of small-angle X-ray scattering for cyclic-AMP dependent protein kinase (Figure 12-2C) and of X-ray solution scattering for aspartate carbamoyltransferase illustrate the use of scattering in comparisons of the **structure of a protein in solution** to its structure in a crystallographic molecular model. In both of these instances, the measurements of scattering were used to adjust appropriately and realistically the structure of the crystallographic molecular model, and there are other instances in which such adjustments have also been required.^{29,31} The fact that reasonable adjustments in the orientations of domains or the disposition of subunits are all that is necessary to

bring errant crystallographic molecular models into coincidence is further evidence that the crystallographic molecular model represents the structure of the protein. The solution scattering curves of helical polymeric proteins can also provide information about the parameters of the helix into which the monomers are assembled.³⁸

The difficulty with measurements of hydrodynamic properties or of small-angle scattering for assessing the shape of a molecule of protein is that they often provide only one unambiguous numerical result, either a frictional coefficient, a Simha factor, or a radius of gyration. If the frictional ratio $f/f_{0,h}$ is less than 1.15, the Simha factor ν is less than 4.0, or the radius of gyration R_G is near a value of $(3/5)^{1/2} R_{0,h}$, it can be concluded that the protein is globular. If the value of one or more of these parameters for a given protein is significantly greater than the values expected for a sphere, it is usually necessary to conclude that the protein has a highly irregular surface, has an extended structure, has a high degree of hydration, or has some combination of these features. It is clear from the foregoing discussion of some of the results that larger values of these parameters are consistent with a large set of particular arrangements of the available mass and values for hydration. The only reason so much is heard about prolate and oblate ellipsoids of revolution is not that molecules of proteins are such geometric solids but that frictional coefficients and radii of gyration can be calculated explicitly for such solids. In using any of these measured parameters in an informative way, other details about the structure of the protein are essential.

One way to observe the shape of a molecule of protein directly is by **electron microscopy**. The three symmetrically protruding regulatory subunits on aspartate carbamoyltransferase and the hollow, water-filled cavity between its two rotationally symmetric, trimeric α_3 subunits (Figure 9–37), which together probably account for its abnormally large frictional coefficient, were first observed in electron micrographs of the protein (Figure 12–4A)^{49,50} before there was a crystallographic molecular model. Another protein with an abnormally large frictional coefficient is fibrinogen. Electron micrographs of fibrinogen (Figure 12–4B),⁵¹ which turned out to be remarkably accurate representations of its structure, were published 20 years before a crystallographic molecular model of the protein (Figure 13–22) became available.⁵²

To prepare it for direct observation in the electron microscope, a protein molecule can either be **negatively stained** by being embedded in a glass of the salt of a heavy metal ion such as uranyl cation or phosphotungstate anion (Figure 12–4C, upper four images)^{53,54} or be positively stained by being **rotary shadowed**⁵⁵ with a layer of platinum that produces a metallic replica of the molecule (Figure 12–4C, lower four images). In the former method, the molecule of protein, because it is less electron dense than the glass, appears as a light image against a dark background; in the latter method, the mol-

ecule of protein coated with the metal appears dark against a light background. Whenever results from such electron microscopic studies are presented, a representative field of molecules (Figure 12–4D)⁵⁶ should be shown so that the reader can judge what fraction of the molecules of protein on the film of carbon or in the metallic replica give images that resemble the images chosen for a gallery of “representative” views (Figures 12–4A–C).

Collagen XII from *Gallus gallus* is a trimer of three polypeptides. All three polypeptides in the trimer are encoded by the same gene, but there are two forms of the polypeptide, one 3100 aa long and the other about 2700 aa long, produced by translation of alternatively spliced versions of the same messenger RNA;⁵³ the shorter translation is missing the amino-terminal 400 aa of the sequence. The carboxy-terminal 380 aa of each polypeptide contains two segments (152 and 103 aa) of collagen repeat, and in this region the three polypeptides of the trimer should form an interrupted triple-helical rope of collagen (Figure 9–33) with two segments 45 and 30 nm in length. This triple-helical rope is the structure holding the three polypeptides together in the oligomer. The amino-terminal 1870 aa of the short splice variant contains 10 fibronectin type III modular domains, strung together in a necklace that should be 32 nm in length,²² and one amino-terminal von Willebrand factor type A modular domain, a globular structure about 4 nm in diameter. The longer splice variant has an additional eight fibronectin type III modular domains (26 nm) and two additional von Willebrand factor type A domains. The electron micrographs of collagen type XII display a structure that is the fulfillment of these expectations (Figure 12–4C).^{53,54} The single, thin collagen tail of 75 nm is kinked about 30 nm from its end.⁵⁴ There is a central globular region from which three significantly thicker arms extend, each either a short or a long variant; the short variant is about 40 nm long with a globular ball at its end,⁵³ and the long variant is about 90 nm long with a globular ball in its middle and two globular balls at its end.⁵⁴ The three polypeptides enter the center wrapped around each other in the collagen tail and leave the center separately as three fibronectin necklaces.

Electron micrographs of activated bovine coagulation Factor Va were instrumental in explaining its unusual behavior upon sedimentation analysis. The protein has a molar mass⁵⁷ of 170,000 g mol⁻¹ and a standard sedimentation coefficient,⁵⁸ $s_{20,w}^0$, of 8.2 S, from which a frictional ratio $f/f_{0,h}$ of 1.6 (based on the assumption that $\delta_{H_2O} = 0.3$) can be calculated. When activated coagulation Factor Va was observed by electron microscopy, it was found to be two globular domains of protein, very similar in diameter, attached together through a narrow neck.⁵⁹ This shape seen in the electron micrographs is responsible for the unusually large frictional ratio of this protein. Although an axial ratio for a prolate ellipsoid was calculated from the earlier results of the sedimentation

analysis,⁵⁸ it was moot when the electron micrographs became available.⁵⁹

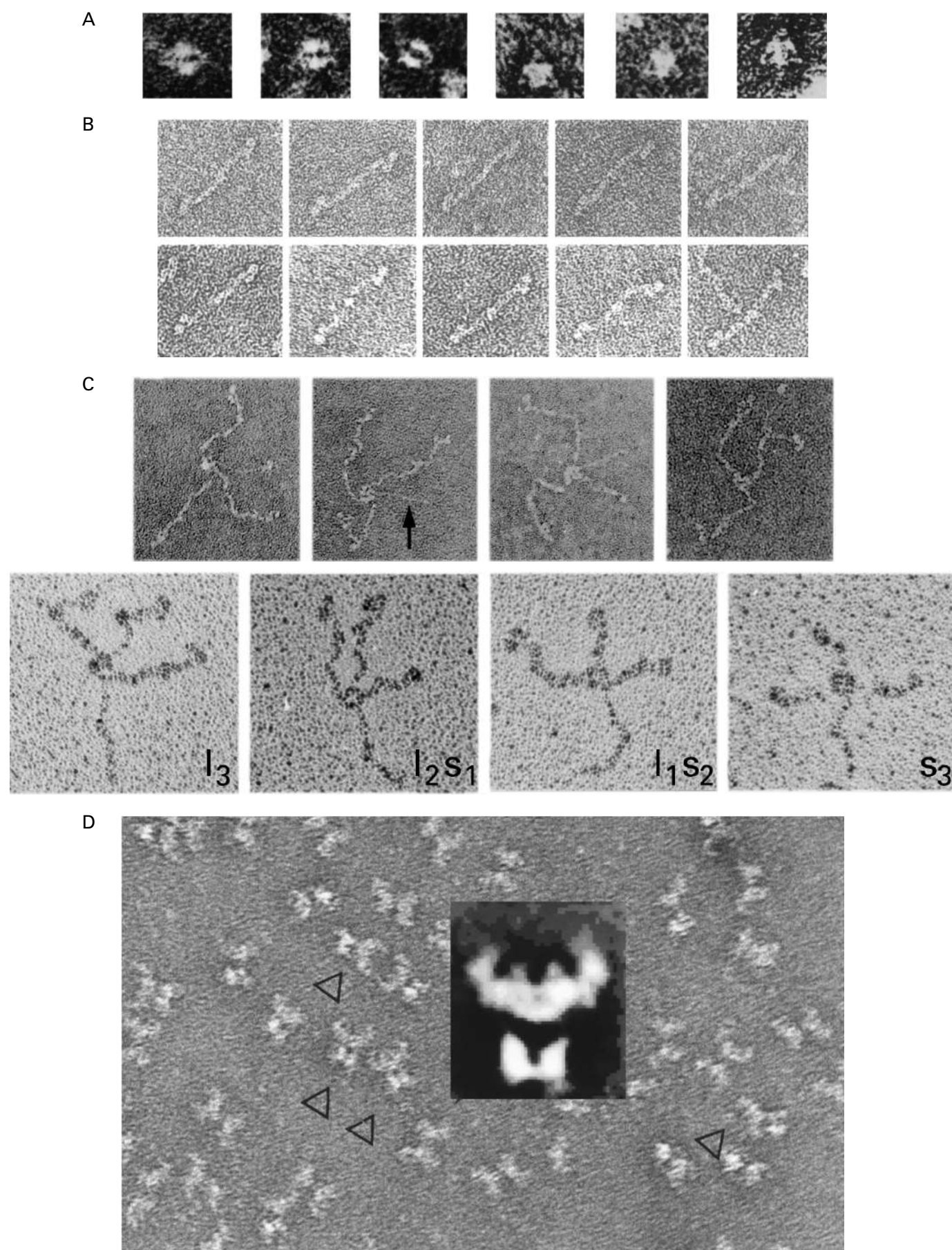
Rotary shadowing is usually used for **estimating the dimensions of extended molecules** such as collagen XII (Figure 12-4C), nidogen,⁶⁰ inversion-specific glycoprotein,⁶¹ fibulin,⁶² and myosin.⁵⁵ The molecule of protein is spread upon a flat sheet of mica before being coated with metal. Consequently, a long, thin, flexible molecule should lie almost flat upon the surface, and the contour length of its replica in the two-dimensional micrograph should be almost as long as its actual contour length in three dimensions. For this reason, rotary shadowing is thought to be the most reliable method to obtain estimates of the length of an extended molecule of protein. For example, the frictional ratio $f/f_{0,h}$ for a fragment of caldesmon (amino acids 166–450 from a polypeptide of 756 aa)⁶³ is 2.2, consistent with a cylinder of the appropriate volume that is 40 nm long. Electron micrographs of this fragment of caldesmon that had been rotary-shadowed with platinum and tungsten displayed elongated molecules of uniform thickness with contour lengths that averaged 35 nm. In rotary-shadowed images, globular domains such as the two heads of myosin or the three globular domains of nidogen appear as dark, unfeathered lumps. Although globular proteins constructed from clusters of globular domains or from globular subunits can appear as clusters of dark lumps upon rotary shadowing,⁶⁴ they usually appear as single undifferentiated and structureless lumps of platinum. Negative staining (Figure 12-4A,D) or embedding in amorphous vitreous ice is required to obtain images displaying details of the structure of a globular protein.

Phosphorylase kinase is a globular protein with a dramatically peculiar shape. Because of its unusual shape, a collection of digitized micrographic images of individual molecules (Figure 12-4D) could be superposed by a computer and stacked one upon the other.⁵⁶

The average of this stack of images could be calculated, and this average represents an enhanced image of the actual molecule (inset, Figure 12-4D). The chalice seen in the **enhanced image** can be imperfectly discerned in each of the individual selected images (Figure 12-4D), and this correspondence fulfills the usual requirement placed upon any reconstruction. A similar procedure has been applied to α_2 -macroglobulin⁶⁵ to obtain enhanced images. This protein has a shape almost as peculiar as that of phosphorylase kinase.

It is also possible to obtain a **three-dimensional reconstruction of the structure** of a macromolecule observed in an electron micrograph. An electron micrograph of a macromolecule is the two-dimensional projection either of the structure of that macromolecule if it is embedded in amorphous vitreous ice or of the mold of that macromolecule if it is embedded in negative stain. It has already been noted that the two-dimensional Fourier transform of the projection of a three-dimensional object is a central section of the three-dimensional Fourier transform of the unprojected object (Equations 9-4 through 9-6). From the complete three-dimensional Fourier transform of the object, the distribution of scattering density within the object, and hence details of its three-dimensional structure, can be calculated by Fourier transformation. To gather the complete three-dimensional Fourier transform of the object, Fourier transforms of projections of the object in a large number of different orientations must be assembled. This is accomplished in a helical polymeric protein by the fact that the helical array positions the monomer in specific, defined orientations, each of which provides a different projection. When molecules are not arranged in such an array but scattered upon the field, as are the molecules of phosphorylase kinase in Figure 12-4D, it is necessary to define the precise orientation of each of them relative to the plane of the micrograph before their individual

Figure 12-4: Asymmetric molecules of protein viewed by electron microscopy. Solutions of the protein of interest ($10 \mu\text{g mL}^{-1}$ to 1.0 mg mL^{-1}) were applied to electron microscopic grids coated with a thin film of carbon⁴⁹ supported by a net of either collodion or formvar. The layer of carbon ($\sim 5 \text{ nm}$) was ionized so that it was hydrophilic enough to accept the aqueous solution. The molecules of protein were adsorbed to this surface and were then negatively stained with either 2% phosphotungstate (panel A) or 1–2% uranyl formate (panels B and D). The water evaporates to leave a glass of the heavy metal salt in which are embedded the molecules of protein. (A) Gallery of selected images⁵⁰ of aspartate carbamoyltransferase from *E. coli* (Figure 9-37) viewed either along one of its 2-fold rotational axes of symmetry (left three images) or along its 3-fold rotational axis of symmetry (right three images) at $480000\times$. Reprinted with permission from ref 50. Copyright 1972 American Chemical Society. (B) Gallery of selected images of bovine fibrinogen⁵¹ at $480000\times$. The elongated molecule has globular domains at each end. Reprinted with permission from ref 51. Copyright 1981 Academic Press. (C) Selected images of collagen type XII from *Gallus gallus*^{53,54} at $240000\times$. The upper four images were negatively stained with uranyl formate. Reprinted with permission from ref 54. Copyright 1992 Blackwell Publishing. The lower four images are molecules of protein that were rotary shadowed.⁵⁵ A solution of the protein was sprayed into an aerosol mist and droplets of the mist were adsorbed to a sheet of mica. A beam of platinum atoms was directed at an angle of 6° onto the surface of the mica as it was rotated at $120 \text{ revolutions min}^{-1}$. The resulting thin film of platinum containing replicas of the molecules of protein was transferred to a copper grid. In the four rotary-shadowed images, selected representatives of a homotrimer of long splice variants (l_3), of a homotrimer of short splice variants (s_3), and of the two heterotrimers (l_2s , ls_2) are presented at $240000\times$. Reprinted with permission from ref 53. Copyright 1995 The Rockefeller University Press. (D) A field of negatively stained molecules of phosphorylase kinase⁵⁶ at $480000\times$. This is an accurate representation of the usual situation. Most of the molecules of protein negatively stained on the grid are featureless asymmetric structures. The minority that present a repeating, definable image (indicated by arrowheads) are selected by the microscopist as representative images of the protein and presented in galleries as in panels A and B. In this instance, the shape of the individual images of phosphorylase kinase was so peculiar that digitized optical densities of a large number of the selected images of individual molecules (62) could be sequentially superposed by a computer to obtain an enhanced image (inset). Reprinted with permission from ref 56. Copyright 1985 Academic Press.



Fourier transforms can be summed together to obtain the complete three-dimensional Fourier transform of the average molecule.

When the objects scattered over the field of the electron micrograph are viruses, the icosahedral symmetry of each of the viral coat proteins permits the exact orientation of each individual virion to be estimated.^{66–70}

The **assignment of an orientation** to each virion permits the Fourier transforms of their projections to be added together to obtain a complete three-dimensional Fourier transform of the average viral particle. In this way, a three-dimensional reconstruction of the structure of the viral coat⁷¹ and other appendages of the virus that are distributed with icosahedral symmetry^{72,73} can be calculated. For such reconstructions, the viral particles are usually suspended in a layer of amorphous vitreous ice.

If it is possible to define somehow the orientation of each member of a population of asymmetric molecules spread upon a grid at random, the same type of summation can be performed. If the molecule has a tendency to lie upon the carbon surface in a preferred orientation, for example, the molecules of phosphorylase kinase that settle on the grid to present a projection in the shape of a chalice (Figure 12–4D), this tendency orients them in one dimension but fortunately only in one dimension. The direction in which each individual oriented molecule faces upon the surface is random, and the direction in which each faces can be defined by direct observation. When the grid is then tilted 50°, a large collection of molecules, each in a completely different three-dimensional orientation but each in a known three-dimensional orientation relative to the others, is created.⁷⁴ From a summation of their individual Fourier transforms in the **tilted image**, a three-dimensional Fourier transform of the structure of the average molecule can be gathered. From this Fourier transform, a molecular model can be calculated. Such reconstructions have been performed for human α_2 -macroglobulin⁷⁵ and the 50S subunit of the ribosome.^{76,77} The resulting molecular model of the 50S subunit of the ribosome, albeit at low resolution, resembled quite closely the crystallographic molecular model that became available 12 years later.⁷⁸

Suggested Reading

Perkins, S.J., Nealis, A.S., Sutton, B.J., & Feinstein, A. (1991) Solution structure of human and mouse immunoglobulin M by synchrotron X-ray scattering and molecular graphics modelling. A possible mechanism for complement activation, *J. Mol. Biol.* 221, 1345–1366.

Mani, R.S., Karimi-Busheri, F., Cass, C.E., & Weinfeld, M. (2001) Physical properties of human polynucleotide kinase: hydrodynamic and spectroscopic studies, *Biochemistry* 40, 12967–12973.

Problem 12–1: Calculate the values of f_{sed} , f_{diff} , and $f_{0,\text{unh}}$ from \bar{v} , $s_{20,w}^0$, $D_{20,w}^0$, and M_{prot} for each protein in Table 12–1. From tabulated values of $f_{\text{av}}/f_{0,h}$ determine a/b for

each protein on the assumption that they are prolate ellipsoids of revolution.

Problem 12–2: Human immunoglobulin G is a protein with a molar mass of 167,000 g mol^{−1} and a partial specific volume of 0.739 cm³ g^{−1}.

- Assume hydration to be 0.3 g of H₂O (g of protein)^{−1} and calculate the minimum frictional coefficient, $f_{0,h}$, for the hydrated hydrodynamic particle at 20 °C in water.
- The standard sedimentation coefficient $s_{20,w}^0$ for immunoglobulin G is 7.0×10^{-13} s, and the standard diffusion coefficient $D_{20,w}^0$ is 4.0×10^{-7} s^{−1}. Calculate the frictional coefficient, first from the standard sedimentation coefficient and then from the standard diffusion coefficient.
- From the average of these two estimates of the frictional coefficient and from the estimate of the minimum frictional coefficient of the hydrated hydrodynamic particle, estimate the axial ratio a/b upon the assumption that the molecule is a prolate ellipsoid of revolution.
- The shape of an immunoglobulin G is displayed in Figure 7–13. How does this structure compare with your estimate of its shape?

Problem 12–3: Thiosulfate sulfurtransferase is a monomeric enzyme. The polypeptide from bovine liver is 296 amino acids in length and has a molar mass of 33,160 g mol^{−1}. In water at 20 °C the standard sedimentation coefficient of the native protein is 3.00×10^{-13} s, and its standard diffusion coefficient is 7.50×10^{-7} cm² s^{−1}. The partial specific volume of the protein is 0.742 cm³ g^{−1}.

- Calculate the frictional coefficient of the protein.
- Calculate the frictional ratio for the hydrodynamic particle $f/f_{0,h}$, with the assumption that the hydration of the protein is 0.3 g of H₂O (g of protein)^{−1}.
- What would be the axial ratio of an ellipsoid of revolution with this frictional ratio?
- How does this estimation compare to the crystallographic molecular model of the protein (Figure 9–18)?

Problem 12–4: The standard sedimentation coefficient of human fibrinogen is 7.63×10^{-13} s, its molar mass is 344,000 g mol^{−1}, and its partial specific volume is 0.725 cm³ g^{−1}.

- Assume $\delta_{\text{H}_2\text{O}} = 0.3$ and determine the volume of the hydrodynamic particle, the frictional coefficient of fibrinogen, its frictional ratio, and its axial ratio and dimensions on the basis of the assump-

- tion that it is a prolate ellipsoid of revolution or a cylindrical rod.
- (B) The length of the fibrinogen molecule has been determined to be 45 nm by electron microscopy (Figure 12–4B) and 45 nm by direct measurement of its crystallographic molecular model (Figure 13–22A). Calculate the dimensions of a prolate ellipsoid or a cylindrical rod with the same volume as the hydrodynamic particle and a major axis of length 22.4 nm.
- (C) The intrinsic viscosity of fibrinogen⁷⁹ is 27 cm³ g^{−1}. Calculate its Simha factor ν and estimate its axial ratio and molecular dimensions on the basis of the assumption that it is a prolate ellipsoid of revolution or a cylindrical rod.

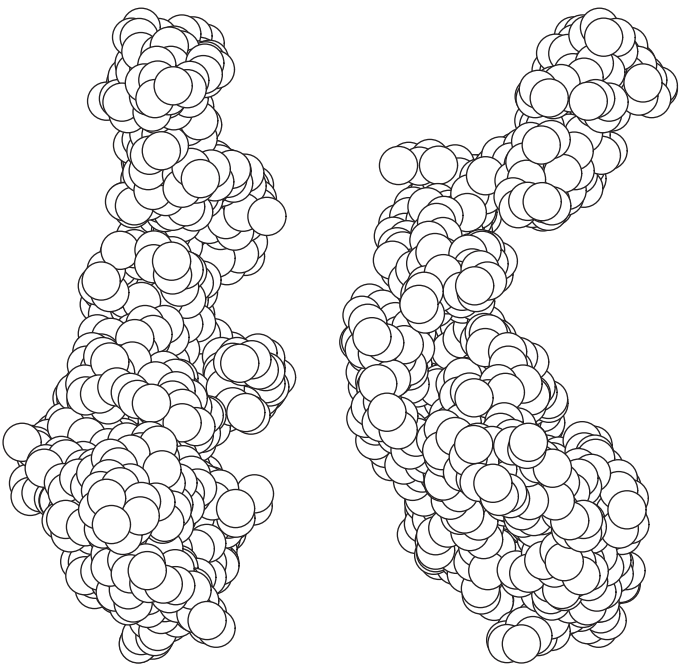
Problem 12–5: The heads of myosin (Figure 13–30A) can be detached from the intact protein by mild treatment with the endopeptidase papain. The detachable domain that was one of the heads and that is produced by the digestion with papain is referred to as the S1 fragment. It is a complex of three folded polypeptides. The S1 fragment from chicken muscle contains the amino-terminal 845 aa of the α polypeptide, or heavy chain, of myosin and two shorter polypeptides, or light chains, of lengths 149 and 163 aa.

- (A) Estimate the molar mass of the S1 fragment from the lengths of its three constituent polypeptides.

The following table lists physical properties of the S1 fragment.⁸⁰

property	value
$s_{20,w}^0$	5.8×10^{-13} s
$D_{20,w}^0$	4.6×10^{-7} cm ² s ^{−1}
\bar{v}	0.73 cm ³ g ^{−1}
$[\eta]$	6.4 cm ³ g ^{−1}

- (B) Calculate the frictional coefficient of the S1 fragment.
- (C) Assume a hydration of 0.3 g g^{−1}, and calculate the frictional ratio for the S1 fragment.
- (D) Assume a hydration of 0.3 g g^{−1}, and calculate the Simha factor for the S1 fragment.
- (E) Estimate the axial ratio that the S1 fragment would have if it were a prolate ellipsoid of revolution.
- (F) The following are two orthogonal views of a space-filling representation of the crystallographic molecular model of the S1 fragment of myosin.⁸¹ Estimate its axial ratio from the figures.



- (G) What value would δ_{H_2O} have to have for both the intrinsic viscosity and the frictional ratio to give the axial ratio you measured from the crystallographic molecular model?

Problem 12–6: Tropomyosin is a protein formed from identical folded polypeptides each 284 amino acids in length ($M_{\text{prot}} = 32,680$ g mol^{−1}). It has a partial specific volume of 0.72 cm³ g^{−1}. The molar mass of tropomyosin was determined by osmotic pressure. Each measurement was extrapolated to $\gamma_{\text{prot}} = 0$. The osmotic pressure at 0 °C as a function of ionic strength is presented in the following table.⁸²

ionic strength (M)	$\lim_{\gamma_{\text{prot}} \rightarrow 0} \frac{\Pi}{\gamma_{\text{prot}}}$ (mmHg cm ³ g ^{−1})
0.10	126
0.20	154
0.27	193
0.30	236
0.60	254
1.10	264

- (A) What are the values for the apparent molar masses of tropomyosin at the several ionic strengths? Show in detail the calculation of molar mass at ionic strength of 0.10 M.
- (B) How many polypeptides compose the major form of tropomyosin present in solution at high ionic strength? What is the exact molar mass of just this major form? Why does the number-average molar mass increase as the ionic strength is lowered?

- (C) The following data⁸² were gathered from solutions of tropomyosin at an ionic strength of 1.1 M:

γ_{prot} [g (100 mL) ⁻¹]	η'/η
1.210	0.33
1.299	0.44
1.466	0.64
1.588	0.76
1.793	0.96
2.223	1.29
2.972	1.72
4.603	2.14

where η' is the viscosity of the solution of protein, η is the viscosity of the solvent, and γ_{prot} is the concentration of protein. Determine the intrinsic viscosity $[\eta]$ at this ionic strength.

- (D) Assume that $\delta_{\text{H}_2\text{O}} = 0.3 \text{ g of H}_2\text{O (g of protein)}^{-1}$ and calculate the Simha factor ν . From ν determine the axial ratio for tropomyosin if it were a prolate ellipsoid of revolution by using Figure 12-1B,D.
- (E) From spectroscopic measurements it is known that, at all ionic strengths, tropomyosin is >90% α -helical. It is a coiled coil in which the two α helices wrap around each other as the strands in a two-stranded rope. The length of an α helix for each of its amino acids is 0.15 nm. Calculate the length of a molecule of tropomyosin at high ionic strength and, assuming it to be a cylindrical rod, calculate its diameter from its hydrated molecular volume. What is its actual axial ratio? Compare this to the axial ratio obtained from ν .
- (F) The intrinsic viscosities of solutions of tropomyosin also vary with ionic strength:⁸²

$[\eta]$	ionic strength (M)
1.00	1.1
1.03	0.6
1.23	0.3
1.75	0.2
2.45	0.1

Plot molar mass against specific viscosity and explain the correlation in terms of structures that could form as the ionic strength is lowered.

Problem 12-7: The following are a set of data for the light scattering of a series of solutions (0.2–0.8 g L⁻¹) of myosin.⁸³ The wavelength of light used for the observations was 436 nm (in a vacuum). The solutions were examined at a temperature of 20 °C, and the refractive

index of the solvent in which the myosin was dissolved was 1.34. The refractive index increment $(\partial\tilde{n}/\partial\gamma_{\text{prot}})_{P,\mu}$ for myosin is 0.208 cm³ g⁻¹, and the molar mass of myosin is 527,000 g mol⁻¹.

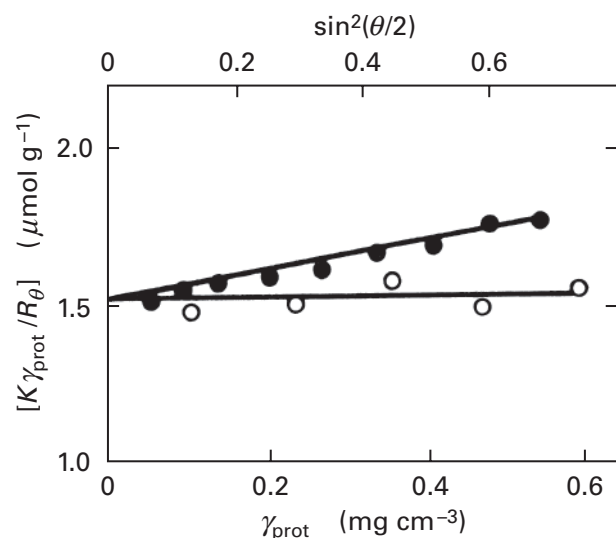
θ (deg)	$\lim_{\gamma_{\text{prot}} \rightarrow 0} \frac{\gamma_{\text{prot}}}{R_\theta}$ (g cm ⁻²)
30	0.74
35	0.74
40	0.78
45	0.80
50	0.83
55	0.86
60	0.89
70	0.95
90	1.07
110	1.17
140	1.24

- (A) What was the wavelength of the light in the solutions?
- (B) What is the radius of gyration of the myosin under these circumstances?
- (C) What would be the length of a rod with this radius of gyration?
- (D) What is the length of the molecules of myosin in Figure 13-30A? The magnification stated in the legend for Figure 13-30A is for the figure in the text.

Problem 12-8: Triskelion is a protein that assembles into spherical structures known as clathrin coats. These clathrin coats are the structures that surround the coated vesicles formed from the invagination of the plasma membrane of an animal cell at sites known as coated pits. Bovine triskelion is formed from three identical heavy polypeptides ($n_{\text{aa}} = 1675$ polypeptide⁻¹, $M_{\text{prot}} = 191,590 \text{ g mol}^{-1}$) and three identical light polypeptides ($n_{\text{aa}} = 228$ polypeptide⁻¹, $M_{\text{prot}} = 25,080 \text{ g mol}^{-1}$). Its partial specific volume, calculated from its amino acid composition, is 0.744 cm³ g⁻¹.

- (A) Calculate the unhydrated volume of triskelion.
- (B) What would be the unhydrated radius ($R_{0,\text{unh}}$) and the unhydrated radius of gyration ($R_{\text{G,sph}}$) of triskelion if it were a sphere?

The light scattering of triskelion dissolved in a buffered solution was monitored either as a function of the concentration of protein (γ_{prot}) or as a function of the scattering angle θ .⁸⁴ Reprinted with permission from ref 84. Copyright 1991 American Chemical Society.



Open and closed circles show the plots of $[K\gamma_{\text{prot}}/R_\theta]_{\theta \rightarrow 0}$ against the concentration of clathrin and $[K\gamma_{\text{prot}}/R_\theta]_{\gamma_{\text{prot}} \rightarrow 0}$ against $\sin^2(\theta/2)$, respectively. The units on the vertical axis are moles gram^{-1} . The authors of this study have used an optical constant K that incorporates the increment of the refractive index so that its value is $2\pi^2\bar{n}_0^2(\partial\bar{n}/\partial\gamma_{\text{prot}})^2/N_A\lambda_0^4$. The refractive indices (\bar{n}) of the two solutions were both 1.34. The laser used in the experiment emitted polarized light of wavelength 632.8 nm in the vacuum.

- (C) How well does the molar mass of the protein observed in the light scattering experiment agree with that calculated from the sequences of the constituent polypeptides of triskelion?
- (D) From the slope of the appropriate line in the figure, calculate the radius of gyration for triskelion.

(One way to approach this problem is to multiply both sides of Equation 12-23 by the optical constant K .)

- (E) Is triskelion a globular protein?

Problem 12-9: The definition of Rayleigh's ratio for the scattering of unpolarized light is

$$R_\theta \equiv \frac{r^2 i_\theta}{I_0 (1 + \cos^2 \theta)}$$

where i_θ is the intensity of the scattered light due only to the protein at an angle θ to the incident beam. At very low angles ($\theta \leq 5 \times 10^{-2}$ rad), $\cos^2 \theta \approx 1.00$. In this situation, by Equation 12-19

$$P(\theta) = \lim_{\gamma_{\text{prot}} \rightarrow 0} \frac{i_\theta}{\gamma_{\text{prot}}} \left[\frac{r^2}{2I_0 K M_{\text{prot}}} \left(\frac{\partial \bar{n}}{\partial \gamma_{\text{prot}}} \right)_{T, P, \mu}^{-2} \right]$$

and

$$\ln P(\theta) = \lim_{\gamma_{\text{prot}} \rightarrow 0} \left(\ln \frac{i_\theta}{\gamma_{\text{prot}}} - A \right)$$

where A is a constant determined by the values of all of the fixed parameters in the brackets. At very low angles, the approximation of Equation 12-29 is also valid, and it follows that

$$\lim_{\gamma_{\text{prot}} \rightarrow 0} \ln \frac{i_\theta}{\gamma_{\text{prot}}} = A - \frac{16\pi^2 R_G^2}{3\lambda^2} \sin^2 \frac{\theta}{2}$$

If ω is expressed in radians

$$\sin \omega = \omega - \frac{\omega^3}{3!} + \frac{\omega^5}{5!} - \dots$$

- (A) Show that

$$\lim_{\substack{\theta \rightarrow 0 \\ \gamma_{\text{prot}} \rightarrow 0}} \ln \frac{i_\theta}{\gamma_{\text{prot}}} = A - \frac{16\pi^2 R_G^2}{3\lambda^2} \left(\frac{\theta}{2} \right)^2$$

It is more convenient to take a series of measurements at varying scattering angles, θ (in radians), of a single solution of protein than to measure the scattering at a fixed angle for several solutions of protein. Therefore, what is usually done is to determine the slope of $\ln(i_\theta/\gamma_{\text{prot}})$ as a function of θ^2 at various fixed values of γ_{prot} and then extrapolate to $\gamma_{\text{prot}} = 0$. If, however, γ_{prot} is held constant, as is done in such an experiment, then

$$\lim_{\theta \rightarrow 0} \ln \frac{i_\theta}{\gamma_{\text{prot}}} = \lim_{\theta \rightarrow 0} \ln i_\theta + \ln \gamma_{\text{prot}}$$

and at constant γ_{prot}

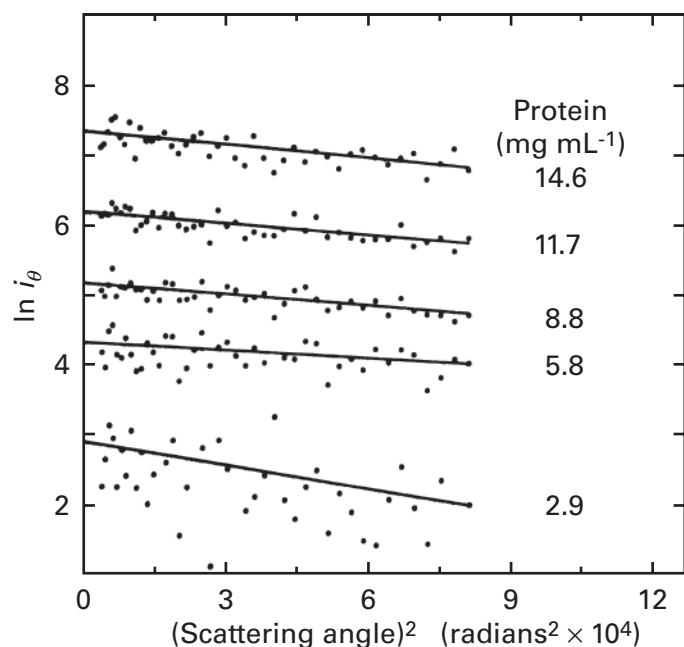
$$\lim_{\theta \rightarrow 0} \ln i_\theta = A' - \frac{16\pi^2 R_G^2}{3\lambda^2} \left(\frac{\theta}{2} \right)^2$$

where A' is a constant equal to $A - \ln \gamma_{\text{prot}}$.

- (B) What should be the slope of the line for a plot of $\ln i_\theta$ against θ^2 at low concentrations of protein and low scattering angle?

Immunity protein is a protein produced by strains of the bacterium *E. coli* producing colicin E₃. The intensities of the scattered X-rays ($\lambda_0 = 0.154$ nm; $\bar{n} = 1.33$; $\lambda = 0.116$ nm) as a function of the square of the scattering angle were measured for a series of solutions of immu-

nity protein.²³ Reprinted with permission from ref 23. Copyright 1983 *Journal of Biological Chemistry*.



The natural logarithm of the intensity of the scattered X-rays, $\ln i_\theta$, is presented as a function of the square of the scattering angle θ in $\text{radians}^2 \times 10^4$. (A value of 6.0 on the abscissa is equal to $6.0 \times 10^{-4} \text{ rad}^2$.) The values for the concentration of protein (γ_{prot} in milligrams milliliter⁻¹) are indicated to the right. The slopes of the lines in the plot are

γ_{prot} (mg mL ⁻¹)	slope (rad ⁻²)
2.9	1130
5.8	420
8.8	540
11.7	550
14.6	630

- (C) Calculate the apparent radius of gyration, $R_{G,\text{app}}$, for each concentration of protein.
- (D) Extrapolate to $\gamma_{\text{prot}} = 0$ and obtain the actual radius of gyration, R_G .
- (E) The molar mass of immunity protein is 9770 g mol^{-1} , and its partial specific volume, \bar{v}_{imm} , is $0.73 \text{ cm}^3 \text{ g}^{-1}$. Calculate its unhydrated volume and the radius a of a sphere with that volume.
- (F) What would be the radius of gyration of immunity protein if it were this sphere?

Absorption and Emission of Light

A valence electron in any molecule occupies an atomic orbital or a molecular orbital that has **energy levels** associated with it (Figure 12-5).⁸⁵ These energy levels have discrete magnitudes because of the quantum theory, and the steps between any two energy levels are also of discrete magnitude or quantized. The energy levels that have the smallest steps between them are the rotational energy levels. These energy levels correspond to the quantized kinetic and potential energy associated with the rotations around the bond in which a particular electron resides and with the bonds that are coupled to it. The steps between successive rotational energy levels are normally $0.5\text{--}50 \text{ J mol}^{-1}$ in magnitude, corresponding to the energy in a photon of wavelength 200 to 2 mm. The energy levels that have the next larger steps between successive stages are the **vibrational energy levels**. These energy levels reflect the quantization of the energy of the vibrations of the bond in which a particular electron resides and of neighboring bonds that are coupled to it. The steps between vibrational energy levels are normally $5\text{--}50 \text{ kJ mol}^{-1}$ in magnitude, corresponding to the energy in a photon of wavelength 20 to $2 \mu\text{m}$. The energy levels that have the next larger steps among them are the **electronic energy levels** of the molecule. The electronic energy levels relevant to these transitions are those of the atomic orbital or molecular orbital in which a particular electron resides and of the vacant atomic orbitals or molecular orbitals that are accessible to it. Steps between two of these electronic energy levels are normally $50\text{--}500 \text{ kJ mol}^{-1}$ in magnitude corresponding to the energy in a photon of wavelength 2000 to 200 nm. These three types of energy levels form nested sets (Figure 12-5). Within a given electronic energy level there are a series of associated vibrational energy levels, and within a given vibrational energy level there are a series of associated rotational energy levels.

In a particular molecule, at a given instant, a discrete set of atomic and molecular orbitals will be occupied by the valence electrons to produce the σ bonds, the π molecular orbital systems, and the lone pairs. For each atomic or molecular orbital occupied by an electron, a particular vibrational energy level will also be occupied, and the rotational motions within the molecule will determine which particular rotational energy levels are also occupied. Because the differences in energy between electronic energy levels are so large, the **equilibrium constants governing the occupation** by electrons of the successive electronic energy levels at temperatures experienced by living organisms are also large, and only the lowest electronic energy levels are significantly occupied. The electrons in a molecule in the ground state are always ($>99.99999\%$) distributed so as to fill in succession the levels of lowest electronic energy. As the differences in energy between vibrational energy levels are significant, the equilibrium constants govern-

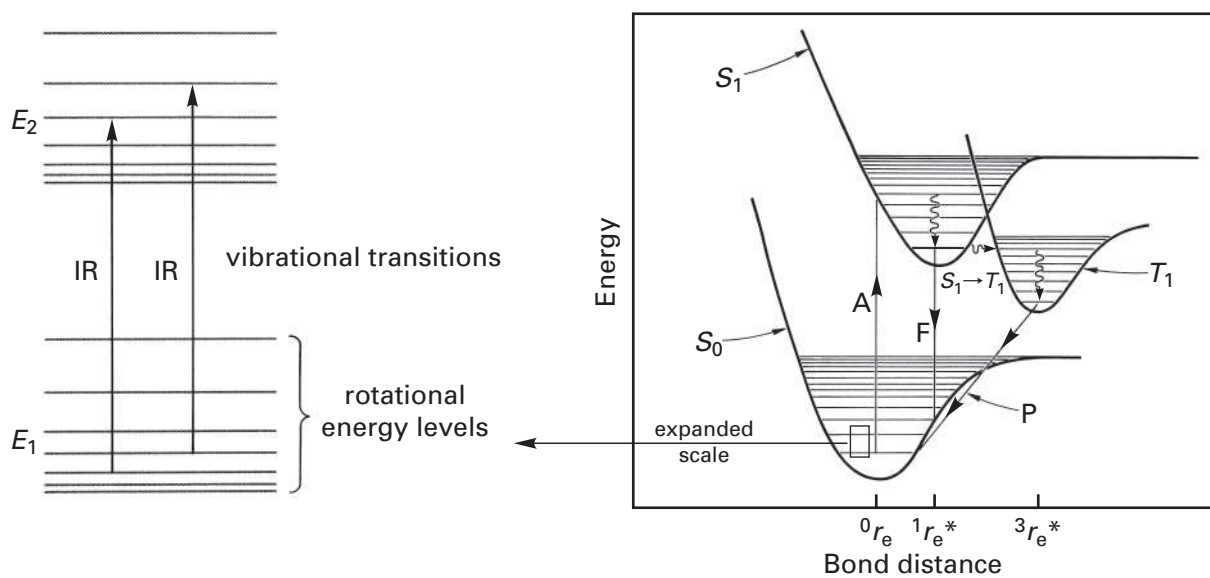


Figure 12-5: Photophysical processes experienced by an electron in a covalent bond between two atoms.⁸⁵ The smooth curve S_0 is the potential energy of the molecular orbital of the covalent bond in which the electron resides as a function of the distance r between the two nuclei. The smooth curve S_1 is the potential energy that would be experienced if the electron were transferred to a particular unoccupied antibonding molecular orbital between the two atoms as a function of the distance between the two nuclei. Each well of potential energy has levels of vibrational energy (the parallel lines within each well) and levels of rotational energy (see expanded scale of the potential energy of the ground state to the left) associated with it. Absorptions by electrons of photons of energy equivalent to the differences in energy between vibrational energy levels (process **IR**) produces an infrared spectrum. When an electron in its occupied molecular orbital, the ground state, absorbs a quantum of electromagnetic energy sufficient to boost its energy high enough to enter the unoccupied molecular orbital, the excited state is created (process **A**). As the excited state relaxes, some of the absorbed energy is lost as heat. When the relaxed excited state emits light as the electron returns to the ground state (process **F**), the quantum of emitted fluorescent light has less energy (longer wavelength) than that of the quantum of light originally absorbed. If the spin of the electron inverts while it is in the excited state (process $S_1 \rightarrow T_1$), the electron enters a triplet excited state. The smooth curve T_1 is the potential energy of the electron in the bond in the triplet state as a function of bond length. The triplet excited state also relaxes by giving off heat. The phosphorescent light emitted from the relaxed triplet state (process **P**) has even less energy (even longer wavelength) than the fluorescence from the initial excited state, and the triplet excited state has an even longer lifetime. The bond lengths of the ground state, the excited singlet state, and the excited triplet state are indicated as 0r_e , ${}^1r_e^*$, and ${}^3r_e^*$. Adapted with permission from ref 85. Copyright 1977 W.A. Benjamin.

ing the occupation of the successive levels at temperatures experienced by living organisms are also significant. Consequently, in the ground state, the vibrational energy level that is occupied is usually (>90%) the lowest for each particular vibration. Because, however, their differences in energy are so small, rotational energy levels are widely occupied by the bonds in the different molecules in the solution.

A photon can encounter an electron in a molecule in such a way that its energy is absorbed. If the electron absorbs the energy of the oscillating electric field and then immediately emits the same photon back again without retaining any of its energy, the direction of the electromagnetic wave is altered so that its new direction of propagation bears no relationship to its incident direction while its wavelength remains the same. This is **elastic scattering**, and it is the phenomenon mainly responsible for X-ray diffraction, low-angle X-ray scattering, and light scattering.

If the electron that has absorbed the photon is in a bond that happens to change its vibrational energy level during the instant that it is excited, the subsequently scattered photon will have a wavelength that is shorter or

longer than that of the incident photon by a difference equivalent to the energy of the transition between the two vibrational energy levels (expanded scale in Figure 12-5). If the photon is absorbed by an electron in the excited state of a vibrational mode, it can carry away the energy of the transition to the ground state and be scattered with higher energy. If it is scattered by an electron in the ground state that enters an excited state during its residence, the photon will provide the energy for this transition and be scattered with lower energy. As a result, the energies of the scattered light vary symmetrically about the energy of the incident light. The intensities of the bands of scattered light with the longer wavelengths are greater than those of shorter wavelength because most vibrations are in the ground state before a photon is absorbed. The **Raman effect** that results is a set of small changes in wavelength that are experienced by a small percentage of the photons that are scattered by the electrons in the sample. As with light scattering itself, the Raman effect on scattered light is usually measured by sampling the light emitted by a sample perpendicular to the incident light. The incident light is from a laser, and it is intense and monochromatic. It is the spectrum of the

wavelengths of the scattered light that is determined. Although almost all of the scattered light is the same wavelength as the incident light, scattered light of other specific, sharply defined wavelengths is also present, and a Raman spectrum of this scattered light provides a catalogue of many of the transitions among the vibrational energy levels of the molecule.

The decision as to whether the photon is immediately scattered from the electron, in an event that is essentially instantaneous, or is absorbed by the electron for a longer period of time depends on how closely the energy of the photon matches one of the differences between the quantized energy levels available to the electron. If the energy of the photon that has just been absorbed by the electron is equal to the difference in energy between the vibrational energy level its bond occupied at the instant the photon was absorbed and a higher vibrational energy level accessible to it, the photon can be absorbed completely, and the vibrational energy of the bond occupied by the electron or that of a bond vibrationally coupled to it will increase by that step in energy (process **IR** in Figure 12-5). Most of the energy absorbed will not be emitted back radiatively as light but will be dissipated nonradiatively by intermolecular collision or by exciting coupled rotational motions the energy levels of which bridge the gap between the vibrational energy levels of the ground state and the excited state (Figure 12-5). Any light emitted radiatively due to a direct transition from the excited state back to the ground state has the same wavelength as the absorbed light but an altered direction and becomes distinguishable from elastically scattered light only by the delay in its reemission. The **absorption of infrared light** (in the range from 20,000 to 2000 nm, or 500 to 5000 cm^{-1})* produces transitions among vibrational energy levels, and a spectrum of infrared absorption has discrete maxima the energies of which correspond to transitions between pairs of vibrational energy levels.

If the energy of the photon absorbed by the electron is equal to the difference in energy between the molecular orbital or atomic orbital the electron occupies in the ground state and an unoccupied molecular or atomic orbital of higher energy, the photon can be absorbed (process **A** in Figure 12-5). The electron enters the unoccupied orbital, and an **electronically excited state** of the molecule is created. Because the excited state differs from the ground state in the distribution of its electrons among molecular and atomic orbitals, it should be thought of as a distinct, albeit similar, molecule. The

formation of the excited state can be followed by monitoring the disappearance of the absorption of light by the ground state⁸⁶ because the excited state, being a new molecule, has a different absorption spectrum. At the very least, in its most stable structure, this **new molecule**, the excited state, will have some bond lengths, bond angles, and bond energies that are different from those of the ground state because its bonding differs from that of the ground state. The instant the electron enters the new orbital, however, the molecule has the structure of the ground state. As the excited state relaxes in energy to the most stable structure available to it, the distance in energy between excited state and ground state shortens, and the excited electron loses energy. Because of the overlap required for excitation, the excited electron usually enters the excited state through one of its higher vibrational energy levels, and it simultaneously loses energy by a nonradiative passage to the lowest vibrational energy level. The net result of these relaxations is that the excited electron very rapidly (10^{-13} to 10^{-11} s) finds itself at an energy considerably below the energy it had achieved immediately after the light was absorbed.

Because the rotational and vibrational energy levels of the electronic excited state usually overlap rotational and vibrational energy levels of the ground state, the electron usually reenters the molecular or atomic orbital of the ground state by pursuing a path among the rotational and vibrational energy levels of excited state and ground state. In this case, the energy originally absorbed is dissipated nonradiatively as heat, and only the absorption of the light is detected. The absorption of ultraviolet and visible light (in the range between 200 and 2000 nm) produces transitions among electronic energy levels, and the result is a **spectrum of ultraviolet or visible absorption** that has maxima the energies of which correspond to the energies of electronic transitions in the molecule.

If, however, the energy levels of the ground state and the excited state overlap weakly, the excited electron can become trapped in the lowest vibrational energy level of the excited state long enough ($>10^{-9}$ s) to reenter the ground state with a bang rather than a whimper. The reentry of the excited electron into the ground state in such a single step requires that the energy it loses be emitted as a photon. This emission is either fluorescence or phosphorescence.

If it came from a covalent bond or a lone pair of electrons, then at the instant of excitation, the excited electron entering the new orbital has a spin opposite to the spin of the partner it left behind in its previous orbital. As it relaxes into the lowest vibrational energy of the excited state and as the excited state rearranges, the excited electron remains coupled to its old partner, and the excited state remains a singlet excited state. From a singlet excited state, the electron can rapidly return to the ground state because the excited electron can readily reenter its old orbital with a spin compatible with the single electron still there (process **F** in Figure 12-5) The

* It is customary for investigators using Raman spectroscopy or infrared spectroscopy to present absorption as a function of the inverse of the wavelength, referred to as the wavenumber (in centimeters^{-1}), which is directly proportional to the energy of the absorption. One advantage of this convention is that the two symmetrical displacements of the Raman effect for the same vibrational mode have the same numerical values when expressed in terms of wavenumber.

reentry is spin-allowed and rapid ($<10^{-7}$ s), and the emitted photon is **fluorescence**.

The energy of a photon of fluorescent light is necessarily less than the energy of the photon absorbed by the electron during excitation because of the nonradiative relaxations of the excited state that have occurred. **Fluorescent light** is light of a longer wavelength (usually visible light) emitted shortly after (within 10^{-7} s) a molecule has absorbed light of a shorter wavelength (usually ultraviolet light). The spectrum of the light absorbed is the **absorption spectrum**; the spectrum of the light emitted is the **emission spectrum**. Fluorescent light, as with scattered light and for the same reasons, is emitted in all directions relative to the incident light unless intramolecular interference occurs. It is usually measured perpendicular to the direction of the incident light. It can be measured under continuous excitation, or the excitation can be a flash ($<10^{-9}$ s in length), and the rate of decay of the fluorescence following the flash, indicative of its lifetime, can be measured.

If the electronically excited state is structured in such a way that the excited electron can become unpaired with the electron it left behind, it can enter a triplet excited state by **intersystem crossing**. The ground state of the triplet state is usually of lower energy than that of the singlet state. Once the triplet excited state has been occupied, the electron can return to the ground state only through a spin-disallowed process that is very slow (on the order of microseconds to seconds). The emitted light, or **phosphorescence** (process **P** in Figure 12-5), emerges from the solution over a relatively long period of time and has an even longer wavelength than the rapidly emitted fluorescence. Fluorescence is light emitted from singlet excited states; phosphorescence, from triplet excited states.

Both the direct and the Raman **infrared spectra of proteins** display absorptions of energy resulting from transitions between the quantized energy levels of molecular vibrations. The intensities of these absorptions are determined by **selection rules** that govern which transitions in vibrational energy are permitted and to what extent these transitions are able to absorb infrared light. Because the selection rules for direct absorption are different from the selection rules governing the Raman effect, direct infrared absorption spectra provide data that are complementary to Raman infrared spectra.

The most obvious and securely assigned absorptions in the infrared or Raman spectrum of a protein arise from excitations of the vibrations of the **amides** in the polypeptide backbone. A secondary amide such as *N*-methylacetamide, a simple model for the peptide bond, absorbs infrared energy of wavelength around 6000 nm (1650 cm^{-1}) into its C=O stretching vibration and of wavelengths around 6500 nm (1540 cm^{-1}) and around 8000 nm (1250 cm^{-1}) into a coupled C–N stretching and N–H bending vibration.⁸⁷ These three peaks of

absorbance are referred to as the **amide I band**, the **amide II band**, and the **amide III band**, respectively. The amide I band of absorbance is the strongest of the three and is the only one that is located in a region of the spectrum that does not contain significant absorptions from other groups in a protein (Figure 12-6).⁸⁸

Direct infrared spectroscopy of proteins in aqueous solution is severely compromised by the strong absorption of infrared light by water and other solutes. An infrared spectrum registered by the Raman effect, however, avoids these drawbacks. A **Raman infrared spectrum** is monitored as small differences in wavelength relative to the wavelength of the incident light. Consequently, the actual light registering each of the bands in the Raman spectrum is within the visible range, not the infrared range, and the problems of the absorption of infrared light by water and other solutes and by the container are avoided. Although the water in the solvent also produces Raman bands in the regions of its absorptions, they are much weaker than their direct absorptions of infrared light,⁸⁹ and the subtraction of background from the spectrum of the protein is much less drastic.

Raman infrared spectroscopy, however, has its own disadvantages. Two of those disadvantages are that the signals registering the Raman infrared spectrum are weak and that these small signals can be swamped by fluorescence. In addition, the presence of large particles such as fragments of membrane, by increasing the scattering from the solution, makes Raman spectroscopy even more difficult. Direct infrared absorption is unaffected by this latter problem, and infrared spectra of sus-

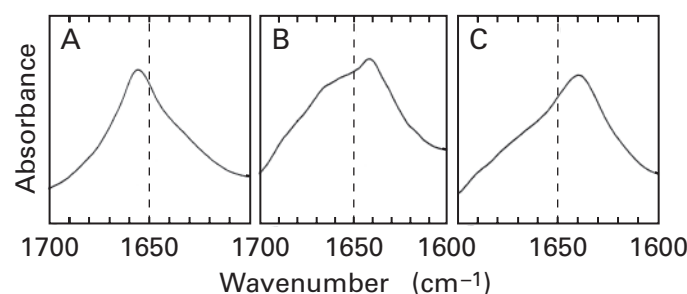


Figure 12-6: Direct infrared spectra in the amide I region of proteins with different mixtures of secondary structure.⁸⁸ (A) Human hemoglobin with its hemes in complex with CO at 130 mg mL^{-1} in 10 mM sodium phosphate, pH 7.4. (B) Bovine ribonuclease A at 50 mg mL^{-1} in 1% NaCl, pH 6.5. (C) Bovine immunoglobulin G at 50 mg mL^{-1} in 1% NaCl, pH 6.5. In each solution, the protein itself served to buffer the pH. The spectra were measured in a Fourier transform infrared spectrophotometer. The spectrum of 10 mM sodium phosphate or 1% NaCl, respectively, was subtracted from each infrared spectrum to obtain the spectrum of the protein alone. Even at these high concentrations of protein, the absorption of the water in the sample accounted for 96% of the total absorption at the wavelength where each of the proteins absorbed most strongly. The vertical dashed line at 1650 cm^{-1} is to aid in comparing the curves. Reprinted with permission from ref 88. Copyright 1990 American Chemical Society.

pensions of membranes can be readily measured.^{90–92} Measurements of the direct infrared spectrum of solid dehydrated protein or even a crystal of protein can also be made.^{93,94}

When a solution of protein is excited with a He–Ne laser, the Raman infrared spectrum of the scattered light (Figure 12–7)⁹⁵ displays a maximum arising from the amide I band of the folded polypeptide with a wavenumber of around 1650 cm^{-1} less than the wavenumber of the majority of the scattered light, which has the same wavenumber as the incident light ($15,802\text{ cm}^{-1}$, 632.8 nm). The amide III maximum at a wavenumber 1250 cm^{-1} less than that of the elastically scattered light and other maxima that can be assigned to vibrational transitions in some of the amino acids, such as phenylalanine, tyrosine, and methionine, are also observed.⁹⁶ By using difference spectra between a selectively deuterated protein and the same protein undeuterated, absorption bands in the Raman infrared spectrum from other amino acids, such as leucine, isoleucine, valine, alanine, glutamate, and aspartate, can be dissected out of the complete spectrum.⁹⁷ Difference Raman infrared spectra between proteins selectively labeled with ^{18}O oxygen and their unlabeled counterpart have also been reported,⁹⁸

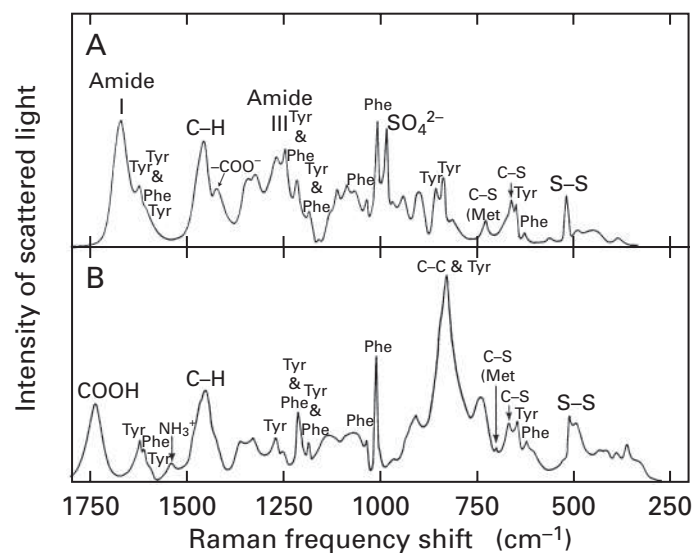


Figure 12–7: Raman spectra of the intensity of the light scattered from (A) a solution of ribonuclease at 200 mg mL^{-1} and (B) a solution of amino acids at the same ratio that they are present in ribonuclease.⁹⁵ The samples were excited with a He–Ne laser ($\lambda = 632.8\text{ nm}$), and the intensity of the light scattered was measured as a function of wavenumber (centimeter^{-1}) in the neighborhood of the wavenumber of the elastically scattered light, which had a wavenumber identical to that of the incident light ($15,802\text{ cm}^{-1}$). The intensity of the scattered light is presented as a function of the difference between the wavenumber of the measured light and the wavenumber of the incident and elastically scattered light. As in the direct infrared spectrum (Figure 12–6), the amide I absorption is the most obvious, but other absorptions that can be assigned to various vibrational modes of the side chains of the amino acids, as well as the partially obscured amide III band, are clearly seen in the spectrum. Reprinted with permission from ref 95. Copyright 1970 Academic Press.

which can permit the observation of the vibrational transitions for a single bond among the thousands within a particular protein.

If the protein contains a functional group that absorbs the exciting visible light in an electronic transition, as does, for example, the heme in hemoglobin,⁹⁹ bands in the Raman infrared spectrum resulting from the absorption of energy by vibrations of the atoms within or adjacent to that functional group will be enhanced, and this enhanced spectrum is referred to as a **resonance Raman infrared spectrum**.¹⁰⁰ The maxima in a resonance Raman infrared spectrum can often be assigned to vibrations of particular bonds, such as an iron–dioxygen stretching vibration in oxygenated hemoglobin,⁹⁹ the oxygen–oxygen stretching vibration of the peroxy form of hemocyanin,¹⁰¹ the iron–oxygen stretching vibration of the ferryl intermediate of cytochrome *d* ubiquinol oxidase,¹⁰² or the copper–sulfur stretching vibrations in halocyanin.¹⁰³ When light of wavelength 200 or 206.5 nm , which is in the range where peptide bonds absorb strongly, is used to produce a resonance Raman infrared spectrum, the amide I, amide II, and amide III bands are selectively enhanced.^{104,105}

The amide I band in the direct infrared spectrum of a solution of a particular protein registers its **secondary structure**.^{88,90,106–108} The amide I band in the direct infrared spectra of a protein containing mainly α helix, for example, hemoglobin (87% α -helical), has a maximum at around 1655 cm^{-1} ; that of a protein rich in β sheet, for example, immunoglobulin G (67% β structure), has a maximum around 1635 cm^{-1} ; and that of a protein with a mixture of these two secondary structures, for example, ribonuclease A (23% α helix and 46% β structure), has a spectrum that seems to register this mixture (Figure 12–6).⁸⁸ Various algorithms have been derived for estimating the percentages of α helix, β structure, and β turn in a protein from the shape of the amide I band in its direct infrared spectrum.^{88,107} The amide I band in the direct infrared spectrum of a coiled coil of α helices also has a characteristic shape, diagnostic of this structure.¹⁰⁹

Unfortunately, as mentioned above, the amide I band falls in a region of the direct infrared spectrum where water absorbs strongly, and this **strong absorption by the solvent** and its vapor must be subtracted to obtain only the amide I band of the protein.⁸⁸ Deuterium oxide does not absorb strongly between 1700 and 1600 cm^{-1} , and direct infrared spectra of proteins in deuterium oxide rather than water display a readily measured amide I band.^{90,106} Unfortunately, it is difficult to exchange the protons with deuteriums on the amide nitrogens of the peptide bonds through the entire protein,¹⁰⁷ because those in the interior are inaccessible to solvent. As a result of this incomplete exchange and the fact that the amide I bands of deuterated peptide bonds are shifted by at least 5 cm^{-1} relative to those of undeuterated peptide bonds,¹⁰⁸ the resulting spectrum

of the mixture of exchanged and unexchanged peptide bonds is difficult to dissect accurately into the components arising from the different secondary structures.

An infrared spectrum registered by the Raman effect avoids these drawbacks, and the amide I band in such a spectrum also registers secondary structure.¹¹⁰ The wavelengths of the absorptions, however, differ from those in a direct infrared spectrum. α -Helical proteins have amide I bands with maxima around 1645 cm^{-1} while proteins composed entirely of β structure have amide I bands with maxima around 1670 cm^{-1} .⁸⁹ In resonance Raman infrared spectra produced by excitation at 205.6 nm , the amide III band is prominent and also registers secondary structure (λ_{max} for α helix at 1299 cm^{-1} and λ_{max} for β sheet at 1235 cm^{-1}).¹⁰⁵ Because of the various drawbacks of infrared spectroscopy with aqueous solutions of proteins, however, circular dichroism is more widely used to estimate percentages of secondary structure.

Circular dichroism is a consequence of the absorption of visible or ultraviolet light by a chiral solute such as a protein. As such, it relies on the excitation of electrons from occupied molecular orbitals into unoccupied molecular orbitals. The most widely used absorptions in spectroscopic studies of proteins by circular dichroism are the electronic absorptions of the amides of the polypeptide backbone between wavelengths of 180 and 240 nm. In this region, two electronic transitions account for the absorption of light.¹¹¹ One is a transition ($n \rightarrow \pi^*$) in which an electron leaves one of the lone pairs on the acyl oxygen (designated by n) and enters the vacant antibonding π molecular orbital of the amide (designated by π^*). This orbital is the one of highest energy of the three molecular orbitals of the π molecular orbital system (Figure 2-3). This $n \rightarrow \pi^*$ transition is responsible for the absorption of light at a wavelength of about 220 nm. The other transition ($\pi^\circ \rightarrow \pi^*$) is one in which an electron leaves the highest occupied nonbonding π molecular orbital of the amide (designated by π°) and enters the vacant antibonding π molecular orbital (Figure 2-3). This $\pi^\circ \rightarrow \pi^*$ transition is responsible for the absorption of light at a wavelength of about 200 nm.

Plane-polarized light is light characterized by an electric vector that oscillates within a plane. One way to describe plane-polarized light mathematically is to assume that it is produced by the sum of two electric vectors of equal amplitude emanating from a single point that is traveling at the speed of light in a straight line. While propagating, these two electric vectors spin in opposite directions, clockwise and counterclockwise at the same frequency (in revolutions second^{-1}) as the frequency of the light (Figure 12-8A).¹¹² The sum of these two spinning vectors produces an electric vector that oscillates in a plane containing the line of propagation to produce the plane-polarized light. These two components that spin in opposite directions around the axis defined by the line of propagation are called **circular polarizations**.

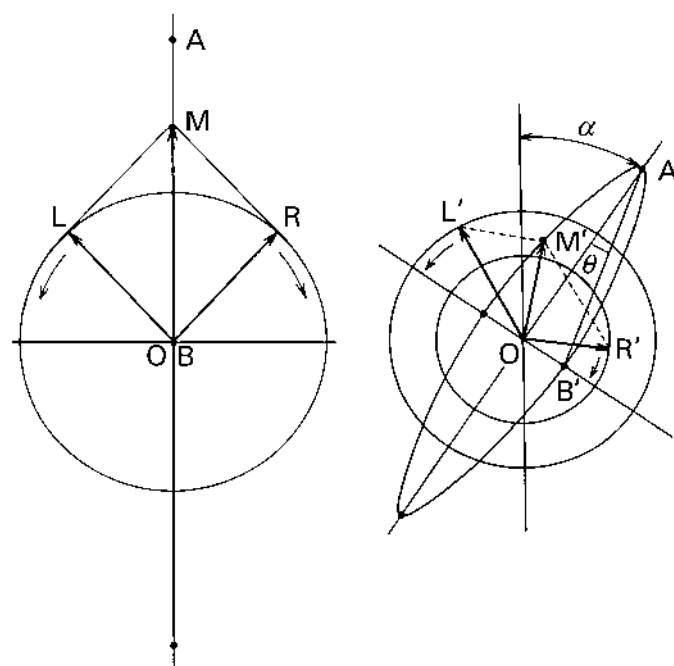


Figure 12-8: Principles of circular dichroism and optical rotation.¹¹² (A) View down a ray of light plane-polarized in the vertical direction looking from the source. The plane-polarized light can be considered to be the sum of two electric vectors, respectively, of right (R) and left (L) circularly polarized light. The electric vector of the plane-polarized light (M) remains fixed in orientation but oscillates in amplitude with a maximum at the point A. The electric vectors of each of the circular polarizations of the ray remain fixed in amplitude but circle in opposite directions at the same angular velocity. (B) Alteration of plane-polarized light during its passage through a solution containing a chiral solute. If the index of refraction of the solution for the left circular polarization (L') of the plane-polarized light is greater than that for the right circular polarization (R'), the left component will have a slower angular velocity than the right component, and the plane of the polarized light will be rotated to the right by an angle α . If the right circular polarization (R') of the plane-polarized light is absorbed more than the left circular polarization component (L'), the plane of the polarized light will broaden into an ellipse because when the two electric vectors are at 180° to each other (at B'), they no longer cancel. The ellipse is created by a composite electric vector (M') that rotates to the left if the absorption of the right circular polarization is greater and to the right if the absorption of the left circular polarization is greater. The maximum of the amplitude of the harmonic oscillation of this electric vector within the rotated plane is at the point A' and its maximum in the dimension at 90° to the rotated plane is at the point B'. The angle θ (Equation 12-33) is indicated. Adapted with permission from ref 112. Copyright 1967 Marcel Dekker.

When these two circular polarizations encounter a chiral object such as one of the functional groups of a protein in a solution, they are absorbed and retarded unequally. If only the amplitude of one component were decreased more than the amplitude of the other while the two components remained in phase, the plane of polarization of the emerging light would retain the same orientation, but its electric vector, which is the sum of the two components, would trace in cross section an ellipse rather than a flat, linear segment (Figure 12-8B). If the only the phase between the two components were

shifted while their amplitudes remained the same, the plane of polarization of the emerging light would rotate by an angle α , but the electric vector would still trace in cross section a flat, linear segment (Figure 12-8B). The first effect is circular dichroism; the second effect, **optical rotation**. Both effects are required to occur simultaneously in any circumstance, and as polarized light is rotated, it necessarily becomes elliptical and vice versa. This obligatory connection permits the spectrum of optical rotation as a function of wavelength to be calculated from the spectrum of circular dichroism as a function of wavelength and vice versa.¹¹²

The degree to which the emerging light has become elliptical can be measured, and it is expressed as an angle

$$\theta = \tan^{-1} \frac{OB'}{OA'} \quad (12-33)$$

where the ratio OB'/OA' is the ratio of the minor and major axes of the ellipse (Figure 12-8B). The **molar ellipticity** at a given wavelength λ , $[\theta]_\lambda$, is defined by the relationship

$$[\theta]_\lambda \equiv \frac{\theta}{l [\text{chromophore}]} \quad (12-34)$$

where $[\text{chromophore}]$ is the molar concentration of the functional group absorbing the light, referred to as the **chromophore**, and l is the path length of the sample chamber. By convention, the units of $[\theta]_\lambda$ are chosen to be degrees centimeter² (decimole of chromophore)⁻¹. A **circular dichroic spectrum** is a display of the amplitude of $[\theta]_\lambda$ as a function of wavelength.

The optical rotation α (Figure 12-8B) produced by the sample can be registered with a spectropolarimeter. It can also be normalized by the concentration of the chromophore responsible for it to produce the specific rotation $[\alpha]_\lambda$. A spectrum of the **optical rotatory dispersion** is simply the amplitude of $[\alpha]_\lambda$ plotted as a function of the wavelength of the polarized light. Because optical rotation arises from a shift in the relative phases of the two circularly polarized components (Figure 12-8B), it is proportional to the derivative with respect to wavelength of the electronic absorption from which it arises. This has the practical disadvantages of both turning each peak of absorption into two peaks, a positive one and a negative one distributed around the wavelength of maximum absorption, and broadening the signal. In a spectrum of optical rotatory dispersion arising from several maxima of absorption, the individual components are difficult to resolve.

A circular dichroic spectrum, however, is usually simpler to interpret. Unless excitonic coupling between two apposed chromophores of similar wavelengths of absorption is occurring, the individual bands in a circular dichroic spectrum of a protein are unsplit peaks that

coincide with absorption maxima in the absorption spectrum of the same protein. In uncomplicated situations, the circular dichroic spectrum simply registers the optical activity of each chiral contributor to the absorption spectrum. For example, most of the peaks in the absorption spectrum of cytochrome c_1 have only one corresponding **negative or positive peak** at the same wavelength in its circular dichroic spectrum (Figure 12-9).¹¹³ The peak of absorption from the pyridoxal phosphate in glycine hydroxymethyltransferase at 422 nm corresponds to a prominent peak of positive polarization at the same wavelength and of the same width in the circular dichroic spectrum, and corresponding peaks in the two corresponding spectra shift to 343 nm upon the addition of the substrate serine to the solution.¹¹⁴ Because adjacent bands of absorption often have different polarities, the circular dichroic spectrum can often reveal details in the absorption spectrum. For example, the two overlapping peaks at 480 and 530 nm in the absorption spectrum of cytochrome- c oxidase from *Thermus thermophilus* correspond to peaks at the same wavelengths of positive polarization and negative polarization, respectively.¹¹⁵ If excitonic coupling between two or more chromophores is occurring, however, the resulting bands in the circular dichroic spectrum will each be split into two or more components of both positive and negative amplitude, and this splitting complicates the situation.¹¹⁶

A polypeptide folded entirely as an α helix has a circular dichroic spectrum that is distinct from that of a polypeptide folded entirely in β structure. Both of these spectra are distinct from that of a polypeptide unfolded as a random coil (Figure 12-10).¹¹⁷ In the circular dichroic spectrum of a polypeptide folded as an α helix, the amido $\pi^\circ \rightarrow \pi^*$ transition at about 200 nm is split into a positive component ($\lambda_{\text{max}} = 191$ nm) and a negative component ($\lambda_{\text{max}} = 205$ nm). This splitting arises from the fact that each amide is held in the same orientation relative to the axis of the α helix.¹¹⁸ There is also the additional band of negative ellipticity at 225 nm from the $n \rightarrow \pi^*$ transition of the peptide bond, which, in combination with the band of negative ellipticity at 205 nm, gives the circular dichroic spectrum of the α helix its characteristic double minimum. The $\pi^\circ \rightarrow \pi^*$ transition from a polypeptide in either **β structure** or random coil is unsplit.

The tyrosines, phenylalanines, and tryptophans in a polypeptide absorb light of wavelength between 180 and 240 nm and have characteristic circular dichroic spectra.¹¹⁷ The contributions of the tyrosines, phenylalanines, and tryptophans in a protein to its circular dichroic spectrum can be numerically subtracted to reveal the circular dichroic spectrum of the amides of polypeptide backbone alone. Because the polypeptide is the main contributor to the circular dichroic spectrum between wavelengths of 180 and 240 nm, the unit in which the molar ellipticity is usually presented is decimolarity of peptide bonds (Figure 12-10).

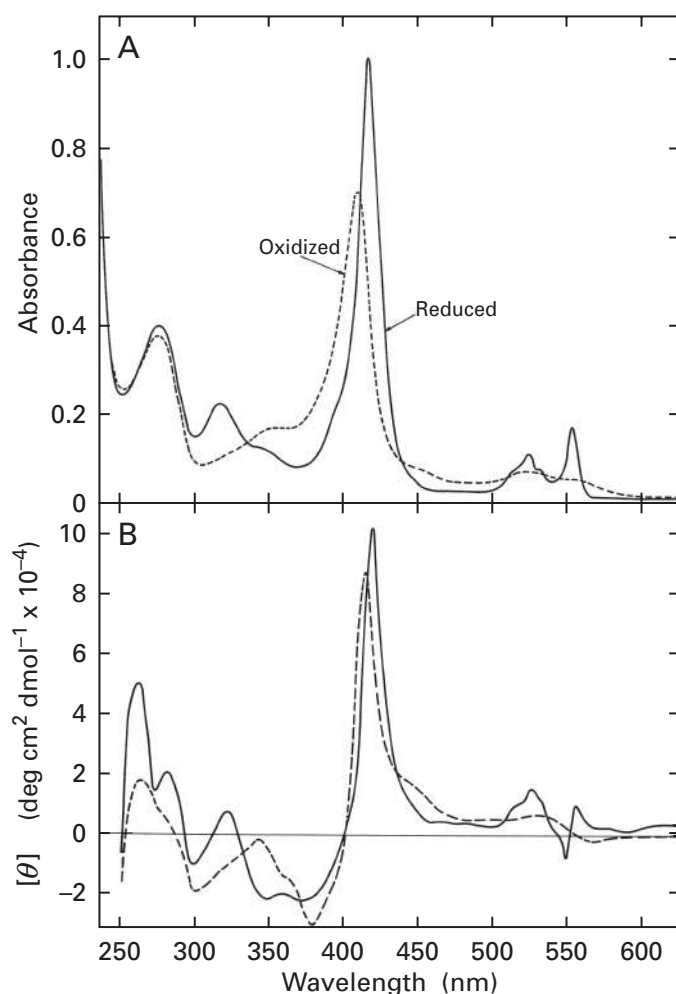


Figure 12-9: Correlation between an optical absorption spectrum (A) and the corresponding circular dichroic spectrum (B).¹¹³ Cytochrome c_1 was purified to homogeneity from bovine heart muscle. Solutions of the hemoprotein were prepared in its oxidized Fe(III) and reduced Fe(II) forms. (A) Optical absorption (absorbance) of the oxidized and reduced proteins as a function of wavelength (nanometers). In each case, the absorbance at wavelengths greater than 300 nm is due entirely to the heme of the hemoprotein. The intense bands of absorption at 400–420 nm are characteristic of hemes. (B) Molar ellipticities ($[\theta] \times 10^{-4}$) of the two solutions of the same two forms of the cytochrome c_1 , oxidized and reduced, as a function of wavelength. The molarities of the solutions were expressed as moles of peptide bond in each liter of solution, which seems inappropriate since the majority of the absorption arises from the heme. Nevertheless, the units for molar ellipticities are degrees centimeter² (decimole of peptide bond)⁻¹. For each band in the absorption spectrum, each of which has a positive value, there is a corresponding band in the circular dichroic spectrum, which is either positive or negative. For example, the band of absorption at 350 nm in the spectrum of the absorbance of the reduced cytochrome c_1 corresponds to a band of negative molar ellipticity in the circular dichroic spectrum. Reprinted with permission from ref 113. Copyright 1971 Academic Press.

The experimentally measured circular dichroic spectrum of the folded polypeptide in a protein can always be resolved numerically into three component spectra as similar as possible to those of pure α helix, pure β structure, and pure random coil. If it is assumed

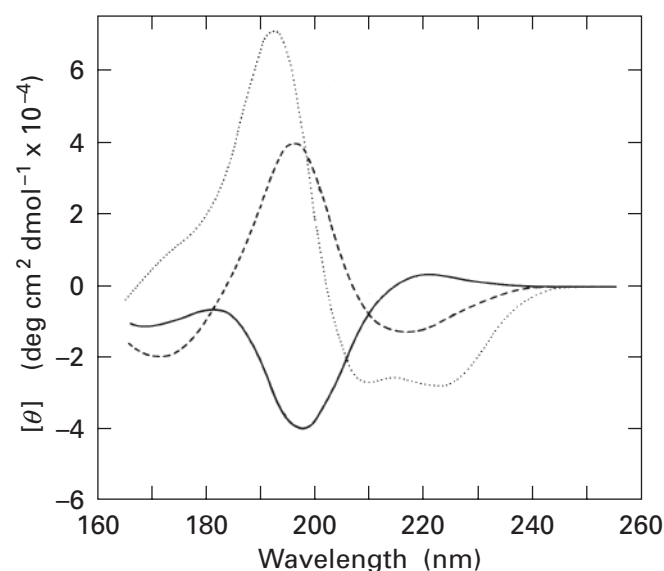


Figure 12-10: Circular dichroic spectra that are used as reference spectra for α helix (dotted line), β structure (dashed line), and random meander (solid line).¹¹⁷ Molar ellipticity, $[\theta] \times 10^{-4}$ [degree centimeter² (decimole of peptide bond)⁻¹], is presented as a function of wavelength (nanometers). Myoglobin from *Physeter catodon*, dissolved in 0.1 M NaF at pH 7, is a protein that is almost entirely α -helical (Figure 4-18). It was used as a reference compound for α helix (dotted line). Poly(Lys-Leu-Lys-Leu) in 0.5 M NaF at pH 7 was used as a polypeptide that is purely β structure (dashed line). Poly(Pro-Lys-Leu-Lys-Leu) in a salt-free solution is completely structureless because of the prolines and the strong repulsions of the lysines. It is not, however, a typical random coil because of both of these features. Nevertheless, it was used as a model for a polypeptide that is purely random meander (solid line). Reprinted with permission from ref 117. Copyright 1980 Academic Press.

that these component spectra, obtained only by numerical analysis, do, nevertheless, accurately represent the contributions of α helix, β structure, and random meander to the entire spectrum, their relative amplitudes should provide the relative amounts of these three components in the actual molecule of protein.¹¹⁹

This simple expectation is diminished by several difficulties. Small peptides that assume particular types of **β turn** show significantly different circular dichroic spectra, that are each unique from that of a random coil, and dissecting out the contribution of each type of β turn to the spectrum of a particular protein is difficult,¹¹⁷ if not impossible. A related shortcoming is the confounding of random coils and random meanders. A random coil is an unfolded polypeptide continuously changing its structure by rotation around its various covalent bonds. **Random meander** is the path assumed by the backbone of a folded polypeptide that is neither an α helix, a β structure, nor a β turn. Random meander is static and respectively identical in all of the folded polypeptides in a solution of a given protein. The random coil of an unfolded polypeptide used as the standard in circular dichroism, unlike the α helix or β structure used as the standard, bears no relationship to the random meander

in a particular folded polypeptide. The random meander in a particular protein will produce a specific circular dichroic spectrum that is distinct from the common circular dichroic spectrum produced by all random coils and also distinct from the unique spectra produced by random meanders in other proteins.

Nevertheless, a least-squares method has been developed to fit the experimental circular dichroic spectrum of a protein, from which the contributions of tyrosine, tryptophan, and phenylalanine have been subtracted, to a calculated spectrum (Figure 12-11).¹¹⁷ The parameters of the **fitting procedure** are the fraction of α helix, the fraction of β structure, the fraction of random meander, and the fraction of each type of β turn. Because each measurement of molar ellipticity is based on decimoles of peptide bonds, it is assumed that the sum of these fractions is unity, that each point on the experimental spectrum is the sum of the molar ellipticity at that wavelength of the appropriate reference spectrum for the respective secondary structure multiplied by the fraction for that particular secondary structure, and that the reference spectrum for random meander is that of random coil. The least-squares procedure gives the respective values for the fractions of the four types of secondary structure that produce a calculated curve most closely reproducing the experimental curve. The fractions for each type of secondary structure estimated in this way for a set of proteins agreed quite closely with the fractions for each type of secondary structure in the respective crystallographic molecular models of these proteins.

One of the more important and informative uses of circular dichroism is to provide evidence that the structure of the protein has changed under particular circumstances. A **conformational change** is a change in the structure of the protein between two states of similar stability. For example, the conformational change of aspartate carbamoyltransferase that occurs upon the binding of its substrates and that is detected both crystallographically and as a change in sedimentation coefficient is also accompanied by significant changes in the circular dichroic spectrum of the protein.¹²⁰ Such changes in circular dichroic spectra coincident with a conformational change of a protein are commonly encountered. This fact increases the concern over the accuracy of secondary structural dissections by numerical analysis of circular dichroic spectra because crystallographic descriptions of conformational changes of proteins rarely involve significant changes in the content of α helix, β structure, β turns, or random meander or changes in their disposition over the sequence of the folded polypeptide. The changes in the circular dichroic spectrum of Na^+/K^+ -transporting ATPase during a conformational change caused by binding of its substrates are consistent with the transformation of 7% of its amino acids from α helix into β structure.¹²¹ When crystallographic molecular models of the two conformations between which the homologous conformational change occurs in

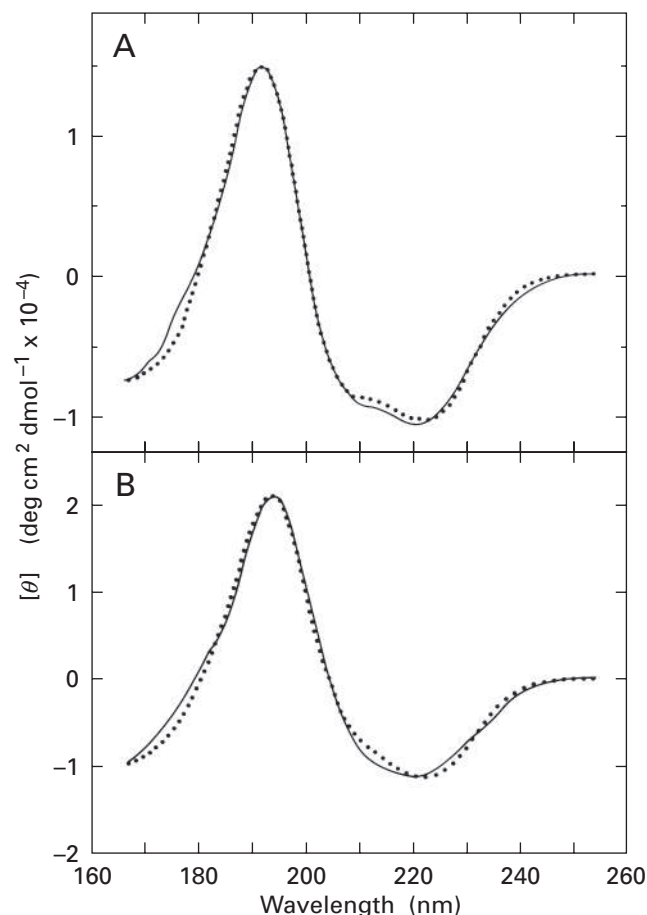


Figure 12-11: Circular dichroic spectra¹¹⁷ of (A) glyceraldehyde-3-phosphate dehydrogenase (phosphorylating) in 0.1 M NaF, pH 7, and (B) subtilisin in 0.2 M NaF, pH 7. Molar ellipticities, $[\theta] \times 10^{-4}$ [degree centimeter² (decimole of peptide bond)⁻¹], are presented as a function of wavelength (nanometers). The spectra were either directly measured (solid lines) or duplicated (dotted lines) by adding together spectra for α helix, β structure, β turn, and random meander (Figure 12-10). In the procedure used to duplicate the experimental spectrum, it was assumed that the proteins contain only α helix, β structure, β turn, and random meander. If f_{α} , f_{β} , f_T , and f_{RM} are the fractions of each of these secondary structures, it is assumed that the sum of these four numbers is 1 and that $f_{\alpha}(\theta^{\circ}_{\alpha}) + f_{\beta}(\theta^{\circ}_{\beta}) + f_T(\theta^{\circ}_T) + f_{RM}(\theta^{\circ}_{RM})$ is equal to the measured value of θ at every wavelength, where the θ° values are the molar ellipticities of the standard curves (Figure 12-10) at the same wavelength. A least-squares method was used to obtain the best values for f_{α} , f_{β} , f_T , and f_{RM} , and these four values were then used to construct the calculated curves presented in the panels. Note that f_{α} , f_{β} , f_T , and f_{RM} are parameters determined only by the structure of the protein and must have the same values for all wavelengths. For the spectrum of glyceraldehyde-3-phosphate dehydrogenase (phosphorylating), the best values of f_{α} , f_{β} , f_T , and f_{RM} were 0.31, 0.30, 0.22, and 0.17; for the spectrum of subtilisin, 0.30, 0.21, 0.21, and 0.28. Reprinted with permission from ref 117. Copyright 1980 Academic Press.

Ca^{2+} -transporting ATPase are examined,⁴⁰³⁻⁴⁰⁸ the content of α helix does change in the correct direction, but only by 2-3%. This rather small change in the amount of α helix is more consistent with the absence of a measurable shift in the amide I absorption in the infrared spectrum under the same circumstances.⁹⁰

As noted previously, most short **peptides** are structureless in water. The formation of α helices by those peptides synthesized to promote this secondary structure is routinely monitored by circular dichroism. It is also possible, by difference circular dichroic spectroscopy, to follow the assumption of a fixed structure by an otherwise structureless peptide when it binds to a protein.¹²²

Ultraviolet absorption spectra of proteins at wavelengths greater than 240 nm are dominated by the absorption of **phenylalanine** ($\lambda_{\text{max}} = 258 \text{ nm}$; $\epsilon_{258} = 197$), **cystine** ($\epsilon_{260} = 280$), **tyrosine** ($\lambda_{\text{max}} = 275 \text{ nm}$; $\epsilon_{275} = 1420$), and **tryptophan** ($\lambda_{\text{max}} = 280 \text{ nm}$; $\epsilon_{280} = 5600$).^{123,124} Tryptophan has the largest extinction coefficient and longest wavelength of maximum absorbance. Because of the strong absorption of tryptophan, the spectra of most proteins between 260 and 310 nm have the same shape as the spectrum of tryptophan alone with its maximum at 280 nm and its pronounced shoulder at 289 nm. Proteins with little or no tryptophan, however, have maxima of absorption shifted toward or coincident with the 275 nm maximum characteristic of tyrosine. The absorption of a protein at its particular maximum, somewhere between 275 and 280 nm, when properly corrected for the absorption due to the scattering of light by the solution, can be used as a rapid measurement of its concentration. Proteins that are posttranslationally modified with chromophores such as flavin, pyridoxal phosphate, or heme or bind noncovalently chromophores such as flavin, heme, metallic cations, coenzyme b_{12} , chlorophyll, pheophytin, or carotenoid, display absorption spectra that are characteristic of those chromophores (Figure 12–9). If one or more of the accessible tyrosines on the protein have been nitrated, their absorption spectra are shifted into the visible range ($\lambda_{\text{max}} = 430 \text{ nm}$) and their acid dissociation constants are increased ($\text{p}K_{\text{a}} = 6.5$), so that at neutral pH they are present mainly as the nitrophenolate, which absorbs strongly.¹²⁵

Either tryptophan or nitrotyrosine can be used as a spectral **reporter group**, the spectrum of which registers its environment or can monitor a conformational change of the protein.¹²⁶ For example, the absorption spectrum of nitrated Tyrosine 115 in micrococcal nuclease indicates that it is in a nonpolar environment in the absence of substrate but a polar environment in the presence of substrates.¹²⁵ This change in environment is also reflected in its accessibility to nitration by tetranitromethane. The conformational change of aspartate carbamoyltransferase that occurs on the binding of substrates can be detected by an upfield shift in the wavelength of the absorption of tryptophans in the protein¹²⁷ or of nitrated tyrosine side chains in its regulatory β subunits.¹²⁸

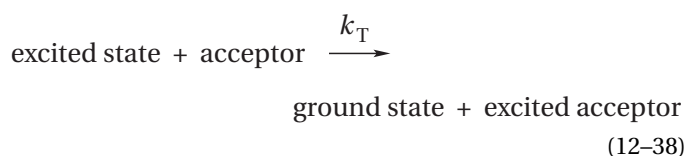
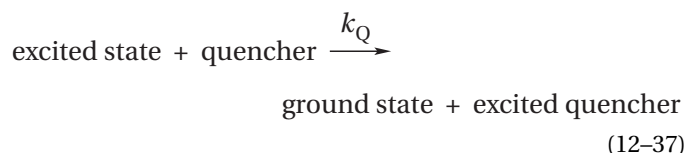
In addition to absorbing ultraviolet light, tryptophan is also both fluorescent and phosphorescent. The wavelength of the maximum **emission of fluorescence from tryptophan** varies from 300 to 350 nm.¹²⁹ The emis-

sion of fluorescence from indole itself varies between these limits systematically as a function of the polarity of the solvent; the more polar the solvent, the longer the wavelength of the emission, and the tryptophans in a protein that are the more buried display shorter wavelengths of maximum emission.¹²⁹ Consequently, the wavelength of its emission is used as a measure of the degree to which a tryptophan is buried within a protein. In the case of the absorption spectrum of tryptophan as opposed to its emission spectrum, the situation is reversed. The most buried tryptophans, in the most nonpolar environments, have been found to absorb light of the longest wavelength, on the red edge of the absorption band for tryptophan in the ultraviolet.¹³⁰

If a protein contains no posttranslationally or experimentally added chromophore, the emission of fluorescence from the protein will be dominated by that of its tryptophans. By the systematic removal of its tryptophans through site-directed mutation, the contribution of each of them to the total emission of fluorescence from the protein can be ascertained.^{131,132} A tryptophan can also be inserted into a particular location in a protein by site-directed mutation to monitor local changes in conformation.¹³³

The fluorescence from each of the tryptophans in a protein displays a characteristic wavelength of maximum emission and a characteristic intensity.¹³² When a protein is unfolded, its emission of fluorescence usually shifts to longer wavelengths as its buried tryptophans become exposed,^{131,134} but the intensity of the fluorescence can either increase¹³¹ or decrease.¹³⁴ These changes can be followed for individual tryptophans in appropriate mutants.¹³¹ Because all of the tryptophans in a protein are fully exposed to solvent upon unfolding, this observation states that the enclosing of a tryptophan by the native structure can either quench or enhance its fluorescence. Consequently, it is the particularity of the **local environment** around each tryptophan in the native protein that governs the intensity of its emission. For example, Tryptophan 94 in folded, native ribonuclease from *Bacillus amyloliquifaciens* has very little emission of fluorescence because one of its immediate neighbors is Histidine 18, which strongly quenches it.¹³⁵ The lowest intensity of emission (by a factor of greater than 3-fold) from the three tryptophans in lysozyme from T4 bacteriophage is that of Tryptophan 158, which is surrounded by a cystine and two methionines, the sulfurs of which are also efficient quenchers.¹³⁶ The amides of glutamine and asparagine are also efficient quenchers. This sensitivity to the particularity of the surroundings explains why tryptophans with the shortest wavelengths of maximum emission are not always the ones with the lowest intensity of emission.¹³²

If intersystem crossing is not significant, the excited state of a fluorescent functional group such as tryptophan can decay to the ground state by at least four **separate pathways**.¹³⁷



where the k_i are the rate constants for the respective processes. Equation 12-35 describes a radiationless decay of the energy of excitation through migration among rotational and vibrational energy levels or other piecemeal transfers to its surroundings as heat. Equation 12-36 describes the release of a portion of the energy of excitation as a photon of fluorescent light. Equation 12-37 describes the transfer of the energy of excitation to another molecule, the quencher. Although there are some molecules or ions, in particular those containing an unpaired electron, that can quench at distances beyond their van der Waals radii, most quenchers must collide with the fluorescent functional group when it is in the excited state to quench it. Equation 12-38 describes the radiationless transfer of the energy of excitation through space by resonance between the excited state and a nearby functional group capable of absorbing the energy. The excited electron is the donor, and the functional group to which the excitation is transferred is the acceptor.

The **quantum yield**, Q_0 , of a fluorescent chromophore is the number of photons appearing as fluorescence for every photon absorbed. When neither quencher nor acceptor is present,¹³⁷

$$Q_0 = \frac{k_F}{k_L + k_F} = k_F \tau_0 \quad (12-39)$$

The time over which 50% of the excited state disappears, or the half-life of the excited state, would be $(\ln 2)(k_L + k_F)^{-1}$, but the **lifetime of the fluorescence**, τ_0 , is defined as $(k_L + k_F)^{-1}$, the time in which the intensity decreases to $\exp(-1)$ of the initial intensity.

If a **collisional quencher** is added to the solution, it affects the quantum yield and lifetime of the excited state because every time a molecule of quencher collides with a molecule of excited state, there is a specific probability that the excitation energy will be transferred from the

excited state to the quencher. Each quencher has a unique efficiency for quenching a particular fluorescent chromophore. The result of this transfer of energy upon collision is that¹³⁷ the quantum yield of the fluorescent chromophore in the presence of the quencher, Q_Q , is decreased:

$$Q_Q = \frac{k_F}{k_L + k_F + k_Q[\text{quencher}]} \quad (12-40)$$

and

$$\frac{Q_Q}{Q_0} = \frac{1}{1 + \tau_0 k_Q [\text{quencher}]} \quad (12-41)$$

The observed lifetime of the quenched fluorescence, τ_Q , is

$$\tau_Q = \frac{\tau_0}{1 + \tau_0 k_Q [\text{quencher}]} \quad (12-42)$$

Phosphorescence, which is simply fluorescence from a triplet state, is also subject to quenching.

The ratio F_Q/F_0 is the ratio between the fluorescence observed in the presence of the quencher and the fluorescence observed in its absence. This ratio is necessarily equal to Q_Q/Q_0 . The ratio F_Q/F_0 can be readily measured with a fluorometer. When its reciprocal, F_0/F_Q , is plotted as a function of [quencher], a linear relationship is obtained, the slope of which is equal to $k_Q \tau_0$ (Figure 12-12).¹²⁹ The **bimolecular rate constant k_Q for the collision** of the quencher with the fluorescent functional group on a protein is the slope of this line divided by the lifetime τ_0 of the fluorescence of the unquenched excited state.

The fluorescence and phosphorescence of tryptophan can be quenched by large inorganic anions, such as Γ^- and NO_2^- ; by molecular oxygen; by unsaturated amides, such as acrylamide; and by ketones, such as 2-oxobutane.^{129,130,137,138} The bimolecular rate constant k_Q for the quenching reflects the accessibility of the quencher to the tryptophans.¹³⁸ If the quencher is a polar molecule confined to the aqueous phase, then the greater the rate constant for quenching, the more exposed is the tryptophan to that phase. On the basis of the observed rate constants k_Q for polar quenchers, the tryptophans in most proteins can be divided into three classes.¹³⁰

The tryptophans in the first class are **fully accessible** to the aqueous phase, and their fluorescence is readily quenched. The rate constants for the collisions between their singlet excited states and various polar quenchers are $2\text{--}10 \text{ M}^{-1} \text{ ns}^{-1}$ as expected from a diffusion-controlled process.

The fluorescence of the tryptophans in the third

class cannot be quenched ($k_Q < 0.01 \text{ M}^{-1} \text{ ns}^{-1}$)¹³⁶ because no collisions with the quenchers can occur within the lifetime of their excited states. Presumably, this is due to the fact that they are **buried**. This conclusion follows from the facts that such unquenchable tryptophans are observed in proteins that contain buried tryptophans in their crystallographic molecular models. For example, Tryptophan 138 in lysozyme from bacteriophage T4 is poorly quenched ($k_Q = 0.009 \text{ M}^{-1} \text{ ns}^{-1}$), and it is well buried in the molecular model of the protein.¹³⁶ Also, such unquenchable tryptophans are optimally excited by light of a longer wavelength.¹³⁰ The phosphorescence of these buried tryptophans, however, has a sufficiently long lifetime ($\tau_0 \cong 1 \text{ s}$) that it can be quenched. The bimolecular rate constants k_Q for the quenching of the phosphorescence of the tryptophans in the buried class are relatively small ($< 0.001 \text{ M}^{-1} \text{ ns}^{-1}$), and the quenching registered by these rate constants seems to result from extensive and momentary unfoldings of the folded polypeptide that occasionally provide access to the interior, but only for a short time.¹³⁰ These observations provide support for the concept that most parts of a protein are conformationally active and continuously expand and contract.

The intermediate, second class of tryptophans, and probably the most numerous, are those that are **partially buried** and have intermediate rate constants for quenching ($0.01\text{--}2 \text{ M}^{-1} \text{ ns}^{-1}$). Examples are Tryptophan 59 in ribonuclease T1 from *Aspergillus oryzae* ($k_Q = 0.3 \text{ M}^{-1} \text{ ns}^{-1}$),¹³⁸ Tryptophan 126 from lysozyme of T4 bacteriophage ($k_Q = 0.3 \text{ M}^{-1} \text{ ns}^{-1}$),¹³⁶ and Tryptophan 333 of phosphoglycerate kinase from *Saccharomyces cerevisiae* ($k_Q = 0.8 \text{ M}^{-1} \text{ ns}^{-1}$).¹³⁹

Oxygen provides an interesting exception to the behavior of most quenchers. The difference in its ability to quench accessible and buried tryptophans is much less than that observed with larger more polar quenchers (Figure 12-12).^{129,140} On the basis of this observation, it has been proposed that oxygen is small enough to insinuate its way through a molecule of protein in liquidlike diffusion among the tightly packed amino acids.

Changes in the accessibility of tryptophans to polar quenchers dissolved in the aqueous phase have been used to monitor conformational changes in the structure of a protein. In the case of succinate-CoA ligase (ADP-forming) from *E. coli*, the binding of ATP to the α subunit of the enzyme causes significant decreases in the accessibility of the tryptophans in the β subunit to acrylamide dissolved in the solution.¹⁴¹ This suggests that a conformational change propagated throughout the whole protein occurs upon the binding of ATP. The implication that both the α and β subunits change their structure in concert when ATP binds is consistent with the observation that they are intimately associated in the oligomeric structure of the protein (Figure 8-22).¹⁴² To follow the conformational change that occurs when the carboxy-terminal portion of colicin E1 from *E. coli* inserts into a

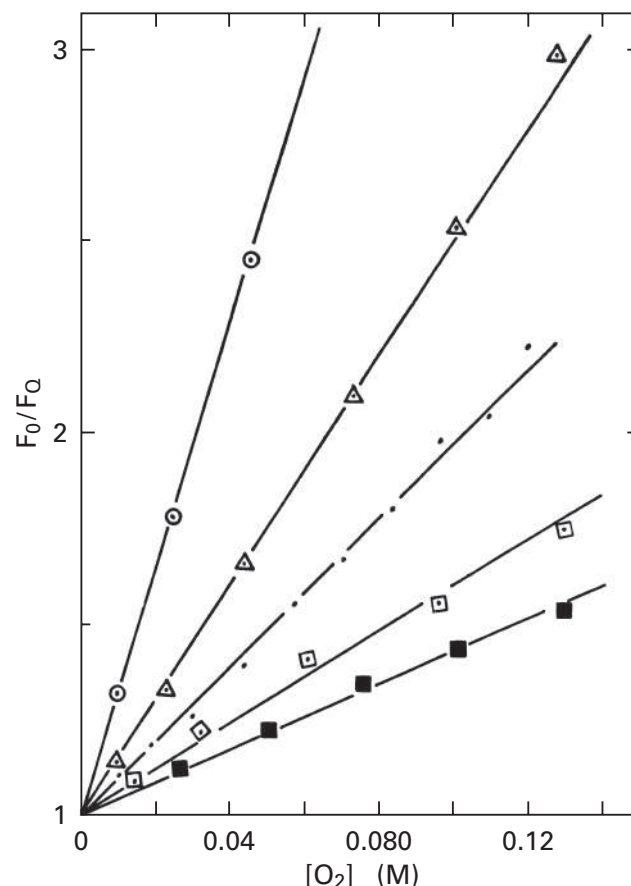


Figure 12-12: Collisional quenching of the fluorescence of tryptophans in several proteins by oxygen.¹²⁹ Solutions of the various proteins were placed in cuvettes in a fluorometer and excited with light of wavelength 280 nm. Fluorescence at 90° to the exciting beam was monitored at the wavelength of maximum emission for each protein (325–350 nm). The high concentrations of oxygen (molar) were produced by enclosing the cuvette in a chamber that could be pressurized to 105 kg cm⁻² O₂ gas and allowing the gas to equilibrate with the solution at various pressures. The proteins used were bovine α -chymotrypsin (■), rabbit fructose-bisphosphate aldolase (□), bovine immunoglobulin G (●), and bovine serum albumin (Δ). A solution of tryptophan (○) was used as an example of a fully exposed side chain. Lines were drawn on the basis of the expectation that F_0/F_0^{-1} as a function of [quencher] would be linear (Equation 12-41). Reprinted with permission from ref 129. Copyright 1973 American Chemical Society.

membrane, tryptophans were inserted at various positions in its amino acid sequence by site-directed mutation, and changes in the rate constants of quenching for these tryptophans were measured before and after insertion.¹⁴³

It is also possible for the energy of the relaxed excited state in excess over the energy of the ground state, which otherwise would be emitted as fluorescence, to be transferred intact through space by resonance to another chromophore in a radiationless process (Equation 12-38). This **fluorescence resonance energy transfer** (FRET) discharges the electronically excited state of the functional group that originally absorbed the

photon, the **donor**, and produces an electronically excited state in the functional group that receives the energy, the **acceptor**. Because this transfer of energy between donor and acceptor occurs by resonance, there must be a matching of energy and a matching of orientation between donor and acceptor. The energies are matched in the region of overlap between the absorption spectrum of the acceptor and the emission spectrum of the donor; the greater the overlap, the greater the probability that the energy will be transferred. The orientations are matched in the coincidence between the orientation of the transition dipole of the donor and the transition dipole of the acceptor; the greater the coincidence, the greater the probability that the energy will be transferred.

If the new excited state of the acceptor created by this transfer normally returns to its respective ground state by radiationless processes—in other words, if its quantum yield is zero—no fluorescent photon is emitted, and the only observations made are that the fluorescence of the donor is quenched and its lifetime is decreased. If the acceptor is also a fluorescent functional group, its excited state will release a fluorescent photon, consistent with its quantum yield. The photon released from the acceptor, however, will be of even longer wavelength than the one that would have been released from the donor (Figure 12–13)¹⁴⁴ because the excited state of the acceptor immediately following the transfer relaxes to a stable excited state, the energy of which, relative to the ground state of the acceptor, is less than the energy that was passed from donor to acceptor during the transfer. At those wavelengths in the emission spectrum of the donor that do not overlap the emission spectrum of the acceptor and are therefore not contaminated by the fluorescent emission of the acceptor, the fluorescence measured in the presence of the acceptor will be less than the fluorescence measured in its absence. The observed, uncontaminated fluorescence of the donor will be quenched in the presence of the acceptor by the ratio Q_A/Q_0 .

Suppose that a molecule of protein has a fluorescent chromophore that can act as a donor covalently attached or noncovalently bound to a particular location in its tertiary structure and a different chromophore that can act as an acceptor covalently attached or noncovalently bound to a different location. The various fluorescent compounds used to modify rhodopsin (Figure 12–14)¹⁴⁵ are typical of the donors and acceptors that are covalently or noncovalently attached to specific locations in a protein. In such a situation, the rate of the transfer of energy from the excited donor to the acceptor by resonance (Equation 12–38) is equal to $k_T f_A$ [excited donor], where f_A is the fraction of the sites on the protein for the acceptor that are occupied and [excited donor] is the molar concentration of the excited donor. As the acceptors are fixed to the molecules of protein that contain an excited donor at a constant fractional occupancy, the pseudo-first-order rate constant for the decay in the

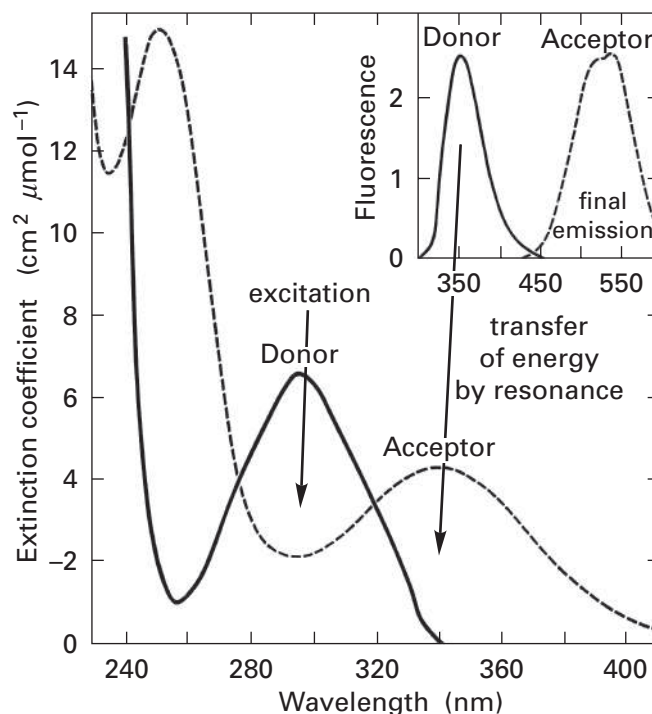


Figure 12–13: Absorption spectra and emission spectra (inset) of 1-acetyl-4-(1-naphthyl)semicarbazide (solid lines), a typical fluorescent donor for observing transfer of energy by resonance, and a matched acceptor, dansyl-L-prolylhydrazide (dashed lines), both dissolved in ethanol.¹⁴⁴ The measurements of absorbance are made by monitoring the intensity of the light of continuously varied wavelength that passes through each solution. The measurements of emission are made by following the intensity of the light emitted at 90° to an incident beam as a function of wavelength while the chromophore is excited with light of wavelength equal to that of its maximum of absorption. The spectrum of the amount of light absorbed is expressed as an extinction coefficient (centimeter² micromole^{−1}) as a function of wavelength (nanometers). The amount of light emitted is expressed as fluorescence (in relative units) as a function of wavelength (nanometers). When donor and acceptor are located near each other, a portion of the excited states of the donor would have their energy of emission at 350 nm transferred radiationlessly by resonance to the overlapping absorption band of the acceptor, and this transfer would quench the fluorescence of the donor. The transferred energy would be emitted as fluorescence at 540 nm from the excited acceptors. Adapted with permission from ref 144. Copyright 1967 National Academy of Sciences.

concentration of excited donor through the transfer of energy to the acceptor is $k_T f_A$.

The **efficiency of transfer** E_T is defined as

$$E_T \equiv \frac{k_T f_A}{k_L + k_F + k_T f_A} \quad (12-43)$$

This is the fraction of the decay of the excited state due to transfer of energy by resonance, or the ratio between the quanta transferred and the quanta absorbed by the donor.¹⁴⁶ The quantum yield Q_A of the fluorescence of the donor in the presence of the acceptor is governed by a relationship analogous to Equation 12–40

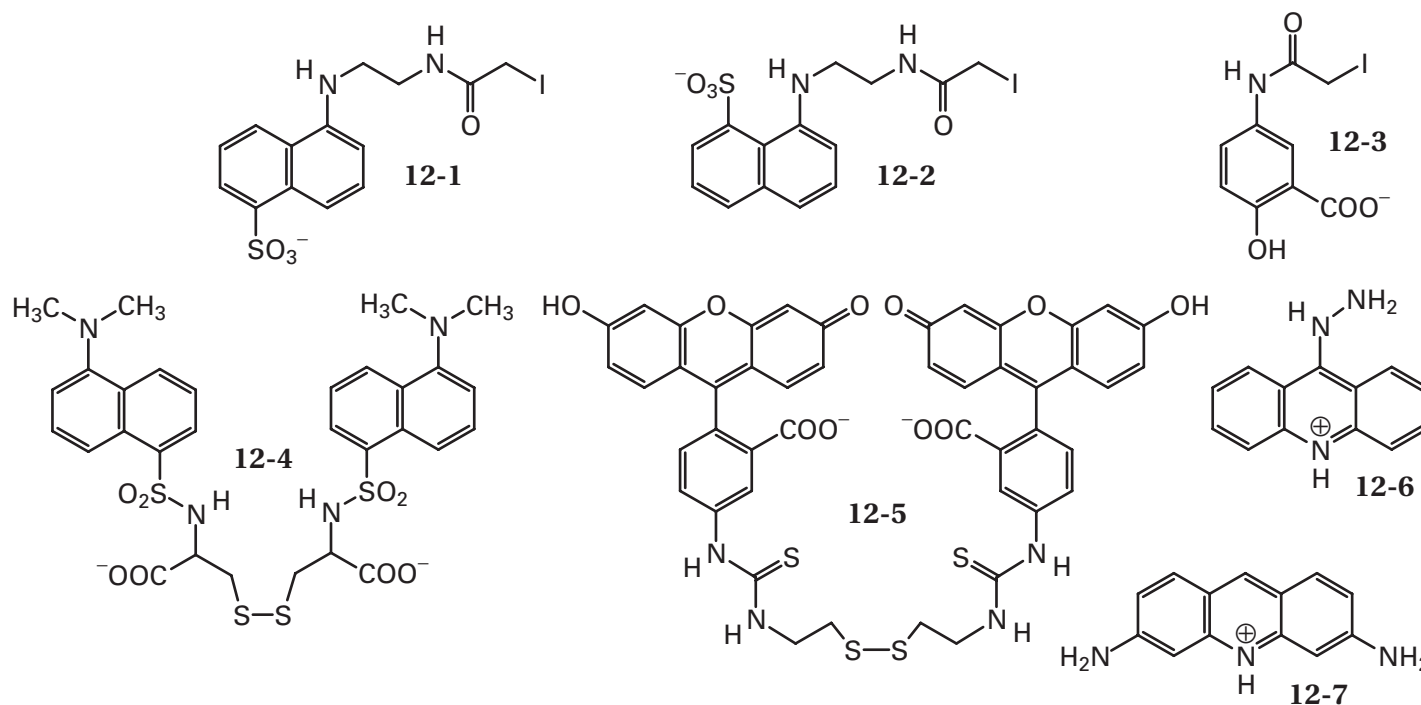


Figure 12-14: Fluorescent, electrophilic reagents used to modify, covalently or noncovalently, three sites on rhodopsin.¹⁴⁵ *N*-[(iodoacetamido)ethyl]-1-aminonaphthalene-5-sulfonate anion **12-1** ($\lambda_{\text{abs}} = 350$ nm; $\lambda_{\text{emit}} = 495$ nm), *N*-[(iodoacetamido)ethyl]-1-aminonaphthalene-8-sulfonate anion **12-2** ($\lambda_{\text{abs}} = 350$ nm; $\lambda_{\text{emit}} = 495$ nm), and 5-(iodoacetamido)salicylate anion **12-3** ($\lambda_{\text{abs}} = 323$ nm; $\lambda_{\text{emit}} = 405$ nm) were used to modify a particular cysteine in the protein by alkylation. *N,N'*-Bis[1-(dimethylamino)naphthalene-5-sulfonato]-L-cystine **12-4** ($\lambda_{\text{abs}} = 350$ nm; $\lambda_{\text{emit}} = 520$ nm) and *N,N'*-bis[fluoresceinyl(isothiocarbamido)]cystamine **12-5** ($\lambda_{\text{abs}} = 495$ nm; $\lambda_{\text{emit}} = 518$ nm) were used to modify a different cysteine in the protein by disulfide exchange. 9-Hydrazinoacridine **12-6** ($\lambda_{\text{abs}} = 440$ nm; $\lambda_{\text{emit}} = 470$ nm) and proflavin **12-7** ($\lambda_{\text{abs}} = 470$ nm; $\lambda_{\text{emit}} = 512$ nm) were used as ligands for a particular site on the protein with a high affinity for aromatic cations. All wavelengths (λ) are wavelengths of maximum absorption or maximum emission. In each instance the fluorescent functional group selectively attached to the protein was used as a donor of resonant energy to 11-*cis*-retinal, a natural, covalent posttranslational modification (Table 3-1) of the protein that absorbs maximally at 500 nm.

$$Q_A = \frac{k_F}{k_L + k_F + k_T f_A} \quad (12-44)$$

from which it follows¹⁴⁵ that

$$E_T = 1 - \frac{Q_A}{Q_0} = 1 - \frac{F_A}{F_0} \quad (12-45)$$

where F_A/F_0 is the ratio between the fluorescence observed in the presence and that observed in the absence of the acceptor.

The lifetime of the excited state τ_A in the presence of acceptor is governed by a relationship analogous to Equation 12-42

$$\tau_A = \frac{\tau_0}{1 + \tau_0 k_T f_A} \quad (12-46)$$

from which it follows that

$$E_T = 1 - \frac{\tau_A}{\tau_0} \quad (12-47)$$

The efficiency of transfer can be assessed by measuring either the decrease in steady-state fluorescence (F_A/F_0 ; Equation 12-45) or the **decrease in the lifetime of the excited state** produced by the acceptor (τ_A/τ_0 ; Equation 12-47), but the latter measurement is more accurate than the former. The efficiency is calculated from the emission or the lifetime of the donor in the presence of the acceptor and the emission or the lifetime of the donor in the protein that has not been modified with the acceptor or in which the acceptor has been bleached.¹⁴⁵

The efficiency of the transfer of energy, E_T , is determined by the distance r between the center of the transition dipole of the donor and the center of the transition dipole of the acceptor by the relationship

$$E_T = \frac{f_A r^{-6}}{f_A r^{-6} + R_0^{-6}} \quad (12-48)$$

where R_0 is the distance at which the efficiency of transfer would be 50% if the site for the acceptor were fully occupied. Consequently, the **distance between the centers of the dipoles of the donor and the acceptor**

$$r = R_0 \left[\frac{f_A(1 - E_T)}{E_T} \right]^{1/6} \quad (12-49)$$

In theory^{146,147}

$$R_0^6 = \frac{9 (\ln 10) K^2 Q_0 J}{128 \pi^5 \tilde{n}^4 N_A} \quad (12-50)$$

where Q_0 is the quantum yield of the donor (dimensionless), \tilde{n} is the refractive index of the medium between the donor and the acceptor (dimensionless), and N_A is Avogadro's number (moles⁻¹). The **overlap integral** J (centimeters⁶ mole⁻¹)^{145,146} is that between the fluorescence emission spectrum $I(\lambda)$ of the donor (in relative units) and the spectrum of the extinction coefficient $\varepsilon(\lambda)$ of the acceptor (in liters mole⁻¹ centimeters⁻¹) normalized by the total fluorescence of the donor

$$J = \frac{\int I(\lambda) \varepsilon(\lambda) \lambda^4 d\lambda}{\int I(\lambda) d\lambda} \quad (12-51)$$

where λ is the wavelength.* This integral quantifies the match between the energies of the donor and acceptor required for the resonance. The integral J is calculated numerically from the absorption spectrum of the acceptor and the emission spectrum of the donor (Figure 12-13), which should be gathered from donor and acceptor when they are in solution attached individually to the protein.

The **orientation factor** K^2 (dimensionless) is defined by

$$K^2 \equiv (\cos \theta_T - 3 \cos \theta_D \cos \theta_A) \quad (12-52)$$

where θ_T is the angle between the transition dipoles of donor and acceptor and θ_D and θ_A are the angles between the transition dipoles of the donor and acceptor, respectively, and the vector between the centers of those dipoles.¹⁴⁸ The transfer of energy is between these transition dipoles of the donor and acceptor, and this factor quantifies the match between the orientations of the donor and acceptor required for the resonance. If the orientations of the transition dipoles are not fixed

because the chromophores are free to adopt a number of different orientations, K^2 is the average of Equation 12-52 over those orientations.

The first requirement for every application of the transfer of energy by resonance is to have a protein in which a donor of energy and an acceptor of that energy are both located at defined positions within its structure. It turns out that measuring the transfer of energy is easy but placing the donors and acceptors at unique and exclusive locations on the protein is difficult. Often, either an **intrinsic donor or an intrinsic acceptor** or both, placed by evolution either covalently or noncovalently at a unique location on the protein, is relied upon to circumvent at least half of the difficulty.

Examples of such evolutionarily positioned donors and acceptors are fluorescent substrates or fluorescent analogues of substrates that bind to the active site of an enzyme, fluorescent ligands that bind to a specific site on a protein, and posttranslationally incorporated functional groups such as coenzymes that happen to be fluorescent or transition metal ions the complexes of which absorb at convenient wavelengths.¹⁴⁹ Tryptophans are often used as donors of resonant energy. Those found naturally in the protein can be used one by one as unique donors by preparing the respective site-directed mutants, each of which retains only one of them.¹⁵⁰ Nitrotyrosine¹⁵¹ or kynurenine,¹⁵² a covalent modification of tryptophan, can be used as acceptors of resonant energy from a tryptophan.

Often cysteines in the protein are used as convenient nucleophiles to be covalently modified by **fluorescent electrophilic reagents** (Figure 12-14). Sometimes, one of the cysteines in a protein, because of its peculiar reactivity, can be selectively modified with one fluorescent reagent and another cysteine can then be modified with another.^{153,154} Cysteines can be placed at specific positions in a protein by site-directed mutation and then selectively modified with appropriate fluorescent reagents.¹⁵⁵ A fluorescent reagent can be attached to a specific glutamine on the surface of a protein by use of the enzyme protein-glutamine γ -glutamyltransferase, which exchanges the ammonia of the glutamine with a primary amine on the reagent.^{24,156} It is also possible to use synthetically produced fluorescent α amino acids and in vitro systems for incorporating unnatural amino acids at specific positions in its amino acid sequence to produce a protein with a fluorescent functional group located at a single, designated point in its native structure.^{157,158}

The **bilayer of phospholipid** in which membrane-bound proteins are embedded also provides a location in which to locate a fluorescent donor or acceptor. The bilayer can be turned into a sheet of fluorescent donors or fluorescent acceptors by dissolving hydrophobic fluorophores^{159,160} in the liquid hydrocarbon at its center or by covalently attaching fluorophores to the phospholipids from which it is formed.¹⁶¹ Because the bilayer

* In the equation presented by Latt et al.,¹⁴⁶ there is a misleading factor of 1000 that serves to correct liters mole⁻¹ to centimeters³ mole⁻¹, a correction that would automatically be made during the cancellation of units. This is an excellent example of the absolute necessity of including units and making sure that they cancel properly whenever any calculation is performed in the physical sciences.

forms a sheet of hydrocarbon and because the molecules of a particular protein all float at the same depth within this sheet of hydrocarbon, the molecules of donor or acceptor dissolved within it end up in a fixed location relative to the rest of the protein. A matched acceptor or donor, respectively, can then be attached covalently to a specific location on the protein, and transfer of energy between the molecules of donor or acceptor within the bilayer and the acceptor or donor on the protein can be monitored.

The orientation factor K^2 is the most uncertain parameter in Equation 12–50.¹⁴⁵ In any given situation, K^2 has a specific numerical value but its value cannot be measured directly. If both donor and acceptor were free enough to assume all possible relative orientations with equal probability, K^2 would be $\frac{2}{3}$.¹⁴⁵ If one of the two were fixed and the other could assume all possible relative orientations, K^2 would have a value between $\frac{1}{3}$ and $\frac{4}{3}$.¹⁴⁵ If both donor and acceptor, however, are fixed in their relative orientations, for example, both rigidly bound to a molecule of protein, K^2 can have a value anywhere between 0 and 4.0. Because R_0 , and hence r , depends for its value on $(K^2)^{1/6}$, the uncertainty of K^2 affects the value of r by more than $\pm 12\%$ only when it is greater than $\frac{4}{3}$ but far more dramatically when it is less than $\frac{1}{3}$.

The more freedom the donor, the acceptor, or both of them have to assume different orientations by rotation around unhindered bonds between them and the rigid portion of the protein, the closer the value of K^2 comes to $\frac{2}{3}$. An estimate of the orientational freedom of donor or acceptor can be made from the rate and extent of the depolarization of its fluorescent emission. If either the donor or acceptor is excited with linearly polarized light, the light emitted as fluorescence immediately, that is, before the chromophore has had time to reorient, will also be polarized. The polarity of the emitted light, however, will decay over the lifetime of the excited state as it reorients. The rate of this decay and the final residual polarity of its fluorescence provide an estimate of the orientational freedom of the donor or acceptor.

From these **estimates of orientational freedom**, a distribution of the probability for particular values of K^2 can be calculated.¹⁶² For example, from the depolarization of the fluorescence from the 5-(*N,N*-dimethylamino)naphthalenesulfonyl group attached to rhodopsin as a donor to the retinal rigidly fixed in the center of the protein, the value for K^2 could be estimated to fall between 0.08 and 1.8 with a confidence of 90%,¹⁵⁶ and from the depolarization of the fluorescence from the pyrene group attached covalently as a donor to the active site of acetylcholinesterase and the depolarization of the fluorescence from the propidium bound as an acceptor at another site on the protein, the value for K^2 could be estimated to fall between 0.25 and 2.2.¹⁴⁸ For donor or acceptor or both to display depolarization, however, they must be reorienting fairly freely anyway, and such estimates of ranges for K^2 may not be significantly more

accurate than simply using $\frac{2}{3}$ for the value of K^2 . For example, predicted distances between four pairs of donors and acceptors positioned on specific amino acids in phosphoglycerate kinase from yeast, a protein for which a crystallographic molecular model is available, were no more reliable when K^2 was estimated from depolarizations than when $\frac{2}{3}$ was used for K^2 .¹⁵⁵

If both donor and acceptor are rigidly bound by the protein at a **fixed orientation** relative to each other, then neither will have any orientational freedom and no limits can be placed on K^2 other than from 0 to 4.¹⁶³ For example, in the crystallographic molecular model of deoxyribodipyrimidine photo-lyase from *Anacystis nidulans*, the angle between the transition dipoles of the flavin adenine dinucleotide and the 8-hydroxy-5-deazaflavin bound to the protein is 36° . From this angle and the angles of the dipoles to the vector between the two chromophores (Equation 12–52), a value of K^2 of 1.6 could be calculated.¹⁶⁴ For the same protein from *E. coli*, however, the angle between the transition dipoles of the flavin adenine dinucleotide and the methylene tetrahydrofolic acid, which takes the place of the 8-hydroxy-5-deazaflavin in the latter crystallographic molecular model, is almost 90° , causing K^2 to be almost 0. Even though the distances between the chromophores in these two proteins are the same (1.7 nm), the efficiency of the transfer of energy by resonance for the protein from *A. nidulans* is 97% while that from *E. coli* is 62%. When K^2 of $\frac{2}{3}$ was used to calculate the distance between the flavin adenine dinucleotide and the tetrahydrofolic acid in the protein from *E. coli*, in the absence of a crystallographic molecular model, the value obtained was 2.2 nm instead of the actual distance of 1.7 nm.¹⁶⁵

The efficiency of the transfer of energy by resonance between Tyrosine 14 and Tyrosine 55 in steroid Δ -isomerase from *Pseudomonas testosteroni* is less than 25% even though these tyrosines are only 0.6 nm apart in the crystallographic molecular model. Consequently, K^2 must be less than 0.003, a fact from which it was concluded that these two tyrosines were held rigidly by the protein in a relative orientation incompatible with efficient transfer even over such a small distance.¹⁶⁶ Had a value of $\frac{2}{3}$ been used for K^2 in the calculation, the distance between these two tyrosines would have been estimated to be greater than 1.5 nm.

The original enthusiasm for measurements of the transfer of energy by resonance was the potential it offered for **measuring the distance between two locations in a protein** the crystallographic molecular model for which is not available.¹⁴⁵ For example, if the donor were attached by the unique stoichiometric covalent modification of a particular amino acid in the sequence of the protein and the acceptor were specifically attached by the unique modification of another amino acid in the sequence, it would be possible to estimate the distance between the donor and acceptor in the folded polypeptide and hence the distance between the two modified

amino acids. In support of this intention, Equations 12–48 and 12–50 have been shown to be consistent with the observed transfers of resonant energy between a donor and an acceptor at the two ends of short synthetic peptides of proline.¹⁴⁴ The distance between donor and acceptor was varied by varying the number of prolines in the peptides to demonstrate that the dependence of efficiency upon distance was as the sixth power. If K^2 was assumed to be $\frac{2}{3}$, the calculated distances between donor and acceptor agreed fairly well (within 25%) with the distances measured from molecular models of these modified peptides. Many estimates of distances between locations in proteins have been made from measurements of the transfer of energy by resonance.

One way to evaluate the **reliability of such estimates** is to compare a distance estimated in this way with the distance observed in a subsequently obtained crystallographic molecular model. The distance between Cysteine 199 and Cysteine 343 in cyclic AMP-dependent protein kinase was estimated to be 3.1–5.2 nm on the basis of the transfer of energy by resonance between two different pairs of donor and acceptor,¹⁵⁴ but in the subsequently reported crystallographic molecular model¹⁶⁷ the distance between the sulfurs of these two cysteines is only 2.12 nm. The distance between the two Cysteines 283 in dimeric creatine kinase from rabbit muscle was estimated to be 4.8–6.0 nm from measurements of the efficiency of transfer for five different pairs of donor and acceptor,¹⁶⁸ but in the subsequently reported crystallographic molecular model of the protein,¹⁶⁹ these cysteines are only 3.33 nm apart. The distance between Lysine 84 on one of the subunits in an α_3 catalytic trimer and the closest Lysine 84 on a subunit in the other α_3 catalytic trimer in aspartate carbamoyltransferase (Figure 9–37) was estimated to be 3.3 nm on the basis of transfer of energy by resonance between a pyridoxamine phosphate and a pyridoxal phosphate attached to the respective side chains.^{170,171} This estimate conveniently splits the difference between the distances of 2.1 and 3.8 nm observed in subsequent crystallographic molecular models of the two respective conformations of the protein.^{172–174} The distance between the binding site for acetylcholine on acetylcholine receptor and the bilayer of phospholipids of the membrane in which it is located was estimated to be 3.0–4.0 nm from measurements of the transfer of energy by resonance between a donor covalently attached to choline and two different acceptors dissolved in the hydrocarbon of the bilayer,¹⁶⁰ and the distance to the closest surface of the bilayer estimated crystallographically is 3.0 nm.¹⁷⁵

A more suspect evaluation of the reliability of distances estimated from the transfer of energy by resonance are comparisons of them with those observed in a crystallographic molecular model available at the time the measurements were made. The distances estimated from the transfer of energy to chloramphenicol bound at the active site of chloramphenicol *O*-acetyltransferase

from Tryptophan 86 and Tryptophan 152, respectively, were both 1.5 nm when K^2 was set at $\frac{2}{3}$, and the distances in the crystallographic molecular model are 1.72 and 1.66 nm, respectively.¹⁵⁰ The distances estimated from the transfer of energy between the amino terminus and Lysines 15, 26, 41, and 46 in bovine pancreatic trypsin inhibitor were 3.4, 2.2, 2.1, and 2.3 nm, respectively, and the distances in the crystallographic molecular model are 3.17, 1.68, 1.80, and 2.17 nm, respectively.¹⁷⁶ The distance between Tyrosine 99 and Tyrosine 138 in the complex between calmodulin and four calcium ions was estimated from the transfer of energy to be between 1.4 and 1.9 nm,¹⁵¹ and the distance in the crystallographic molecular model is 1.2 nm.¹⁷⁷ The distance between a cysteine substituted for Phenylalanine 239 and Cysteine 343 in cyclic-AMP dependent protein kinase was estimated from the transfer of energy to be 4.1 nm,¹⁷⁸ and the distance in the crystallographic molecular model is 3.7 nm.¹⁷⁹

Aside from the uncertainty of the values of K^2 , one of the main difficulties in measuring distances by transfer of energy is that the donors and acceptors are often attached covalently to the protein by using reagents that end up placing the chromophore on a **flexible tether** a significant distance from the amino acid to which it is attached. For example, the reagents used to modify rhodopsin (Figure 12–14) place the centers of the chromophores 0.4–1.2 nm away from the electrophilic carbon or sulfur that is directly attached to the nucleophilic amino acid that has been modified. The fluorescent (5-sulfonaphthalen-1-yl) amino group and the fluorescent (7-nitrobenz-2-oxa-1,3-diazol-4-yl) amino group that were used as donor and acceptor, respectively, in estimates of distances between locations in the complex between DNA, deoxymononucleotide, and the Klenow fragment of DNA-directed DNA polymerase from *E. coli*, were attached to various atoms in the complex by tethers that were each about 1.2 nm in length.¹⁸⁰ If both donor and acceptor are attached through such long tethers, the distance between them can be significantly different from the actual distance between the two amino acids to which they are attached. One approach to adjusting the estimates of distance for these added lengths is to assume that the fluorescent functional group and its tether extend unrestrained outwards from the surface of the protein and correct for this extra distance geometrically.¹⁸¹ The difficulty with this approach is that the fluorescent functional group may adsorb to the surface of the protein or insert into a crevice on the surface.¹⁵³

Many of the failures of the measurements of distance to agree very closely with subsequent or even prior crystallographic determinations may result from technical shortcomings. Too frequently the necessary parameters such as quantum yield and spectral overlap are not measured directly but are based on prior published values. Measurements in the ultraviolet are often compromised

by contaminants in the solutions. It has already been mentioned that steady-state measurements of fluorescence are much less accurate than direct measurements of lifetimes of the fluorescence. Nevertheless, because of the uncertainties concerning the orientations of the dipoles of donor and acceptor, the degree of their orientational freedom, and the relationship of the distance between them and the distance between their points of attachment, because of the modest success of such estimates, and because crystallographic molecular models have become far more common, measurements of the transfer of energy by resonance are used infrequently to estimate distances. They are, however, still widely used for other purposes, because they have the advantage of providing information about a protein while it is in solution.

The transfer of energy by resonance is used to detect **conformational changes** in a protein. For example, the efficiency of the transfer of energy by resonance between Tryptophan 133 and a (5-sulfonaphthalen-1-yl)amino group attached through a tether to Cysteine 93 in dolichylphosphate β -D-mannosyltransferase increased from 42% to 66% upon the binding of the substrate dolichyl phosphate.¹⁸² From this observation, it was concluded that a conformational change occurs in the protein upon the binding of the substrate, and from the magnitude of the change in the efficiency of the transfer of energy, it was estimated that the distance between Tryptophan 133 and Cysteine 93 decreased about 0.3 nm during this conformational change. From similar observations, conformational changes producing shifts in the apparent positions of donors and their acceptors of 0.3, 0.3, 0.5, and 1.2 nm have been observed for the binding of DNA to transcription factor AP-1,¹⁸³ the exchange of Na^+ for K^+ in the active site of Na^+/K^+ -exchanging ATPase,¹⁸⁴ the exchange of Ca^{2+} cations for Mg^{2+} cations in troponin,¹⁸⁵ and the binding of a bisubstrate analogue to adenylate kinase, respectively.¹⁸⁶

For all of the same reasons, associating a **change in the distance between two locations** on a protein with the change in the transfer of energy by resonance is just as uncertain as estimating a distance between them. Upon the binding of the $\beta(1\rightarrow4)$ trimer of *N*-acetylglucosamine to lysozyme from chicken, the distance between a kynurenine at position 62 and Tryptophan 108 appeared to decrease by 0.5 nm when calculated from the change in the efficiency of transfer.¹⁵² It is unlikely, however, that such a large conformational change occurs on the binding of the oligosaccharide because no such differences (<0.04 nm) in the distance between these two amino acids is observed between the crystallographic molecular models of unliganded lysozyme and lysozyme to which the $\beta(1\rightarrow4)$ tetramer of *N*-acetylglucosamine is bound.¹⁸⁷ It is possible that crystal packing constrains both the liganded and unliganded conformations to the same structure in the crystal even though their structures are so different in solution. It is more likely, however, that the relative orientations of the donor or the acceptor or both are changed upon the association of the ligand, not

the distance between them. In the case of the sliding clamp of bacteriophage T4 DNA-directed DNA polymerase, however, it is thought that the ring of subunits composing the protein must split open so that a molecule of DNA can enter the hole in its center, and changes in the efficiency of the transfer of energy by resonance between donors and acceptors on different subunits equivalent to changes in distances of up to 1.5 nm are thought to reflect real changes in distance upon the opening of the ring and its subsequent intimate embrace of the DNA.¹⁸⁸

The transfer of energy by resonance is also used to monitor the association between a molecule of protein modified with a donor and another molecule of protein modified with an acceptor. The catalytic α subunit of cyclic AMP-dependent protein kinase has been modified with a tethered fluorescein and the regulatory β subunit with a tethered rhodamine. During the heterologous association of the two subunits, the fluorescence of the fluorescein decreases by about 30% as the rhodamine, an acceptor of the energy of its excited state, is brought into its vicinity.¹⁸⁹ Such assays based on the decrease in the fluorescence of a donor produced by an acceptor upon formation of a complex have been used to follow the association of cytochrome *c* and cytochrome-*c* oxidase¹⁹⁰ and the association of myosin and actin.¹⁹¹

Changes in the efficiency of the transfer of energy by resonance resulting from changes in the molar concentrations of the participants can also be used to measure the **dissociation constant** for a complex between two proteins or the complex between a protein and a nucleic acid. The dissociation constant between cytochrome *c* and cytochrome-*c* oxidase¹⁹⁰ and the dissociation constant between Rho-GDP dissociation inhibitor and GTP-binding protein Cdc42¹⁹² have been determined by monitoring changes in the efficiency of the transfer of energy, as has the equimolar stoichiometry of the complex between the sliding clamp and the clamp holder of DNA-directed DNA polymerase from bacteriophage T4.¹⁹³ Both the dissociation constant and the kinetics of the association between DNA and transcription factor AP-1^{183,194} have been monitored by changes in the transfer of energy by resonance.

Dissociation constants are also measured by monitoring changes in fluorescence that do not involve transfer of energy by resonance. For example, the enhancement of the fluorescence of poly(deoxy-1,*N*⁶-ethenoadenylic acid), a fluorescent analogue of poly(adenylic acid), upon the association of RecA protein from *E. coli* has been used to determine the dissociation constant and the kinetics of the association between the protein and the nucleic acid.¹⁹⁵

The transfer of energy by resonance has also been used to monitor changes in the structure of DNA produced by a protein. By modifying the immediately adjacent 3' end of one strand and 5' end of the other strand of a molecule of double-stranded DNA with a donor and an

acceptor, respectively, the increase in fluorescence of the donor when the two strands are dissociated has been used to monitor the unwinding of DNA catalyzed by ATP-dependent DNA helicase Rep from *E. coli*¹⁹⁶ and the exchange of one strand in a duplex of DNA for another catalyzed by RecA protein from *E. coli*.¹⁹⁷ By labeling short, double-stranded molecules of DNA with a donor at one end and an acceptor at the other, the decrease in the distance between donor and acceptor during the bending of the DNA caused by the binding of high mobility group protein Z from *Drosophila melanogaster* could be monitored.¹⁹⁸

Another manifestation of fluorescence that can be used to assess the proximity of two sites in a molecule of protein is the formation of an excited-state dimer or **excimer**. If upon its formation, the excited state of a suitable chromophore, for example, a pyrenyl group, is immediately adjacent to another one of the same chromophore, for example, another pyrenyl group, the excited state will form a dimer with its unexcited twin. Such a dimeric excited state, because of the greater opportunities for dissipation of the energy of the excited state radiationlessly, emits light of longer wavelength. For example, a pyrenyl group when excited at 344 nm emits light of wavelengths 378, 398, and 417 nm, but its excimer emits light of wavelength 470 nm. The observation of excimer fluorescence from a protein modified at two different amino acids with an appropriate chromophore is evidence that in the native structure of the protein those two amino acids are immediately adjacent to each other, even though they are distant from each other in its sequence.^{199,200}

Suggested Reading

Chi, Z., Chen, X.G., Holtz, J.S., & Asher, S.A. (1998) UV Resonance Raman-selective amide vibrational enhancement: quantitative methodology[sic] for determining protein secondary structure, *Biochemistry* 37, 2854–2864.

Pober, J.S., Iwanij, V., Reich, E., & Stryer, L. (1978) Trans-glutaminase-catalyzed insertion of a fluorescent probe into the protease-sensitive region of rhodopsin, *Biochemistry* 17, 2163–2169.

Problem 12–10: Hepatitis B virus produces severe inflammation of the liver. The mature, infectious virion is composed of three spherical, concentric layers. The outer layer is a continuous envelope of membrane the phospholipid bilayer of which came from the plasma membrane of the cell out of which the virion budded and the protein of which was encoded by viral DNA. The inner layer is a sphere of condensed, double-stranded viral DNA (3.2 kb) encoding the viral genome. The middle layer is a capsid enclosing the viral DNA. This viral capsid is an oligomeric protein composed of multiple copies of the same folded polypeptide, 183 amino acids in length. The amino acid sequence of this polypeptide is

```
MDIDPYKEFGATVELLSFLPSDFFPVSRDLLDTAAALYRD
ALESPEHCSPHHTALRQAILCWGDLMTLATWVGTNLEDPA
SRDLVVSYVNTNVGLKFRQLLWFHISCLTFGRETVLEYLV
SFGVWIRTPPAYRPPNAPILSTLPETTIVRRRGRSPRRRT
PSPRRRRSQSPRRRRSQSRESQC
```

The first 145 amino acids of the polypeptide fold to form the compact subunit that creates the viral capsid, and the last 40 amino acids interact with the viral DNA to provide counterions for the phosphodiester of the DNA and assist in condensing and packaging it.

A recombinant form of the gene encoding the viral capsid has been constructed for studies of its assembly. This recombinant gene, carried on the plasmid ptacHBc144, is under the control of the lacUV5 promoter and directs the expression in *Escherichia coli* of a polypeptide 154 amino acids long. The amino-terminal sequence of this polypeptide is TMITDSLEFH–, and the carboxy-terminal sequence is –IS. Between these two terminal sequences, which result from the cloning strategy, is the sequence of the polypeptide of the viral capsid from Isoleucine 3 to Proline 144. The expressed recombinant polypeptide lacks the carboxy-terminal portion of the subunit that interacts with the viral DNA. Nevertheless, it folds as it is being expressed in *E. coli*, and the resulting subunit then assembles, also within the bacterium, to form an oligomer similar to the viral capsid in the intact, infectious virion. This recombinant, empty, unenveloped viral capsid can be purified, and the purified protein is composed of only the one polypeptide, which is not post-translationally modified. This native, empty oligomeric protein is referred to as the HBe viral capsid.²⁰¹

The molar mass of the polypeptide of the HBe viral capsid at a net charge number of zero, calculated from the amino acid sequence of the expressed protein, is 17,369.9 g mol^{–1}.

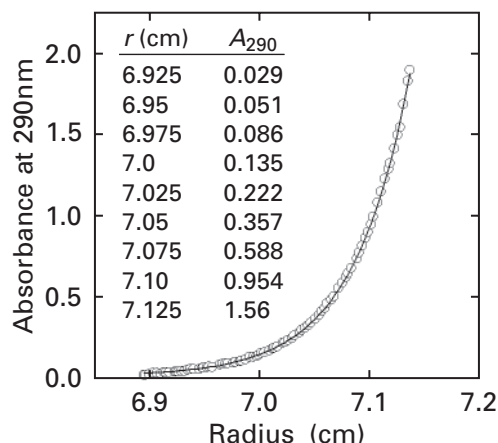
- The actual molar mass of a folded polypeptide differs from its molar mass at zero net charge depending on the conditions. Give reasons other than posttranslational modification for this difference.
- Estimate the magnitude of the effect of these factors (plus or minus how many grams mole^{–1}) on the molar mass of the subunit of the HBe viral capsid. What would be the appropriate choice of significant figures for expressing the molar mass of the subunit?

The HBe viral capsid was submitted to sedimentation equilibrium at 2800 rpm, 20 °C, in 0.15 M NaCl and 50 mM tris(*N*-hydroxymethyl)aminomethane hydrochloride, pH 7.0 ($\rho_{\text{sol}} = 1.00494 \text{ g cm}^{-3}$).

- Why was the sodium chloride present?

At sedimentation equilibrium, the absorbance at 290 nm (A_{290}) within the sample chamber showed the following dependence on the radial distance (r) from

the center of rotation.²⁰¹ Reprinted with permission from ref 201. Copyright 1995 American Chemical Society.



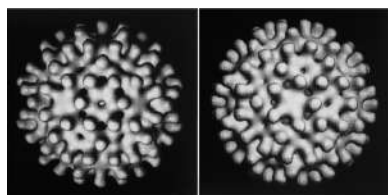
- (D) Calculate the molar mass of the HBe viral capsid. In this calculation, use the approximation of Equation 8-21. The partial specific volume of the protein, calculated from its amino acid composition, is $0.743 \text{ cm}^3 \text{ g}^{-1}$. Note that

$$\frac{d \ln C_{\text{prot}}}{dr^2} = \frac{d \ln \epsilon_{290}}{dr^2} + \frac{d \ln A_{290}}{dr^2} = \frac{d \ln A_{290}}{dr^2}$$

where ϵ_{290} is the extinction coefficient for the protein at 290 nm.

- (E) On what measurements of the concentration of protein does this calculation of molar mass rely?

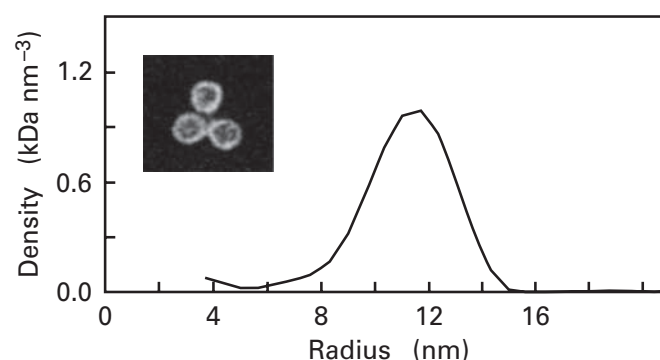
A similar recombinant version of the viral capsid of hepatitis B virus has been embedded in amorphous ice, and electron micrographs of that protein have been submitted to image reconstruction. The following are two views of that reconstruction.²⁰² Reprinted with permission from ref 202. Copyright 1994 Elsevier B.V.



- (F) What is the symmetry of the capsid of hepatitis B virus? How many folded polypeptides are in each protomer?
- (G) Exactly how many folded polypeptides are there in the HBe viral capsid? What is its exact molar mass?
- (H) The standard sedimentation coefficient at 20 °C in water, $s_{20,w}^0$, of the HBe viral capsid is 44 S. Calculate its frictional coefficient.

- (I) On the basis of this frictional coefficient, if the viral capsid were a smooth, unhydrated sphere, what would be its radius?

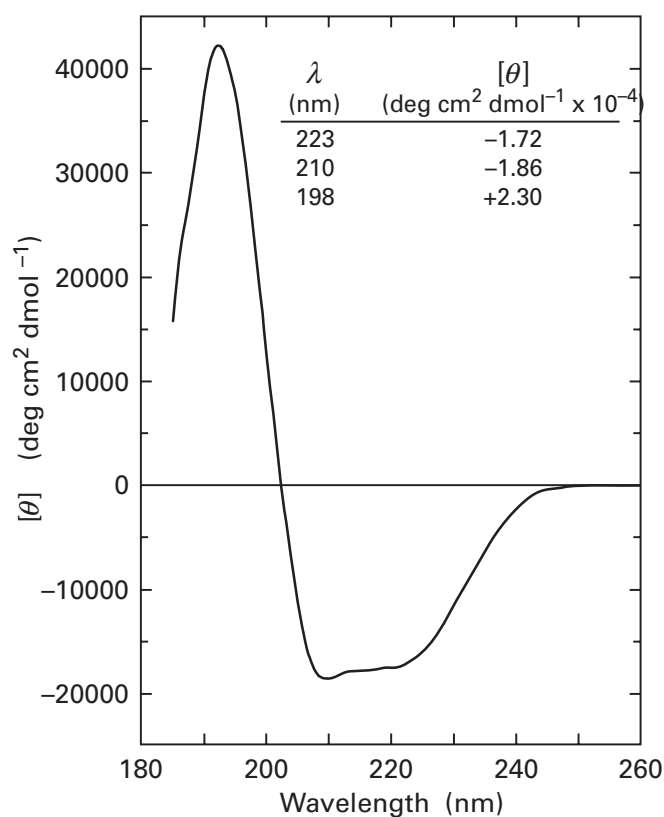
The HBe viral capsid was examined by scanning transmission electron microscopy. The inset in the figure is a selected field from one of these micrographs containing three HBe viral capsids.²⁰¹ Reprinted with permission from ref 201. Copyright 1995 American Chemical Society.



Each image obtained in this procedure is a projection of the three-dimensional molecule onto a two-dimensional plane. The image is in negative contrast so the protein appears as a light object on a dark background, and the degree of contrast at any point in the plane is directly proportional to the amount of protein contributing to that projected point. These images can be scanned, and the two-dimensional distribution of contrast can be converted into a radial distribution of density if it is assumed that the viral capsid is spherically symmetric. A graphical representation of this radial distribution of density in kilodaltons nanometer⁻³ is presented in the figure.

- (J) Why does the density of protein decrease gradually beyond a radius of 12 nm rather than abruptly as it would if the protein were a smooth sphere?
- (K) On the basis of this distribution of protein density, what is the maximum value for the radius of the HBe viral capsid?
- (L) What would be the frictional coefficient for a spherical molecule of protein with this radius?
- (M) What reasons could explain why the actual frictional coefficient of the HBe viral capsid is larger than this maximum theoretical frictional coefficient?

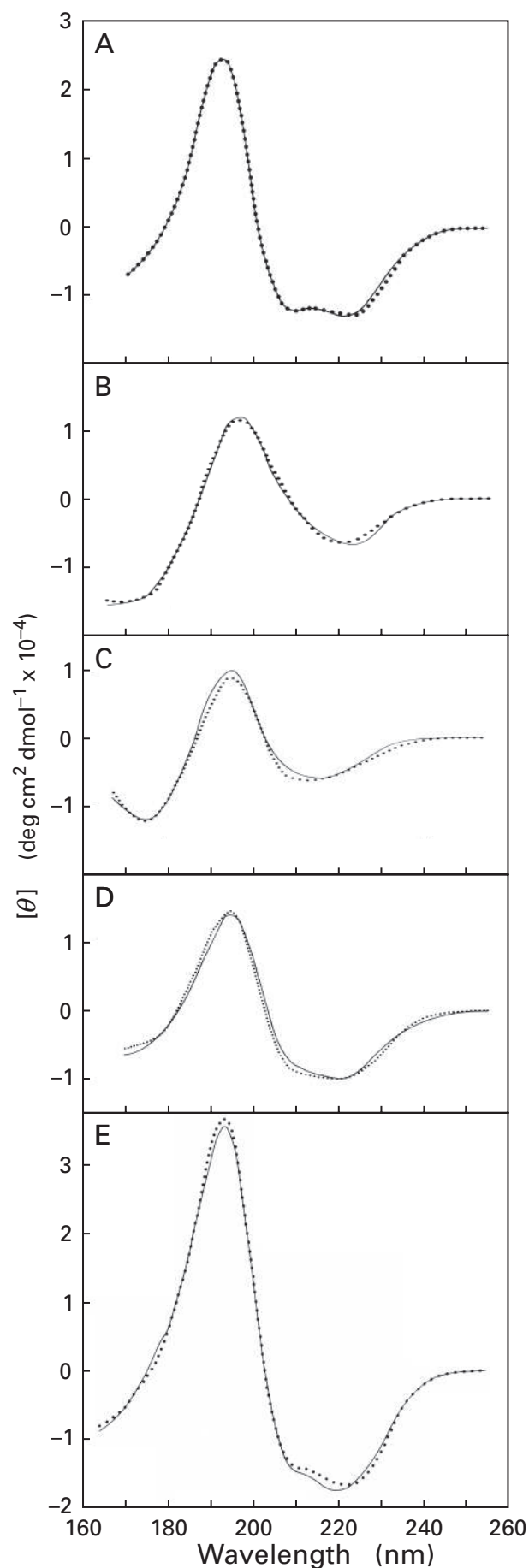
The following is the circular dichroic spectrum of the HBe viral capsid and a list of the values of the molar ellipticity [degrees centimeter² (decimole of peptide bond)⁻¹] at three selected wavelengths.²⁰¹ Reprinted with permission from ref 201. Copyright 1995 American Chemical Society.



The values for the molar ellipticities at these same three wavelengths for the three standard curves in Figure 12-10, which represent pure α helix, pure β structure, and pure random meander, are

λ (nm)	$[\theta]_{\alpha}$ (deg cm ² dmol ⁻¹ × 10 ⁻⁴)	$[\theta]_{\beta}$ (deg cm ² dmol ⁻¹ × 10 ⁻⁴)	$[\theta]_{RM}$ (deg cm ² dmol ⁻¹ × 10 ⁻⁴)
223	-2.75	-0.97	+0.31
210	-2.67	-0.78	-0.78
198	+3.85	+3.85	-3.96

- (N) Assume that the folded polypeptide of the HBe viral capsid contains only α helix, β structure, and random meander, and formulate a set of three simultaneous equations relating the three unknowns f_{α} , f_{β} , and f_{RM} , the respective fractions of each type of secondary structure in the protein. Do not use as an equation the assumption that the sum of these three fractions is 1.
- (O) Solve this set of equations for f_{α} , f_{β} , and f_{RM} .
- (P) Why is it so surprising that the values of f_{α} , f_{β} , and f_{RM} add up to 1 anyway?
- (Q) On the basis of these numerical values of f_{α} , f_{β} , and f_{RM} , why is one required to conclude that the viral capsid of southern bean mosaic virus (Figure 9-27) and the viral capsid of hepatitis B virus do not share a common ancestor even though their overall structures are similar?



Problem 12–11: The circular dichroic spectra of several proteins are shown to the left.¹¹⁷ The solid lines are the observed spectra; the dotted lines are theoretical fits to the data. Reprinted with permission from ref 117. Copyright 1980 Academic Press.

- Using the letters in each panel to designate each protein, rank them in order from the one with the most α helix and the least β structure to the one with the most β structure and least α helix.
- Consider the protein with the most α helix. What would be the maximum percentage of α helix it could have?
- Consider the protein with the least α helix. What would be the maximum percentage of β structure it could have?

Problem 12–12: Suppose that $R_0 = 1.7$ nm for the transfer of energy by resonance between a donor and an acceptor on a protein and that the efficiency of the energy transfer between donor and acceptor is 0.79. When a ligand that binds to the protein is added, the efficiency of energy transfer decreases to 0.64. What is the apparent change in the distance between donor and acceptor that occurs upon binding of the ligand?

Nuclear Magnetic Resonance^{203,204}

Many atomic nuclei display rotational motion known as **nuclear spin**. Because nuclear spin is quantized, its angular velocities can assume only those magnitudes dictated by **spin quantum numbers**. Among many other atomic nuclei, those of ^1H , ^{13}C , ^{15}N , ^{19}F , and ^{31}P have only two spin quantum numbers, $+\frac{1}{2}$ and $-\frac{1}{2}$. These dictate two specific angular velocities of the same magnitude but of opposite polarity. These two angular velocities are the two **spin states** of these nuclei. As any one of these nuclei is a charged particle by virtue of its protons, either of these angular velocities creates a magnetic field of the respective polarity aligned with the axis of the nuclear spin. When such a nucleus is placed in an **external, homogeneous magnetic field** of a given polarity, its axis tends to align with the direction of the applied field, and its spin states, because they are of opposite polarity to each other, become different in energy. This **difference in energy**, ΔE , is directly proportional to the **magnetic flux density** B_i (tesla) at the location of nucleus i ; and as in optical spectroscopy, the difference in energy determines the frequency ν_i (hertz) of electromagnetic energy that is absorbed by nucleus i :

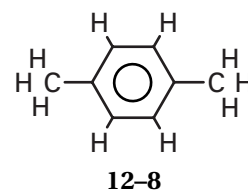
$$\Delta E = \frac{\gamma_i h B_i}{2\pi} = h\nu_i \quad (12-53)$$

where h is Planck's constant and γ_i is the magnetogyric ratio (radians tesla⁻¹ second⁻¹) for nucleus i , which is

determined only by the type of nucleus, ^1H , ^{13}C , ^{15}N , ^{19}F , or ^{31}P , that it is (Table 12–2). The frequency ν_i at which nucleus i absorbs in an applied external field is its **Larmor frequency**.

At readily accessible magnetic flux densities (<25 T), the difference in energy between the two spin states of one of these nuclei is less than 0.5 J mol^{-1} , which is the energy contained in a photon of wavelength greater than 3×10^8 nm and frequency less than or equal to 1000 MHz. This is in the **radiofrequency range of electromagnetic energy**.

A solution of molecules contains discrete **populations of atomic nuclei** in which each and every nucleus is chemically identical. For example, if the naturally present deuterium is ignored, a solution of *p*-xylene



uniformly and completely labeled with ^{13}C carbon would contain one discrete population of ^1H hydrogen nuclei composed from the four hydrogens attached to the phenyl ring in each molecule of *p*-xylene, a discrete population of ^1H hydrogen nuclei composed from the six hydrogens attached to the methyl groups in each molecule of *p*-xylene, a discrete population of ^{13}C carbon nuclei composed from the carbons of the methyl groups, a discrete population of ^{13}C carbon nuclei composed from the para carbons of the ring, and a discrete population of ^{13}C carbon nuclei composed from the meta carbons of the ring. The nuclear spin of each nucleus in a given population of nuclei can be represented as a vector of unit length parallel to the axis of its spin. The **net magnetization** of a given population of nuclei is the vector sum of its individual nuclear spins. In an applied magnetic field, although the individual nuclear spins show only a tendency to align with the field, the net magnetization of each population of nuclei is aligned exactly with the direction of the field. In an applied magnetic field, each of these discrete populations of nuclei in the solution has a corresponding Larmor frequency for the nuclear magnetic resonance absorption associated with it.

When a population of chemically identical fundamental particles, such as electrons or atomic nuclei, is exposed to electromagnetic radiation of a wavelength equivalent in energy to the difference between two energy levels accessible to the particles, the electromagnetic radiation catalyzes the movement of the members of that population of particles between these two energy levels. The reason is that the electromagnetic radiation is in **resonance** with the transition between the two energy levels. Photons with this energy will be absorbed during this process only if, at the time of irradiation, the popu-

lation of particles occupying the lower energy level is greater than the population occupying the higher energy level. The absorption of photons, however, necessarily increases the population in the higher energy level at the expense of the population in the lower energy level. When the populations in the two resonating energy levels become equal to each other, absorption can no longer occur, and a state of **saturation** is reached.*

In the electronic transitions and vibrational transitions of electrons in atomic and molecular orbitals (Figure 12-5), the energy levels are sufficiently different that almost all unexcited electrons are in the state of lower energy, and relaxation back to the state of lower energy is sufficiently rapid ($>10 \text{ ns}^{-1}$)⁸⁶ that absorption of a particular wavelength of light by a given population of chemically identical electrons rarely displays saturation. In nuclear magnetic resonance, however, the energy difference between the two spin states that can be achieved with the available magnetic flux densities is so small ($<0.5 \text{ J mol}^{-1}$) that the equilibrium constant K_{sp} between the two spin states for a population of identical nuclei at normal temperatures (300 K) is very close to 1 ($1 < K_{\text{sp}} < 1.0002$). This means that the difference in the concentrations of the nuclei in the two spin states at equilibrium will be less than 200 ppm. The difference between the populations in the two energy levels set in resonance is small enough and the **rate of relaxation** of a nucleus in the level of higher energy back to the level of lower energy is slow enough ($\geq 1 \text{ s}^{-1}$) that saturation occurs readily. This causes the amplitude of the observed absorption of the electromagnetic energy in nuclear magnetic resonance spectroscopy to be sensitive to the rate of relaxation of the populations of individual nuclei from the state of saturation to the state of equilibrium. The faster the population relaxes, the more energy it can absorb. For this relaxation to occur, the excess energy that has been absorbed has to be dissipated.

The **chemical shift** of the nuclear magnetic resonance absorption of a population of nuclei is a measure of the frequency of the electromagnetic energy at which the absorption appears in the spectrum. The chemical shift of the absorption of a particular population of nuclei is determined by the chemical environment of the chemically identical nuclei that compose the population. The electrons surrounding a given nucleus circulate in response to the applied magnetic field as a current would in a copper coil. This current decreases the local magnetic flux density B_i experienced by the nucleus and establishes its characteristic Larmor frequency (Equation 12-53). The chemical shift δ_i of a nuclear magnetic resonance absorption from a population of chemically iden-

tical nuclei i is the normalized difference between its Larmor frequency ν_i , the frequency at which it absorbs, and the frequency ν_{std} at which the population of a standard nucleus absorbs:

$$\delta_i = \frac{\nu_{\text{std}} - \nu_i}{\nu_{\text{std}}} = \frac{B_{\text{std}} - B_i}{B_{\text{std}}} \quad (12-54)$$

The units used for chemical shift are **parts per million** (ppm) relative to the absorption of the standard nucleus in a particular reference compound because the local differences in magnetic flux density are never greater than about 0.0002 (200 ppm) of the applied field. Chemical shift cannot be expressed in absolute units of energy because the energy difference between a particular absorption and that of the standard varies with the magnitude of the applied field (Equation 12-53). The magnitude of the chemical shift provides chemical information about the disposition of the electrons in the environment surrounding the nucleus, in other words, the molecular structure in its vicinity.

Nuclear magnetic resonance spectroscopy measures the same phenomenon as optical spectroscopy. In an external magnetic field every nucleus of spin $\frac{1}{2}$ has two energy levels. Depending on the flux density of the applied magnetic field and the type of nucleus, electromagnetic energy of a discrete wavelength (frequency) somewhere between $2 \times 10^{10} \text{ nm}$ (15 MHz) and $3 \times 10^8 \text{ nm}$ (1000 MHz) will be absorbed by a particular population of identical nuclei in the process of exciting nuclei in the population of the spin state with lower energy to the spin state with higher energy. In theory, these absorptions of energy by each discrete population could be recorded as a function of wavelength to obtain a spectrum, as is done with an optical absorption spectrum. **Continuous wave (CW) nuclear magnetic spectrometers** approximate this ideal. They measure the absorption of energy of a fixed frequency as the flux density of the applied magnetic field is varied slowly and continuously, and they record the variation in the intensity of the radiofrequency signal as it is absorbed by the sample. This measurement produces a scan of absorption as a function of the flux density of the magnetic field. The direct proportionality between magnetic flux density and frequency (Equation 12-53) permits the spectrum of absorption to be presented as a function of frequency. Maxima of absorption appear in the spectrum at the Larmor frequencies of the different populations of nuclei in the sample. Nuclei of the different elements differ dramatically in their ability to absorb radiofrequency energy (Table 12-2) and hence the intensities of their maxima.

Almost all instruments used today, however, are **Fourier transform (FT) nuclear magnetic resonance spectrometers**. In such a spectrometer there is a radio-transmitter that generates radiowaves of a set frequency, for example 600 MHz, referred to as the **carrier fre-**

* Because the absorption of electromagnetic energy is the consequence of resonance and because the resonance is so much closer to equilibrium in nuclear magnetic resonance spectroscopy than in optical spectroscopy, nuclear magnetic resonance spectroscopists often use the word "resonance" in place of the word "absorption".

Table 12-2: Nuclear Properties²⁰³ of Nuclei with Spin Quantum Number $\frac{1}{2}$

nucleus	magnetogyric ratio ($\text{rad T}^{-1} \text{s}^{-1} \times 10^8$)	frequency of maximum absorption at 1 T (MHz)	relative amplitude of absorption at constant field
^1H	2.675	42.577	1.000
^{13}C	0.673	10.705	0.016
^{15}N	-0.271	4.315	0.001
^{19}F	2.517	40.055	0.834
^{31}P	1.083	17.231	0.066

quency ν_0 . The radiowaves generated by the radiotransmitter are propagated in a direction x perpendicular to the direction z of the fixed magnetic field so that the magnetic fields of the radiowaves oscillate in the dimension y at the carrier frequency. The net magnetization of a given population of nuclei is precisely aligned with the z axis.

The magnetic field of the instrument is adjusted (Equation 12-53) so that the span of the Larmor frequencies of the various populations of nuclei to be examined, for example, all of the populations of ^1H hydrogen in the sample, is centered on the carrier frequency. The sample is excited with a strong (50 W) pulse of radiowaves at the carrier frequency. The pulse, however, is so short (5–50 μs) that a set of radiowaves is produced of almost equal intensity within a continuous range of frequencies encompassing the Larmor frequencies of the various populations of nuclei. If the proper length of the pulse is chosen by trial and error, the net magnetization of each of the populations of nuclei will be diverted from its alignment with the applied field in the z direction to an alignment with the oscillating magnetic field of the pulse in the y direction. A pulse of the proper duration is a **90° pulse**, referring to the fact that it has moved all of the net magnetizations by 90°. The 90° pulse is strong enough to saturate by resonance each population of nuclei as well as divert the alignment of its net magnetization.

At the end of the pulse, when only the applied field in the z direction remains, the net magnetization of each population of nuclei begins to precess in the x,y plane around the z axis at its Larmor frequency. As they lose their saturation, each of the various populations of these **precessing nuclei** emits a radio signal at its Larmor frequency, just as a population of excited electrons emits fluorescent light of a frequency equivalent to the difference in energy between the excited state and the ground state upon its relaxation. Because this emission decays as the saturation is lost both through the emission itself and through normal relaxation, it is referred to as a **free induction decay**. The free induction decay, which is registered from the conclusion of the 90° pulse until it has relaxed to nothing, is the only direct measurement made

by the spectrometer and must contain all of the data necessary to produce a spectrum. The free induction decay is registered by a radio receiver that is tuned to the carrier signal so that the output produced by the receiver is the modulations of the frequency of the carrier signal that each population of nuclei produces. Consequently, the signal that is registered from each population by the receiver is at the frequency of the difference between its Larmor frequency and the carrier frequency, which happens to be proportional to its chemical shift from the carrier frequency (Equation 12-54). After it has been submitted to a suitable linear combination of its real and imaginary terms, the **Fourier transform** of the total free induction decay from the receiver, which is the sum of all the free induction decays from all of the populations of nuclei, has positive maxima at the chemical shifts of each of the populations of nuclei relative to the carrier frequency. This Fourier transform reproduces a continuous wave spectrum of absorption from the same sample.

The absorption of a population of chemically identical nuclei is usually split into a series of peaks producing a symmetrical pattern around its mean resonant frequency. This splitting is due to **spin-spin coupling**, also referred to as J coupling. It arises from the fact that any adjacent spinning nucleus A, covalently linked to the nucleus X being observed, acts as a small magnet that induces the electrons between it and the nucleus X to circulate. This induced current alters the local magnetic field B_i at the nucleus X. Within the whole population of molecules, the various nuclei A assume both of their spin states randomly and almost equivalently, but each particular nucleus X in the population can be spin-coupled to a particular nucleus A in only one of those two spin states. This divides the population of nuclei X into two different groups, each group having each of its corresponding nuclei A in only one of its two available spin states.

The spin-spin **coupling constant**, J_{AX} , is the magnitude of the magnetic effect of nucleus A on nucleus X. Because spin-spin coupling is a function only of the intrinsic magnetic fields of the neighboring nuclei, coupling constants are not a function of the magnitude of the applied field, and their values (which are invariant differences in energy) are expressed quantitatively as the number of hertz by which the magnetic field of nucleus A splits the frequency of the absorption of nucleus X into two peaks of different frequency. Because spin-spin coupling is relayed by the electrons in the covalent bonds connecting the nuclei, it decreases in magnitude with the number of bonds separating them. For example, the values of J for the spin-spin coupling of ^1H hydrogen nuclei through two bonds can be as large as 15 Hz, and for coupling of ^1H hydrogen nuclei through three bonds, as large as 12 Hz, but for coupling of ^1H hydrogen nuclei through four bonds, the values are less than 1 Hz.

In addition to being determined by the number of bonds separating the two coupled nuclei, the value of the

spin-spin coupling constant J also depends on the angles at which two nuclei are held with respect to each other. For example, for two ^1H nuclei coupled through two bonds, the range of values is 10–15 Hz if the bond angle is close to the sp^3 tetrahedral angle of 109° , but the value of the coupling constant falls to 2–3 Hz at the sp^2 angle of 120° . In three-bond coupling, the value of the spin-spin coupling constant depends on the **dihedral angle** between the two nuclei along the bond connecting the two neighboring atoms to which they are attached:



12-9

The coupling constant J_{AX} is at its maximum when the dihedral angle is 0° or 180° and at its minimum when the dihedral angle is 90° or 270° . At these latter angles, the spin-spin coupling constant can be almost zero. The maxima at 0° and 180° for such coupling between the nuclei of two ^1H hydrogens is about 10 Hz, and between the nuclei of a ^1H hydrogen and a ^{13}C carbon, about 8 Hz.²⁰³

Two populations of nuclei, X and A, can also be coupled by a nuclear Overhauser effect. A **nuclear Overhauser effect** of the population of nuclei A on the population of nuclei X is any change in the net spin state of the population of nuclei X produced by a change in the net spin state of the population of nuclei A. For example, a nuclear Overhauser effect can be the consequence of either an alteration in the relaxation rate between the two spin states accessible to the members of the population of nuclei X or the consequence of an alteration in the levels of occupation of the two spin states within the population of nuclei X caused by a change in the net spin state of the population of nuclei A. A change in the net spin state of the population of one nucleus produced by a change in the net spin state of the population of another nucleus results from dipolar interactions between the two respective nuclei. A **dipolar interaction** is a function of, among other things, the distance between the two nuclei, r , and its magnitude is proportional to r^{-3} . The change in the spin state of nucleus X by nucleus A caused by a dipolar interaction is a second order perturbation, and hence it is proportional to r^{-6} . The transfer of energy between two transition dipoles by resonance is also a dipolar interaction and has the same dependence on distance (Equation 12-48). Because of the inverse dependence on the sixth power of the distance, nuclear Overhauser effects are significant only if the nucleus X and the nucleus A in a particular molecule are close to each other.

Because a nuclear Overhauser effect is not transmitted by changes in the static, local magnetic field of nucleus X brought about by a change in the spin state of

nucleus A, there is no requirement that the two nuclei be associated by covalent bonds as there is with spin-spin coupling. Nuclear Overhauser effects can indicate that the two nuclei involved are adjacent to each other in the tertiary structure of a protein even though they may be distant from each other in its primary structure. As with the transfer of energy by resonance, however, there are factors other than the distance between the nuclei associated with the dipolar interactions producing nuclear Overhauser effects, causing the intensity of a nuclear Overhauser effect not to be directly proportional to the inverse of the sixth power of this distance.²⁰⁵

Because nuclear Overhauser effects are manifested only as changes in the net spin state of the population of one nucleus under the influence of a change in the net spin state of the population of the other, no change occurs in the chemical shift of either nucleus involved in the nuclear Overhauser coupling as there was with spin-spin coupling. Rather, a nuclear Overhauser effect is registered as a **change in the intensity of the absorption** for nucleus X. For example, if the net rate of relaxation of the population of nucleus X is increased by the change that has occurred in the net spin state of the population of nuclei A, then the amplitude of the absorption for nucleus X will increase. If the occupation of the two spin states available to the members of the population of nuclei X is caused to become more equal by the change that has occurred in the net spin state of the population of nuclei A, then the amplitude of the absorption for nucleus X will decrease.

In large, relatively rigid macromolecules such as proteins, nuclear Overhauser effects between ^1H hydrogen nuclei are usually a consequence of the transfer of saturation.²⁰⁶ **Transfer of saturation** is the transfer of a portion of the saturation of one population of nuclei, nuclei A, to another population of nuclei, nuclei X. For transfer of saturation to occur, each of the individual nuclei X must be adjacent in space to a nucleus A. Transfer of saturation results from the summation of a large number of individual exchanges of spin state between a nucleus A and a nucleus X. The two adjacent nuclei simultaneously and reciprocally exchange their spin states in opposite directions with essentially zero change in the total energy of the two exchanging nuclei. During each exchange, the spin state of that particular nucleus X becomes what was the spin state of the particular nucleus A and vice versa. As a result of a large number of such exchanges of spin state at the atomic level, a portion of the saturation of the population of nuclei A is transferred to the population of nuclei X. The driving force is the resulting increase in entropy. A sequence of such transfers of saturation among a number of populations of adjacent nuclei can cause the saturation of one population of nuclei to spread outward over populations of nearby nuclei.

Spin diffusion is this outward spread of saturation from the saturated population of nuclei A. For spin diffu-

sion to occur, the populations of the nuclei in the vicinity of the nuclei A must be unsaturated so they are able to assume the saturation transferred from the population of nuclei A. It is the saturation of only the population of nuclei A that permits the diffusive force to be observed, just as the creation of a gradient of concentration permits the diffusive force to be observed.

Spin diffusion by transfer of saturation can be observed in an experiment analogous to the transfer of energy by resonance. The population of nuclei A is irradiated at the radio frequency with which its spins resonate and with sufficient amplitude to saturate its absorption, which equalizes the number of nuclei in its two spin states. The stimulating radiation is then turned off. The population of nuclei A will slowly relax back to its equilibrium distribution by losing the excess energy it has gained. One of the ways the population of nuclei A may relax is by transferring saturation to the population of nuclei X, if within the molecule nucleus X is close to nucleus A. If the absorption of the population of nuclei X is measured after a time, t_m , sufficient for some of the saturation in the population of nuclei A to be transferred to the population of the nuclei X, the absorption of the population of nuclei X will have decreased relative to its absorption in the absence of transfer of saturation because the population of nuclei X will have been moved closer to saturation.

An example of a nuclear Overhauser effect that was the consequence of spin diffusion was observed during a spectroscopic study of cytochrome *c* from *Katsuwonus pelamis*.²⁰⁷ The heme in cytochrome *c*, as it is a large aromatic ring (2-4), produces a substantial toroidal magnetic field when its π electrons circulate as a **ring current** in the presence of the applied magnetic field. The δ_2 methyl group of Leucine 68 (Figure 7-9C) resides adjacent to the heme in the region of this local field that is opposed to the applied field, and the chemical shift (-2.7 ppm) of the absorption of its three equivalent ^1H hydrogens is even less than that of the reference absorption. This substantial displacement isolates this peak of absorption from the absorptions of the rest of the methyl ^1H hydrogens in the protein. When the population of the δ_2 methyl ^1H hydrogens on Leucine 68 was saturated by preirradiation at the frequency of its chemical shift, the absorptions of four other populations of ^1H hydrogens were found to decrease. These were assigned to ^1H hydrogens neighboring the δ_2 methyl group of Leucine 68 in the crystallographic molecular model of the protein.

The initial nuclear magnetic resonance spectra of proteins were of ^1H hydrogen nuclei in molecules dissolved in $[^2\text{H}]\text{H}_2\text{O}$. They were **one-dimensional spectra of absorption** as a function of chemical shift. Even a small protein of 100 amino acids has more than 700 hydrogens in it, most of them unique and most of their absorptions split by spin-spin coupling. It is not surprising, therefore, that such spectra contain, by and large, several broad, unresolved absorptions, each resulting from the overlap

of hundreds of individual absorptions.²⁰⁸ The ranges in which these overlapping absorptions occur can be assigned to particular classes of nuclei: those of methyl ^1H hydrogens on leucines, isoleucines, valines, alanines, and threonines ($\delta = 0.9\text{--}1.5$ ppm); methylene ^1H hydrogens ($\delta = 1.5\text{--}3.5$ ppm); α ^1H hydrogens on each amino acid ($\delta = 3.5\text{--}5.5$ ppm); the ^1H hydrogens on the peripheries of the aromatic rings of tryptophans, phenylalanines, histidines, and tyrosines ($\delta = 6.4\text{--}7.4$ ppm); and the unexchanged amido ^1H hydrogens of the peptide bonds and glutamines and asparagines ($\delta = 7.0\text{--}9.0$ ppm).

The central difficulty in nuclear magnetic resonance spectroscopy of even a small protein is that in a one-dimensional spectrum of absorption as a function of chemical shift, regardless of the nucleus chosen, the peaks of absorption from the individual nuclei overlap and cannot be distinguished from each other, let alone assigned. It has been possible, however, to dissect the nuclear-magnetic resonance spectra of small molecules of protein^{209,210} into their individual components by using the **two-dimensional spectroscopy** that has been developed by Ernst and his colleagues^{211,212} and the **three-dimensional spectroscopy** that has been developed by Bax and his colleagues (Table 12-3).

All of the techniques of multidimensional nuclear magnetic resonance spectroscopy rely upon the technique of **frequency labeling**, which is a direct elaboration of Fourier transform nuclear magnetic spectroscopy. To label the absorption of a population of nuclei *i* with its Larmor frequency, two successive 90° pulses are used. Following the first 90° pulse applied in the *x* direction, the net magnetization of the population of nuclei *i* is aligned with the *y* axis and then begins to precess in the *xy* plane around the *z* axis at its Larmor frequency. After a period of time t_1 , a second 90° pulse in the *x* direction is applied. The second pulse is in phase at the carrier frequency ν_0 with the first pulse to insure that

$$\nu_0 t_1 = n \quad (12-55)$$

where *n* is an integer. In this way, the carrier frequency acts as an internal clock. This second 90° pulse diverts only the *y* component of the precessing net magnetization at that instant into the *z* direction but leaves the *x* component at that instant in the *xy* plane. Consequently, the amplitude of the remaining net magnetization in the *xy* plane after the second 90° pulse is equal to $M_i \sin 2\pi \nu_i t_1$ where M_i is the amplitude of the net magnetization from the population of nuclei *i* before the second pulse and ν_i is its Larmor frequency. Because

$$M_i \sin(2\pi \nu_i t_1) = M_i \sin[2\pi(\nu_i t_1 - n)] = M_i \sin[2\pi(\nu_i - \nu_0)t_1] \quad (12-56)$$

Table 12-3: Couplings Giving Rise to a Peak in a Three-Dimensional or Four-Dimensional Nuclear Magnetic Resonance Spectrum

acronym	couplings ^a	acronym	coupling ^a
HNCO ²¹³		HNHB ²²²	
HNCA ²¹³		HCA(CO)N ^{e 213}	
HCACO ²¹³		HCA(CO)NH ^{e 224}	
HNCACO ²¹⁴		HN(CO)CA ^{e 225,226}	
HNCOCA ²¹⁴		CBCA(CO)NH ^{e 227}	
CBCANH ^{b 215} HNCACB ^{b 216}		HN(CO)HAHB ^{f 228} HBHA(CO)NH ^{f 229}	
CBCACO(CA)HA ^{b 217}			
HCCH ^{c 218,219}		HCB(CGCD)HD ²³⁰	
HNHA ^{d 213,220,221}		HCB(CGCDCE)HE ²³⁰	
HACAHB ^{222,223}			

^aBoxes enclose the three or four coupled atoms that both produce the peak or peaks and determine the three or four chemical shifts, one for each of the three or four atoms, in the three or four respective dimensions. ^bThe three-dimensional peaks from the α ¹³C carbon and the β ¹³C carbon are separate peaks on the same field, each coupled respectively to the same amide ¹⁵N nitrogen and amide ¹H hydrogen or to the same combination of acyl ¹³C carbon and α ¹H hydrogen, respectively, in the other two dimensions. ^cThe three dimensions are ¹³C carbon, ¹³C carbon, and ¹H hydrogen. The ¹H hydrogens coupled to one or the other of the two ¹³C carbons appear on the three-dimensional field at the chemical shifts of the other ¹³C carbon and their own ¹³C carbon. ^dThe coupling between the amide ¹H hydrogen and the α ¹H hydrogen is the usual strong three-bond *J* coupling between adjacent ¹H hydrogens. This coupling can be relayed through the α ¹³C carbon or can be enhanced by using HOHAHA. ^eThe coupling between the α ¹³C carbon and amide ¹⁵N nitrogen is relayed through the acyl ¹³C carbon. ^fThe three-dimensional peaks from the α ¹H hydrogen and the β ¹H hydrogens are separate peaks each coupled respectively to the same amide ¹⁵N nitrogen and amide ¹H hydrogen through the acyl ¹³C carbon.

after the second 90° pulse, the amplitude of the net magnetization of the population of nuclei i precessing at its Larmor frequency in the xy plane has become a function of the length of the interval t_1 between the pulses. Furthermore, as t_1 is varied, this amplitude will vary harmonically with a frequency $\nu_i - \nu_0$, which is directly proportional to the chemical shift (Equation 12-54) of the population of nuclei i if the applied magnetic flux density B_{app} is such that the carrier frequency ν_0 is equal to the frequency at which the standard nuclei absorb, ν_{std} .

Immediately following the second 90° pulse, the free induction decay of the excited sample is gathered in the usual way. The Fourier transform of the output of the radio receiver produces a nuclear magnetic spectrum. The amplitude of each peak in the spectrum, however, because it was derived only from the net magnetization remaining in the xy plane, has become a harmonic function of t_1 with a frequency $\nu_i - \nu_0$. If spectra are gathered at systematically increasing values of t_1 , the amplitude of each peak will vary with a frequency proportional to its chemical shift (Figure 12-15).²⁰³ Each peak has become **labeled with its own Larmor frequency**, and this labeling is manifested in the **amplitude modulation** of its peak in the spectrum with a frequency equal to the difference between its Larmor frequency and the carrier frequency. In a spectrometer with a carrier signal of 600 MHz, the values of $\nu_i - \nu_0$ for ^1H hydrogens are less than 6000 Hz, and in a spectrometer with a carrier signal of 150 MHz, the values of $\nu_i - \nu_0$ for ^{13}C carbon are less than 15,000 Hz, so the systematically increasing lengths of the

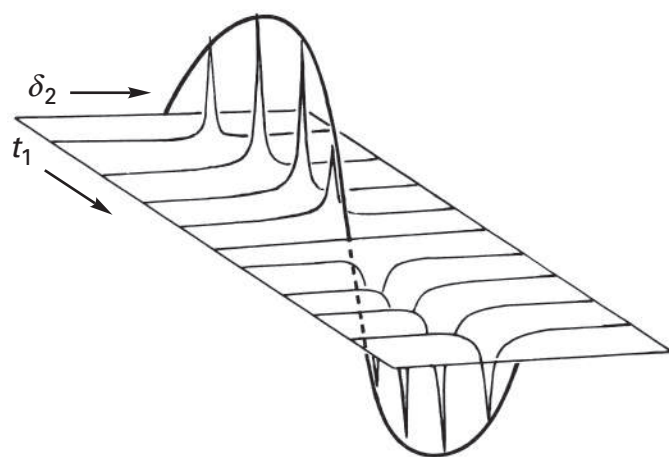


Figure 12-15: Amplitude modulation of the absorption of a nucleus produced during correlated spectroscopy.²⁰³ The absorption of a particular population of identical nuclei produces a peak in the spectrum of absorption as a function of the chemical shift δ_2 after $f(t_1, t_2)$ has been submitted to Fourier transformation only in the second dimension. Each trace is this spectrum of absorption as a function of the chemical shift δ_2 at a different t_1 . The amplitude of the peak of absorbance records the harmonic precession of the nuclear spin relative to the carrier frequency during the interval t_1 . Reprinted with permission from ref 203. Copyright 1995 John Wiley and Sons Ltd.

intervals t_1 must be in the range of tens of microseconds to milliseconds to produce reliable modulations of the amplitudes.

A simple two-dimensional spectrum is an extension of this procedure of frequency labeling. A series of free induction decays from the sample are gathered at systematically increasing intervals t_1 . The time dimension of each free induction decay is designated as t_2 . The complete set of this series of free induction decays defines a function that is two-dimensional in time, $f(t_1, t_2)$. The information in the first dimension of this function is encoded in the modulations of the amplitudes (**AM**) of the signals from the individual populations of nuclei, and the information in the second dimension is encoded in the modulations of the frequency (**FM**) of the free induction decays. A two-dimensional Fourier transform of this function extracts the frequencies of these modulations in the two dimensions. The two-dimensional Fourier transform of the function $f(t_1, t_2)$ is a two-dimensional function in frequency, which when divided by the carrier frequency (Equation 12-54) is a two-dimensional function in chemical shift, $f(\delta_1, \delta_2)$.

If none of the populations of nuclei in the sample is spin-spin-coupled to any other, the amplitude of the signal from each population of nuclei is modulated only by its own Larmor frequency, and $f(\delta_1, \delta_2)$ has peaks only when $\delta_1 = \delta_2$. In such a case, the **diagonal of the two-dimensional spectrum** replicates the one-dimensional nuclear magnetic resonance spectrum, and nothing has been gained. If, however, one population of nuclei is spin-spin-coupled to another population of nuclei, the modulations of the amplitudes of their precessions are transferred between themselves during the second 90° pulse, and each of their precessions becomes labeled not only with its own Larmor frequency but also with the Larmor frequency of the other population of nuclei to which it is spin-spin-coupled.

The Fourier transform picks out these coupled frequencies, and on the two-dimensional field, in addition to the one-dimensional spectrum along the diagonal, there are **off-diagonal cross-peaks**. Each of these cross peaks is located on the field at a chemical shift δ_1 of one population of nuclei and a chemical shift δ_2 of another population of nuclei to which the first population is spin-spin-coupled. Because spin-spin coupling is fully reciprocal, these cross-peaks are distributed symmetrically about the diagonal of the two-dimensional field. Correlated spectroscopy (COSY) is the technique that has just been described. A two-dimensional **correlated spectrum** is a two-dimensional spectrum in which the off-diagonal cross-peaks arise from spin-spin couplings between different populations of nuclei and identify populations of spin-spin-coupled nuclei by their chemical shifts. It is these off-diagonal cross-peaks that pull out individual absorptions from the one-dimensional spectrum, spread them into two dimensions, and permit them to be observed individually.

A two-dimensional correlated spectrum (Figure 12-16)²¹⁰ is a presentation of absorption as a function of two values of chemical shift (in parts per million), δ_1 and δ_2 . Each off-diagonal cross-peak in the spectrum has the same value of the chemical shift δ_2 as a peak buried in the one-dimensional spectrum and the same value of the chemical shift δ_1 as another peak buried elsewhere in the one-dimensional spectrum but connected to the first by spin-spin coupling. The result is that two individual absorptions unresolved in the one-dimensional spectrum are simultaneously drawn out of it and placed in isolation from all of the other absorptions otherwise overlapping them. This provides the **resolution**. The **information** provided by an off-diagonal cross-peak is that the two nuclei responsible for these two now isolated absorptions are connected through covalent bonds that mediate spin-spin coupling.

The off-diagonal region displayed in Figure 12-16 has a range for δ_2 (6.5–10.6 ppm) that includes the chemical shifts for the **amido ^1H hydrogens** of the polypeptide backbone and a range for δ_1 (1.7–6 ppm) that includes the chemical shifts of the **^1H hydrogens on the α carbons** of the amino acids in the protein. The diagonal, one-

dimensional spectrum is just beyond the lower right hand corner of the figure. Each cross-peak within the panel arises from the spin-spin coupling between the amido ^1H hydrogen of one of the amino acids in the protein and its own α ^1H hydrogen. Each cross-peak has pulled the absorption of each amido ^1H hydrogen and the absorption of its adjacent α ^1H hydrogen out of the unresolved one-dimensional spectrum so that they can be individually observed. Each cross-peak also assigns numerical values for the individual chemical shifts (δ_1 and δ_2) of the two nuclei of each of these pairs of spin-spin-coupled ^1H hydrogens and states that the two ^1H hydrogens with these two chemical shifts are connected to each other by three covalent bonds. This region of a two-dimensional (^1H - ^1H) correlated spectrum is a **fingerprint** for the protein because almost every amino acid in its sequence is represented by a single cross-peak,* and the distribution of the cross-peaks on the field is unique to that protein.

* Glycines, because they have two diastereotopic α ^1H hydrogens, usually produce two cross-peaks of the same chemical shift in the amido ^1H hydrogen dimension, and prolines, because they have no α ^1H hydrogens, produce none.

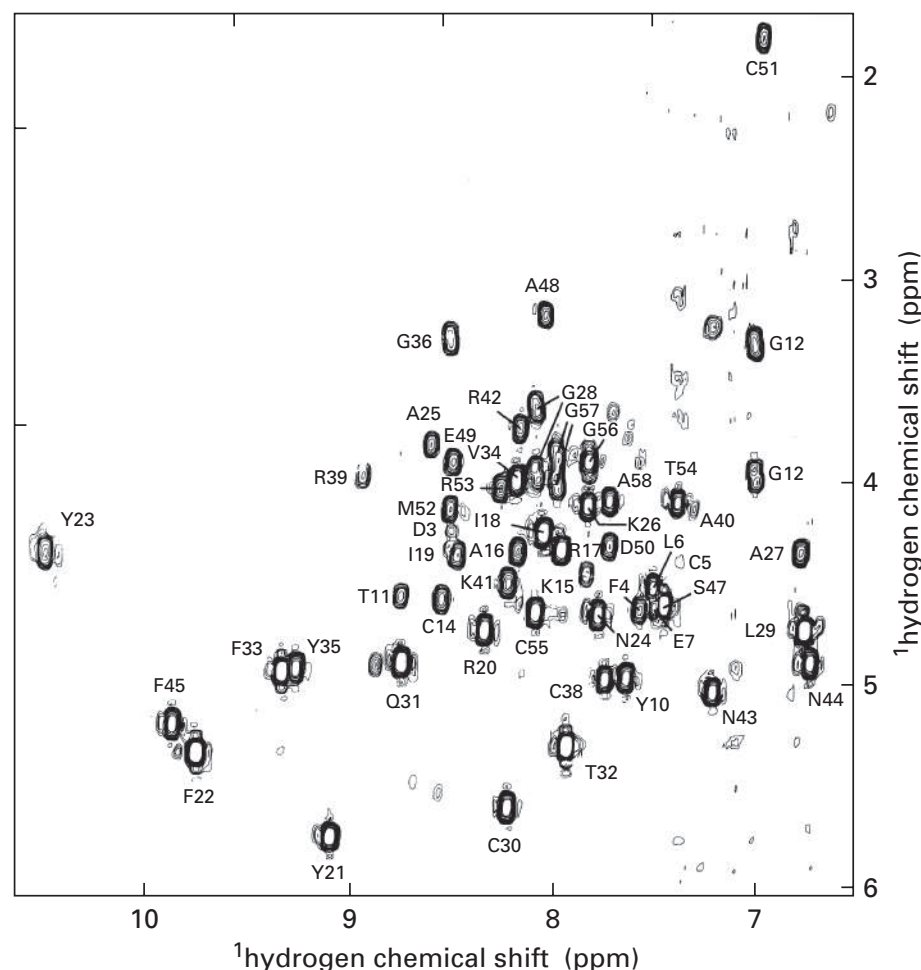


Figure 12-16: Two-dimensional (^1H - ^1H) correlated nuclear magnetic resonance spectrum of a 20 mM solution of basic pancreatic trypsin inhibitor ($n_{\text{aa}}=58$) in $^1\text{H}_2\text{O}$ at pH 4.6 and 68°C .²¹⁰ The spectrum is presented on a two-dimensional field with axes of the two respective chemical shifts. Each cross-peak is represented topographically as mountains are represented topographically on a map. Each cross-peak is a set of closed curves within closed curves. As in a map of electron density (Figure 4-12), each curve connects points of equal amplitude in the Fourier transform $f(\delta_1, \delta_2)$; the more closed curves, the greater the amplitude of the peak. The spectrum has three dimensions, the two chemical shifts δ_1 and δ_2 , in the x and y dimensions in the plane of the page, and the amplitude of the Fourier transform, in the z dimension normal to the page and represented in the contours. The region of the spectrum presented contains the cross-peaks created by spin-spin couplings between the ^1H hydrogen on each α carbon (vertical axis) and the ^1H hydrogen on the respective, immediately adjacent amido nitrogen (horizontal axis). The spectrum is a fingerprint for the protein, and each cross-peak assigns the respective chemical shifts of the two coupled nuclei. Every peak in this region of the spectrum has been assigned to one of the amino acids in the sequence of the protein. All of the amino acids in the sequence are represented, with the exception of the four prolines (which have no α hydrogens), Glycine 37, and Arginine 1. Lysine 46, because the chemical shift of its absorption coincides with that of $[^1\text{H}]\text{H}_2\text{O}$ and is suppressed along with that of the $[^1\text{H}]\text{H}_2\text{O}$, is also missing from the spectrum. Reprinted with permission from ref 210. Copyright 1982 Academic Press.

Many **improvements** have been made to the original correlated spectrum. There are many sophisticated and intricate elaborations of the sequence of the pulses of oriented radiowaves that narrow the cross-peaks, eliminate background, and enhance dramatically the signals from weakly absorbing nuclei such as ^{13}C and ^{15}N . Each of these elaborations is identified by its own acronym (Table 12–4). Because the nucleus of ^{12}C has no magnetic moment and the nucleus of ^{14}N has a spin quantum number of 1, the proteins examined are now almost always modified so that all of their nitrogens are ^{15}N and all of their carbons are ^{13}C by expressing them in bacteria grown on $[\text{N}^{15}]\text{NH}_4^+$ as their sole source of nitrogen and on a $[\text{C}^{13}]$ nutrient as their sole source of carbon. In this way, cross-peaks from these enriched proteins produced by heteronuclear spin–spin coupling can be observed. For example, the heteronuclear spin–spin coupling between an **amido ^{15}N nitrogen** and its own **amido ^1H hydrogen**, enhanced by an HSQC pulse sequence (Figure 12–17),²⁵¹ provides an alternative fingerprint of the protein in which every amino acid is also represented. Two-dimen-

sional correlated spectroscopy has been expanded to three dimensions (Table 12–3). The cross-peaks in these spectra (Figure 12–18)²¹³ are produced by spin–spin coupling among three nuclei, usually of two or three different elements. By expanding the dimensions, these procedures are able to resolve absorptions that overlap in two dimensions just as two dimensions separate absorptions that overlap in one dimension. Each cross-peak in such spectra assigns chemical shifts simultaneously to three or four individual nuclei.

Each of the cross-peaks in the two-dimensional (^1H – ^1H) correlated spectrum of basic pancreatic trypsin inhibitor (Figure 12–16) and the two-dimensional (^{15}N – ^1H) HSQC correlated spectrum of dihydrofolate reductase (Figure 12–17) has been labeled with the position in the sequence of the protein of the amino acid containing the two nuclei that produced it. The spectra themselves do not come with labels, and each of these **assignments** has been performed by tracing connections among cross-peaks that affiliate nuclei in the polypeptide backbone through spin–spin coupling and by assigning the type of each amino acid from the pattern of

Table 12–4: Methods for Improving Cross-Peaks in Two-Dimensional and Three-Dimensional Nuclear Magnetic Resonance Spectra

acronym	full name	improvement
HMQC ^{231,232}	heteronuclear multiple-quantum coherence	increases amplitude of cross-peaks arising from coupling involving nuclei with small magnetogyric ratios such as ^{13}C and ^{15}N (Table 12–2)
HSQC ²³³	heteronuclear single-quantum coherence	
HSMQC ²³⁴	heteronuclear single–multiple-quantum coherence	
DQF ^{235,236} MQF ²³⁷	double quantum filtered multiple quantum filtered	removes unwanted noise from spectrum
HOHAHA ^{238–240}	homonuclear Hartmann–Hahn	increases amplitude of cross-peaks and extends, by relaying coherence, the number of bonds through which coupling can produce a cross-peak
TOCSY ^{238,241}	total correlation spectroscopy (same method as HOHAHA)	
HMBC ^{242,243}	heteronuclear multiple bond correlation	extends the number of bonds through which coupling can produce a cross-peak
PS ²⁴⁴	pseudo single quantum	narrows the line widths of the cross-peaks
random fractional deuteration ^{245,246}		replacement of 50% of the ^1H hydrogens in a protein with ^2H hydrogens at random narrows the widths of the cross peaks in correlated spectroscopy and increases the amplitude and the number of the detectable nuclear Overhauser effects from larger proteins
TROSY ²⁴⁷	transverse relaxation-optimized spectroscopy	narrows the line widths of the cross-peaks
CRINEPT ²⁴⁸	cross relaxation insensitive nuclei enhanced polarization transfer	increases amplitude of cross-peaks arising from coupling involving nuclei with small magnetogyric ratios such as ^{13}C and ^{15}N
SBC ^{249,250}	single bond correlation	permits spin–spin coupling between ^{13}C and ^{15}N to be used for two-dimensional spectrum



Figure 12-17: Two-dimensional (^1H - ^{15}N) HSQC correlated nuclear magnetic resonance spectrum of a 2 mM solution of human dihydrofolate reductase ($n_{\text{aa}} = 186$) expressed in *E. coli* grown on [^{15}N]NH $_4$ Cl as its sole source of nitrogen and dissolved at pH 6.5 and 25 °C.²⁵¹ Each of the spin-spin couplings between an amido ^{15}N and an amido ^1H at each of the positions in the backbone of the polypeptide creates one of the cross-peaks in the spectrum. The range of chemical shift for ^{15}N (vertical axis) spans the chemical shifts for the amido ^{15}N in the protein, and the range of chemical shift for ^1H (horizontal axis) spans the chemical shifts for amido ^1H in the protein. The two-dimensional spectrum is a fingerprint for the protein, and each cross-peak assigns pairs of chemical shifts to the respective coupled amido ^{15}N and amido ^1H hydrogens. Each cross-peak is labeled by the amino acid in the sequence of the protein to which it has been assigned. The pairs of peaks each having the same chemical shift in the ^{15}N dimension (connected by horizontal lines) are those arising from the spin-spin coupling of the two respective distinct ^1H hydrogens on each amido nitrogen in the primary amides of the glutamines and asparagines in the protein. They are also labeled by the amino acid to which they have been assigned. Reprinted with permission from ref 251. Copyright 1992 American Chemical Society.

affiliations among cross-peaks that trace connections out into each side chain.

Connections among nuclei along the polypeptide backbone are usually traced^{252–255} in a systematic sequence of sections through three-dimensional correlated spectra (Figure 12–18). The large set of three-dimensional spectra available for **tracing the connections among nuclei** (Table 12–3) is redundant, so that when connections are obscured by the overlap of peaks or when a particular cross-peak is missing (for example, Lysine 46 in Figure 12–16), the trace can take an alternative path. The result of such a trace is that all of the nuclei—¹hydrogens, ¹³carbons, and ¹⁵nitrogens—in long segments of polypeptide backbone are consecutively connected to each other and each individually assigned a chemical shift.

The respective positions of these segments of connected nuclei in the amino acid sequence of the protein are then established by tracing connections from the nuclei out into each side chain (Figures 12–19 through 12–22). Each of these paths of **connections out into a side chain** usually starts at one of the cross-peaks in a fingerprint of the protein. For example, the relayed

spin-spin coupling among the ^1H hydrogens of each side chain registered in a two-dimensional (^1H - ^1H) TOCSY spectrum (Figure 12-19)²⁵⁶ begins at the cross-peak between the amido ^1H hydrogen and the α ^1H hydrogen (Figure 12-16). The coupling between the α ^{13}C carbon and the β ^{13}C carbon of each side chain registered in a three-dimensional (^{13}C - ^{15}N - ^1H) CBCA(CO)NH correlated spectrum (Figure 12-20)²⁵⁷ begins at the cross-peak between the amido ^{15}N nitrogen and the amido ^1H hydrogen (Figure 12-17) of the next amino acid in the sequence (Table 12-3). Connections among ^1H hydrogens of a side chain (Figure 12-19) or ^{13}C carbons of a side chain (Figure 12-20) can be extended to their own ^{13}C carbons or ^1H hydrogens, respectively, with two-dimensional (^1H - ^{13}C) HSQC correlated spectra (Figure 12-21).²⁵⁸ When two-dimensional spectra (Figure 12-19) become too crowded to trace connections, they can be expanded in a third dimension (Figure 12-22)²⁵⁹ to resolve the individual cross-peaks.

Each side chain has a characteristic **pattern of connections** among nuclei of characteristic chemical shifts that identifies it in the spectrum. For example, glutamate has two β^1 hydrogens that have smaller chemical shifts than its two α^1 hydrogens (Figure 12–19). Isoleucine has

^1H hydrogens on a δ methyl group and a γ methyl group as well as two ^1H hydrogens on a γ methylene, all in the aliphatic range of chemical shifts (Figures 12–19 and 12–22). Lysine has ^1H hydrogens on a β methylene, a γ methylene, and a δ methylene in the aliphatic range but two ^1H hydrogens on an ε methylene with chemical shifts around 3 ppm (Figures 12–19 and 12–22). Threonine and serine have β ^1H hydrogens with larger chemical shifts (around 4 ppm), and threonine has γ ^1H hydrogens with chemical shifts characteristic of a methyl group (Figures 12–19 and 12–22). From these patterns, from the sequence of the connections along the backbone (Figure 12–18), and from the amino acid sequence of the protein, it is usually possible to identify the long, unbroken segments of connections among the atoms of the backbone that run through the spectra with segments in the amino acid sequence of the protein and thereby assign the cross-peaks to specific positions in the amino acid sequence, much as the pattern of protrusions from the polypeptide backbone in a map of electron density allows segments of the amino acid sequence to be identified.

The final results of this process are that each cross-peak on the various two- and three-dimensional correlated spectra has been assigned to the two or three nuclei in the amino acid sequence of the protein that produce it and that the nucleus of almost every ^1H hydrogen, ^{13}C carbon, and ^{15}N nitrogen in the protein has been assigned a specific chemical shift. By themselves these assignments are not very informative. They are, however, an indispensable prelude to using nuclear magnetic resonance to provide insight into the dynamics of a protein, to produce its molecular model, to measure the acid dissociation constants of its side chains, and to follow the rates of exchange of its protons with protons in the solution.

When the spin state of a particular population of chemically identical nuclei is saturated by the absorption of electromagnetic energy at its Larmor frequency and then allowed to relax back to its equilibrium distribution, the rate of its relaxation contains information about the **dynamics of the structure** in which each of these nuclei is contained. For example, the relaxation of a particular population of identical nuclei of amido ^{15}N nitrogens, each at the same position in the polypeptide backbones of the identical molecules of a protein in a solution, is dominated by the dipolar interactions between the nuclei of the ^{15}N nitrogens in that population and the nuclei of the directly attached ^1H hydrogens on each of the individual amido nitrogens. The rate at which the bonds between the ^{15}N nitrogens and the ^1H hydrogens in those two particular populations reorient relative to the magnetic field determines the rate of the relaxation of the population of ^{15}N nitrogen nuclei. From an analysis of this rate of relaxation, information about the rate of this reorientation can be extracted.^{260,261} This information is expressed as an **order parameter** S^2 , which assumes a value of 1 when

the ^{15}N – ^1H bond is fixed rigidly in the protein so that it reorients at the same rate at which the entire molecule of protein reorients by its normal rotational and translational diffusion and which assumes a value of 0 when it is so loosely attached to the protein that it reorients completely independently of the reorientation of the molecule of protein. Therefore, the order parameter S^2 is a measure of the flexibility of a particular position within a molecule of the protein.

Two-dimensional spectra of proteins in which each amido ^{15}N nitrogen in the polypeptide backbone has been assigned its position in the amino acid sequence can be used to measure relaxation rates of these individual ^{15}N nitrogens and calculate values of the order parameter S^2 for each. Most of the values of the order parameter S^2 for these ^{15}N nitrogens are near 1 (≥ 0.8) because most of the polypeptide backbone of a molecule of protein is rigidly fixed within the tertiary structure, but there are **flexible segments** that can be identified by the smaller values (0.4–0.6) of the order parameter S^2 of the amido ^{15}N nitrogens they contain.^{262,263} These segments are often flexible loops on the surface of the molecule of protein and correspond to segments in a crystallographic map of electron density that are so flexible that they do not appear in the map or to segments the atoms of which have high B -factors.²⁶³ High values of these B -factors indicate that the segment is also flexible in the crystal. The informative exceptions are those in which the order parameters are low but the segment appears to be rigid in the crystallographic molecular model and its constituent atoms have low B -factors. These exceptions indicate that the crystal packing has confined an otherwise flexible segment of polypeptide.

Order parameters can also be obtained for bonds between ^{15}N nitrogens and ^1H hydrogens in side chains. For example, the order parameters for bonds between ^{15}N nitrogens and ^1H hydrogens in the side chains of tryptophans buried in the core of the protein are usually greater than 0.8, but those for the bonds between ^{15}N nitrogens and ^1H hydrogens in the side chains of arginines on the surface of a protein can be as small as 0.05.²⁶³ Unfortunately, there is no simple correlation between values of the order parameter S^2 and the rates at which the flexible segments or two side chains are fluctuating relative to the entire molecule, and motions slower than the rotational diffusion of the entire molecule of protein do not register in the order parameter.

In the rare instances in which a flexible segment of the folded polypeptide assumes only two significantly occupied conformations of about equal occupancy, each nucleus in the flexible segment will have a different chemical shift in each conformation. If these chemical shifts are different enough, each pair of nuclei that is spin–spin-coupled will produce two cross-peaks in a two-dimensional spectrum, each with the chemical shifts of the respective nuclei in the **two respective conformations**.²⁶⁴ In such cases, it is possible to obtain a rate

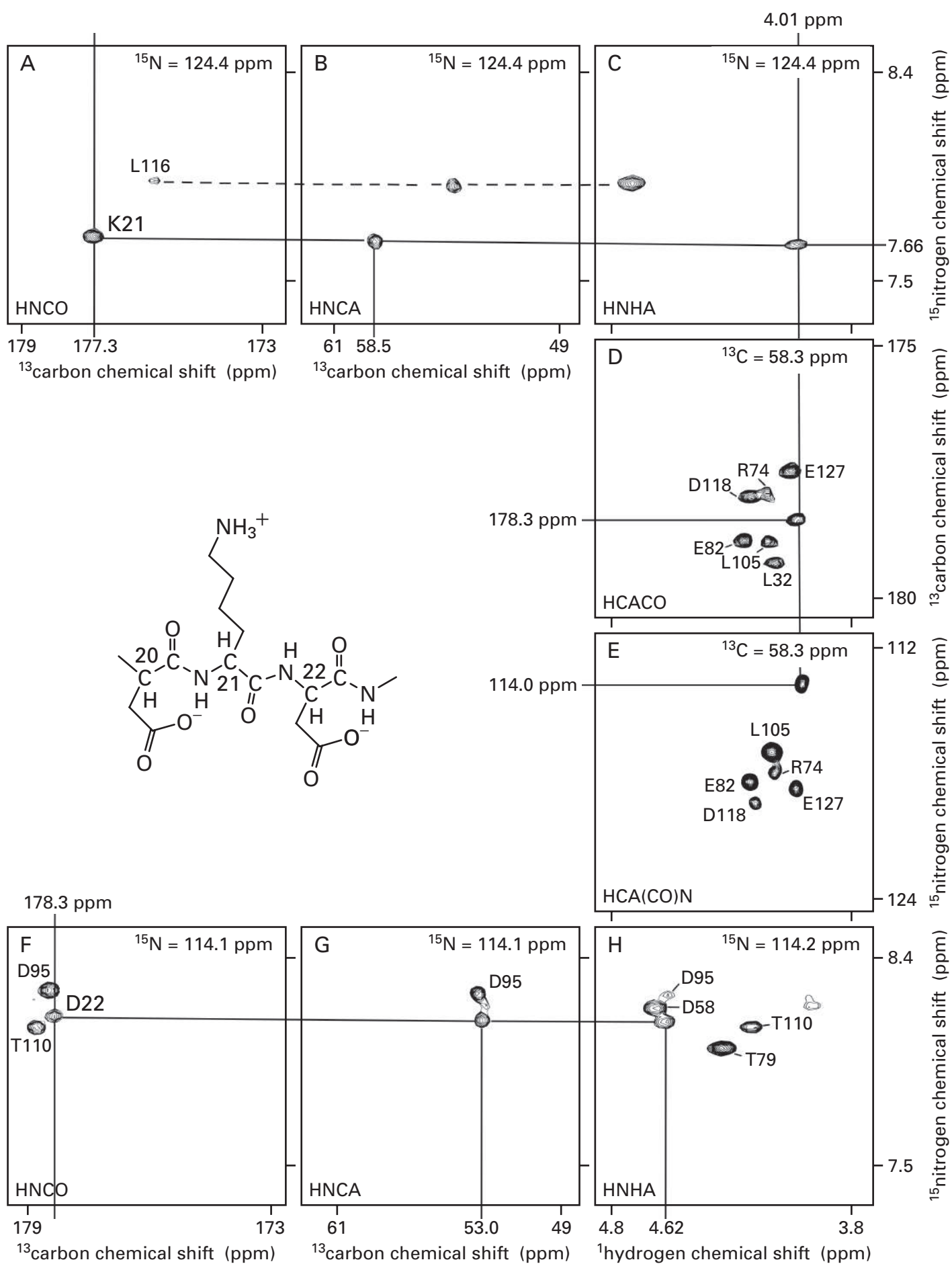


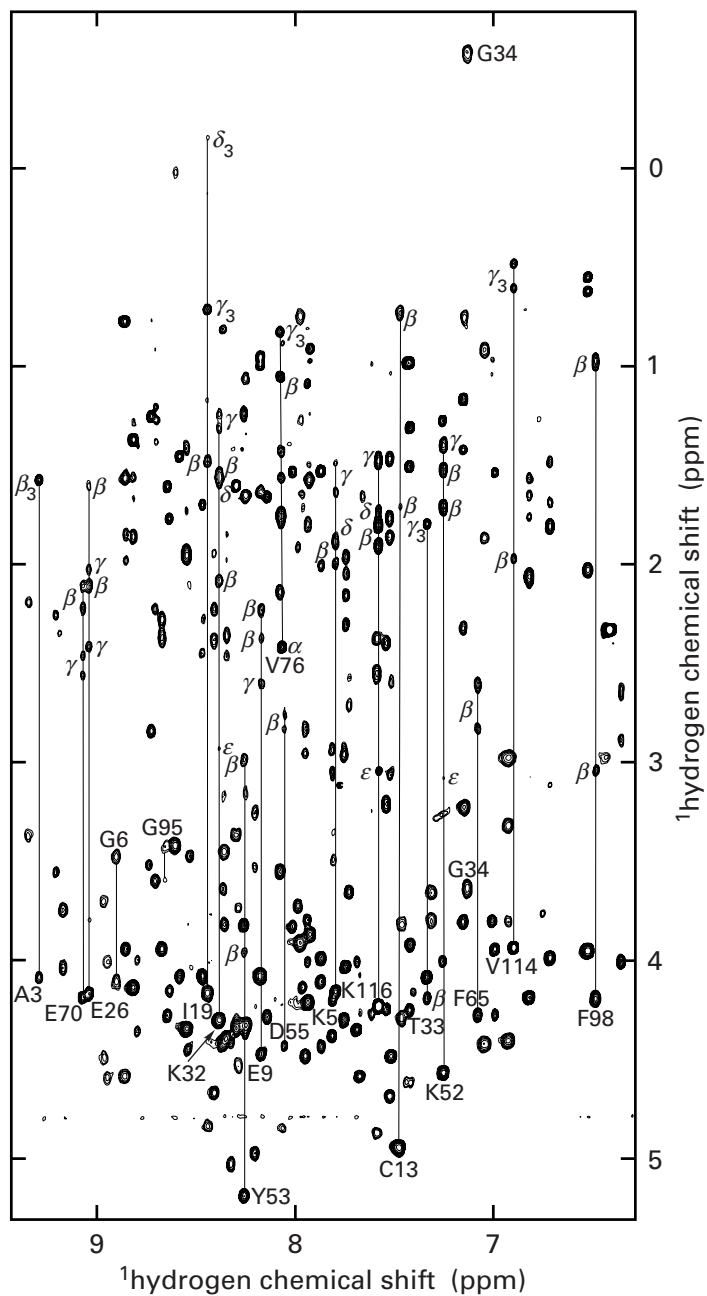
Figure 12–18: Sequential assignment of the chemical shifts for the ^1H hydrogens, ^{13}C carbons, and ^{15}N nitrogens in a polypeptide backbone by use of three-dimensional nuclear magnetic resonance spectra.²¹³ The complementary DNA encoding calmodulin from *D. melanogaster* ($n_{\text{aa}} = 148$) was expressed in *E. coli* grown on $^{15}\text{N}[\text{NH}_4\text{Cl}]$ and $^{13}\text{C}_6[\text{glucose}]$ as sole nitrogen and carbon sources, so that the protein was uniformly and completely (>95%) labeled with ^{15}N nitrogen and ^{13}C carbon. Spectra were recorded from a 1.5 mM solution of calmodulin in a 93:7 mixture of $^1\text{H}_2\text{O}$ to $^2\text{H}_2\text{O}$ at pH 6.3 and 47 °C. Individual panels (A–H) are two-dimensional sections through a series of three-dimensional nuclear magnetic resonance spectra (Table 12–3) of the protein. In panels A, B, and C, the sections through the respective three-dimensional spectra are only wide enough to contain cross-peaks from ^{15}N nitrogens the chemical shifts of which are 124.4 ± 0.2 ppm, which includes the chemical shift of the amido ^{15}N nitrogen of Lysine 21. Each of these three sections has as its vertical axis the chemical shift of ^1H hydrogen between 7.4 and 8.5 ppm, the region covering the chemical shifts of ^1H hydrogens on amido nitrogens. (A) Section containing the cross-peak produced by the spin–spin coupling connecting the amido ^1H hydrogen of Lysine 21, the amido ^{15}N nitrogen of Lysine 21, and the acyl ^{13}C carbon of Aspartate 20. This section has as its horizontal axis the chemical shift for ^{13}C carbon between 172 and 180 ppm, the region covering the chemical shifts of acyl ^{13}C carbons. This cross-peak is located by the chemical shift of the acyl ^{13}C carbon of Aspartate 20 (177.3 ppm) and assigns the chemical shift of the amido ^1H hydrogen of Lysine 21 as 7.66 ppm. (B) Section containing the cross-peak produced by the spin–spin coupling connecting the amido ^1H hydrogen, the amido ^{15}N nitrogen, and the α ^{13}C carbon of Lysine 21. The section has as its horizontal axis the chemical shift for ^{13}C carbon between 47 and 64 ppm, the region covering the chemical shifts of the α ^{13}C carbons. The position of this cross-peak, located by the value for the chemical shift of its amido ^1H hydrogen (horizontal line), assigns the chemical shift of the α ^{13}C carbon of Lysine 21 as 58.5 ppm. (C) Section containing the cross-peak produced by the spin–spin coupling connecting the amido ^1H hydrogen, the amido ^{15}N nitrogen, and the α ^1H hydrogen of Lysine 21. This section has as its horizontal axis the chemical shift for ^1H hydrogen between 3.7 and 4.9 ppm, the region covering the chemical shifts of α ^1H hydrogens. The position of this cross-peak, located by the value of the chemical shift of its amido ^1H hydrogen (horizontal line), assigns the chemical shift of the α ^1H hydrogen of Lysine 21 as 4.01 ppm. In panels D and E, the sections through the respective three-dimensional spectra are only wide enough to contain cross-peaks from ^{13}C carbons the chemical shifts of which are 58.3 ± 0.3 ppm, which includes the chemical shift of the α ^{13}C carbon of Lysine 21 (assigned in panel B). Each of these two sections has as its horizontal axis the chemical shift for ^1H hydrogen between 3.7 and 4.9 ppm, the region covering the chemical shifts of α ^1H hydrogens. (D) Section containing the cross-peak produced by the spin–spin coupling connecting the α ^1H hydrogen, the α ^{13}C carbon, and the acyl ^{13}C carbon of Lysine 21. This section has as its vertical axis the chemical shift for ^{13}C carbon between 174 and 181 ppm, the region covering the chemical shifts of acyl ^{13}C carbons. The position of this cross-peak, located by the value of the chemical shift of its α ^1H hydrogen (vertical line), assigns the chemical shift of the acyl ^{13}C carbon of Lysine 21 as 178.3 ppm. (E) Section containing the cross-peak produced by the spin–spin coupling connecting the α ^1H hydrogen of Lysine 21, the α ^{13}C carbon of Lysine 21, and the amido ^{15}N nitrogen of Aspartate 22. This section has as its vertical axis the chemical shift for ^{15}N nitrogen between 111 and 125 ppm, the region covering the chemical shifts of amido ^{15}N nitrogens. The position of the cross-peak for Lysine 21, located by the value of the chemical shift of its α ^1H hydrogen (vertical line), assigns the chemical shift of the amido ^{15}N nitrogen of Aspartate 22 as 114.0 ppm. (F–H) Sections through three-dimensional spectra containing cross-peaks for Aspartate 22 corresponding respectively to the sections in panels A–C that contain cross-peaks for Lysine 21. The sections in panels F, G, and H are only wide enough to contain cross-peaks from ^{15}N nitrogens, the chemical shifts of which are 114.1 ± 0.2 ppm, which includes the chemical shift of the amido ^{15}N nitrogen of Aspartate 22 assigned in panel E. The value of the chemical shift of the amido ^{15}N nitrogen of Lysine 21 (124.4 ppm) that was used to set the position of the slabs in panels A, B, and C was assigned with a section corresponding to panel E but with the section for the sequential assignment of the chemical shifts of the nuclei in Aspartate 20 rather than Lysine 21. The cross-peak in panel F produced by spin–spin couplings connecting the amido ^1H hydrogen of Aspartate 22, the amido ^{15}N nitrogen of Aspartate 22, and the acyl ^{13}C carbon of Lysine 21 was located with the chemical shift of the acyl ^{13}C carbon of Lysine 21 assigned in panel D. The position of the cross-peak from Lysine 21 in panel A was located with the chemical shift of the acyl ^{13}C carbon of Aspartate 20 assigned in a section for the sequential assignment of the nuclei in Aspartate 20 corresponding to that in panel D for Lysine 21. Cross-peaks from Leucine 116 appear in panels A, B, and C because the chemical shift of its amido ^{15}N nitrogen is 124.2 ppm. Cross-peaks from Leucine 32, Arginine 74, Glutamate 82, Leucine 105, Aspartate 118, and Glutamate 127 appear in panels D and E because the chemical shifts of their α ^{13}C carbons are 58.2, 58.1, 58.2, 58.5, 58.5, and 58.5 ppm, respectively. Cross-peaks from Aspartate 58, Threonine 79, Aspartate 95, and Threonine 110 appear in panels F, G, and H because the chemical shifts of their amido ^{15}N nitrogens are 113.9, 114.0, 114.2, and 114.4 ppm, respectively. Reprinted with permission from ref 213. Copyright 1990 American Chemical Society.

constant for the exchange of the flexible segment between the two conformations. For example, a loop between Alanine 9 and Leucine 24 in dihydrofolate reductase from *E. coli* exchanges between its two conformations²⁶⁵ at a rate of 35 s^{-1} . The heterologous association between two molecules of protein also causes changes in the chemical shifts of nuclei that end up in the interface. These changes produce pairs of cross-peaks, one from the unassociated protein and one from the associated protein. These changes in chemical shift identify the amino acids involved in the interface,²⁶⁶ and rates of exchange of the participants between free and bound states can be calculated from such spectra.²⁶⁷

A nuclear magnetic resonance molecular model of a protein is produced from a list of the individual nuclear Overhauser effects between its ^1H hydrogens. The thousands of nuclear Overhauser effects that occur between pairs of the thousands of unique ^1H hydrogens in a protein

are resolved from each other by using two-dimensional and three-dimensional nuclear Overhauser enhanced spectroscopy (NOESY) just as the individual absorptions of each of the thousands of nuclei in a protein are resolved by using two-dimensional and three-dimensional correlated spectroscopy.

A two-dimensional **nuclear Overhauser enhanced spectrum** is a two-dimensional spectrum in which the off-diagonal cross-peaks arise from nuclear Overhauser effects between two different populations of ^1H hydrogen nuclei and identify by their chemical shifts those pairs of ^1H hydrogen nuclei that are connected by those respective nuclear Overhauser effects. A two-dimensional nuclear Overhauser enhanced spectrum (Figure 12–23)^{268,269} is produced in the same way as a two-dimensional correlated spectrum, except that after the second 90° pulse has labeled the precession of the net magnetization of each population of nuclei with its Larmor frequency by



modulating the amplitude of its precession in the xy plane (Figure 12–15), there is a fixed delay or **time of mixing**, t_m , of 50 ms to several hundred milliseconds to allow the saturation of each population of nuclei, which has been labeled with its own characteristic Larmor frequency, to diffuse outward, mixing with the spin states of nuclei in its vicinity. After this fixed delay, a third 90° pulse initiates the collection of the free induction decay from the sample during t_2 .

In the two-dimensional spectrum that results, there is a cross-peak produced by the transfer of some of the saturation from the population of nuclei A, which has been labeled with its own Larmor frequency, to the population of nuclei X, each of which is adjacent to a nucleus A in a molecule of the protein. As a result, the

Figure 12-19: Two-dimensional (^1H - ^1H) TOCSY nuclear magnetic resonance spectrum of a 2 mM solution of ferrocycytochrome c_2 from *Rhodobacter capsulatus* ($n_{\text{aa}} = 116$) dissolved in a 90:10 mixture of $^1\text{H}_2\text{O}$ and $^2\text{H}_2\text{O}$, respectively, at pH 6 and 30 °C.²⁵⁶ The spectrum contains peaks resulting from the relay of spin-spin coupling from the amido ^1H hydrogen of each amino acid out along the atoms in its side chain. Each cross-peak is labeled with the letter of the Greek alphabet designating the carbon of the side chain on which the ^1H hydrogen producing it resides. Each sequence of cross-peaks resulting from these relayed spin-spin couplings (vertical lines) is anchored on the cross-peak connecting the spins of the amido ^1H hydrogen and the α ^1H hydrogen of an amino acid (Figure 12-16). Each sequence identifies cross-peaks arising from relayed coupling between the amido ^1H hydrogen and other hydrogens in the side chain of that amino acid and assigns chemical shifts (the values of the ordinate of each cross-peak) to each of those other hydrogens. For example, the chemical shifts of the α , β , γ , and δ_3 ^1H hydrogens of Isoleucine 19 are 4.12, 1.42, 0.66, and -0.22 ppm, respectively, and those of the α ^1H hydrogen and the two β ^1H hydrogens of Cysteine 13 are 4.90, 1.67, and 0.67 ppm, respectively. Each cross-peak between an amido ^1H hydrogen and an α ^1H hydrogen anchoring each of the relayed connections is labeled with the amino acid on which they reside. Reprinted with permission from ref 256. Copyright 1990 American Chemical Society.

population of nuclei X becomes labeled with the Larmor frequency of the population of nuclei A as well as its own Larmor frequency, and in the two-dimensional nuclear Overhauser enhanced spectrum that results, there is an off-diagonal cross-peak at chemical shift δ_2 of nucleus X and chemical shift δ_1 of nucleus A. Transfer of saturation, however, is reciprocal, and the saturation reciprocally and coincidentally transferred from the population of nuclei X to the population of nuclei A, which is labeled with the Larmor frequency of nuclei X, produces an off-diagonal peak at chemical shift δ_2 of nucleus A and chemical shift δ_1 of nucleus X. Each of these two symmetrically displayed peaks connects nucleus A and nucleus X by a nuclear Overhauser effect and identifies the two nuclei connected by their chemical shifts. Each of the hundreds of symmetrically displayed peaks in a nuclear Overhauser enhanced spectrum (Figure 12-23) makes its own respective connection between two ^1H hydrogens, each identified by its chemical shift.

A specific example illustrates the spin diffusion resulting from these individual transfers of saturation. A cross section through a two-dimensional nuclear Overhauser enhanced spectrum of bovine acrosin inhibitor IIA, at the chemical shift δ_1 equivalent to the Larmor frequency of the amido ^1H of Alanine 37 (8.45 ppm), contains cross-peaks at the chemical shifts δ_2 of the amido ^1H s of Asparagine 34, Cystine 36, and Phenylalanine 38; of the α ^1H s of Cystine 36 and Alanine 37; and of the β ^1H s of Cystine 36 and Alanine 37 (Figure 12-24)²⁷⁰ because the saturation transferred to each of these populations of ^1H nuclei retains the amplitude modulation, labeling it with the Larmor frequency of the population of the nuclei of the amido ^1H s of Alanine 37. Consequently, in the dimension of chemical shift δ_1 cross-peaks are located at the chemical shifts of these other populations

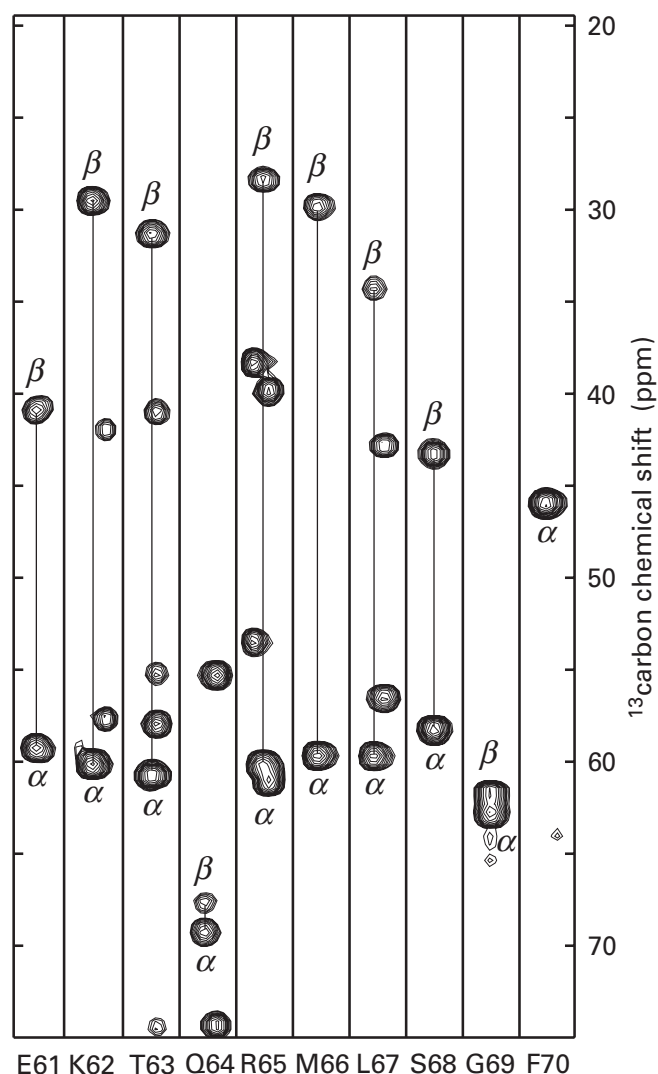
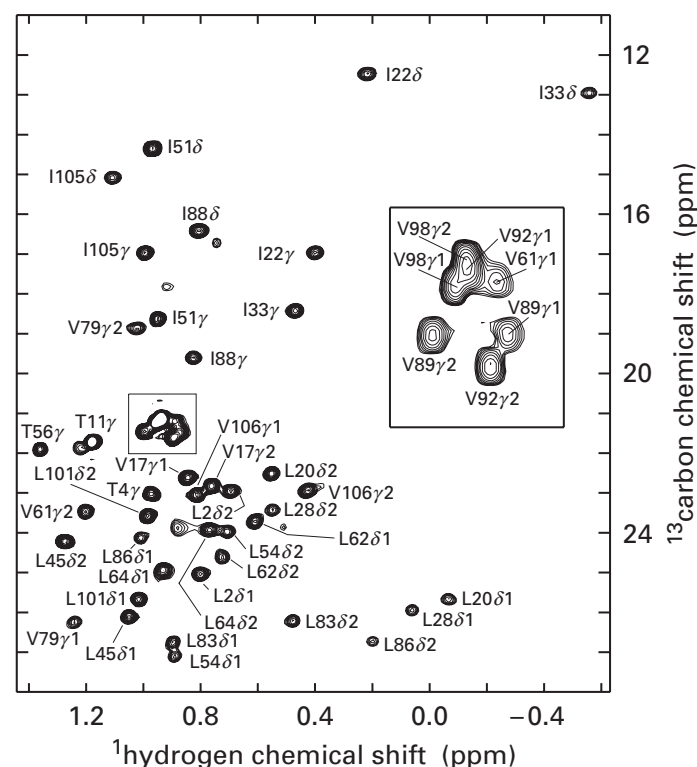


Figure 12-21: Two-dimensional (^1H - ^{13}C) HSQC correlated nuclear magnetic resonance spectrum of a solution of human transforming growth factor $\beta 1$ (homodimer of subunits 112 aa in length) expressed in Chinese hamster ovarian cells grown on a mixture of [^{14}C]amino acids and dissolved in a 95:5 mixture of [^1H]H $_2\text{O}$ and [^2H]H $_2\text{O}$, respectively, at pH 4.2 and 45 °C.²⁵⁸ The region of the spectrum presented covers the range of chemical shift for ^1H hydrogen (horizontal axis) and ^{13}C carbon (vertical axis) of the methyl groups of threonine, valine, leucine, and isoleucine. Each cross-peak is produced by the spin-spin coupling between a methyl ^{13}C carbon and its three ^1H hydrogens. Each is labeled with the amino acid in the sequence of the protein to which it has been assigned and the Greek letter designating the position of the methyl group within that amino acid. The inset is the boxed region within the full spectrum expanded in the dimension of the chemical shift of ^{13}C carbon to resolve peaks that overlapped in the full spectrum. Reprinted with permission from ref 258. Copyright 1996 American Chemical Society.

Figure 12-20: Strips from sections of a three-dimensional (^{13}C - ^{15}N - ^1H) CBCA(CO)NH nuclear magnetic resonance spectrum of a 1 mM solution of human interleukin-13 ($n_{\text{aa}} = 113$) expressed in *E. coli* grown on [^{15}N](NH $_4$) $_2$ SO $_4$ and [$^{13}\text{C}_6$]glucose as sole sources of nitrogen and carbon and dissolved in a 90:10 mixture of [^1H]H $_2\text{O}$ and [^2H]H $_2\text{O}$, respectively, at pH 6.0 and 25 °C.²⁵⁷ Each section through the three-dimensional spectrum from which each strip is taken is centered on the chemical shift of the amido ^{15}N nitrogen of the respective amino acid. A strip is then taken from the resulting two-dimensional section. Each strip is aligned vertically with its neighbors and is designated at its bottom by the amino acid from which the pair of cross-peaks (connected by vertical lines) arises. Each strip is centered on the chemical shift of the amido ^1H hydrogen of the next amino acid in the sequence of the protein (Table 12-3) so the horizontal axis of each strip is the chemical shift of ^1H hydrogen. Each strip is wide enough to include the cross-peak produced by the spin-spin coupling connecting the α ^{13}C carbon of the amino acid, the amido ^{15}N nitrogen of the next amino acid and the amido ^1H hydrogen of the next amino acid as well as the cross-peak produced by the spin-spin coupling connecting the β ^{13}C carbon of the amino acid, the amido ^{15}N nitrogen of the next amino acid, and the amido ^1H hydrogen of the next amino acid. The two cross-peaks in each strip appear in the same section and have the same value on the horizontal axis because they are both coupled to the same respective amido ^{15}N nitrogen and amido ^1H hydrogen on the next amino acid. Therefore, each strip is anchored on a cross-peak in the two-dimensional (^1H - ^{15}N) HSQC correlated nuclear magnetic resonance spectrum (Figure 12-17) providing the fingerprint of the protein, albeit on the cross-peak of the next amino acid, and each cross-peak assigns a value for the chemical shifts of the α ^{13}C carbon and the β ^{13}C carbon of the amino acid preceding the amino acid producing the cross-peak in the fingerprint. Reprinted with permission from ref 257. Copyright 2001 Elsevier B.V.



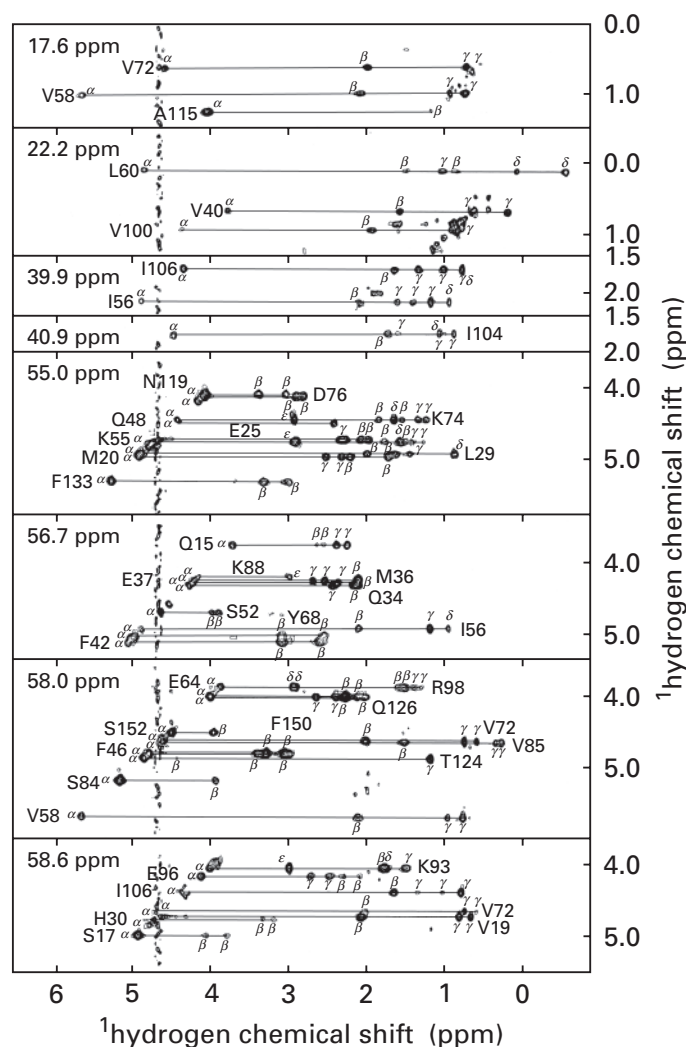


Figure 12-22: Sections from three-dimensional (^1H - ^{13}C - ^1H) HCCH-TOCSY nuclear magnetic resonance spectra of a 2 mM solution of human interleukin-1 β ($n_{\text{aa}} = 153$) expressed in *E. coli* grown on ^{15}N and ^{13}C as the sole sources of nitrogen and carbon and dissolved in $^2\text{H}_2\text{O}$ at pH 5.4 and 36 °C.²⁵⁹ Each section through the three-dimensional spectrum is 0.6–0.8 ppm in width centered on the ^{13}C chemical shift noted in its upper left corner. These chemical shifts serve to identify each section. Each cross-peak is produced by the spin–spin coupling of two hydrogens on the side chain of a particular amino acid often relayed through multiple bonds. One of the two hydrogens is on the ^{13}C providing the third dimension. The sections through the three-dimensional spectrum centered on chemical shifts for ^{13}C of 17.6 and 22.2 ppm, respectively, contain cross-peaks arising from the absorptions of the ^{13}C carbons in the δ and γ positions of leucines, the β positions of valines, and the β position of an alanine with chemical shifts in each of these ranges. The cross-peak produced by the self-coupling of the ^1H hydrogen on each of these selected ^{13}C carbons lies on the diagonal. Every cross-peak along each horizontal line is coupled, respectively, to this ^{13}C and ^1H , each labeled with its eventual assignment. Each cross-peak is labeled by the position of the other coupled hydrogen in the side chain, and each cross-peak assigns the chemical shift of that other hydrogen. The sections through the three-dimensional spectrum centered on chemical shifts for ^{13}C of 39.9 and 40.9 ppm, respectively, contain cross-peaks arising from the absorptions of the ^{13}C carbons in the β positions of isoleucines with chemical shifts in each of these ranges. The cross-peaks produced by the self-coupling of the ^1H hydrogens on each of these selected β carbons lie on the diagonal. Every cross-peak along each horizontal line is coupled, respectively, to this β ^{13}C and β ^1H , each labeled with its eventual assignment. Each cross-peak is labeled by the position of the other coupled hydrogen in the side chain, and each cross-peak assigns the chemical shift of that other hydrogen. The sections through the three-dimensional spectrum centered on chemical shifts for ^{13}C of 55.0, 56.7, 58.0, and 58.6 ppm, respectively, contain cross-peaks arising from the absorptions of the ^{13}C carbons in the α positions of amino acids with chemical shifts in each of these ranges. The cross-peaks produced by the self-coupling of the α ^1H hydrogens on each of these selected α ^{13}C carbons lie on the diagonal. Every cross-peak along each horizontal line is coupled, respectively, to this α ^{13}C and α ^1H , each labeled with its eventual assignment. Each cross-peak is labeled by the position of the other coupled hydrogen in the side chain, and each cross-peak assigns the chemical shift of that other hydrogen. Reprinted from ref 259. Copyright 1990 American Chemical Society.

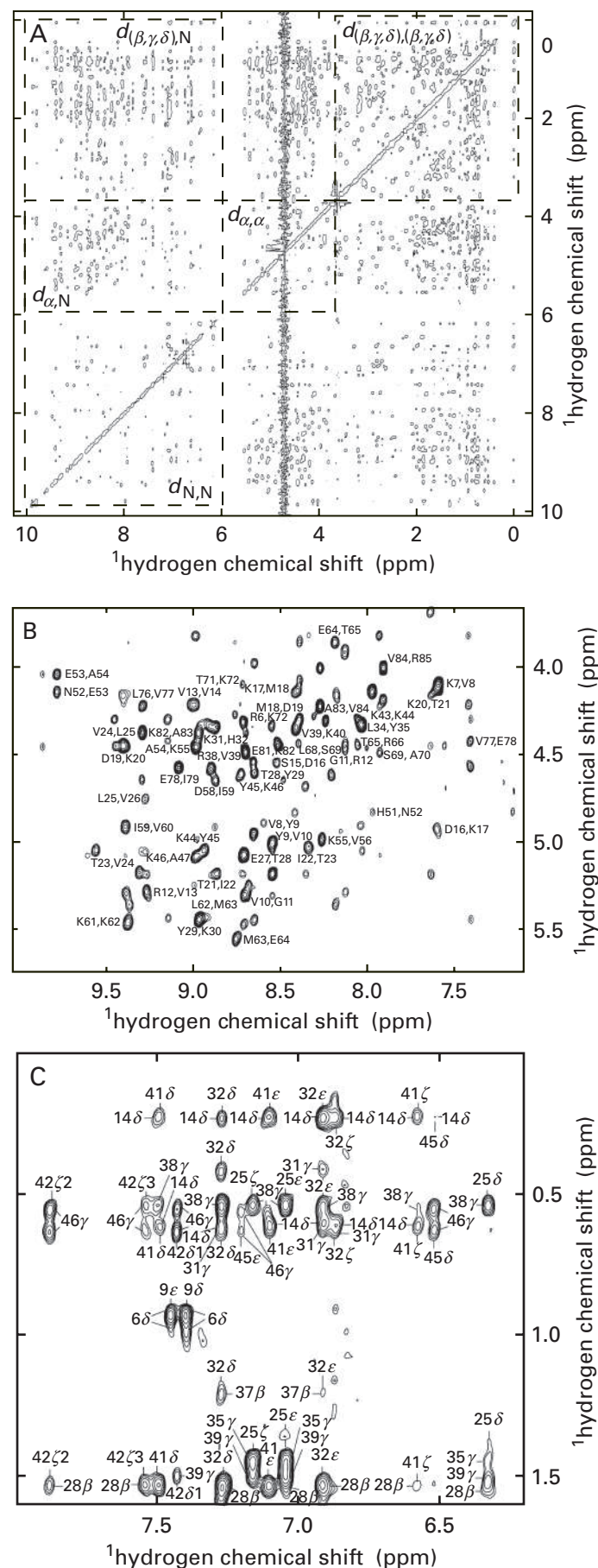
of ^1H hydrogens. The existence of these cross-peaks states that all of these ^1H hydrogens are in the vicinity of the amido ^1H hydrogen of Alanine 37 in the tertiary structure of the protein. As the length of the fixed delay t_m was increased, the intensity of the cross-peaks increased as more and more of the amplitude modulation of the population of the nuclei of amido ^1H hydrogens of Alanine 37 was transferred to the populations of neighboring nuclei.

Two **problems with nuclear Overhauser effects** that are the consequence of spin diffusion are that they are complicated by the spectral density function²⁰⁵ inherent to the dipolar interaction and that they are usually not confined to nuclei immediately adjacent to the source of the diffusing spin but spread outward from the source in rather complex pathways that cannot be delineated unless the detailed structure of the molecule is already known.^{206,271,272} The time t_m between the second and the third 90° pulses must be chosen by trial and

error to maximize the amount of transfer to immediately adjacent nuclei while minimizing the spread to more distant locations (Figure 12-24). Because the nuclear Overhauser effect results from a dynamic, inhomogeneous process, no reliable absolute measurements of particular distances between nuclei can be made. An intuition of relative distances between the nuclei can be gained, however, by following the changes in the intensity of the nuclear Overhauser effects as a function of the time interval t_m . If a nuclear Overhauser effect is one that develops early in the progress of spin diffusion, the two nuclei connected by that nuclear Overhauser effect are presumed to be close to each other (<0.5 nm) in the folded polypeptide.²⁷⁰

In the full two-dimensional nuclear Overhauser spectrum of a protein (Figure 12-23A), the one-dimensional spectrum of the individual absorptions of the ^1H hydrogens lies along the diagonal. The nuclear

Figure 12-23: Two-dimensional (^1H - ^1H) nuclear Overhauser enhanced spectra. (A) Full spectrum of a solution of ribosomal protein S17 from *Bacillus stearothermophilus* ($n_{\text{aa}} = 86$) dissolved in a 90:10 mixture of $^1\text{H}_2\text{O}$ and $^2\text{H}_2\text{O}$, respectively, at pH 6.5 and 25°C .²⁶⁹ In this spectrum, a cross-peak appears whenever the absorptions of two ^1H hydrogens are connected by a nuclear Overhauser effect. The two chemical shifts of each cross-peak are those of the two respective ^1H hydrogens. Because most ^1H hydrogens on adjacently bonded atoms are close enough to be connected by a nuclear Overhauser effect as well as spin-spin coupled through the bonds, most of the cross-peaks in a correlated spectrum of the protein are also present here. There are, however, more cross-peaks. The additional ones connect ^1H hydrogens on atoms adjacent in space but not connected by covalent bonds. Several regions are highlighted within boxes on the full spectrum: $d_{\text{N,N}}$, connections between ^1H hydrogens on different amide nitrogens; $d_{\alpha,\text{N}}$, connections between α ^1H hydrogens and amide ^1H hydrogens; $d_{(\beta,\gamma,\delta),\text{N}}$, connections between ^1H hydrogens on β , γ , or δ carbons and amide ^1H hydrogens; $d_{\alpha,\omega}$, connections between ^1H hydrogens on two different α carbons; and $d_{(\beta,\gamma,\delta),(\beta,\gamma,\delta)}$, connections between two different β , γ , or δ ^1H hydrogens. Reprinted from ref 269. Copyright 1996 American Chemical Society. (B) Expansion of the $d_{\alpha,\text{N}}$ region of spectrum A. Each cross-peak is a connection produced by a nuclear Overhauser effect between the α ^1H hydrogen on one amino acid and the amide ^1H hydrogen on another. The identity of the amino acids on which the paired ^1H hydrogens are located is determined by the chemical shifts at which the cross-peak is situated. Those that connect consecutive amino acids in the sequence of the protein are labeled to illustrate the fact that most but not all of the nuclear Overhauser effects in this region are local and uninformative. Reprinted from ref 269. Copyright 1996 American Chemical Society. (C) Spectrum of a 2 mM solution of the γ -carboxyglutamate-rich domain of human factor IX ($n_{\text{aa}} = 47$) dissolved in a 90:10 mixture of $^1\text{H}_2\text{O}$ and $^2\text{H}_2\text{O}$, respectively, at pH 5.3 and 35°C .²⁶⁸ The region of the spectrum displayed contains cross-peaks between ^1H hydrogens on aromatic rings (horizontal axis) and ^1H hydrogens on methyl groups (vertical axis). Because aromatic amino acids have no methyl groups, most of the cross-peaks in this region, unlike those in panel B, connect amino acids distant from each other in the sequence of the protein. The cross-peaks are identified in the respective dimensions by the number of and position in the amino acids containing the two ^1H hydrogens. Reprinted with permission from ref 268. Copyright 1995 American Chemical Society.



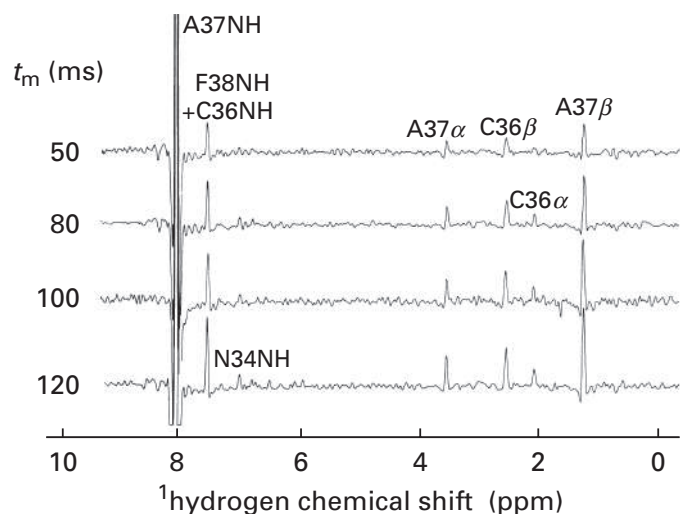


Figure 12-24: Diffusion of saturation, labeled with its Larmor frequency, from a nucleus of ^1H hydrogen into surrounding nuclei of ^1H hydrogen as a function of the length of the fixed delay t_m .²⁷⁰ Each trace is a cross section through a two-dimensional (^1H - ^1H) nuclear Overhauser enhanced spectrum of a 16 mM solution of acrosin inhibitor IIA from bovine seminal plasma ($n_{\text{aa}} = 57$) in $[\text{H}]_2\text{O}$ at pH 5.3 and 47 °C. Each cross section cuts through one of the two-dimensional spectra at a chemical shift δ_1 of 8.45 ppm, which is the chemical shift of the amido ^1H hydrogen on Alanine 37. Other ^1H hydrogens connected to this hydrogen by nuclear Overhauser effects are represented by cross-peaks in the dimension of ^1H hydrogen chemical shift δ_2 (horizontal axis). They are labeled by the ^1H hydrogen to which they have been assigned by the individual values of their chemical shifts. The large peak labeled A37NH is the self-connection of the amido ^1H hydrogen of Alanine 37. Each cross section is from a two-dimensional spectrum gathered with a different fixed delay t_m , noted to its left in milliseconds. Reprinted with permission from ref 270. Copyright 1985 Academic Press.

spin-spin-coupled to a ^{13}C carbon or a ^{15}N nitrogen the chemical shift of which falls within a narrow range of values are registered. For example, only the ^1H hydrogens connected by nuclear Overhauser effects to ^1H hydrogens on ^{13}C carbons with chemical shifts of 13.3 ppm are registered in the left strip in Figure 12-25. The nuclear Overhauser effects in such spectra are assigned to particular pairs of hydrogens on the basis of the two chemical shifts of each cross-peak (Figure 12-23B,C) or the two chemical shifts of the cross-peak and the chemical shift of a ^{13}C carbon or ^{15}N nitrogen to which one or the other of the ^1H hydrogens is spin-spin-coupled. These chemical shifts were determined during the initial assignments of chemical shifts to all of the nuclei in the protein.

As many pairs of hydrogens coupled by nuclear Overhauser effects as possible are identified and catalogued. For example, 531 pairs of ^1H hydrogens were identified as being connected by nuclear Overhauser effects in spectra of the major cold-shock protein ($n_{\text{aa}} = 70$) from *E. coli*;²⁷⁶ 1281 pairs, in spectra of glutaredoxin 2 ($n_{\text{aa}} = 215$) from *E. coli*;²⁷⁷ and 3125 pairs, in spectra of phosphoglycerate mutase ($n_{\text{aa}} = 205$) from *Schizosaccharomyces pombe*.²⁷⁸ As is usually the case, these connections were spread unevenly over the amino

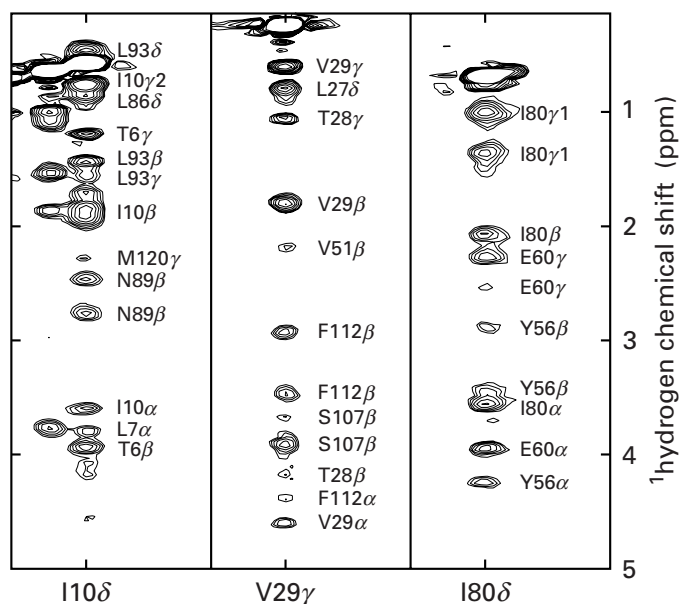


Figure 12-25: Three-dimensional (^1H - ^{13}C - ^1H) NOESY-HMQC spectra of a 2 mM solution of human interleukin-4 ($n_{\text{aa}} = 129$) that had been expressed in *E. coli* grown on $[\text{H}_4]\text{SO}_4$ and $[\text{H}_3]\text{glycerol}$ as sole sources of nitrogen and carbon and that was dissolved in $[\text{H}_2]\text{O}$ at pH 4.5 and 20 °C.²⁷³ The three strips are from two-dimensional sections through the three-dimensional spectra. The three two-dimensional sections, cut in the ^{13}C carbon dimension, contain cross-peaks from ^{13}C carbons with chemical shifts of 13.3, 16.6, and 6.7 ppm, respectively, which are the chemical shifts of the ^{13}C carbons in the δ methyl group of Isoleucine 10, one of the γ methyl groups of Valine 29, and the δ methyl group of Isoleucine 80. Each strip from each of these two-dimensional sections is centered on the chemical shift (horizontal axis) of the ^1H hydrogens of the respective methyl group. The most intense cross-peak on each strip is the self-connection of those hydrogens and is not labeled. The vertical axis defines the chemical shifts of the other hydrogens connected to the respective methyl hydrogens by nuclear Overhauser effects. Each of these cross-peaks, identified by its chemical shift, is labeled with the amino acid and the position in that amino acid at which the ^1H hydrogen producing the nuclear Overhauser effect is located. Reprinted from ref 273. Copyright 1994 Elsevier B.V.

acid sequences of these proteins with as few as 5–10 involving ^1H hydrogens on one particular amino acid to as many as 160 involving ^1H hydrogens on another.²⁷⁷ The latter values are extraordinarily impressive because no one hydrogen in a protein can have more than about 20–25 hydrogens within 0.5 nm of it, and the ^1H hydrogens on methyl groups are indistinguishable from each other in chemical shift and do not register as separate ^1H hydrogens. It is from this **catalogue of pairs of connected ^1H hydrogens** that a molecular model is built.

It is the nuclear Overhauser effects between ^1H hydrogens that are on amino acids two or more positions away from each other in the amino acid sequence of a protein that provide the information on which the molecular model is based. Nuclear Overhauser effects between ^1H hydrogens within the same amino acid or on immediately adjacent amino acids are usually uninformative because the covalent structure requires that they occur.

Usually, only about 50% of the assigned nuclear Overhauser effects arise from ^1H hydrogens that are on amino acids two or more positions away from each other in the amino acid sequence of a protein.^{276–279} Certain regions of a two-dimensional nuclear Overhauser enhanced spectrum, such as the one containing connections between α ^1H hydrogens and amido ^1H hydrogens, are dominated by pairs of ^1H hydrogens that are on amino acids at adjacent positions in the amino acid sequence (labeled cross-peaks in Figure 12–23B), while other regions, such as the one containing connections between aromatic ^1H hydrogens and aliphatic ^1H hydrogens (labeled cross-peaks in Figure 12–23C), are dominated by pairs of ^1H hydrogens distant from each other in the primary structure but adjacent to each other in the tertiary structure of the protein. It is these latter types of nuclear Overhauser effects that draw together two distant hydrogens as the covalent structure of the protein is folded into the molecular model.

A **nuclear magnetic resonance molecular model** is a molecular model of the covalent structure of the protein folded by the builder of the model into a tertiary structure in which the maximum number of ^1H hydrogens observed to be connected by short-range nuclear Overhauser effects end up close (≤ 0.5 nm) to each other. It is possible to start with a molecular model of the extended polypeptide in a computer and use molecular dynamics and simulated annealing, modified so that the potential function includes the constraints of the nuclear Overhauser effects, to produce a preliminary molecular model, similar to the preliminary crystallographic molecular model that results from inserting the molecular model of the polypeptide into the map electron density.

The validity of this initial molecular model can be assessed by examining its secondary structure. Just as α helices and β structure can be recognized in a map of electron density, segments of the polypeptide that are α helices or β structure in the actual molecule of protein can be recognized by **patterns in the observed nuclear Overhauser effects** (Figure 12–26).²⁷⁰

The most dominant pattern is that of the ^1H hydrogens in an **α helix**. In an α helix, nuclear Overhauser effects systematically connect hydrogens in each amino acid to those in amino acids three positions and four positions (Figure 6–6) away from it in the amino acid sequence.^{280–282} An α helix holds the consecutive nitrogen–hydrogen bonds of the amides of the backbone close to each other, and these short distances promote transfer of saturation. For example, in the two-dimensional nuclear Overhauser enhanced spectrum of the anaphylatoxin from human complement factor 3a, 36 of the 44 nuclear Overhauser effects observed between amido ^1H hydrogens on spatially adjacent amino acids were those for amino acids in α helices, while only 46 of the 77 amino acids in the protein are in α helices.²⁸³

In **parallel β structure**, ^1H hydrogens in a string of successive amino acids in the sequence are connected in

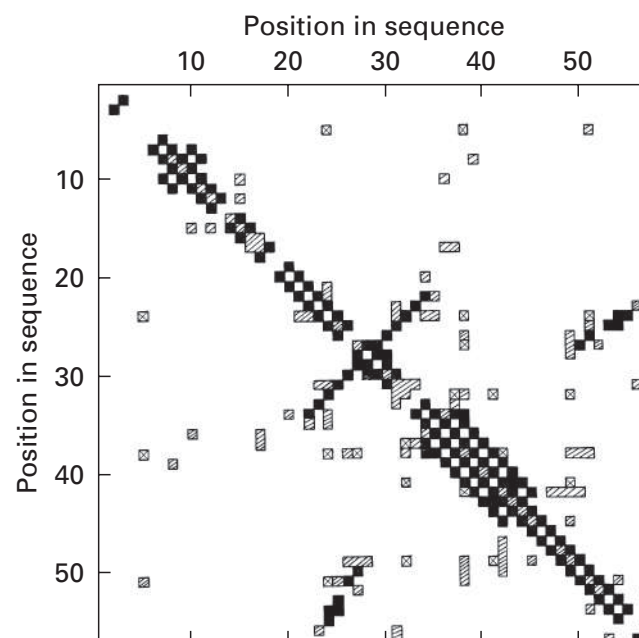
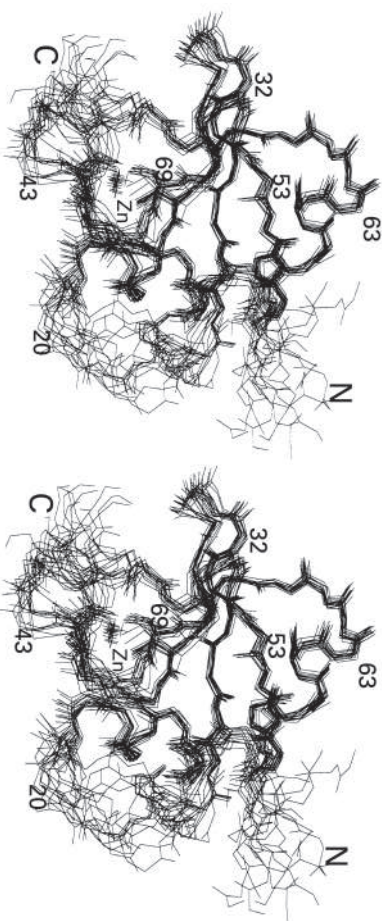


Figure 12–26: Diagonal plot of the nuclear Overhauser effects observed between different amino acids in the amino acid sequence of acrosin inhibitor IIA from bovine seminal plasma.²⁷⁰ Each nuclear Overhauser effect was established by the existence of a cross-peak in a two-dimensional nuclear Overhauser enhanced spectrum of the protein that had two chemical shifts identical to the respective chemical shifts of particular ^1H hydrogens on the two different amino acids. The two axes are the numbering of the amino acids in the sequence of the protein, and a square represents a connection between the two positions in the sequence. Solid squares represent nuclear Overhauser effects between ^1H hydrogens on α carbons or amido ^1H hydrogens from the respective amino acids; hatched squares, between a ^1H hydrogen on an α carbon or amido ^1H hydrogen on one amino acid and a ^1H hydrogen on the side chain of the other; squares with \times , between ^1H hydrogens on the side chains of both amino acids. Patterns of connections can be recognized in the plot that define three turns of α helix from positions 34 to 45 and one turn of α helix from positions 8 to 11, and three segments from positions 52 to 55, 27 to 23, and 29 to 33 define three strands of antiparallel β structure in a pleated sheet. Reprinted with permission from ref 270. Copyright 1985 Academic Press.

pairs to ^1H hydrogens in another string of successive amino acids in the order in which those spectrally connected pairs of amino acids occur in the sequence. In **antiparallel β structure**, the pairs of amino acids are connected in the reverse order to the order in which they occur in the sequence. For example, connections between amino acids at positions 20 and 17, 21 and 16, 22 and 15, 23 and 14, 24 and 13, and 25 and 12 defined an antiparallel β hairpin in α -amylase inhibitor HOE-467A from *Streptomyces tendae*.²⁸⁴

Other patterns indicating the organization of the tertiary structure of the actual molecule of protein can also be recognized in the spectra. For example, patterns similar to those for antiparallel β structure, in which ^1H hydrogens in one segment of amino acids are connected in reverse order to ^1H hydrogens in another segment of amino acids, can identify two adjacent, antiparallel

Figure 12-27: Nuclear magnetic resonance molecular model of the amino-terminal domain ($n_{\text{aa}} = 92$) of the ADP regulatory protein of *E. coli*.^{293,294} The nuclear Overhauser effects observed in spectra of the protein were used as constraints in a computing program written to build a tertiary structure from a model of the polypeptide with optimal bond lengths and bond angles by first bringing as many hydrogens connected by nuclear Overhauser effects to within 0.5 nm of each other and then adjusting the structure to obtain the optimal length for designated hydrogen bonds and optimal dihedral angles and to eliminate overlap of atomic volumes. An ensemble of superposed structures is presented in stereo, each of which is equally compatible with the constraints imposed. The positions of a structural Zn^{2+} cation (Zn) in the various superposed structures are indicated by adjacent crosses. Reprinted with permission from ref 293. Copyright 1996 American Chemical Society.



α helices. In this instance, however, the patterns are discontinuous because only ^1H hydrogens within the interface between the two α helices are connected by nuclear Overhauser effects.²⁸⁵⁻²⁸⁷

Just as the preliminary crystallographic molecular model is then submitted to refinement against the data set, the preliminary nuclear magnetic resonance molecular model is submitted to **refinement** with the nuclear Overhauser effects as constraints. In addition to nuclear

Overhauser effects, other constraints can be applied to the process of refinement. Designated donors and acceptors of hydrogen bonds in α helices and β structure can be assigned ideal lengths.²⁸⁸ Dihedral angles within the structure can be constrained to particular ranges by values of observed coupling constants.^{289,290} If the protein contains a paramagnetic metallic cation, the effect of that cation on the relaxation rates of ^1H hydrogens in the protein can provide estimates of the distances between each of those ^1H hydrogens and the metallic cation.²⁹¹

In the final refined nuclear magnetic resonance molecular model, segments of **random meander** connecting clearly defined segments of secondary structure will often be less well defined even though there is crystallographic or chemical evidence that they do assume specific structures.^{292,293} This problem is emphasized by the practice of presenting a nuclear magnetic resonance molecular model as an ensemble of structures, each of which satisfies the constraints of the nuclear Overhauser effects (Figure 12-27).^{293,294} In such representations, the certainty of the structures of α helices and β structure is set in sharp contrast to the uncertainty of the structure of the random meander. Although such a representation implies that these poorly defined segments are more flexible and less rigidly confined than the central regions of regular secondary structure, crystallographic molecular models of the same protein often show no evidence of such flexibility.²⁹⁵ It is possible to distinguish whether or not the poor definition of these regions of random meander results from dynamic flexibility by examining the rates of relaxation of the ^{15}N nitrogens in these regions. These rates of relaxation are sensitive to thermal motion and can be used to identify segments of polypeptide that are dynamically flexible in the actual molecule of protein. In the absence of evidence for flexibility, it must be assumed that the poor definition of random meander is a consequence of an insufficiency of constraints in the data.

In refined nuclear magnetic resonance molecular models, it is those regions of the protein that are within or sandwiched between α helices or β structure that are the most precisely defined. There are, however, regions of the secondary structure that are poorly defined by nuclear magnetic resonance. Hydrogens known to be within short segments of secondary structure, such as **β turns** or 3_{10} helix, participate in so few nuclear Overhauser effects that those that are observed are often inadequate to define the structure of these segments.²⁹⁶

Molecules of **water** confined to particular locations on the surface of the protein can be incorporated into the molecular model on the basis of the nuclear Overhauser effects between their ^1H hydrogens and ^1H hydrogens of the amino acids by using rotating-frame Overhauser enhanced spectroscopy (ROESY).²⁹⁷ Molecules of water, however, at locations buried within the structure of the molecule of protein have residence times long enough to be observed directly by their nuclear Overhauser effects.

These two types of locations for molecules of water, exterior and interior, are usually found to occupy the same positions in the nuclear magnetic resonance molecular model that they do in the crystallographic molecular model of the same protein.²⁹⁸ The position of metallic cations in the molecular model can be established by substituting the natural cation with a cation of nuclear spin $\frac{1}{2}$. For example, the Zn^{2+} cations normally bound to the amino-terminal domain of regulatory protein GAL4 from *S. cerevisiae* ($n_{\text{aa}} = 62$) were replaced with $^{113}\text{Cd}^{2+}$ cations, and spin-spin couplings between the $^{113}\text{Cd}^{2+}$ cations and the β ^1H hydrogens on the cysteines that covalently bind them (6–19) produced a two-dimensional correlated spectrum.²⁹⁹

The fundamental problem with building a molecular model from nuclear Overhauser effects is that those nuclear Overhauser effects do not define a distance between two ^1H hydrogens because the relative rates of spin diffusion are too dependent on the character of the unique surroundings around each nucleus to obtain reliable estimates of distances. The distances in the final refined model between pairs of ^1H hydrogens connected by nuclear Overhauser effects are always quite different even though almost all of them can be made less than 0.5 nm.²⁸⁸ When distances between ^1H hydrogens connected by nuclear Overhauser effects are measured in a crystallographic molecular model of the same protein,²⁹⁶ there is usually little correlation between the actual distance between the hydrogens and the strength of the nuclear Overhauser effect at the optimal mixing time t_m . Those nuclear Overhauser effects observed only after extended mixing times do arise from hydrogens that are farther apart, but the range of those longer distances is broad, and consequently they are not very useful.²⁹⁶ If a nuclear Overhauser effect is observed between two ^1H hydrogens after an optimal mixing time t_m , it can only be assumed that they are less than 0.5 nm apart,^{205,272} but there are usually notable exceptions even to this limit.^{296,300–302}

In effect, the existence of a nuclear Overhauser effect allows the investigator to connect the two hydrogens in a molecular model of the covalent structure of the polypeptide with a rubber band that is elastic enough to stretch to a distance equivalent to about 0.5 nm. If enough of these rubber bands were inserted into a molecular model of the polypeptide and the regions of the amino acid sequence identified as β structure and α helix (Figure 12–26) had already been locked into these secondary structures, the model would snap into a conformation that resembles the native folded conformation of the polypeptide.

It has been possible in many instances to compare nuclear magnetic resonance molecular models with crystallographic molecular models. It is usually observed that the two molecular models resemble each other, occasionally quite closely.²⁷⁹ The resemblance is strongest in the assignment of α helices and β struc-

ture.²⁷⁹ The arrangement of these regular secondary structures in the tertiary structure, however, often differs significantly, but not dramatically, between the two molecular models.^{303–306} Some of these differences are real and informative.³⁰⁷ They often result from the fact that the protein in question is small and flexible or has flexible domains and the fact that contacts between the molecules of protein in the crystal are able to shift secondary structures relative to each other.³⁰⁶ If a protein is constructed so that in solution it assumes two or more conformations because shifts among these conformations are required for its function, nuclear magnetic resonance can be used to determine which crystallographic molecular models of these various conformations represent the species actually present in solution.³⁰⁸

In **superpositions of nuclear magnetic resonance and crystallographic molecular models** of the same protein, the root mean square deviations between heavy atoms (oxygen, nitrogen, and carbon) are usually the least within the polypeptide backbone of the regular secondary structure in the core, greater for side chains buried between these secondary structures in the core, and greatest for random meander at the periphery.^{300,303,309,310} In the comparison of the two molecular models for α -amylase inhibitor HOE-467A ($n_{\text{aa}} = 74$), the value of the root mean square deviation for the heavy atoms of the polypeptide backbone was 0.105 nm; that for the heavy atoms of the side chains buried in the core was 0.125 nm; and that for all heavy atoms was 0.184 nm.³⁰⁰ This protein, however, is almost entirely β structure with little random meander. In the comparison of the two molecular models of human granulocyte colony-stimulating factor ($n_{\text{aa}} = 174$ aa), a much larger protein with significant amounts of random meander, the root mean square deviation for the heavy atoms of the polypeptide backbone in its four α helices was 0.286 nm; that for all of the heavy atoms in these α helices was 0.333 nm; that for all of the heavy atoms in the entire polypeptide backbone was 0.315 nm; and that for all heavy atoms was 0.370 nm.³⁰³

It is in the **details of the atomic structure** that nuclear magnetic resonance and crystallographic molecular models differ most significantly. For example, the dispositions of the aromatic rings of Tyrosine 3, Tyrosine 45, and Phenylalanine 52 were different in two nuclear magnetic resonance molecular models of the immunoglobulin-binding domain of immunoglobulin G binding protein G from *Streptococcus* ($n_{\text{aa}} = 56$), and those dispositions in turn were both different from the dispositions in the crystallographic molecular model.³¹¹ In nuclear magnetic resonance molecular models it is the conformations of the side chains that are always more uncertain than those of the polypeptide backbone;³⁰⁰ but, unfortunately, it is the conformations of the side chains that usually accomplish the function of the protein. There are, however, some instances in which the conformation of a particular side chain in a nuclear mag-

netic resonance molecular model is more compatible with its known chemical properties than its conformation in the crystallographic molecular model of the same protein³¹² and other instances in which the atomic details of the crystallographic molecular model could be adjusted by nuclear magnetic resonance.³¹³

A nuclear magnetic resonance molecular model is a significantly less accurate²⁰⁵ representation of the actual structure of a molecule of protein than is a crystallographic molecular model for the following reasons. First, a crystallographic data set contains significantly more information than even the most extensive list of nuclear Overhauser effects and coupling constants³¹⁰ because the number of observable nuclear Overhauser effects is always less than the number of observable reflections and because each reflection comes with an amplitude. Second, the heavy reliance on energy minimization in building the nuclear magnetic resonance molecular model assures that unusual local conformations that are excluded by the procedure used for the minimization either intentionally or unintentionally will be missed. Third, it has been demonstrated by calculation that even with an average of 19 nuclear Overhauser effects for each amino acid, a value in excess of the number usually available, the root mean square deviation of the atoms in a nuclear magnetic resonance molecular model from their actual positions in the structure of the protein has to be at least 0.1 nm.²⁰⁵ Even this is an overestimate of the accuracy because the paucity of nuclear Overhauser effects from random meander was not considered when the pairs of hydrogens incorporated into the calculations were chosen. Fourth, because the crystallographic data constrains the crystallographic molecular model much more than the nuclear magnetic resonance data constrains the nuclear magnetic resonance molecular model, upon refinement with a combination of both the crystallographic data set and the observed nuclear Overhauser effects, the crystallographic molecular model of a protein quickly converges with only small changes in its structure to accommodate both sets of data while the nuclear magnetic resonance molecular model undergoes far more extensive changes to reach accommodation.³¹⁰ Fifth, the fact that the number of nuclear Overhauser effects observed between ¹hydrogens in the same amino acid that are inescapably greater than 0.5 nm apart in the final nuclear magnetic resonance molecular model is much greater than the number of nuclear Overhauser effects observed between ¹hydrogens in different amino acids that are greater than 0.5 nm apart in the final nuclear magnetic resonance molecular model indicates that in bringing as many ¹hydrogens connected by nuclear Overhauser effects as possible to within 0.5 nm of each other, the construction of the molecular model has produced a structure significantly different from the actual structure of the molecule of protein.³¹⁴ Sixth, it is usually observed that increasing the number of constraints in the construction of the nuclear

magnetic molecular model causes it to become closer to the crystallographic molecular model of the same protein rather than to assume its own distinct structure.²⁹⁰

There are, however, informative exceptions to this rule and in these instances, nuclear magnetic resonance does reveal differences in the **structure of a protein when it is in solution** and when it is in a crystal.^{307,308,315} As with solution scattering of X-rays, these differences permit the crystallographic molecular model to be adjusted, often minutely,³¹³ to a conformation representing the molecule of protein in solution, which is the goal of all structural studies.

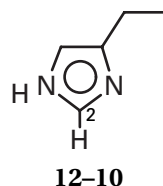
Perhaps the greatest drawback of nuclear magnetic resonance spectroscopy is that it is confined to small proteins. This confinement is due not only to the problem of overlapping cross-peaks on two-dimensional and three-dimensional spectra. Because the rate at which a molecule rotationally diffuses in the solution affects both the ratio of signal to noise in a nuclear magnetic resonance spectrum and the effectiveness of the pulse sequences used for multidimensional and multinuclear spectra, the **size of a molecule of the protein** determines whether or not it will even yield a spectrum. Although methods have been reported that can increase the rate at which a molecule of protein rotationally diffuses in a solution,³¹⁶ this problem has yet to be solved satisfactorily. From 1986 to 2001, the size of the largest asymmetric units and the size of the largest symmetric dimers^{317,318} for which nuclear magnetic resonance provided a molecular model increased from 120 to 220 aa and from 200 to 450 aa, respectively.* The average size of the proteins for which molecular models were reported, however, increased only modestly during the same period, from about 90 to about 125 aa. Unfortunately, most proteins are oligomers with asymmetric units larger than 300 aa, sizes that present no difficulty to crystallography.

Because of their poor definition of the structure of random meander and of the conformation of side chains, because of their inaccuracy, because of their indistinguishability from crystallographic molecular models, and because of their confinement to small proteins, nuclear magnetic resonance molecular models have provided far less structural information than have crystallographic molecular models. Although there are situations in which nuclear magnetic resonance can be applied when crystallography cannot, for example in defining the details of the conformational change that occurs upon the binding of porcine phospholipase A₂ to micelles of dodecyl phosphocholine,³²⁰ and situations in which nuclear magnetic resonance establishes clear and significant differences between a crystallographic molecular model and the structure of a protein in solution, for example in showing that the central unsupported α helix

* A preliminary nuclear magnetic molecular model of malate synthase from *E. coli*, which is a monomer of 723 amino acids, has been reported.³¹⁹

in the crystallographic molecular model of calmodulin does not form in solution,²⁸⁰ the dramatic advantages that nuclear magnetic resonance has over crystallography are that it readily observes hydrogens and it observes proteins while they are in solution. Both of these advantages are exploited when nuclear magnetic resonance is used to monitor acid–base titrations of individual side chains and when it is used to monitor the exchange of specific amido protons or deuterons in the polypeptide backbone with deuterons or protons, respectively, of the water in which the protein has been dissolved.

The first successful application of nuclear magnetic resonance spectroscopy in the study of proteins was the determination of the **acid dissociation constants** of their **histidines**. The π electrons of an aromatic ring (2–24) are induced to circulate in a ring current by an applied magnetic field. This ring current creates a toroidal magnetic field opposite in direction to the applied field in the center of the ring but reinforcing the applied field at the periphery of the ring. This additional local magnetic field at the periphery causes all of the ^1H hydrogens around aromatic rings to absorb at higher frequencies and hence greater chemical shift (Equations 12–53 and 12–54). The nitrogens on either side of carbon 2 of the imidazole of a histidine



are electronegative elements that withdraw electrons from carbon 2, decreasing the shielding provided by the σ electrons in the carbon–hydrogen bond and shifting the absorption of a ^1H hydrogen on carbon 2 further downfield and away from the absorptions of the other aromatic ^1H hydrogens on phenylalanines, tyrosines, and tryptophans. The absorption of the ^1H hydrogen on carbon 2 of the imidazole of a histidine is not divided by spin–spin coupling when the adjacent protons on the two nitrogens have been exchanged with deuterons from the $[\text{D}_2]\text{H}_2\text{O}$ in which the protein is dissolved. For all of these reasons, the absorption from this ^1H hydrogen on each histidine in a protein dissolved in $[\text{D}_2]\text{H}_2\text{O}$ appears as a sharp individual peak in the nuclear magnetic resonance spectrum. One of the first nuclear magnetic spectra displaying these absorptions in a native protein was the spectrum for ribonuclease (Figure 12–28).³²¹ Improvements have been made since these early experiments that sharpen the peaks of absorbance and eliminate peaks in this region of the spectrum from unexchanged amido ^1H hydrogens.³²²

As the imidazole of histidine gains a proton during its acid–base reaction, the nitrogens of the conjugate acid become even more electron-withdrawing than those of the conjugate base, and the absorption of the

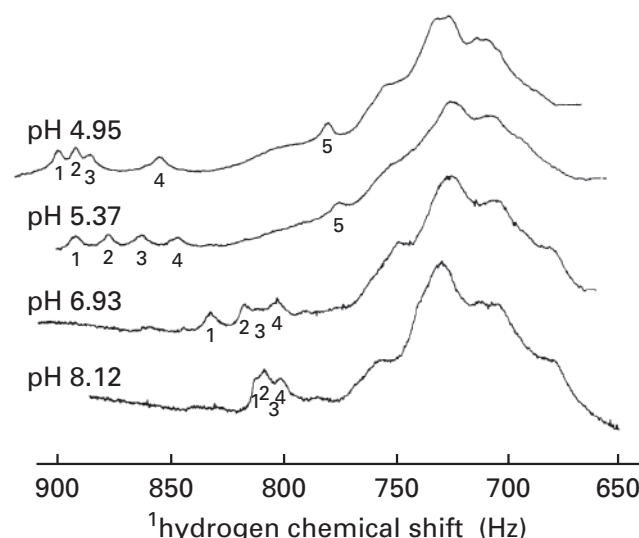


Figure 12–28: Low-resolution (100 MHz) nuclear magnetic resonance spectra covering the region of the unresolved absorptions of the ^1H hydrogens of the six tyrosines and three phenylalanines (aromatic) and the four resolved absorptions of the ^1H hydrogens on the carbons 2 of the four histidines (numbered 1–4) of ribonuclease A.³²¹ Absorption is presented as a function of chemical shift (in hertz from the peak of absorption of an internal standard). As the excitation frequency is 100 MHz, 100 Hz of chemical shift is 1 ppm. The sample was a 12 mM solution of ribonuclease A ($n_{\text{aa}} = 124$) in deuterioacetate buffers in $[\text{D}_2]\text{H}_2\text{O}$ at various values of p^2H (noted on the drawing). Peak 5 is a proton on carbon 4 of one of the histidines. Reprinted with permission from ref 321. Copyright 1967 retained by authors.

^1H hydrogen on carbon 2 assumes an even larger chemical shift (notice the movement of the peaks in Figure 12–28 to higher chemical shift as the pH decreases). Because a specific fraction of the imidazoles in a population of identical histidines is the cationic conjugate acid at a particular pH and because the transfer of protons among the individuals in that population is much faster than the time resolution of nuclear magnetic resonance spectroscopy, the absorption of the population of ^1H hydrogens on carbon 2 of a given population of histidines in a protein assumes a chemical shift, $\delta_{\text{H,obs}}$, that is the weighted mean between that of the neutral conjugate base $\delta_{\text{H,A}}$ and that of the cationic conjugate acid $\delta_{\text{H,HA}}$:

$$\delta_{\text{H,obs}} = f_{\text{A}} \delta_{\text{H,A}} + f_{\text{HA}} \delta_{\text{H,HA}} \quad (12-57)$$

where f_{A} and f_{HA} are the fractions of conjugate base and conjugate acid, respectively, at a particular pH. Therefore, the chemical shift as a function of pH traces the **titration curve** of a particular histidine in a molecule of protein. The first application of this method was the measurement of the titration curves for the four histidines of ribonuclease.³²¹

In a number of proteins, such as ribonuclease,^{323,324} myoglobin,³²⁵ subtilisin,³²⁶ and carbonate dehydratase,^{327,328} the individual peaks of absorption from ^1H hydrogens on the carbons 2, and hence the titrations of

their imidazoles, could be assigned to specific histidines in the sequence of each protein. These assignments are now usually made by mutating each of the histidines consecutively to another amino acid and observing which of the absorptions disappears in each mutant.^{326,329} The acid dissociation constants for particular histidines in native proteins have been used to test computational methods^{324,330} for assessing the effect of electric field and relative permittivity on the acid dissociation constant of a particular histidine in a crystallographic molecular model.^{324,325,330}

The absorptions of the ^1H hydrogens on the carbons 2 of the histidines in a protein can be observed in its one-dimensional nuclear magnetic resonance spectrum. Although the absorptions of the individual ^1H hydrogens on the indole nitrogens of tryptophan³³¹ and the individual ^1H hydrogens on the methyl groups of threonines³³² can also often be observed in a one-dimensional spectrum, these ^1H hydrogens do not register acid dissociations.

The absorptions from nuclei in the side chains of a particular type of amino acid in a protein, however, can also be isolated on a one-dimensional spectrum by expressing that protein in bacteria that are grown on ^{13}C carbon or ^{15}N nitrogen versions of that particular type of amino acid. When this is done, a bacterium auxotrophic for that amino acid is chosen for the expression. For example, endo-1,4- β -xylanase from *Bacillus circulans* could be expressed in a strain of *E. coli* that was missing the three transaminases normally capable of producing glutamate.²⁴⁶ In such cells, only the $[\delta\text{-}^{13}\text{C}]$ glutamate added to the growth medium is incorporated into the expressed protein.³³³ When the xylanase was purified from this expression system, a peak of absorption from the carboxyl ^{13}C carbon of each of its two **glutamates** could be readily observed in a one-dimensional ^{13}C nuclear magnetic resonance spectrum, and the values of the chemical shifts for these two peaks as a function of the pH of the solution traced the acid-base titrations of their side chains (Figure 12–29).³³⁴ Using enrichments such as the one in this example permits acid-base titrations to be performed on fairly large proteins (the xylanase is a monomer of 185 aa) without the necessity of making a full set of assignments.

If the cross-peaks in two- and three-dimensional spectra and the chemical shifts of all of the nuclei in a protein have been assigned, it is possible to follow in those two- or three-dimensional spectra the chemical shift of a particular nucleus in a particular side chain as it changes with pH to produce a titration curve for an acid-base in that side chain. For example, the chemical shifts of the β ^1H hydrogens on the **aspartates** and the γ ^1H hydrogens on the glutamates in murine epidermal growth factor, which could be monitored with two-dimensional (^1H – ^1H) TOCSY spectra (Figure 12–19), decrease in magnitude as the respective adjacent carboxyl groups lose their protons as the pH is increased.³³⁵ The chemical shifts of the nuclei in ribonuclease H from *E. coli* ($n_{\text{aa}} = 155$) that had been enriched in ^{13}C carbon and

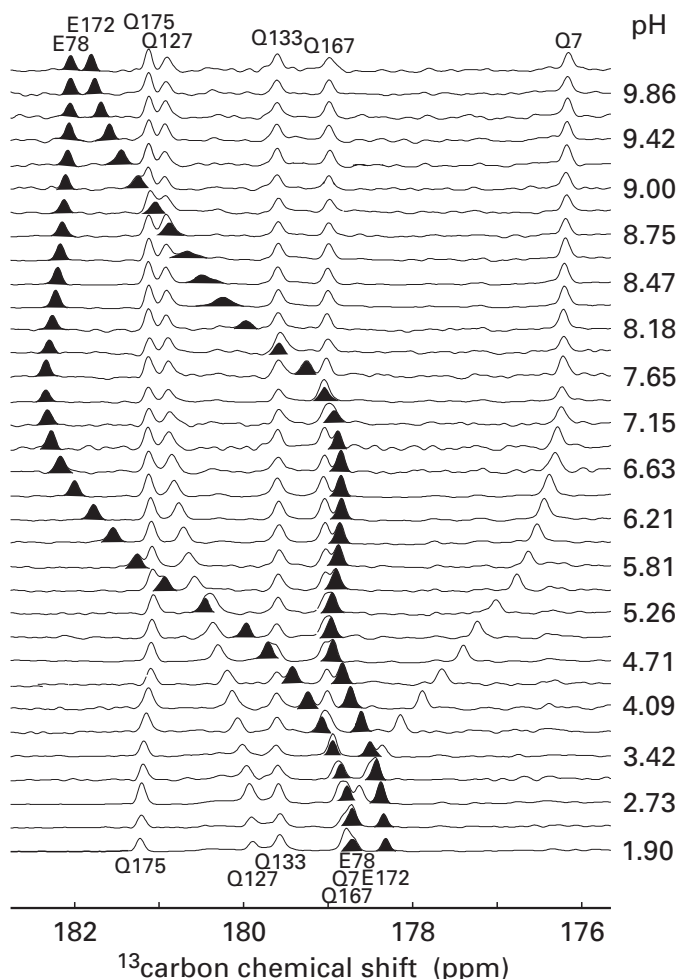
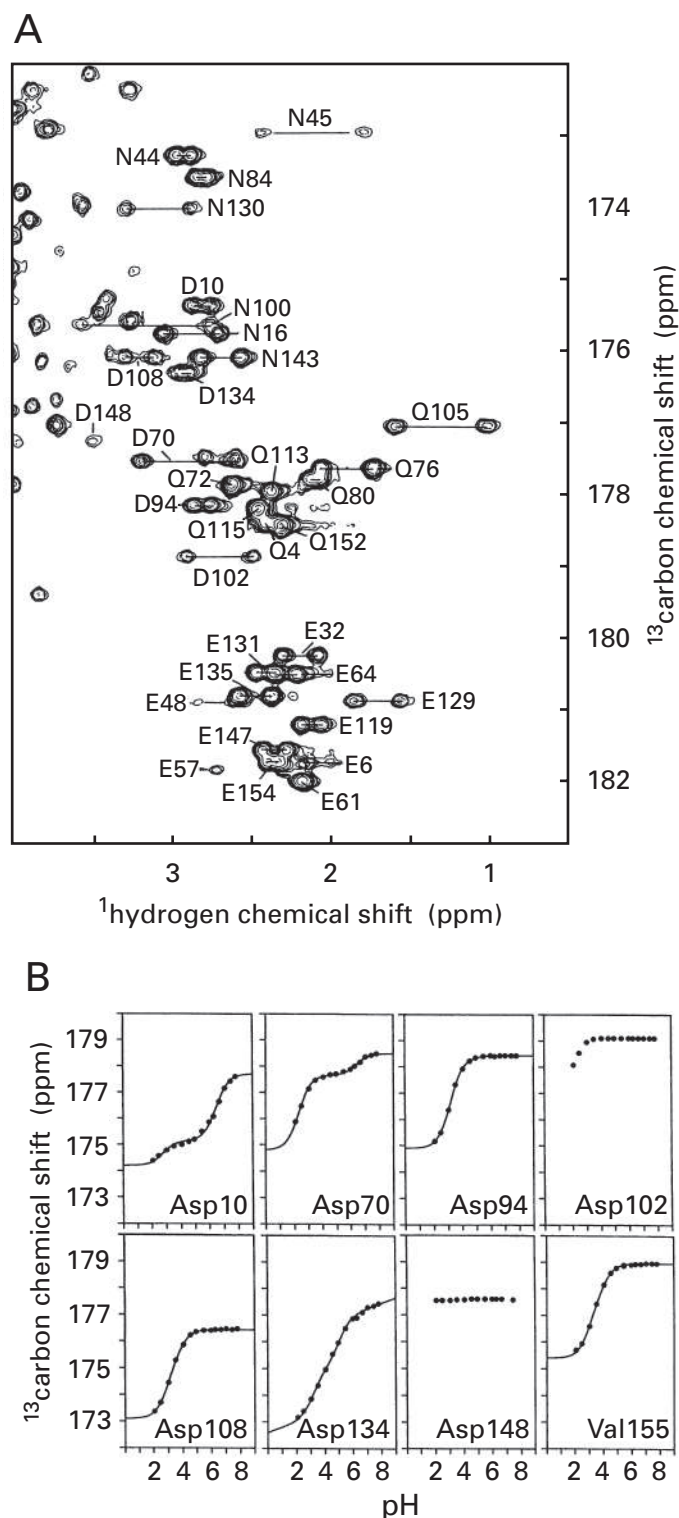


Figure 12–29: Acid–base titration of Glutamate 78 and Glutamate 172 in endo-1,4- β -xylanase ($n_{\text{aa}} = 185$) from *B. circulans*.³³⁴ A strain of *E. coli*, DL39, which is deficient in amino acid transaminases, was used to express the xylanase. It was grown on a medium containing $[\delta\text{-}^{13}\text{C}]$ glutamate as the sole source of this amino acid so that all of the glutamates in the resulting protein were enriched in ^{13}C carbon at the acyl position in their side chains. A series of one-dimensional ^{13}C nuclear magnetic resonance spectra were gathered at various values of pH. Every other value of the pH is noted on the vertical axis. The values for the pH of unlabeled traces are intermediate between those for the labeled traces. The region containing the peaks of absorption from the acyl ^{13}C carbons of glutamate and glutamine of the resulting spectra are presented, one above the other. The peaks assigned by site-directed mutation to Glutamate 78 and Glutamate 172, the only two glutamates in the protein, are highlighted. Peaks arising from the acyl ^{13}C carbons of several glutamines are labeled individually. Their chemical shifts respond to acid–base titrations of side chains in their vicinity. Reprinted with permission from ref 334. Copyright 2000 Elsevier B.V.

^{15}N nitrogen were assigned, and the acyl ^{13}C carbons of individual aspartates and glutamates could be distinguished on a two-dimensional ^{13}C – ^1H HSQC/HSQC correlated spectrum (Figure 12–30A).³³⁶ When the chemical shifts of the acyl ^{13}C carbons of the aspartyl side chains were plotted as a function of pH, titration curves for each aspartate were obtained (Figure 12–30B). Titration curves for the same aspartates could also be obtained by following the chemical shifts of the two ^1H hydrogens on the β carbons

on each aspartate (chemical shifts on the abscissa of Figure 12–30A). Aspartates 94, 108, and 134 had values of pK_a expected for carboxylates exposed on the surface of a protein (3.2, 3.2, and 4.1, respectively). Aspartates 102 and 194 had values of pK_a less than 2, which suggests that in the native structure of the protein they are in electropositive surroundings.

Aspartate 10 and Aspartate 70 are immediately



adjacent to each other in the crystallographic molecular model of ribonuclease H³³⁷, and their acid–base titrations are coupled tautomerically to each other (Figure 12–30B). The two **coupled titration curves** should be complex functions of the microscopic acid dissociation constants (Equation 2–24) and of the four values of the chemical shift of Aspartate 10 and the four values of the chemical shift for Aspartate 70 (the ones when Aspartate 10 and Aspartate 70 are both protonated, the ones when Aspartate 10 is protonated and Aspartate 70 is not, the ones when Aspartate 10 is not protonated and Aspartate 70 is, and the one when both are unprotonated).³³⁸ To obtain exact values for these 12 parameters, additional experiments in which each of the aspartates in turn is mutated to an asparagine would have to be performed. If, however, it is assumed that neither the chemical shift of protonated Aspartate 10 nor the chemical shift of unprotonated Aspartate 10 is perturbed by the ionization of Aspartate 70 and that neither the chemical shift of protonated Aspartate 70 nor the chemical shift of unprotonated Aspartate 70 is perturbed by the ionization of Aspartate 10, then the two respective titration curves register only the fraction of each aspartate that is ionized at a particular pH. If this is the case, as it seems to be, then the equilibrium constant between the concentrations of the two tautomers, the one in which Aspartate 10 is protonated and Aspartate 70 is unprotonated and the one in which Aspartate 10 is unprotonated and Aspartate 70 is protonated, is 3 and the two values for the macroscopic pK_a (Figure 2–7) are $pK_{a1} = 2.9$ and $pK_a = 6.4$.³³⁶ It follows that the microscopic pK_a for Aspartate 10 when Aspartate 70 is protonated is 3.5, that for Aspartate 70 when Aspartate 10 is protonated is 3.0, that for Aspartate 10 when Aspartate 70 is unprotonated is 6.3, and that for Aspartate 70 when Aspartate 10 is unprotonated is 5.8.

The titration curves of the two glutamates of the xylanase from *B. circulans* (Figure 12–29), which are also immediately adjacent to each other in the crystallographic molecular model,³³⁴ cannot be directly registering the macroscopic acid dissociations and the ratio of

Figure 12–30: Acid–base titration of aspartates in isoenzyme I of ribonuclease H from *E. coli*.³³⁶ (A) Two-dimensional (^{13}C - ^1H) HSQC/HSQC nuclear magnetic resonance spectrum of a 0.15 mM solution of the protein in $[\text{D}_2]\text{H}_2\text{O}$ at pH 5.5. The region of the two-dimensional spectrum displayed contains cross-peaks from the acyl ^{13}C carbons of the side chains of glutamate, glutamine, aspartate, and asparagine and the ^1H hydrogens on the adjacent methylene carbon. Each of the side chains produces a pair of cross-peaks because there are two ^1H hydrogens on each adjacent methylene carbon, and each of these ^1H hydrogens is positioned by the tertiary structure of the protein in a chemically different environment. The pairs of cross-peaks from each side chain are labeled with the position in the sequence of the amino acid producing them. (B) Titrations of individual aspartates in the protein. A series of two-dimensional spectra were gathered at different values of pH. The chemical shift of the acyl ^{13}C carbon of each side chain (ordinate of the pairs of cross-peaks in panel A) is plotted as a function of pH. Reprinted with permission from ref 336. Copyright 1994 American Chemical Society.

the two tautomers, because in this example, the peak of absorption from the acyl ^{13}C carbon of Glutamate 78 drifts to a smaller value of the chemical shift rather than to a larger one upon the titration of Glutamate 172, and the peak of absorption from Glutamate 172 also drifts to a smaller value of the chemical shift rather than to a larger one upon the titration of Glutamate 78.

The carboxylates on the side chains of glutamates and aspartates sometimes form hydrogen bonds with amido nitrogen–hydrogens from the polypeptide backbone. These can be identified in nuclear magnetic resonance spectra because the chemical shift of the amido ^1H hydrogen participating in the hydrogen bond will increase in magnitude in concert with the decrease in the chemical shift of the ^1H hydrogens adjacent to the carboxyl group as it loses its proton, becomes a carboxylate, and forms the hydrogen bond with the amido nitrogen–hydrogen.³³⁹

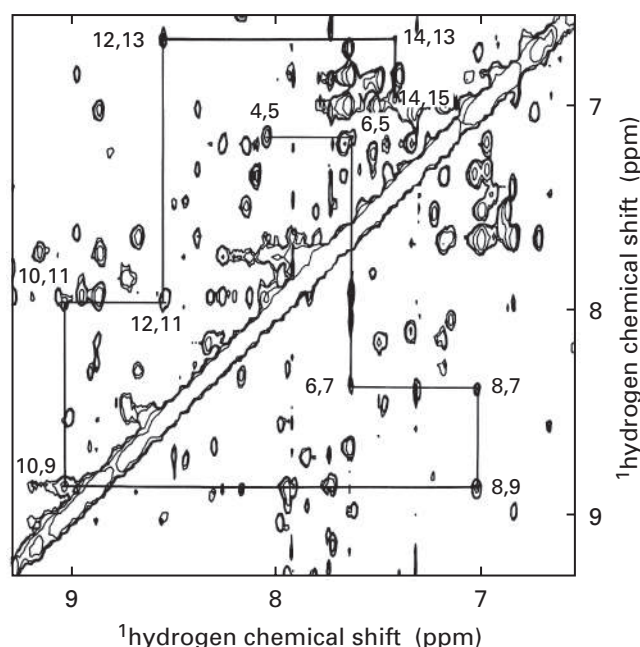
Suggested Reading

Ikura, M., Kay, L.E., & Bax, A. (1990) A novel approach for sequential assignment of ^1H , ^{13}C , and ^{15}N spectra of proteins: heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to calmodulin, *Biochemistry* 29, 4659–4667.

Ikura, M., Spera, S., Barbato, G., Kay, L.E., Krinks, M., & Bax, A. (1991) Secondary structure and side-chain ^1H and ^{13}C resonance assignments of calmodulin in solution by heteronuclear multi-dimensional NMR spectroscopy, *Biochemistry* 30, 9216–9228.

Chou, J.J., Li, S., Klee, C.B., & Bax, A. (2001) Solution structure of $\text{Ca}(2+)\text{-calmodulin}$ reveals flexible hand-like properties of its domains, *Nat. Struct. Biol.* 8, 990–997.

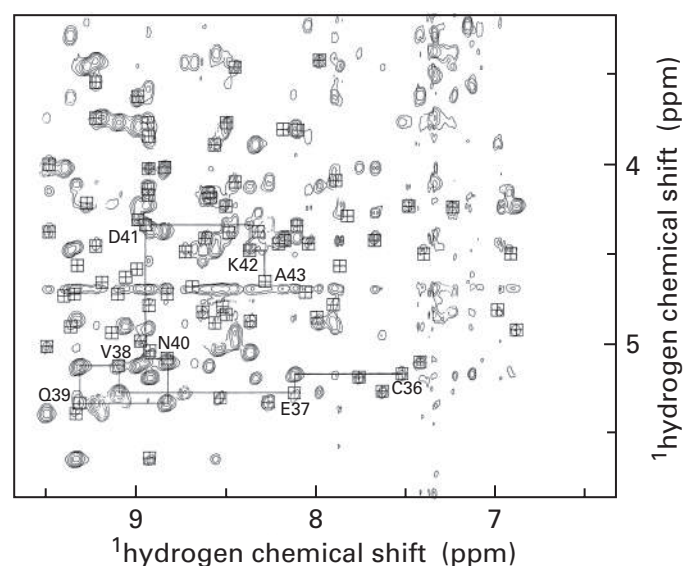
Problem 12–13: This figure is a two-dimensional nuclear magnetic resonance spectrum of the cytochrome *c*-551 from *Pseudomonas aeruginosa*.³⁴⁰ Reprinted with permission from ref 340. Copyright 1990 American Chemical Society.



In the spectrum, the off-diagonal peaks arise from nuclear Overhauser enhancements ($t_m = 150$ ms). The symmetrically displayed peaks on the two sides across the diagonal result from the same nuclear Overhauser effects. You should convince yourself that the patterns are symmetric across the diagonal.

- What hydrogens in a protein produce off-diagonal peaks in this region of the spectrum?
- The peaks connected by the horizontal and vertical lines have been identified with a particular subset of these hydrogens. Each of the peaks connected by these lines is produced by a nuclear Overhauser enhancement between two hydrogens. Draw a polypeptide in the extended conformation as in 2–15. On your drawing show with double arrows only the connections that give rise to those peaks that are highlighted by the horizontal and vertical lines.
- What do the horizontal and vertical lines indicate about these peaks? Why are the peaks labeled with pairs of numbers that increase consecutively?
- What other information was used to assign the numbers to the particular peaks?

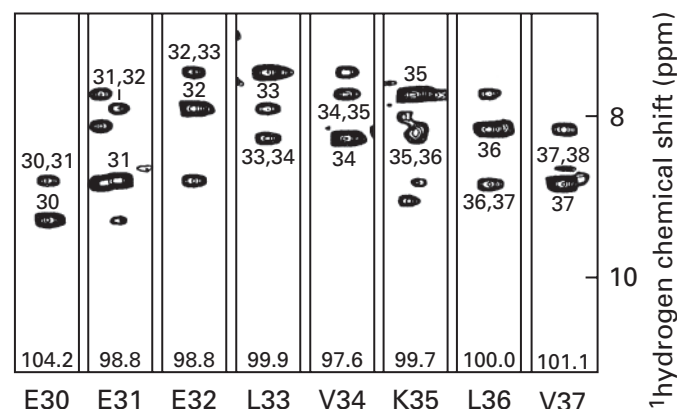
Problem 12–14: The figure is a portion of the two-dimensional nuclear Overhauser enhanced spectrum³⁴¹ of the lipoyl domain from the pyruvate dehydrogenase complex of *B. stearothermophilus*. Reprinted with permission from ref 341. Copyright 1991 Blackwell Publishing.



- What are the two functional groups in the polypeptide on which the hydrogens are located that produce the peaks in the spectrum?

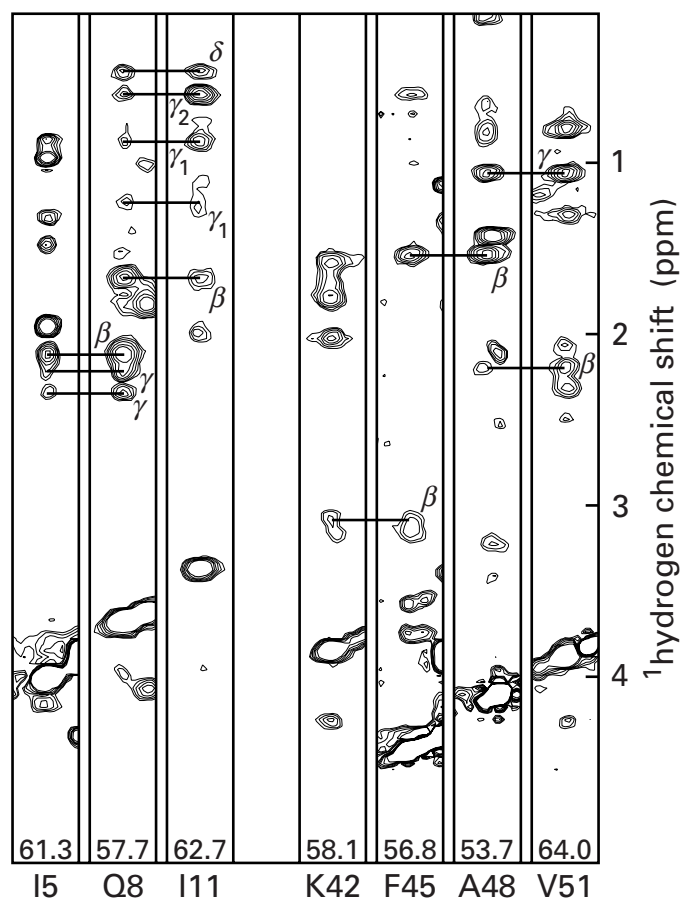
- (B) Draw the polypeptide from Cysteine 36 to Alanine 43, and draw double-headed arrows to indicate the connections between the pairs of hydrogens represented by the horizontal and vertical lines between the different peaks in the spectrum.
- (C) The squares in the spectrum represent the positions of peaks in another type of two-dimensional spectrum of the protein. What is this other type of spectrum, and what type of connections does it record?
- (D) Some of the squares coincide with peaks in the nuclear Overhauser enhanced spectrum and some do not. Why are there peaks in these positions in the other spectrum but not in the nuclear Overhauser enhanced spectrum?

Problem 12–15: Immunity protein Im9 is a folded polypeptide of 86 amino acids. It is responsible for inhibiting the action of colicin E9, an extracellular antibiotic protein produced by *E. coli*. Immunity protein Im9 uniformly labeled with ^{15}N was obtained by growing *E. coli* JM105 cells expressing high levels of the protein on minimal medium made with $[^{15}\text{N}]\text{NH}_4\text{Cl}$. The protein was purified, and the chemical shifts of most of the ^1H hydrogens and the ^{15}N nitrogens in the protein were assigned by the usual procedures. After the chemical shifts had been assigned, a three-dimensional spectrum was taken in which the three dimensions were the chemical shifts of ^1H hydrogen, ^1H hydrogen, and ^{15}N nitrogen.³⁴² The two hydrogen dimensions display nuclear Overhauser connections between pairs of hydrogens. Because the chemical shifts of each of the nitrogens in the polypeptide had been assigned, it was possible to select sections through the three-dimensional spectrum, each fixed at the chemical shift of a particular backbone nitrogen in the dimension of the chemical shifts of the ^{15}N nitrogens. Narrow strips from the resulting successive two-dimensional ^1H – ^1H nuclear Overhauser enhanced spectra are displayed in the figure. Each strip is centered on the horizontal axis at the chemical shift of the hydrogen on the amido nitrogen the chemical shift of which is fixed. The value of the chemical shift of each amido ^{15}N nitrogen at which the section was fixed and the identity of that amido nitrogen are indicated on each strip. Reprinted with permission from ref 342. Copyright 1994 American Chemical Society.



- (A) Peaks in the figure are labeled with pairs of numbers. Draw the polypeptide between Glutamate 30 and Valine 37 as the polypeptide is drawn in 2–15. Draw, however, the actual side chains along the polypeptide to identify each amino acid. Draw double-headed arrows labeled with the same pairs of numbers as the peaks in the strips are labeled and connecting every pair of hydrogens in your drawings that produce a labeled peak in the strips for Glutamate 30 to Valine 37.
- (B) What are the peaks in the spectra that are labeled with only a single number?
- (C) What are the peaks in the spectra that are not in the center of their strip?

Human interleukin-4 is a member of the family of hematopoietic cytokines that modulate cell proliferation and differentiation within the immune system. It is a folded polypeptide of 130 amino acids. A gene encoding human interleukin-4 was inserted into a pTR550 plasmid so that the protein could be expressed in *E. coli*. The human interleukin-4 expressed at high levels by these cells was uniformly labeled with ^{13}C by growing them on minimal medium made with $[^{13}\text{C}]\text{glycerol}$ (99 atom%). The protein was purified and the chemical shifts of most of the ^1H hydrogens and the ^{13}C carbons in the protein were assigned by the usual procedures. After the chemical shifts had been assigned, a three-dimensional spectrum was taken in which the three dimensions were the chemical shifts of ^1H hydrogen, ^1H hydrogen, and ^{13}C carbon.²⁷³ The two hydrogen dimensions display nuclear Overhauser connections between pairs of hydrogens. Because the chemical shifts of each of the carbons in the polypeptide had been assigned, it was possible to choose sections through the three-dimensional spectrum each fixed at the chemical shift of a particular α carbon in the protein in the dimension of the chemical shifts of the ^{13}C carbons. Strips from the resulting successive two-dimensional ^1H – ^1H nuclear Overhauser enhanced spectra are displayed in the figure. Reprinted with permission from ref 273. Copyright 1994 Elsevier B.V.



Each strip is centered on the horizontal axis at the chemical shift of the hydrogen on the α carbon the chemical shift of which is fixed. The value of the chemical shift of each $\alpha^{13}\text{C}$ carbon at which the section was fixed and the identity of that α carbon are indicated on each strip.

- (D) Peaks in the figure are labeled with the position of a hydrogen in an amino acid. The sequence of human interleukin-4 between Lysine 42 and Valine 51 is KETFCRAATV. Draw the polypeptide in this segment as the one is drawn in 2-15. Draw, however, the actual side chains along the polypeptide to identify each amino acid. Draw double-headed arrows labeled with the same numbers and Greek letters as the four peaks in the strips are labeled and connecting the pairs of hydrogens that produce the four labeled peaks in the strips for Phenylalanine 45, Alanine 48, and Valine 51. Peaks in adjacent strips are connected with horizontal lines. Draw double-headed arrows indicating the connections that are represented by each of the four horizontal lines connecting pairs of peaks from two different strips.
- (E) Into what type of secondary structure is the polypeptide between Isoleucine 5 and Isoleucine 11 and between Lysine 42 and Valine 51 folded in human interleukin-4?

Exchange of Protons

An acidic proton at any position in the covalent structure of a protein is subject to exchange with protons in the solution. Protons on the side chains of exposed polar amino acids such as asparagines, glutamines, aspartic acids, glutamic acids, serines, threonines, cysteines, arginines, lysines, and histidines usually exchange with protons in the solution so rapidly that the rates of their individual exchanges cannot be measured. The rate of exchange of a proton on the indole nitrogen of a **tryptophan**, when that proton is sterically hindered from exchanging because the side chain is buried, is, however, often slow enough to be measured.^{343,344} For example, when the constant fragment of human Bence-Jones protein Nag is dissolved in $[^2\text{H}]\text{H}_2\text{O}$, the proton on the indole nitrogen of Tryptophan 150, which is buried in its interior (Figure 6-39), exchanges with deuterons in the solvent at $1.2 \times 10^{-5} \text{ s}^{-1}$ at p^2H 7.1 and 25°C ,³⁴³ which is equal to the rate constant for the global unfolding of the protein at this pH and temperature. It was concluded that only upon the complete, transient unfolding of the protein does the proton on the indole nitrogen become sufficiently exposed to the solvent to exchange. A similar study of the rates of exchange of indole protons on the tryptophans in lysozyme from *G. gallus*, however, demonstrated that they exchanged more rapidly than the protein unfolds, presumably during local fluctuations in its structure that cause them to become exposed in turn to the aqueous phase.³⁴⁴

Almost all of the protons that are so sterically hindered from exchange by the structure of a molecule of protein that their rates of exchange are slow enough to be measured conveniently are **amido protons** in its peptide bonds. The rates of exchange of these amido protons can be monitored by following the rate at which protons are replaced by deuterons when the unmodified protein is transferred to a solution prepared with $[^2\text{H}]\text{H}_2\text{O}$ or the rate at which deuterons or tritons are replaced by protons when the protein that has been equilibrated in solutions made with $[^2\text{H}]\text{H}_2\text{O}$ or $[^3\text{H}]\text{H}_2\text{O}$, respectively, is transferred to a solution prepared with $[^1\text{H}]\text{H}_2\text{O}$. The exchange of a proton is registered when an acidic proton, deuteron, or triton dissociates from a lone pair of electrons on the protein and a deuteron, proton, or proton, respectively, then associates with that same lone pair of electrons. Because the concentrations of water, protons, and hydroxide ions are all constant at any particular pH, the exchange of a proton is a pseudo-first-order process with a pseudo-first-order rate constant.

The **exchange of an amido proton for a deuteron** when an undeuterated peptide is dissolved in $[^2\text{H}]\text{H}_2\text{O}$ can be followed by observing the decrease in the nuclear magnetic absorptions of its amido protons as a function of time.³⁴⁵ The observed pseudo-first-order rate constants, k_{ex} , display both **specific acid and specific base catalysis** (Figure 12-31):³⁴⁵

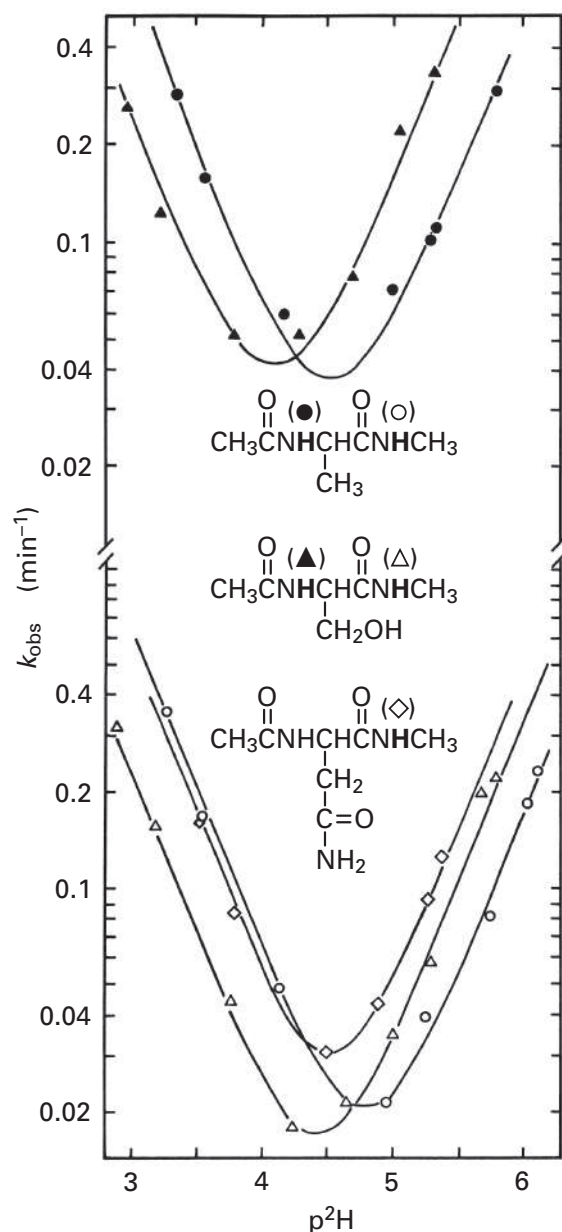
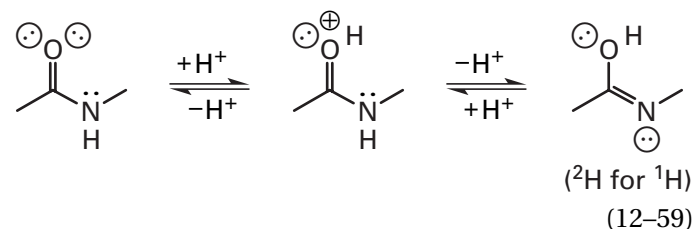


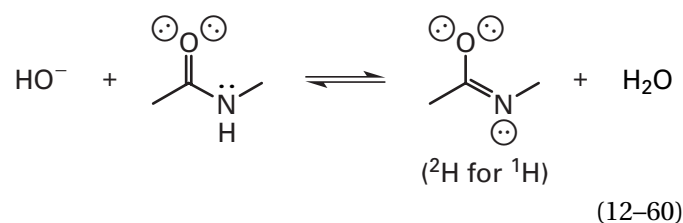
Figure 12-31: Magnitude of the observed first-order rate constants for the exchange of the amido proton on the amino-terminal side (solid symbols) or carboxy-terminal side (open symbols) in the *N*^α-acetyl-*N*-methyl amides of alanine (●, ○), serine (▲, △), and asparagine (◆, ◇).³⁴⁵ The structure of each compound is drawn and the protons that exchange are in boldface type. Solutions of each model compound were prepared in [²H]H₂O at the noted p²H and immediately introduced into a nuclear magnetic resonance spectrometer. The two amido protons in each compound produced the usual splitting into a doublet of the absorption of the spin-spin-coupled ¹hydrogen or ¹hydrogens on the respective, immediately adjacent carbons. As the proton on each nitrogen exchanged for a deuterium, the respective doublet was converted into a singlet. The areas of doublet and singlet were measured as a function of time, and it was observed that the doublet was converted to the singlet in a first-order process. The rate of this process was converted to an observed first-order rate constant, *k*_{obs} (minute⁻¹), and its value is plotted logarithmically as a function of p²H. The curves are fits of Equation 12-58 to the data. Reprinted with permission from ref 345. Copyright 1972 American Chemical Society.

$$k_{\text{ex}} = k_{\text{D}^+}[\text{D}^+] + k_{\text{OD}^-}[\text{OD}^-] \quad (12-58)$$

where D stands for deuterium. At the minimum rate (Figure 12-31), acid catalysis and base catalysis are of equal magnitude and the domination of one over the other inverts. The **catalytic mechanisms** are known to be



and



respectively.^{345,346} In both cases, the removal of the proton is the rate-limiting step in the reaction. The **second-order rate constants**, *k*_{D⁺} and *k*_{OD⁻}, have been tabulated for the amido protons on either side of specific side chains of amino acids in model compounds³⁴⁵ and small peptides.³⁴⁷ The second-order rate constants for acid catalysis, *k*_{D⁺}, vary between 6 and 5000 M⁻¹ min⁻¹ and those for base catalysis, *k*_{OD⁻}, vary between 2 × 10⁸ and 1 × 10¹¹ M⁻¹ min⁻¹.

When a protein such as myoglobin is incubated in tritiated water for an extended period of time at moderate temperature (37 °C), most of its amido protons reach equilibrium with the protons and tritons in the water. The tritiated water can then be replaced with untritiated water by molecular exclusion chromatography. During the chromatography all of the tritons on exposed polar side chains exchange with protons. When the tritium remaining on the protein is measured by liquid scintillation counting as a function of time at low temperature (0 °C), a population of tritiated amides that lose their tritons very slowly can be distinguished (Figure 12-32).³⁴⁸ The rates at which these amides exchange are far slower than the rates observed for small peptides in solution.³⁴⁷

The **amido hydrons on peptide bonds in a protein that exchange slowly** with other hydron isotopes in the solvent are the hydrons that participate in stable hydrogen bonds in the folded polypeptide.³⁴⁹ That the number of slowly exchanging tritons in myoglobin (Figure 12-32) is about equal to the number of amido protons participating in **buried hydrogen bonds** in the crystallographic molecular model (approximately 120) is consistent with this conclusion.³⁴⁸ In the case of lysozyme, the exchange

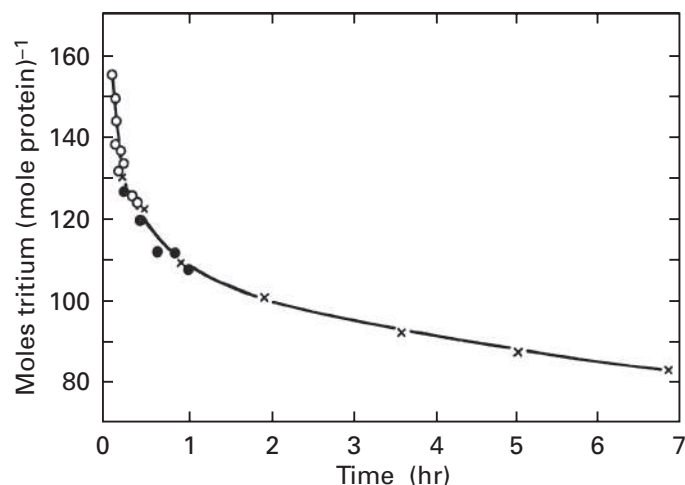


Figure 12-32: Exchange of tritons from myoglobin equilibrated with $[^3\text{H}]\text{H}_2\text{O}$ and then transferred to $[^1\text{H}]\text{H}_2\text{O}$ at pH 5, 0 °C.³⁴⁸ Myoglobin from *P. catodon* ($n_{\text{aa}} = 153$) was incubated at 37 °C and pH 9 with $[^3\text{H}]\text{H}_2\text{O}$ until equilibrium was reached (20 h) at all of its amides. The solution was cooled to 0 °C, and the protein was rapidly transferred by molecular exclusion chromatography to $[^1\text{H}]\text{H}_2\text{O}$. Over these intervals of preequilibrium, only protons on amides and more acidic acid-bases on the protein should exchange with tritons, and only the amides should retain any tritium through the chromatography. The amount of tritium associated with the protein [moles of tritium (mole of protein)⁻¹] was followed as a function of time (hours). Myoglobin contains 162 amido protons, and this number agrees favorably with the 155 tritons found on the protein at the earliest time after the chromatography. Reprinted with permission from ref 348. Copyright 1969 Academic Press.

of protons for deuterons at the amides of the polypeptide could be followed directly by observing the decrease in the absorbance of the amide II vibration in the infrared spectrum relative to the absorbance of the amide I vibration. The agreement between the number of slowly exchanging amido protons (44 moles for every mole of lysozyme) and the number of buried hydrogen bonds involving the amido protons of the polypeptide in the crystallographic molecular model is also quite close.³⁵⁰

It is also possible to follow such global exchange of the protons in a protein by **mass spectrometry**. The protein is diluted into $[^2\text{H}]\text{H}_2\text{O}$, and samples are removed at successive times and submitted to electrospray mass spectrometry to determine the number of deuterons that have been incorporated by exchange during each interval.³⁵¹⁻³⁵³ The samples are usually submitted to liquid chromatography in $[^1\text{H}]\text{H}_2\text{O}$ immediately prior to mass spectrometry to remove the $[^2\text{H}]\text{H}_2\text{O}$. This chromatography is performed at pH 3 and 0 °C to prevent exchange of the amido deuterons with protons but to permit exchange at carboxylic acids, amines, hydroxyls, thiols, and imidazoles, because the intention of such measurements is to monitor the global exchange of amido protons in peptide bonds. In some instances different populations of protons that exchange at different rates can be distinguished. In the case of fructose-bisphosphate aldolase from rabbit, these populations were

thought to represent protons from its different structural domains.³⁵³

A limitation of such global measurements of the exchange of amido protons is that the identity of those amides in the polypeptide that display slow exchange is not established. One way to increase the resolution is to digest quickly samples of the protein removed at different intervals over the time during which exchange is permitted to occur, separate the resulting peptides in each sample chromatographically, and use a mass spectrometer to assess the extent of incorporation of deuterium into each of the peptides.^{351,354} The exchange of protons is quenched at the end of each interval in $[^2\text{H}]\text{H}_2\text{O}$ by dropping the pH to 3 and the temperature to 0 °C to minimize further exchange. The protein, unfolded by the low pH, is digested with pepsin A, an endopeptidase that functions best at low pH; and the resulting peptides are separated by chromatography in $[^1\text{H}]\text{H}_2\text{O}$ at low pH and low temperature. Each peptic peptide is then identified by its mass and its pattern of fragmentation (Figure 3-8). In this way the amount of incorporation of deuterium into the amides of the peptide bonds within a particular segment of the folded polypeptide in the native protein, namely, that segment ending up in the peptic peptide, can be monitored.^{355,356} Still, the protons at the individual amides within that segment cannot be distinguished one from the other.

The advantage of such **endopeptidolytic analyses** is that they can be applied to large proteins³⁵⁵⁻³⁵⁸ such as rabbit fructose-bisphosphate aldolase ($n_{\text{aa}} = 4 \times 363$), the α -catalytic subunit of cyclic AMP-dependent protein kinase ($n_{\text{aa}} = 350$), human dual specificity mitogen-activated protein kinase kinase 1 ($n_{\text{aa}} = 393$), and dihydrodipicolinate reductase from *E. coli* ($n_{\text{aa}} = 4 \times 273$). Such large proteins cannot be analyzed by nuclear magnetic resonance. If, however, the protein is small enough, nuclear magnetic resonance spectroscopy can provide rates of exchange of most of the resolved amido protons along the polypeptide backbone in the native structure.

The cross-peaks in the fingerprint region of a two-dimensional (^1H - ^1H) **nuclear magnetic resonance correlated spectrum** (Figure 12-33)³⁵⁹ or a two-dimensional (^{15}N - ^1H) HSQC correlation spectrum³⁶⁰ arise from spin-spin coupling between an amido proton and its adjacent α ^1H hydrogen or amido ^{15}N nitrogen, respectively. When the protein is transferred from $[^1\text{H}]\text{H}_2\text{O}$ to $[^2\text{H}]\text{H}_2\text{O}$, each of these cross-peaks decreases in intensity as the amido proton exchanges for a deuteron (Figure 12-33).^{359,360} The rate of exchange of each proton is equal to the rate at which its cross-peak decreases in intensity.

It is generally assumed that a proton on a particular amide in the polypeptide backbone of a protein can exchange with a deuteron in the $[^2\text{H}]\text{H}_2\text{O}$ surrounding the protein only when that proton is exposed to the solution. An **unexposed position** in the polypeptide backbone becomes exposed as the result of a conformational change in the protein.^{361,362} Although the details of such conformational changes are unknown, they must differ

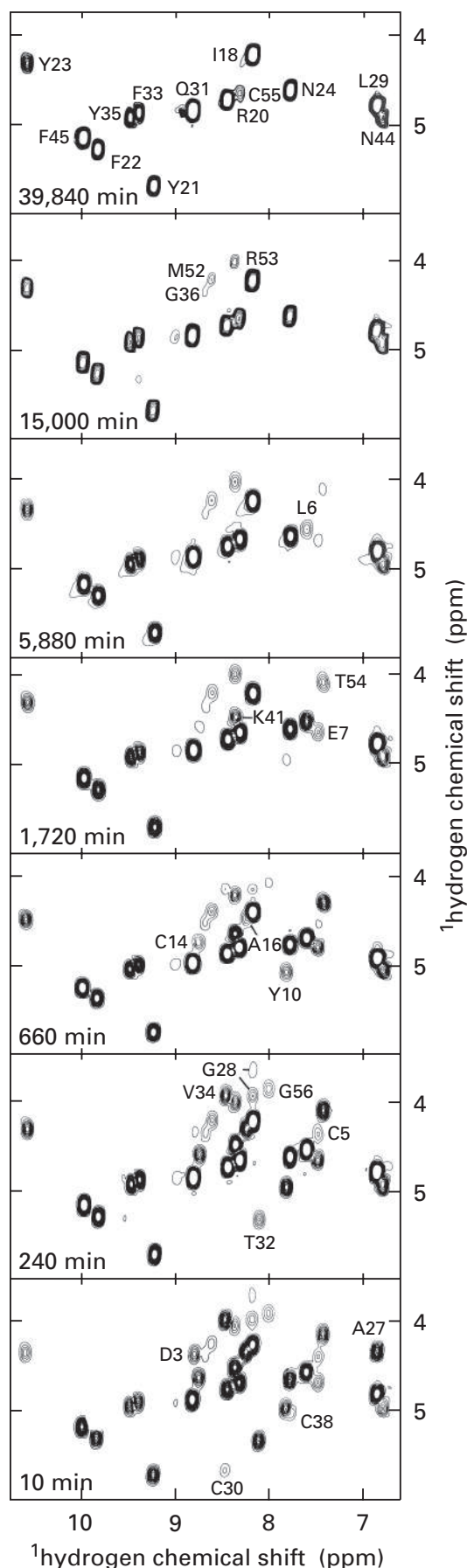


Figure 12-33: Exchange of amide protons in the polypeptide backbone of bovine basic trypsin inhibitor for deuterons in the solution.³⁵⁹ The protein was dissolved at a concentration of 0.02 M in $[^2\text{H}]\text{H}_2\text{O}$ and a p^2H of 3.5 and brought to 36 °C. After the noted times, samples were removed, the temperature was lowered to 25 °C, and a two-dimensional (^1H - ^1H) correlated spectrum was recorded. The fingerprint region containing cross-peaks resulting from spin-spin coupling between amide protons and α ^1H hydrogens (Figure 12-16) is presented. As amide protons in a given population exchange for deuterons, its cross-peak decreases in intensity. Each cross-peak is labeled in the last spectrum in which it can be observed with the position of its α ^1H hydrogen in the sequence of the protein. The cross-peaks from those amide protons remaining after 39,840 min are labeled with their positions in the sequence. Twenty of the original 32 cross-peaks disappear as a result of exchange of amide protons for deuterons in the solvent, and most of the cross-peaks remaining display sufficient decreases in intensity over this interval to give rate constants for the exchange of their amide protons. Reprinted with permission from ref 359. Copyright 1982 Academic Press.

from one location in the protein to another because protons exchange with a wide range of rate constants. **Conformational changes that expose protected protons** to the solution may be relatively rapid movements of loops on the surface, motions opening crevices in the surface of the protein more widely, fraying motions in α helices that unwind them from their ends,^{363,364} unzipping motions in β structure,^{365,366} or the complete, transient unfolding of individual elements of secondary structure,³⁶⁴ domains,³⁵³ or even the entire molecule of the protein.^{343,350,367} Almost all of the amide protons in the polypeptide backbone that exchange slowly enough that their rates can be measured are involved in hydrogen bonds in the native structure of the protein. Those hydrogen bonds must be broken during the conformational change to permit the unretarded³⁶⁸ exchange of the protons to occur.³⁴⁹

Associated with whatever conformational change is responsible for the exchange of a particular proton in a protein is a rate of opening k_{op} and a rate of closing k_{cl} . The **kinetic mechanism** for the process is³⁶⁹



where unexchangeable is the proton in a location where exchange is sterically prohibited, exchangeable is the proton in a location where it is exposed to the solution and exchange is unobstructed, exchanged is the site at which exchange has occurred and a deuteron is occupying the location that was occupied by the original proton, and k_{ex} is the pseudo-first-order rate constant for the exchange of the original proton by the deuteron.

There are two limits to the rate equation for this mechanism. If $k_{\text{ex}} \gg k_{\text{cl}}$ then

$$k_{\text{obs}} = k_{\text{op}} \quad (12-62)$$

where k_{obs} is the observed rate constant of exchange. This condition is the **EX₁ limit**. If $k_{\text{cl}} \gg k_{\text{ex}}$ then

$$k_{\text{obs}} = \frac{k_{\text{op}}}{k_{\text{cl}}} k_{\text{ex}} = K_{\text{conf}} k_{\text{ex}} \quad (12-63)$$

where K_{conf} is the equilibrium constant for the conformational change producing the exchangeable conformation, $[\text{exchangeable}]/[\text{unexchangeable}]$. This condition is the **EX₂ limit**.

It is customary to present the rate of exchange of a particular amido proton i in the polypeptide backbone of a protein in terms of its **protection factor**³⁴⁷ $k_{\text{ex},i}/k_{\text{obs},i}$. The rate constant $k_{\text{ex},i}$ is the reference rate constant for the exchange of that proton in the fully exposed, unfolded polypeptide, and $k_{\text{obs},i}$ is the observed rate constant under a given set of conditions. The rate constant for exchange in the fully exposed conformation, $k_{\text{ex},i}$, is assumed to be equal to the measured rate constant for exchange of an equivalent amido proton in a derivative of the amino acid (Figure 12-31)³⁴⁵ or short peptide³⁴⁷ as a function of pH. It has been shown by measuring the rates of exchange of particular amido protons in an unfolded polypeptide that such estimates are reliable.³⁷⁰

A protection factor for a particular proton is meaningful only when the exchange of the proton occurs at the EX₂ limit, where the protection factor is equal to the inverse of the equilibrium constant for the conformational change, K_{conf}^{-1} (Equation 12-63). When exchange occurs at the EX₁ limit, the observed rate constant is equal to $k_{\text{op},i}$, the rate constant for the opening of the structure that protects proton i (Equation 12-62). Dividing the observed rate constant by $k_{\text{ex},i}$ would be meaningless. Consequently, it is necessary to demonstrate that the exchange process is at the EX₂ limit for the protection factor to be meaningful. Presentations of protection factors in the absence of such a demonstration³⁷¹ are equivocal.

It is possible to make a **distinction between an EX₁ limit and an EX₂ limit** for exchange by evaluating whether the exchange at adjacent amido positions is **concerted** or not.³⁷² When exchange of a proton at a particular location occurs at the EX₁ limit, the open conformation persists long enough ($k_{\text{ex}} \gg k_{\text{cl}}$) that all of the protons at adjacent positions are exchanged before the protein snaps shut. Consequently, only two populations of protons at that location exist, one in which the proton and its neighbors have not yet exchanged and one in which both the proton and its neighbors have exchanged. When exchange of a proton at a particular location occurs at the EX₂ limit, the open conformation has such a short lifetime ($k_{\text{ex}} \ll k_{\text{cl}}$) that at most only one proton in a neighborhood can exchange at each opening. Consequently, in addition to the two populations already described, there are populations in which the proton has exchanged but its neighbors have not and populations in which the proton has not exchanged but its neighbors have. All of these populations can be distinguished by the

ratios of the intensities of cross-peaks arising from nuclear Overhauser effects. In the cases of both bovine basic pancreatic trypsin inhibitor ($n_{\text{aa}} = 58$) and α -amylase inhibitor HOE-467A from *S. tendae* ($n_{\text{aa}} = 74$), it was observed that, below 55 °C, the exchanges of all of the amido protons that could be monitored in the polypeptide backbone of each of these native proteins were unconcerted, consistent with exchange at the EX₂ limit.^{366,373}

Whether the exchange of a particular proton is at the EX₂ limit or the EX₁ limit can also be assessed by following the **effect of pH** on the observed rate of exchange. Above pH 5, $k_{\text{ex},i}$ for an amido proton in a peptide bond is directly proportional to the concentration of OH⁻ (Equation 12-58).^{345,347} If the exchange is at the EX₂ limit (Equation 12-63), it must display the expected increase in rate as the pH of the solution is increased. The observed rates of proton exchange at several of the peptide bonds in bovine basic pancreatic trypsin inhibitor at 68 °C increased by a factor of 10 for each increase of one unit in pH from pH 5 to 7³⁷³ as expected of exchange at the EX₂ limit. Above pH 7, however, the rates no longer increased as the pH was increased, perhaps because k_{ex} for these protons at this high temperature had become so rapid (Figure 12-31) that k_{ex} became greater than k_{cl} . These results suggest that the balance between exchange governed by the EX₂ limit and that governed by the EX₁ limit under certain conditions may lie in the range of physiological pH. In such instances, by following the exchange of a particular proton as a function of pH, the range of pH in which the EX₂ limit governs the exchange and the range of pH in which the EX₁ limit governs the exchange can be distinguished. In the former range, the exchange increases by a factor of 10 for every increase of one unit in pH; in the latter range, the exchange is invariant with pH. From such observations, values of k_{cl} , k_{op} , and K_{conf} for each member of a set of specific peptide bonds can be measured (Equations 12-62 and 12-63).³⁷⁴

It is difficult to make measurements of the exchange of protons at several values of pH. Often an analysis is made of rates of exchange at only two values of pH. In this instance, the logarithms of the observed rate constants of exchange of the amido protons in the polypeptide backbone at one pH are plotted as a function of their logarithms at a different pH.^{360,375} If they fall on a line of slope 1, then the individual conformational changes producing each exchange are not significantly dependent on pH, and if the line has an intercept at the ordinate axis equal to the difference in pH between the two solutions,³⁷⁵ then the process of exchange for each proton has the pH dependence expected of $k_{\text{ex},i}$ and is judged to be at the EX₂ limit. If, however, the intercept of the line is not equal to the difference in pH of the two solutions, the exchanges are not at the EX₂ limit.³⁶⁰ If the intercept of the line is zero, the exchange is probably at the EX₁ limit and the observed rate constants of exchange are equal to the

rate constants k_{op} for the various conformational changes permitting exchange.

In lysozyme from *G. gallus*, there is a group of amido protons in the core of the protein that all exchange their protons 10^6 times more slowly than the same amides in a small peptide.³⁵⁰ This group of amides in the center of the protein may be exposed to the solvent only during **large cooperative unfoldings** of a considerable fraction of the protein that expose many amido protons simultaneously for a short time before the polypeptide snaps shut again. If this is the case, these deeply buried regions of lysozyme spend 10^{-6} of their life in an unfolded state. The individual amido protons in the polypeptide backbone of bovine basic pancreatic trypsin inhibitor, however, have a rather heterogeneous set of rate constants of exchange (Figure 12–33).³⁷⁶ This suggests that **local vibrational modes** producing local unfoldings are responsible, at least in this protein, for performing the disconnections of a particular structure required to expose the amide to the solvent so that its proton can exchange. Whether the motions responsible for the exchange of a particular proton are local, regional, or global, the recurring observation that all or almost all of the amido protons, at least of small proteins, eventually exchange means that a molecule of protein is **continuously breathing** or unfolding throughout its lifetime.

It is also possible to observe the exchange of protons by **neutron diffraction**. Crystals of protein are routinely prepared for neutron diffraction by soaking them in $[^2\text{H}]\text{H}_2\text{O}$ to replace as many protons as possible with deuterons, which scatter neutrons more strongly. Molecules of trypsin³⁷⁷ and ribonuclease³⁷⁸ that had been soaking within crystals in $[^2\text{H}]\text{H}_2\text{O}$ for periods of 1 year retained protons on 54 and 28 of their amides, respectively. All of the sites that remained unexchanged were in the interior of the folded polypeptide. These were mainly on the central strands of β sheets and at the centers of α helices. Most of the sites retaining protons after 1 year were not even partially exchanged with deuterons, while those sites that had deuteriums were almost fully exchanged. The location of these sites and their occurrence in regions with very little thermal motion suggest that, in a crystal, only local motions are responsible for what exchange takes place. Because large regional unfolding or the complete unfolding of the polypeptide cannot occur in a crystal, these observations provide further evidence that the exchange of protons at deep locations in a protein observed in free solution result from such types of extensive unfolding.

Observations of the exchange of protons can be used to examine **heterologous associations**. For example, the rates of exchange of 11 amido protons in the polypeptide backbone of equine cytochrome *c*, which are at unrelated positions in its amino acid sequence but form a continuous region on the surface of the tertiary structure of the protein, decrease significantly when

it is bound by a monoclonal immunoglobulin.³⁷⁹ It was concluded that this region formed the epitope on the surface of the cytochrome *c*. When a synthetic peptide with the amino acid sequence from Alanine 1730 to Leucine 1747 in smooth muscle [myosin-light-chain] kinase from *G. gallus*, which was known to be the site to which calmodulin binds in the intact protein, associates with calmodulin, the rates of exchange of 12 of its amido protons decrease by factors between 10^3 and 10^6 as it forms the α helix embraced by the calmodulin in the complex.³⁶³

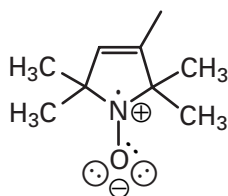
Results of such experiments, however, should be evaluated with caution. Effects on the exchange of amido protons can be felt at **locations distant from the site of interaction** because the association of a protein with a ligand always increases its global stability and consequently decreases the time its entire structure spends in open conformations. For example, the binding of thymidine 3',5'-diphosphate to micrococcal nuclease, causes the rate of exchange of at least 34 of its amido protons to decrease dramatically³⁸⁰ even though the site at which the ligand binds encompasses far fewer amino acids. In this instance, the binding of the ligand stabilizes the protein globally and decreases its conformational fluctuations. Likewise, when NADH is bound to dihydrodipicolinate reductase, the rates of exchange of amido protons in widely different locations showed significant decreases.³⁵⁵

Suggested Reading

Wand, A.J., Roder, H., & Englander, S.W. (1986) Two-dimensional ^1H NMR studies of cytochrome *c*: hydrogen exchange in the N-terminal helix, *Biochemistry* 25, 1107–1114.

Electron Paramagnetic Resonance

If an atomic orbital or a molecular orbital contains a pair of electrons, the magnetic moments of their spins cancel because of the Pauli principle, and that orbital is diamagnetic. If an orbital contains a single unpaired electron, that electron has an uncanceled magnetic moment and that orbital is **paramagnetic**. There are several ways in which a molecule of protein can contain an orbital with an unpaired electron. A **paramagnetic ion** of a transition metal such as Mn^{2+} , Fe^{3+} , Co^{2+} , Ni^{3+} , or Cu^{2+} can be bound to the protein either on its own or within a coenzyme like the heme in ferrimyoglobin (Figure 4–18). A **stable organic radical** like the glycyl radical in formate C-acetyltransferase or the tyrosyl radical in ribonucleoside-diphosphate reductase can be formed by posttranslational modification of the protein (Table 3–1). A coenzyme bound to the protein can contain an organic radical.³⁸¹ The protein can be modified with a reagent in which there is a stable organic radical, such as the one in a 1-oxyl-2,2,5,5-tetramethylpyrrolin-3-yl group:



12-11

Such a functional group can be coupled covalently either by incorporating an unnatural amino acid containing it into the protein in a cell-free translation¹⁵⁷ or by modifying the protein with an electrophilic reagent containing it.^{382,383}

An unpaired electron can have a **spin quantum number** of $+\frac{1}{2}$ or $-\frac{1}{2}$, and these two quantum numbers dictate two respective spin states with two respective angular velocities of the same magnitude but opposite polarity. When an unpaired electron is placed in an external homogeneous magnetic field, the axis of its spin tends to align with the direction of the applied field. The two degenerate energy levels of the spinning electron are split into two distinct energy levels, one for the spin aligned in the direction of the magnetic field and one for the spin aligned in the direction opposed to the magnetic field. The difference in energy between these two spin states for a given population i of identical unpaired electrons, ΔE_i , is directly proportional to the magnetic flux density B_i (tesla) at the location of unpaired electron i . The frequency ν_i (hertz) of electromagnetic energy that is absorbed by the population of electrons i during its transition between these spin states is

$$\Delta E_i = g_e \mu_B B_i = h\nu_i \quad (12-64)$$

where g_e is the g factor of a free electron (2.0023) and μ_B is the Bohr magneton ($9.27 \times 10^{-24} \text{ J T}^{-1}$). At magnetic flux densities normally used for electron paramagnetic resonance ($<2 \text{ T}$) the difference in energy is less than 20 J mol^{-1} , the energy contained in a photon of frequency less than 40 GHz , which is the microwave range of electromagnetic energy.

As with continuous-wave nuclear magnetic resonance spectrometers, an **electron paramagnetic resonance spectrometer** has a microwave generator of a fixed frequency, for example, 35 or 9 GHz , and the magnetic flux density is varied while the absorption of energy is monitored (Figure 12-34).³⁸⁴ Peaks of absorption are observed when Equation 12-64 is satisfied for a given population of identical electrons of unpaired spin. As with nuclear magnetic resonance, saturation of the absorption occurs readily in electron paramagnetic resonance owing to the small difference in the occupancy of the two spin states ($1 < K_{\text{sp}} < 1.008$) and the slow rate of relaxation between them. There are, however, several features of an electron paramagnetic resonance spec-

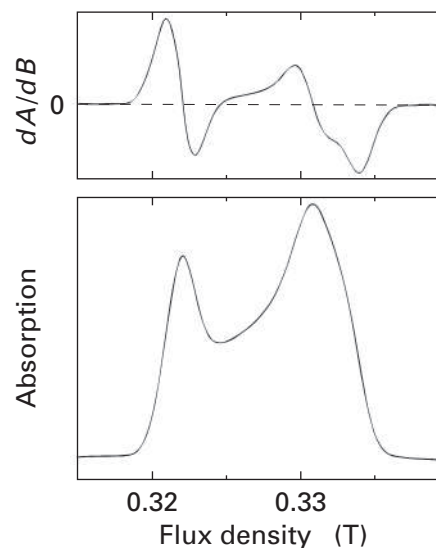


Figure 12-34: Electron paramagnetic resonance spectra of a 0.5 mM solution of CDP-6-deoxy-L-threo-D-glycero-4-hexulose-3-dehydrase in the presence of 10 mM CDP-6-deoxy-L-glycero-4-hexulose and 1 mM NADH frozen at 77 K .³⁸⁴ The bottom trace is the absorption as a function of the flux density of the magnetic field (tesla) at a carrier frequency of 9.05 GHz . The spectrum was obtained by varying the magnetic field while the microwave frequency remained fixed. The top trace is the first derivative (change in absorption/change in magnetic flux density) of the bottom trace; the dotted line sits on a value of zero for the first derivative. The $[2\text{Fe}-2\text{S}]$ cluster in the protein contains an unpaired electron that absorbs at g factors of 2.012 , 1.950 , and 1.932 , producing the three peaks of absorption in the two spectra at 0.321 , 0.331 , and 0.334 T , respectively. Reprinted with permission from ref 384. Copyright 1996 American Chemical Society.

trum that distinguish it from a nuclear magnetic resonance spectrum.

For unpaired electrons on ions of transition metals, the intensity of the electron paramagnetic absorption increases and the width of the peak of absorption decreases dramatically as the **temperature of the sample** is lowered. For unpaired electrons on carbon, nitrogen, or oxygen, the intensity of the absorption also increases as the temperature is lowered. Consequently, electron paramagnetic resonance is often monitored while a sample of the protein containing the unpaired electron is in the frozen state at low temperature. For example, the electron paramagnetic resonance spectrum of aminocyclopropane carboxylate oxidase, the iron in whose heme had been complexed with nitrous oxide, was observed at 8 K ,³⁸⁵ and that of the molybdenum-iron protein of nitrogenase, the metal cluster of which had been complexed with ethene, was observed at 2 K .³⁸⁶ At such low temperatures, molecular motions and chemical reactions are severely limited. The most convenient low temperature is 77 K , the boiling point of liquid nitrogen, but the spectra of many types of organic radicals can be observed at room temperature even though the amplitudes of the peaks of absorption are less than they would be at lower temperatures.

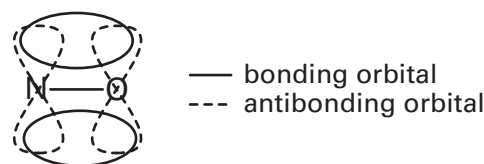
The absorption of microwave energy is usually monitored as its **first derivative with respect to the flux density** of the magnetic field. Consequently, peaks of absorbance appear as pairs of positive and negative deflections representing the positive slope of the rising phase and the negative slope of the declining phase of the absorption itself, and the differential passes through zero at the maximum of each absorption (Figures 12–34 and 12–35).³⁸³

The absorption from a particular population of identical unpaired electrons is often split into two or more separate peaks by **spin–spin coupling**, either between the electron and magnetic nuclei connected to it by covalent bonds or between the electron and a magnetic nucleus on which it happens to reside. As in nuclear magnetic resonance, this spin–spin coupling results from local perturbations to the applied magnetic field. These perturbations are caused by differences in the orientations of the spins of the coupled nuclei, the magnetic fields of which are transmitted through the diamagnetic electrons in the covalent bonds surrounding the unpaired electron. As a result, the magnetic flux density sensed by a given electron i , B_i , is the sum of the flux density of the applied magnetic field, B_{app} , and the flux densities of any local magnetic fields, B_{loc} , created by these coupled magnetic nuclei. The spin–spin splitting in electron paramagnetic resonance is referred to as **hyperfine splitting**; the resulting pattern of peaks, as hyperfine structure; and the spin–spin coupling, as hyperfine coupling.

The 1-oxyl-2,2,5,5-tetramethylpyrrolin-3-yl group (12–11) serves as a simple example of hyperfine splitting (Figure 12–35). In this stable free radical, the nucleus that dominates the local magnetic field is that of the ¹⁴**nitrogen**, which has a spin quantum number of 1. The ¹⁴nitrogen nucleus is quadrupolar and can assume spins of +1, 0, and –1 with equal probability, because the distribution among its energy levels is insignificantly affected by the applied magnetic field. As a result, the local magnetic flux

density, B_{loc} , created by the nitrogen nucleus assumes three values, one of which is zero. The spectrum consists of a central absorption, arising from unpaired electrons coupled to nitrogen nuclei of spin quantum number 0 and from unpaired electrons not coupled to any magnetic nucleus, and two peaks of hyperfine absorption on either side of the central peak, arising from unpaired electrons coupled to nitrogen nuclei of spin quantum number +1 and –1. The hyperfine absorptions are of variable magnitude and are split different distances from the central absorption depending on the quality of the coupling between the nitrogen and the electron, but the central absorption will always be fixed because it is at the position where the contribution of the ¹⁴nitrogen to the local magnetic flux density is zero. Information about environment, rotational diffusion, and anisotropy is contained in the hyperfine absorptions.

The full coupling of the electron to the ¹⁴nitrogen is expressed in aqueous solution (Figure 12–35) because the electronic structure in which the radical occupies a **p orbital over nitrogen**, a distribution which requires a separation of charge, can be readily solvated by the water. In nonpolar environments such as within a molecule of protein, the nitroxyl radical can shift its hybridization to form an ethenyl molecular orbital system



12–12

composed from one of the lone pairs on oxygen and the radical. The unpaired electron occupies the **antibonding molecular orbital** and spends more of its time over oxygen, which is diamagnetic. This delocalization decreases the effect of the quadrupolar nucleus of the ¹⁴nitrogen, and the two hyperfine components decrease accordingly in intensity.

Hyperfine coupling can be used to draw conclusions about the **properties and location of the unpaired electron**. The unpaired electron on the glycy radical in formate C-acetyltransferase is unaffected by the diamagnetic ¹²carbon on which it resides but is split into two peaks by the single ¹hydrogen on that carbon (Figure 12–36).³⁸⁷ When formate C-acetyltransferase is transferred to [²H]H₂O, the ¹hydrogen exchanges with ²hydrogen in the solution in a reaction catalyzed by a nearby cysteine, and the spin–spin splitting disappears. Tryptophan tryptophylquinone is present in amine dehydrogenase as a posttranslational modification (Table 3–1, Figure 3–18). A model compound for tryptophan tryptophylquinone in which the two ⁺H₃N(COO[–])CH– groups of the bis(amino acid) are replaced with hydrogens can be reduced with one electron to produce the semiquinone with an unpaired elec-

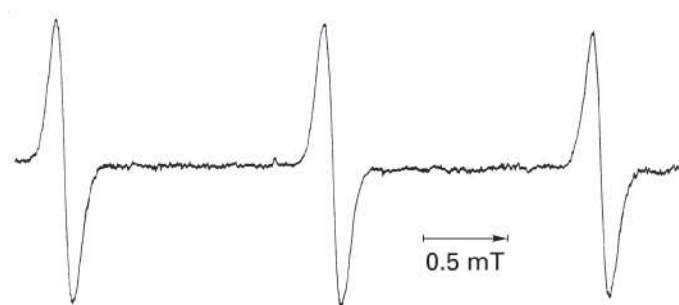


Figure 12–35: Electron paramagnetic spectrum of 3-carbamoyl-1-oxyl-2,2,5,5-tetramethylpyrroline (see 12–11) in water at room temperature.³⁸³ The scale indicates the dimension of the horizontal axis in units of magnetic flux density (tesla). The carrier frequency of the spectrophotometer was 9.5 GHz. The first derivative of the absorption (change in absorption/change in magnetic flux density) is presented. Reprinted with permission from ref 383. Copyright 1965 held by authors.

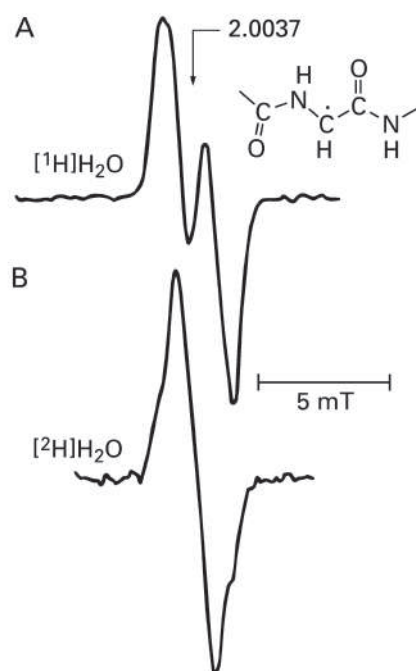


Figure 12-36: Electron paramagnetic resonance spectra of the glycy radical (Table 3-1) in formate C-acetyltransferase.³⁸⁷ The spectra were gathered from solutions of the protein at pH 7.6. (A) A solution (20 mg mL⁻¹) of the protein in $[^1\text{H}]\text{H}_2\text{O}$ was frozen in liquid nitrogen, and the spectrum was recorded at 77 K. (B) The solution was then thawed, mixed with three volumes of $[^2\text{H}]\text{H}_2\text{O}$, allowed to sit for 2 min, and refrozen, and another spectrum was recorded. The carrier frequency of the spectrometer was 9.23 GHz. The vertical axis is the first derivative of the absorption (change in absorption/change in magnetic flux density). The dimension of the magnetic flux density (tesla) along the horizontal axis is indicated by the scale, and the g factor of the absorption (2.0037) is indicated in panel A. Reprinted with permission from ref 387. Copyright 1995 American Chemical Society.

tron. The absorption from the unpaired electron in the semiquinone was split into at least 22 peaks. This splitting was thought to result from four nonequivalent ^1H hydrogens, one set of methyl ^{14}H hydrogens, and two nonequivalent ^{14}N nitrogens. It follows that, in the semiquinone, the electron must be delocalized over both indole rings.³⁸⁸ There is an unpaired electron in formate C-acetyltransferase that has been inactivated by covalent modification with fluoropyruvate. It could be shown to be located on carbon 3 of a defluorinated pyruvyl group because its absorption was split into three peaks when $[^1\text{H}]\text{fluoropyruvate}$ was used for the modification but was unsplit when $[^2\text{H}]\text{fluoropyruvate}$ was used.³⁸⁹

Just as in nuclear magnetic resonance, the set of peaks arising from the hyperfine splitting of the absorption from a particular population of identical unpaired electrons is distributed more or less symmetrically about a central point, which is the point in the spectrum at which that electron would have absorbed were its absorption not split by hyperfine coupling. This central point gives the g factor for that electron. The peak of the

absorption from a population of identical unpaired electrons that is unsplit by hyperfine coupling has its lone peak at its g factor (Figure 12-37).³⁹⁰ The flux density of the applied magnetic field, B_{app} , which is varied to produce the spectrum, can be converted into units of **g factor** (Figure 12-37) with the relationship

$$g = \frac{h\nu_0}{B_{\text{app}}\mu_B} \quad (12-65)$$

where h is Planck's constant (6.626×10^{-34} J s), ν_0 is the carrier frequency of the spectrometer, and μ_B is the Bohr magneton (9.274×10^{-24} J T⁻¹). When the unpaired electrons in a particular population are located on a carbon, a nitrogen, or an oxygen, the g factor of the absorption lies close to **2.0023**, which is that for a **free electron**. For example, the g factor for the absorption of the unpaired electron in a pheophytin radical in photosystem II from spinach is 2.0034;³⁸¹ that for the glycy radical in formate C-acetyltransferase from *E. coli* is 2.0037 (Figure 12-36);³⁸⁷ and that for the unpaired electron localized on ^{14}N nitrogen in **12-11** when it is covalently attached to Cysteine 76 of T4 lysozyme is 2.0066.³⁸²

If the unpaired spins of a particular population of identical electrons are on a paramagnetic transition metal ion such as Mn^{2+} , Fe^{3+} , Co^{2+} , Ni^{3+} , or Cu^{2+} , which affects the local magnetic field significantly, the g factor can vary dramatically depending on the **orientation of the orbital** it occupies relative to the direction of the applied magnetic field. For example, the orbital in which the unpaired electron resides on the ferric ion in a molecule of ferrimyoglobin (Figure 4-18) in a crystal of this protein is held in a specific orientation by the ligand field of the heme, which in turn is held in position by the crystal. In the unit cell of a crystal, there are two molecules of myoglobin, each with a different orientation. In the electron paramagnetic resonance spectrum of a crystal of myoglobin, there are two peaks of absorption, one for each of these orientations (Figure 12-37). When the crystal is rotated in the magnetic field, the g factors of the populations of electrons producing these peaks vary sinusoidally between a maximum of 6.0 and a minimum of 2.4.³⁹⁰ When the molecules of myoglobin are oriented randomly in a frozen solution, the absorption observed is the sum of all of the absorptions from the individual random orientations. As with the g factor of the absorption from an unpaired electron on a paramagnetic ion, the coupling constant for a particular hyperfine splitting from a magnetic nucleus can vary dramatically as the orientation of the orbital of the unpaired electron relative to the applied magnetic field is varied.³⁹¹ Such variations in g factor or coupling constant can be used to draw conclusions about the orientation in a particular sample of molecules in which the unpaired electron resides.

Just as energy transfer between two chromophores can be used to estimate the distance between them

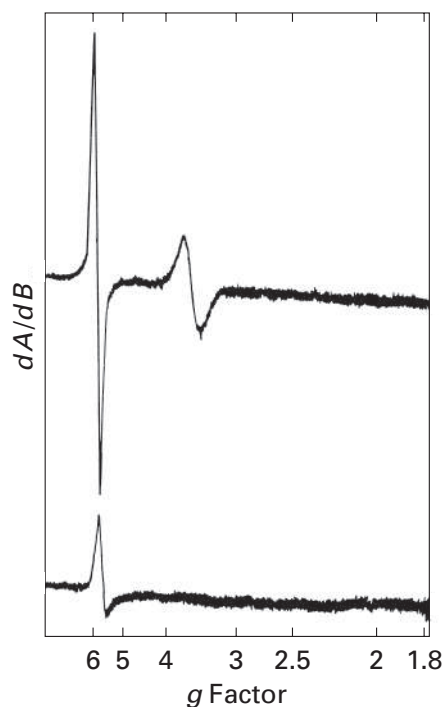


Figure 12-37: Electron paramagnetic spectra of a crystal of myoglobin in the ferric oxidation state.³⁹⁰ A single crystal of ferrimyoglobin from *P. catodon* (1×10^{-8} mol) was oriented in the applied magnetic field so that its *ab* plane was parallel to the direction of the field and its *a* axis was at an angle of about 40° to the direction of the field. After the spectrum (A) of the crystal in this orientation was recorded, it was dissolved in 10 μ L of distilled water, the solution was frozen, and a second spectrum (B) was recorded. Both spectra were recorded at 77 K. The vertical axis, dA/dB , is the first derivative of the absorption with respect to magnetic flux density. The horizontal axis is calibrated in *g* factor, which is inversely proportional to magnetic flux density (Equation 12-65), hence the inverse calibration. Reprinted with permission from ref 390. Copyright 1967 American Society for Biochemistry and Molecular Biology.

(Equation 12-49), so can the magnitude of the **magnetic dipolar interactions** between two unpaired electrons. These dipolar interactions can be between an unpaired electron on a transition metal cation and an unpaired electron on a 1-oxyl-2,2,5,5-tetramethylpyrrolin-3-yl group³⁸² or between two 1-oxyl-2,2,5,5-tetramethylpyrrolin-3-yl groups.³⁹² The magnetic dipolar interactions decrease the intensity of the absorptions from the two unpaired electrons, and under appropriate circumstances an estimate of the distance between them can be made.³⁹² As with transfer of fluorescent energy, however, there are significant and usually uncontrolled **orientation factors** affecting the calculation. Changes in the magnitude of magnetic dipolar interactions between two 1-oxyl-2,2,5,5-tetramethylpyrrolin-3-yl groups inserted at specific positions in the amino acid sequence of T4 lysozyme have been used to monitor a change in its conformation upon binding of a ligand and estimate the change in distance between those positions during that conformational change.³⁹³

In simple situations it is often possible to infer the

identities of the magnetic nuclei producing a particular set of hyperfine splittings from a knowledge of the chemical structure of the radical bearing the unpaired electron. The identities of these inferred magnetic nuclei are usually confirmed by replacing them with one of their **isotopes** and observing the expected change in the hyperfine splitting. For example, the identity of the ^1H producing the hyperfine splitting of the absorption from the unpaired electron in the glycy radical in formate C-acetyltransferase (Figure 12-36) and the identities of the ^1H s around the phenyl ring producing the hyperfine splitting of the absorption from the unpaired electron in the tyrosyl radical of ribonucleoside-diphosphate reductase from *E. coli*³⁹⁴ were confirmed by replacing them with ^2H s.

If the pattern of hyperfine splitting cannot be explained because several unknown magnetic nuclei contribute to it, further information about their identity can be obtained from an **electron nuclear double resonance (ENDOR) spectrum**.^{395,396} The magnetic field of the spectrometer is adjusted so that the population of unpaired electrons is absorbing at the center of one of its hyperfine peaks. That absorption is then saturated by increasing the power of the microwave transmitter. The sample is then irradiated with a radiofrequency transmitter, the frequency of which is varied progressively. When the transmitter reaches the Larmor frequency of a population of magnetic nuclei participating in the particular hyperfine coupling that produced the peak on which the spectrophotometer is poised, the rate of relaxation of the electron increases and its absorption at saturation increases. The output of the spectrophotometer shows peaks of increases in absorption at radio frequencies equal to the Larmor frequencies of the population of nuclei producing the hyperfine splitting. There are a pair of Larmor frequencies for each population of coupled nuclei because the unpaired electron splits their absorption by the same coupling constant with which they split the absorption of the electron. These two peaks are centered on the Larmor frequency the population of coupled magnetic nuclei would have in the absence of the population of unpaired electrons.

Because the resolution of an electron nuclear double resonance spectrum is low and the splitting is large, usually all that can be learned is the **element of the population of coupled nuclei** and the coupling constant between that population of nuclei and the population of unpaired electrons. When $[8-^{15}\text{N}]N^8$ -hydroxyarginine is bound by rat neuronal nitric-oxide synthase, the hyperfine splitting of the absorption from the Fe^{3+} in the heme of the enzyme at a *g* factor of 4.03 could be shown to arise from the ^{15}N of the ligand because in the electron nuclear double resonance spectrum (Figure 12-38) there were a pair of peaks with the proper coupling constant centered on the Larmor frequency of the nucleus of a ^{15}N .³⁹⁷ Likewise, it could be shown that the hyperfine peak at a *g* factor of 1.96 in the electron paramag-

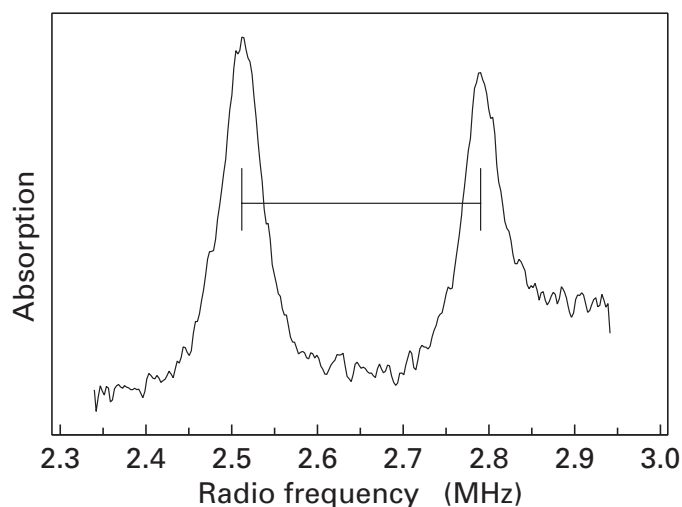


Figure 12-38: Electron nuclear double resonance (ENDOR) spectrum of a solution of 0.5 mM neuronal nitric-oxide synthase from *Rattus norvegicus* and 5 mM $[8-^{15}\text{N}]N^6$ -hydroxyarginine.³⁹⁷ The electron paramagnetic resonance spectrum of the frozen solution at 2 K had three peaks of absorption from the unpaired electron in the heme of the protein with g factors of 7.65, 4.03, and 1.8. The spectrometer was poised on the peak of absorption at the g factor of 4.03 at a microwave carrier frequency of 34.7 GHz, the absorption of microwave energy was saturated by increasing the output of the transmitter, and the absorption of microwave energy was recorded as a function of the applied radiofrequency energy. The absorption is presented as a function of radio frequency (megahertz) centered on the Larmor frequency (2.65 MHz) for ^{15}N at the applied magnetic field. The width of each peak exceeds by many fold the span of the full range of chemical shifts observed for ^{15}N in a protein (200 ppm). Reprinted with permission from ref 397. Copyright 1999 American Chemical Society.

netic spectrum of the complex between $[^{13}\text{C}]$ ethene and the molybdenum-iron protein from nitrogenase was caused by the spin-spin coupling between the unpaired electron in the metal cluster and a ^{13}C in the bound ethene³⁸⁶ and that the hyperfine peak at a g factor of 4.23 in the complex between $[^{15}\text{N}]$ alanine, nitrous oxide, and aminocyclopropane carboxylate oxidase was caused in part by the spin-spin coupling between the ^{15}N of the alanine and the unpaired electron in the nitroxyl heme.³⁸⁵ In situations where electron nuclear double resonance spectra have several peaks because the peak of absorption on which the spectrophotometer is poised has several components, isotopic substitution, for example ^2H for ^1H , can sort out the origin of particular peaks.^{398,399} In complicated electron paramagnetic spectra in which it is unclear on which g factor a particular pattern of hyperfine peaks are centered, the coupling constants obtained from the electron nuclear double resonance spectra can often resolve the problem.

The similarity between electron nuclear double resonance spectra or electron paramagnetic spectra from two different molecules can be used as evidence that the radicals present in each are chemically identical. For example, the similarity between the electron nuclear double resonance spectrum of the radical in bovine cata-

lase to that of the tyrosyl radical in ribonucleoside-diphosphate reductase from *E. coli* confirmed that the former also was a tyrosyl radical.⁴⁰⁰ Likewise, the coincidence of the g factor for synthetic bacteriochlorophyll radical cation with the g factor of the initial photooxidized donor in the photosynthetic reaction center of photosynthetic bacteria and the fact that the absorptions from both of these radicals decreased by exactly the same factor when the bacteriochlorophyll and the bacteria, respectively, were perdeuterated identified the donor in the reaction center as a bacteriochlorophyll. The fact that under all circumstances the width of the absorption in the electron paramagnetic spectrum of the radical formed from the donor in the reaction center was $\sqrt{2}$ that of a monomeric bacteriochlorophyll radical cation demonstrated that the radical cation in the reaction center was actually an electronically coupled dimer of bacteriochlorophyll.⁴⁰¹ These conclusions were validated by the crystallographic molecular model of the protein.⁴⁰²

Suggested Reading

Parast, C.V., Wong, K.K., Kozarich, J.W., Peisach, J., & Magliozzo, R.S. (1995) Mechanism-based inactivation of pyruvate formate-lyase by fluoropyruvate: direct observation of an α -keto carbon radical, *J. Am. Chem. Soc.* 117, 10601-10602.

References

1. Tanford, C. (1961) *Physical Chemistry of Macromolecules*, Wiley, New York.
2. Reynolds, J.A., & Tanford, C. (1976) *Proc. Natl. Acad. Sci. U.S.A.* 73, 4467-4470.
3. Perrin, F. (1936) *J. Phys. Radium* 7, 1-11.
4. De La Torre, J.G., & Bloomfield, V.A. (1977) *Biopolymers* 16, 1779-1793.
5. Rocco, M., Carson, M., Hantgan, R., McDonagh, J., & Hermans, J. (1983) *J. Biol. Chem.* 258, 14545-14549.
6. Graceffa, P., Wang, C.L., & Stafford, W.F. (1988) *J. Biol. Chem.* 263, 14196-14202.
7. Mabuchi, K., Lin, J.J., & Wang, C.L. (1993) *J. Muscle Res. Cell Motil.* 14, 54-64.
8. Eimer, W., Niermann, M., Eppe, M.A., & Jockusch, B.M. (1993) *J. Mol. Biol.* 229, 146-152.
9. Gerhart, J.C., & Schachman, H.K. (1968) *Biochemistry* 7, 538-552.
10. Howlett, G.J., & Schachman, H.K. (1977) *Biochemistry* 16, 5077-5083.
11. Krause, K.L., Volz, K.W., & Lipscomb, W.N. (1985) *Proc. Natl. Acad. Sci. U.S.A.* 82, 1643-1647.
12. Kirschner, M.W., & Schachman, H.K. (1971) *Biochemistry* 10, 1919-1926.
13. Kumosinski, T.F., & Pessen, H. (1985) *Methods Enzymol.* 117, 154-182.
14. Kumosinski, T.F., & Pessen, H. (1982) *Arch. Biochem. Biophys.* 219, 89-100.
15. Potschka, M., Nave, R., Weber, K., & Geisler, N. (1990) *Eur. J. Biochem.* 190, 503-508.
16. Smith, M.H. (1970) in *Handbook of Biochemistry and*

- Selected Data for Molecular Biology* (Sober, H., Ed.) pp C-3 to C-12, CRC Press, Cleveland, OH.
17. Zubrzycki, I.Z., Frankel, L.K., Russo, P.S., & Bricker, *Biochemistry* 37, 13553–13558.
 18. Phillips, M.L., Lembertas, A.V., Schumaker, V.N., Lawn, R.M., Shire, S.J., & Zioncheck, T.F. (1993) *Biochemistry* 32, 3722–3728.
 19. Mani, R.S., Karimi-Busheri, F., Cass, C.E., & Weinfeld, M. (2001) *Biochemistry* 40, 12967–12973.
 20. Simha, R. (1940) *J. Phys. Chem.* 44, 25–34.
 21. Van Holde, K.E. (1985) *Physical Biochemistry*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ.
 22. Rocco, M., Infusini, E., Daga, M.G., Gogioso, L., & Cuniberti, C. (1987) *EMBO J.* 6, 2343–2349.
 23. Levinson, B.L., Pickover, C.A., & Richards, F.M. (1983) *J. Biol. Chem.* 258, 10967–10972.
 24. Lai, C.S., Wolff, C.E., Novello, D., Griffone, L., Cuniberti, C., Molina, F., & Rocco, M. (1993) *J. Mol. Biol.* 230, 625–640.
 25. Kataoka, M., Nishii, I., Fujisawa, T., Ueki, T., Tokunaga, F., & Goto, Y. (1995) *J. Mol. Biol.* 249, 215–228.
 26. Olah, G.A., Mitchell, R.D., Sosnick, T.R., Walsh, D.A., & Trewella, J. (1993) *Biochemistry* 32, 3649–3657.
 27. Trewella, J., Carlson, V.A., Curtis, E.H., & Heidorn, D.B. (1988) *Biochemistry* 27, 1121–1125.
 28. Guinier, A. (1939) *Ann. Phys.* 12, 161–237.
 29. Vachette, P., Koch, M.H., & Svergun, D.I. (2003) *Methods Enzymol.* 374, 584–615.
 30. Chacon, P., Diaz, J.F., Moran, F., & Andreu, J.M. (2000) *J. Mol. Biol.* 299, 1289–1302.
 31. Koch, M.H., Vachette, P., & Svergun, D.I. (2003) *Q. Rev. Biophys.* 36, 147–227.
 32. Perkins, S.J., Nealis, A.S., Sutton, B.J., & Feinstein, A. (1991) *J. Mol. Biol.* 221, 1345–1366.
 33. Shilton, B.H., Flocco, M.M., Nilsson, M., & Mowbray, S.L. (1996) *J. Mol. Biol.* 264, 350–363.
 34. Grossmann, J.G., Neu, M., Pantos, E., Schwab, F.J., Evans, R.W., Townes-Andrews, E., Lindley, P.F., Appel, H., Thies, W.G., & Hasnain, S.S. (1992) *J. Mol. Biol.* 225, 811–819.
 35. Svergun, D., Barberato, C., & Koch, M.H.J. (1995) *J. Appl. Crystallogr.* 28, 768–773.
 36. Grossman, J.G., Hasnain, S.S., Yousafzai, F.K., Smith, B.E., & Eady, R.R. (1997) *J. Mol. Biol.* 266, 642–648.
 37. Mendelson, R.A., Schneider, D.K., & Stone, D.B. (1996) *J. Mol. Biol.* 256, 1–7.
 38. DiCapua, E., Schnarr, M., Ruigrok, R.W., Lindner, P., & Timmins, P.A. (1990) *J. Mol. Biol.* 214, 557–570.
 39. Perkins, S.J., Nealis, A.S., & Sim, R.B. (1991) *Biochemistry* 30, 2847–2857.
 40. Perkins, S.J., Nealis, A.S., & Sim, R.B. (1990) *Biochemistry* 29, 1167–1175.
 41. Capel, M.S., Kjeldgaard, M., Engelman, D.M., & Moore, P.B. (1988) *J. Mol. Biol.* 200, 65–87.
 42. Moore, P.B., & Engelman, D.M. (1979) *Methods Enzymol.* 59, 629–638.
 43. Brodersen, D.E., Clemons, W.M., Jr., Carter, A.P., Wimberly, B.T., & Ramakrishnan, V. (2002) *J. Mol. Biol.* 316, 725–768.
 44. Grossmann, J.G., Sharff, A.J., O'Hare, P., & Luisi, B. (2001) *Biochemistry* 40, 6267–6274.
 45. Svergun, D.I., Barberato, C., Koch, M.H., Fetler, L., & Vachette, P. (1997) *Proteins: Struct., Funct., Genet.* 27, 110–117.
 46. Pickover, C.A., McKay, D.B., Engelman, D.M., & Steitz, T.A. (1979) *J. Biol. Chem.* 254, 11323–11329.
 47. Jin, L., Stec, B., Lipscomb, W.N., & Kantrowitz, E.R. (1999) *Proteins: Struct., Funct., Genet.* 37, 729–742.
 48. Fetler, L., & Vachette, P. (2001) *J. Mol. Biol.* 309, 817–832.
 49. Valentine, R.C., Shapiro, B.M., & Stadtman, E.R. (1968) *Biochemistry* 7, 2143–2152.
 50. Richards, K.E., & Williams, R.C. (1972) *Biochemistry* 11, 3393–3395.
 51. Williams, R.C. (1981) *J. Mol. Biol.* 150, 399–408.
 52. Yang, Z., Kollman, J.M., Pandi, L., & Doolittle, R.F. (2001) *Biochemistry* 40, 12515–12523.
 53. Koch, M., Bohrmann, B., Matthison, M., Hagios, C., Trueb, B., & Chiquet, M. (1995) *J. Cell Biol.* 130, 1005–1014.
 54. Koch, M., Bernasconi, C., & Chiquet, M. (1992) *Eur. J. Biochem.* 207, 847–856.
 55. Shotton, D.M., Burke, B.E., & Branton, D. (1979) *J. Mol. Biol.* 131, 303–329.
 56. Schramm, H.J., & Jennissen, H.P. (1985) *J. Mol. Biol.* 181, 503–516.
 57. Suzuki, K., Dahlbeack, B., & Stenflo, J. (1982) *J. Biol. Chem.* 257, 6556–6564.
 58. Laue, T.M., Johnson, A.E., Esmon, C.T., & Yphantis, D.A. (1984) *Biochemistry* 23, 1339–1348.
 59. Dahlbeack, B. (1986) *J. Biol. Chem.* 261, 9495–9501.
 60. Fox, J.W., Mayer, U., Nischt, R., Aumailley, M., Reinhardt, D., Wiedemann, H., Mann, K., Timpl, R., Krieg, T., Engel, J., et al. (1991) *EMBO J.* 10, 3137–3146.
 61. Ertl, H., Hallmann, A., Wenzl, S., & Sumper, M. (1992) *EMBO J.* 11, 2055–2062.
 62. Sasaki, T., Kostka, G., Gohring, W., Wiedemann, H., Mann, K., Chu, M.L., & Timpl, R. (1995) *J. Mol. Biol.* 245, 241–250.
 63. Wang, C.L., Chalovich, J.M., Graceffa, P., Lu, R.C., Mabuchi, K., & Stafford, W.F. (1991) *J. Biol. Chem.* 266, 13958–13963.
 64. Voss, T., Eistetter, H., Schafer, K.P., & Engel, J. (1988) *J. Mol. Biol.* 201, 219–227.
 65. Boisset, N., Taveau, J.C., Pochon, F., Barray, M., Delain, E., & Lamy, J.N. (1991) *J. Struct. Biol.* 106, 31–41.
 66. Crowther, R.A. (1971) *Philos. Trans. R. Soc. London, Ser. B: Biol. Sci.* 261, 221–230.
 67. Kessel, M., Frank, J., & Goldfarb, W. (1980) *J. Supramol. Struct.* 14, 405–422.
 68. Baker, T.S., Newcomb, W.W., Booy, F.P., Brown, J.C., & Steven, A.C. (1990) *J. Virol.* 64, 563–573.
 69. Fuller, S.D. (1987) *Cell* 48, 923–934.
 70. Baker, T.S., Newcomb, W.W., Olson, N.H., Cowser, L.M., Olson, C., & Brown, J.C. (1991) *Biophys. J.* 60, 1445–1456.
 71. Crowther, R.A., Amos, L.A., Finch, J.T., De Rosier, D.J., & Klug, A. (1970) *Nature* 226, 421–425.
 72. Kolatkar, P.R., Bella, J., Olson, N.H., Bator, C.M., Baker, T.S., & Rossmann, M.G. (1999) *EMBO J.* 18, 6249–6259.
 73. Olson, N.H., Kolatkar, P.R., Oliveira, M.A., Cheng, R.H., Greve, J.M., McClelland, A., Baker, T.S., & Rossmann, M.G. (1993) *Proc. Natl. Acad. Sci. U.S.A.* 90, 507–511.

74. Radermacher, M., Wagenknecht, T., Verschoor, A., & Frank, J. (1986) *J. Microsc.* 141, RP1–2.
75. Boisset, N., Penczek, P., Pochon, F., Frank, J., & Lamy, J. (1993) *J. Mol. Biol.* 232, 522–529.
76. Radermacher, M., Wagenknecht, T., Verschoor, A., & Frank, J. (1987) *EMBO J.* 6, 1107–1114.
77. Carazo, J.M., Wagenknecht, T., Radermacher, M., Mandiyan, V., Boublik, M., & Frank, J. (1988) *J. Mol. Biol.* 201, 393–404.
78. Ban, N., Nissen, P., Hansen, J., Capel, M., Moore, P.B., & Steitz, T.A. (1999) *Nature* 400, 841–847.
79. Shulman, S. (1953) *J. Am. Chem. Soc.* 75, 5846–5852.
80. Lowey, S., Slayter, H.S., Weeds, A.G., & Baker, H. (1969) *J. Mol. Biol.* 42, 1–29.
81. Rayment, I., Rypniewski, W.R., Schmidt-Base, K., Smith, R., Tomchick, D.R., Benning, M.M., Winkelmann, D.A., Wesenberg, G., & Holden, H.M. (1993) *Science* 261, 50–58.
82. Tsao, T.C., Bailey, K., & Adair, G.S. (1951) *J. Biochem. (Tokyo)* 49, 27–36.
83. Holtzer, A., & Lowey, S. (1959) *J. Am. Chem. Soc.* 81, 1370–1377.
84. Yoshimura, T., Kameyama, K., Maezawa, S., & Takagi, T. (1991) *Biochemistry* 30, 4528–4534.
85. Roberts, J.D., & Caserio, M.C. (1977) *Basic Principles of Organic Chemistry*, 2nd ed., W. A. Benjamin, Menlo Park, CA.
86. Kandori, H., Yoshihara, K., & Tokutomi, S. (1992) *J. Am. Chem. Soc.* 114, 10958–10959.
87. Susi, H. (1972) *Methods Enzymol.* 26C, 455–472.
88. Dong, A., Huang, P., & Caughey, W.S. (1990) *Biochemistry* 29, 3303–3308.
89. Bussian, B.M., & Sander, C. (1989) *Biochemistry* 28, 4271–4277.
90. Chetverin, A.B., & Brazhnikov, E.V. (1985) *J. Biol. Chem.* 260, 7817–7819.
91. Challou, N., Goormaghtigh, E., Cabiaux, V., Conrath, K., & Ruysschaert, J.M. (1994) *Biochemistry* 33, 6902–6910.
92. Fahmy, K., Weidlich, O., Engelhard, M., Sigrist, H., & Siebert, F. (1993) *Biochemistry* 32, 5862–5869.
93. Potter, W.T., Houtchens, R.A., & Caughey, W.S. (1985) *J. Am. Chem. Soc.* 107, 3350–3352.
94. Sage, J.T., & Jee, W. (1997) *J. Mol. Biol.* 274, 21–26.
95. Lord, R.C., & Yu, N.T. (1970) *J. Mol. Biol.* 51, 203–213.
96. Benevides, J.M., Kukolj, G., Autexier, C., Aubrey, K.L., DuBow, M.S., & Thomas, G.J., Jr. (1994) *Biochemistry* 33, 10701–10710.
97. Overman, S.A., & Thomas, G.J., Jr. (1999) *Biochemistry* 38, 4018–4027.
98. Whiting, A.K., & Peticolas, W.L. (1994) *Biochemistry* 33, 552–561.
99. Duff, L.L., Appelman, E.H., Shriver, D.F., & Klotz, I.M. (1979) *Biochem. Biophys. Res. Commun.* 90, 1098–1103.
100. Carey, P.R., Schneider, H., & Bernstein, H.J. (1972) *Biochem. Biophys. Res. Commun.* 47, 588–595.
101. Ling, J., Nestor, L.P., Czernuszewicz, R.S., Spiro, T.G., Fraczekiewicz, R., Sharma, K.D., Loehr, T.M., & Sanders-Loehr, J. (1994) *J. Am. Chem. Soc.* 116, 7682–7691.
102. Kahlow, M.A., Zuberi, T.M., Gennis, R.B., & Loehr, T.M. (1991) *Biochemistry* 30, 11485–11489.
103. Hildebrandt, P., Matysik, J., Schrader, B., Scharf, B., & Engelhard, M. (1994) *Biochemistry* 33, 11426–11431.
104. Wang, Y., Purrello, R., Georgiou, S., & Spiro, T.G. (1991) *J. Am. Chem. Soc.* 113, 6368–6377.
105. Chi, Z., Chen, X.G., Holtz, J.S., & Asher, S.A. (1998) *Biochemistry* 37, 2854–2864.
106. Susi, H., & Byler, D.M. (1986) *Methods Enzymol.* 130, 290–311.
107. Lee, D.C., Haris, P.I., Chapman, D., & Mitchell, R.C. (1990) *Biochemistry* 29, 9185–9193.
108. Holloway, P.W., & Mantsch, H.H. (1989) *Biochemistry* 28, 931–935.
109. Heimbürg, T., Schuenemann, J., Weber, K., & Geisler, N. (1996) *Biochemistry* 35, 1375–1382.
110. Garfinkel, D., & Edsall, J.T. (1958) *J. Am. Chem. Soc.* 80, 3818–3822.
111. Holzwarth, G., & Doty, P. (1965) *J. Am. Chem. Soc.* 87, 218–228.
112. Beychok, S. (1967) in *Poly- α -Amino Acids: Protein Models for Conformational Studies* (Fasman, G.D., Ed.) pp 293–337, Marcel Dekker, New York.
113. Yu, C.A., Yong, F.C., Yu, L., & King, T.E. (1971) *Biochem. Biophys. Res. Commun.* 45, 508–513.
114. Angelaccio, S., Pascarella, S., Fattori, E., Bossa, F., Strong, W., & Schirch, V. (1992) *Biochemistry* 31, 155–162.
115. Slutter, C.E., Sanders, D., Wittung, P., Malmstrom, B.G., Aasa, R., Richards, J.H., Gray, H.B., & Fee, J.A. (1996) *Biochemistry* 35, 3387–3395.
116. MacColl, R., Williams, E.C., Eisele, L.E., & McNaughton, P. (1994) *Biochemistry* 33, 6418–6423.
117. Brahms, S., & Brahms, J. (1980) *J. Mol. Biol.* 138, 149–178.
118. Moffitt, W. (1956) *Proc. Natl. Acad. Sci. U.S.A.* 42, 736–746.
119. Greenfield, N., & Fasman, G.D. (1969) *Biochemistry* 8, 4108–4116.
120. Griffin, J.H., Rosenbusch, J.P., Weber, K.K., & Blout, E.R. (1972) *J. Biol. Chem.* 247, 6482–6490.
121. Gresalfi, T.J., & Wallace, B.A. (1984) *J. Biol. Chem.* 259, 2622–2628.
122. Renzoni, D.A., Pugh, D.J., Siligardi, G., Das, P., Morton, C.J., Rossi, C., Waterfield, M.D., Campbell, I.D., & Ladbury, J.E. (1996) *Biochemistry* 35, 15646–15653.
123. Edelman, H. (1967) *Biochemistry* 6, 1948–1954.
124. Fasman, G.D. (1975–1977), *Handbook of Biochemistry and Molecular Biology*, 3rd ed., Vol. I, pp 186, CRC Press, Cleveland, OH.
125. Cuatrecasas, P., Fuchs, S., & Anfinsen, C.B. (1968) *J. Biol. Chem.* 243, 4787–4798.
126. Kirtley, M.E., & Koshland, D.E., Jr. (1972) *Methods Enzymol.* 26C, 578–601.
127. Jacobson, G.R., & Stark, G.R. (1973) *J. Biol. Chem.* 248, 8003–8014.
128. Wang, C., Yang, Y.R., Hu, C.Y., & Schachman, H.K. (1981) *J. Biol. Chem.* 256, 7028–7034.
129. Lakowicz, J.R., & Weber, G. (1973) *Biochemistry* 12, 4171–4179.
130. Calhoun, D.B., Vanderkooi, J.M., & Englander, S.W. (1983) *Biochemistry* 22, 1533–1539.
131. Szpikowska, B.K., Beechem, J.M., Sherman, M.A., & Mas, M.T. (1994) *Biochemistry* 33, 2217–2225.
132. Nishimura, J.S., Mann, C.J., Ybarra, J., Mitchell, T., & Horowitz, P.M. (1990) *Biochemistry* 29, 862–865.

133. Nelson, S.W., Iancu, C.V., Choe, J.Y., Honzatko, R.B., & Fromm, H.J. (2000) *Biochemistry* 39, 11100–11106.
134. Divita, G., Rittinger, K., Restle, T., Immendorfer, U., & Goody, R.S. (1995) *Biochemistry* 34, 16337–16346.
135. Willaert, K., Loewenthal, R., Sancho, J., Froeyen, M., Fersht, A., & Engelborghs, Y. (1992) *Biochemistry* 31, 711–716.
136. Harris, D.L., & Hudson, B.S. (1990) *Biochemistry* 29, 5276–5285.
137. Lehrer, S.S. (1971) *Biochemistry* 10, 3254–3263.
138. Eftink, M.R., & Ghiron, C.A. (1975) *Proc. Natl. Acad. Sci. U.S.A.* 72, 3290–3294.
139. Varley, P.G., & Pain, R.H. (1991) *J. Mol. Biol.* 220, 531–538.
140. Calhoun, D.B., Vanderkooi, J.M., Woodrow, G.V.d., & Englander, S.W. (1983) *Biochemistry* 22, 1526–1532.
141. Prasad, A.R., Nishimura, J.S., & Horowitz, P.M. (1983) *Biochemistry* 22, 4272–4275.
142. Wolodko, W.T., Fraser, M.E., James, M.N., & Bridger, W.A. (1994) *J. Biol. Chem.* 269, 10883–10890.
143. Merrill, A.R., Palmer, L.R., & Szabo, A.G. (1993) *Biochemistry* 32, 6974–6981.
144. Stryer, L., & Haugland, R.P. (1967) *Proc. Natl. Acad. Sci. U.S.A.* 58, 719–726.
145. Wu, C.W., & Stryer, L. (1972) *Proc. Natl. Acad. Sci. U.S.A.* 69, 1104–1108.
146. Latt, S.A., Cheung, H.T., & Blout, E.R. (1965) *J. Am. Chem. Soc.* 87, 995–1003.
147. Forster, T. (1948) *Ann. Phys. (Berlin) [6 Folge]* 2, 55–75.
148. Berman, H.A., Yguerabide, J., & Taylor, P. (1980) *Biochemistry* 19, 2226–2235.
149. Hansen, J.E., Longworth, J.W., & Fleming, G.R. (1990) *Biochemistry* 29, 7329–7338.
150. Ellis, J., Bagshaw, C.R., & Shaw, W.V. (1995) *Biochemistry* 34, 3513–3520.
151. Steiner, R.F., Albaugh, S., & Kilhoffer, M.C. (1991) *J. Fluoresc.* 1, 15–22.
152. Yamashita, S., Nishimoto, E., Szabo, A.G., & Yamasaki, N. (1996) *Biochemistry* 35, 531–537.
153. Hu, L., & Colman, R.F. (1997) *Biochemistry* 36, 1635–1645.
154. First, E.A., Johnson, D.A., & Taylor, S.S. (1989) *Biochemistry* 28, 3606–3613.
155. Lillo, M.P., Beechem, J.M., Szpikowska, B.K., Sherman, M.A., & Mas, M.T. (1997) *Biochemistry* 36, 11261–11272.
156. Pober, J.S., Iwanij, V., Reich, E., & Stryer, L. (1978) *Biochemistry* 17, 2163–2168.
157. Cornish, V.W., Benson, D.R., Altenbach, C.A., Hideg, K., Hubbell, W.L., & Schultz, P.G. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 2910.
158. Steward, L.E., Collins, C.S., Gilmore, M.A., Carlson, J.E., Ross, J.B.A., & Chamberlin, A.R. (1997) *J. Am. Chem. Soc.* 119, 6–11.
159. Carraway, K.L.r., Koland, J.G., & Cerione, R.A. (1990) *Biochemistry* 29, 8741–8747.
160. Fernando Valenzuela, C., Weign, P., Yguerabide, J., & Johnson, D.A. (1994) *Biophys. J.* 66, 674–682.
161. Ahn, T., Guengerich, F.P., & Yun, C.-H. (1998) *Biochemistry* 37, 12860–12866.
162. Dale, R.E., Eisinger, J., & Blumberg, W.E. (1979) *Biophys. J.* 26, 161–193.
163. Wu, P., & Brand, L. (1992) *Biochemistry* 31, 7939–7947.
164. Tamada, T., Kitadokoro, K., Higuchi, Y., Inaka, K., Yasui, A., de Ruiter, P.E., Eker, A.P., & Miki, K. (1997) *Nat. Struct. Biol.* 4, 887–891.
165. Kim, S.T., Heelis, P.F., Okamura, T., Hirata, Y., Mataga, N., & Sancar, A. (1991) *Biochemistry* 30, 11262–11270.
166. Wu, P., Li, Y.K., Talalay, P., & Brand, L. (1994) *Biochemistry* 33, 7415–7422.
167. Knighton, D.R., Zheng, J.H., Ten Eyck, L.F., Ashford, V.A., Xuong, N.H., Taylor, S.S., & Sowadski, J.M. (1991) *Science* 253, 407–414.
168. Grossman, S.H. (1989) *Biochemistry* 28, 4894–4902.
169. Rao, J.K., Bujacz, G., & Wlodawer, A. (1998) *FEBS Lett.* 439, 133–137.
170. Kempe, T.D., & Stark, G.R. (1975) *J. Biol. Chem.* 250, 6861–6869.
171. Hahn, L.H., & Hammes, G.G. (1978) *Biochemistry* 17, 2423–2429.
172. Jin, L., Stec, B., & Kantrowitz, E.R. (2000) *Biochemistry* 39, 8058–8066.
173. Kosman, R.P., Gouaux, J.E., & Lipscomb, W.N. (1993) *Proteins: Struct., Funct. Genet.* 15, 147–176.
174. Gouaux, J.E., & Lipscomb, W.N. (1990) *Biochemistry* 29, 389–402.
175. Brejc, K., van Dijk, W.J., Klaassen, R.V., Schuurmans, M., van Der Oost, J., Smit, A.B., & Sixma, T.K. (2001) *Nature* 411, 269–276.
176. Amir, D., & Haas, E. (1987) *Biochemistry* 26, 2162–2175.
177. Babu, Y.S., Sack, J.S., Greenhough, T.J., Bugg, C.E., Means, A.R., & Cook, W.J. (1985) *Nature* 315, 37–40.
178. Li, F., Gangal, M., Juliano, C., Gorfain, E., Taylor, S.S., & Johnson, D.A. (2002) *J. Mol. Biol.* 315, 459–469.
179. Knighton, D.R., Bell, S.M., Zheng, J., Ten Eyck, L.F., Xuong, N.H., Taylor, S.S., & Sowadski, J.M. (1993) *Acta Crystallogr. D* 49, 357–361.
180. Allen, D.J., & Benkovic, S.J. (1989) *Biochemistry* 28, 9586–9593.
181. McWherter, C.A., Haas, E., Leed, A.R., & Scheraga, H.A. (1986) *Biochemistry* 25, 1951–1963.
182. Xing, J., Forsee, W.T., Lamani, E., Maltsev, S.D., Danilov, L.L., Shibaev, V.N., Schutzbach, J.S., Cheung, H.C., & Jedrzejewski, M.J. (2000) *Biochemistry* 39, 7886–7894.
183. Patel, L.R., Curran, T., & Kerppola, T.K. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 7360–7364.
184. Lin, S.H., & Faller, L.D. (1996) *Biochemistry* 35, 8419–8428.
185. Tao, T., Gowell, E., Strasburg, G.M., Gergely, J., & Leavis, P.C. (1989) *Biochemistry* 28, 5902–5908.
186. Bilderback, T., Fulmer, T., Mantulin, W.W., & Glaser, M. (1996) *Biochemistry* 35, 6100–6106.
187. Hadfield, A.T., Harvey, D.J., Archer, D.B., MacKenzie, D.A., Jeenes, D.J., Radford, S.E., Lowe, G., Dobson, C.M., & Johnson, L.N. (1994) *J. Mol. Biol.* 243, 856–872.
188. Alley, S.C., Abel-Santos, E., & Benkovic, S.J. (2000) *Biochemistry* 39, 3076–3090.
189. Adams, S.R., Harootunian, A.T., Buechler, Y.J., Taylor, S.S., & Tsien, R.Y. (1991) *Nature* 349, 694–697.
190. Hall, J., Moubarak, A., O'Brien, P., Pan, L.P., Cho, I., & Millett, F. (1988) *J. Biol. Chem.* 263, 8142–8149.
191. Takashi, R., & Kasprzak, A.A. (1987) *Biochemistry* 26, 7471–7477.

192. Nomanbhoy, T.K., Erickson, J.W., & Cerione, R.A. (1999) *Biochemistry* 38, 1744–1750.
193. Latham, G.J., Pietroni, P., Dong, F., Young, M.C., & von Hippel, P.H. (1996) *J. Mol. Biol.* 264, 426–439.
194. Kohler, J.J., & Schepartz, A. (2001) *Biochemistry* 40, 130–142.
195. Chabbert, M., Cazenave, C., & Helene, C. (1987) *Biochemistry* 26, 2218–2225.
196. Bjornson, K.P., Amaratunga, M., Moore, K.J., & Lohman, T.M. (1994) *Biochemistry* 33, 14306–14316.
197. Gumbs, O.H., & Shaner, S.L. (1998) *Biochemistry* 37, 11692–11706.
198. Lorenz, M., Hillisch, A., Payet, D., Buttinelli, M., Travers, A., & Diekmann, S. (1999) *Biochemistry* 38, 12150–12158.
199. Jung, K., Jung, H., Wu, J., Prive, G.G., & Kaback, H.R. (1993) *Biochemistry* 32, 12273–12278.
200. Zhan, H., Choe, S., Huynh, P.D., Finkelstein, A., Eisenberg, D., & Collier, R.J. (1994) *Biochemistry* 33, 11254–11263.
201. Wingfield, P.T., Stahl, S.J., Williams, R.W., & Steven, A.C. (1995) *Biochemistry* 34, 4919–4932.
202. Crowther, R.A., Kiselev, N.A., Bottcher, B., Berriman, J.A., Borisova, G.P., Ose, V., & Pumpens, P. (1994) *Cell* 77, 943–950.
203. Gunther, H. (1995) *NMR spectroscopy: basic principles, concepts, and applications in chemistry*, 2nd ed., Wiley, New York.
204. Derome, A.E. (1987) *Modern NMR techniques for chemistry research*, Vol. 6, Pergamon Press, Oxford, England.
205. Zhao, D., & Jardetzky, O. (1994) *J. Mol. Biol.* 239, 601–607.
206. Gordon, S.L., & Wuethrich, K. (1978) *J. Am. Chem. Soc.* 100, 7094–7096.
207. Williams, G., Moore, G.R., Porteous, R., Robinson, M.N., Soffe, N., & Williams, R.J. (1985) *J. Mol. Biol.* 183, 409–428.
208. Mandel, M. (1965) *J. Biol. Chem.* 240, 1586–1592.
209. Wagner, G., Kumar, A., & Wuethrich, K. (1981) *Eur. J. Biochem.* 114, 375–384.
210. Wagner, G., & Wuethrich, K. (1982) *J. Mol. Biol.* 155, 347–366.
211. Aue, W.P., Bartholdi, E., & Ernst, R.R. (1976) *J. Chem. Phys.* 64, 2229–2246.
212. Jeener, J., Meier, B.H., Bachmann, P., & Ernst, R.R. (1979) *J. Chem. Phys.* 71, 4546–4553.
213. Ikura, M., Kay, L.E., & Bax, A. (1990) *Biochemistry* 29, 4659–4667.
214. Yang, D., & Kay, L.E. (1999) *J. Am. Chem. Soc.* 121, 2571–2575.
215. Grzesiek, S., & Bax, A. (1992) *J. Magn. Reson.* (1969–1992) 99, 201–207.
216. Wittekind, M., & Mueller, L. (1993) *J. Magn. Reson. Ser. B* 101, 201–205.
217. Kay, L.E. (1993) *J. Am. Chem. Soc.* 115, 2055–2057.
218. Bax, A., Clore, G.M., & Gronenborn, A.M. (1990) *J. Magn. Reson.* (1969–1992) 88, 425–431.
219. Fesik, S.W., Eaton, H.L., Olejniczak, E.T., Zuiderweg, E.R.P., McIntosh, L.P., & Dahlquist, F.W. (1990) *J. Am. Chem. Soc.* 112, 886–888.
220. Van Doren, S.R., Kurochkin, A.V., Ye, Q.Z., Johnson, L.L., Hupe, D.J., & Zuiderweg, E.R. (1993) *Biochemistry* 32, 13109–13122.
221. Vuister, G.W., & Bax, A. (1993) *J. Am. Chem. Soc.* 115, 7772–7777.
222. Archer, S.J., Ikura, M., Torchia, D.A., & Bax, A. (1991) *J. Magn. Reson.* (1969–1992) 95, 636–641.
223. Grzesiek, S., Kuboniwa, H., Hinck, A.P., & Bax, A. (1995) *J. Am. Chem. Soc.* 117, 5312–5315.
224. Boucher, W., Laue, E.D., Campbell-Burk, S., & Domaille, P.J. (1992) *J. Am. Chem. Soc.* 114, 2262–2264.
225. Bax, A., & Ikura, M. (1991) *J. Biomol. NMR* 1, 99–104.
226. Grzesiek, S., & Bax, A. (1992) *J. Magn. Reson.* (1969–1992) 96, 432–440.
227. Grzesiek, S., & Bax, A. (1992) *J. Am. Chem. Soc.* 114, 6291–6293.
228. Grzesiek, S., Ikura, M., Clore, G.M., Gronenborn, A.M., & Bax, A. (1992) *J. Magn. Reson.* (1969–1992) 96, 215–221.
229. Grzesiek, S., & Bax, A. (1993) *J. Biomol. NMR* 3, 185–204.
230. Yamazaki, T., Forman-Kay, J.D., & Kay, L.E. (1993) *J. Am. Chem. Soc.* 115, 11054–11055.
231. Morris, G.A., & Freeman, R. (1979) *J. Am. Chem. Soc.* 101, 760–762.
232. Mueller, L. (1979) *J. Am. Chem. Soc.* 101, 4481–4484.
233. Bodenhausen, G., & Ruben, D.J. (1980) *Chem. Phys. Lett.* 69, 185–189.
234. Zuiderweg, E.R.P. (1990) *J. Magn. Reson.* (1969–1992) 86, 346–357.
235. Bax, A., Freeman, R., & Kempsell, S.P. (1980) *J. Am. Chem. Soc.* 102, 4849–4851.
236. Rance, M., Soerensen, O.W., Bodenhausen, G., Wagner, G., Ernst, R.R., & Wuethrich, K. (1983) *Biochem. Biophys. Res. Commun.* 117, 479–485.
237. Piantini, U., Soerensen, O.W., & Ernst, R.R. (1982) *J. Am. Chem. Soc.* 104, 6800–6801.
238. Braunschweiler, L., & Ernst, R.R. (1983) *J. Magn. Reson.* (1969–1992) 53, 521–528.
239. Davis, D.G., & Bax, A. (1985) *J. Am. Chem. Soc.* 107, 2820–2821.
240. Eich, G., Bodenhausen, G., & Ernst, R.R. (1982) *J. Am. Chem. Soc.* 104, 3731–3732.
241. Bax, A., & Davis, D.G. (1985) *J. Magn. Reson.* (1969–1992) 65, 355–360.
242. Westler, W.M., Kainosho, M., Nagao, H., Tomonaga, N., & Markley, J.L. (1988) *J. Am. Chem. Soc.* 110, 4093–4095.
243. Bax, A., Sparks, S.W., & Torchia, D.A. (1988) *J. Am. Chem. Soc.* 110, 7926–7927.
244. Bax, A., Kay, L.E., Sparks, S.W., & Torchia, D.A. (1989) *J. Am. Chem. Soc.* 111, 408–409.
245. Nietlispach, D., Clowes, R.T., Broadhurst, R.W., Ito, Y., Keeler, J., Kelly, M., Ashurst, J., Oschkinat, H., Domaille, P.J., & Laue, E.D. (1996) *J. Am. Chem. Soc.* 118, 407–415.
246. LeMaster, D.M., & Richards, F.M. (1988) *Biochemistry* 27, 142–150.
247. Pervushin, K., Riek, R., Wider, G., & Wuethrich, K. (1997) *Proc. Natl. Acad. Sci. U.S.A.* 94, 12366–12371.
248. Riek, R., Wider, G., Pervushin, K., & Wuethrich, K. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 4918–4923.
249. Oh, B.H., Mooberry, E.S., & Markley, J.L. (1990) *Biochemistry* 29, 4004–4011.
250. Mooberry, E.S., Oh, B.H., & Markley, J.L. (1989) *J. Magn. Reson.* (1969–1992) 85, 147–149.
251. Stockman, B.J., Nirmala, N.R., Wagner, G., Delcamp,

- T.J., DeYarman, M.T., & Freisheim, J.H. (1992) *Biochemistry* 31, 218–229.
252. Zhang, X., Gonnella, N.C., Koehn, J., Pathak, N., Ganu, V., Melton, R., Parker, D., Hu, S.I., & Nam, K.Y. (2000) *J. Mol. Biol.* 301, 513–524.
253. Allain, F.H., Gilbert, D.E., Bouvet, P., & Feigon, J. (2000) *J. Mol. Biol.* 303, 227–241.
254. Mosbah, A., Belaich, A., Bornet, O., Belaich, J.P., Henrissat, B., & Darbon, H. (2000) *J. Mol. Biol.* 304, 201–217.
255. Arora, A., Abildgaard, F., Bushweller, J.H., & Tamm, L.K. (2001) *Nat. Struct. Biol.* 8, 334–338.
256. Gooley, P.R., Caffrey, M.S., Cusanovich, M.A., & MacKenzie, N.E. (1990) *Biochemistry* 29, 2278–2290.
257. Moy, F.J., Diblasio, E., Wilhelm, J., & Powers, R. (2001) *J. Mol. Biol.* 310, 219–230.
258. Hinck, A.P., Archer, S.J., Qian, S.W., Roberts, A.B., Sporn, M.B., Weatherbee, J.A., Tsang, M.L., Lucas, R., Zhang, B.L., Wenker, J., & Torchia, D.A. (1996) *Biochemistry* 35, 8517–8534.
259. Clore, G.M., Bax, A., Driscoll, P.C., Wingfield, P.T., & Gronenborn, A.M. (1990) *Biochemistry* 29, 8172–8184.
260. Lipari, G., & Szabo, A. (1982) *J. Am. Chem. Soc.* 104, 4546–4559.
261. Sorensen, M.D., Bjorn, S., Norris, K., Olsen, O., Petersen, L., James, T.L., & Led, J.J. (1997) *Biochemistry* 36, 10439–10450.
262. Wolf-Watz, M., Grundstrom, T., & Hard, T. (2001) *Biochemistry* 40, 11423–11432.
263. Buck, M., Boyd, J., Redfield, C., MacKenzie, D.A., Jeenes, D.J., Archer, D.B., & Dobson, C.M. (1995) *Biochemistry* 34, 4041–4055.
264. Otting, G., Liepinsh, E., & Wuthrich, K. (1993) *Biochemistry* 32, 3571–3582.
265. Falzone, C.J., Wright, P.E., & Benkovic, S.J. (1994) *Biochemistry* 33, 439–442.
266. Jones, D.D., Stott, K.M., Reche, P.A., & Perham, R.N. (2001) *J. Mol. Biol.* 305, 49–60.
267. Yi, Q., Erman, J.E., & Satterlee, J.D. (1994) *1H NMR J. Am. Chem. Soc.* 116, 1981–1987.
268. Freedman, S.J., Furie, B.C., Furie, B., & Baleja, J.D. (1995) *Biochemistry* 34, 12126–12137.
269. Jaishree, T.N., Ramakrishnan, V., & White, S.W. (1996) *Biochemistry* 35, 2845–2853.
270. Williamson, M.P., Havel, T.F., & Wuthrich, K. (1985) *J. Mol. Biol.* 182, 295–315.
271. Dubs, A., Wagner, G., & Wuethrich, K. (1979) *Biochim. Biophys. Acta* 577, 177–194.
272. Clore, G.M., Robien, M.A., & Gronenborn, A.M. (1993) *J. Mol. Biol.* 231, 82–102.
273. Redfield, C., Smith, L.J., Boyd, J., Lawrence, G.M., Edwards, R.G., Gershater, C.J., Smith, R.A., & Dobson, C.M. (1994) *J. Mol. Biol.* 238, 23–41.
274. Kay, L.E., Clore, G.M., Bax, A., & Gronenborn, A.M. (1990) *Science* 249, 411–414.
275. Powers, R., Garrett, D.S., March, C.J., Frieden, E.A., Gronenborn, A.M., & Clore, G.M. (1993) *Biochemistry* 32, 6744–6762.
276. Feng, W., Tejero, R., Zimmerman, D.E., Inouye, M., & Montelione, G.T. (1998) *Biochemistry* 37, 10881–10896.
277. Xia, B., Vlamis-Gardikas, A., Holmgren, A., Wright, P.E., & Dyson, H.J. (2001) *J. Mol. Biol.* 310, 907–918.
278. Uhrinova, S., Uhrin, D., Nairn, J., Price, N.C., Fothergill-Gilmore, L.A., & Barlow, P.N. (2001) *J. Mol. Biol.* 306, 275–290.
279. Gargaro, A.R., Soteriou, A., Frenkiel, T.A., Bauer, C.J., Birdsall, B., Polshakov, V.I., Barsukov, I.L., Roberts, G.C., & Feeney, J. (1998) *J. Mol. Biol.* 277, 119–134.
280. Ikura, M., Spera, S., Barbato, G., Kay, L.E., Krinks, M., & Bax, A. (1991) *Biochemistry* 30, 9216–9228.
281. Moy, F.J., Glasfeld, E., Mosyak, L., & Powers, R. (2000) *Biochemistry* 39, 9146–9156.
282. Drohat, A.C., Baldisseri, D.M., Rustandi, R.R., & Weber, D.J. (1998) *Biochemistry* 37, 2729–2740.
283. Chazin, W.J., Hugli, T.E., & Wright, P.E. (1988) *Biochemistry* 27, 9139–9148.
284. Kline, A.D., Braun, W., & Wuethrich, K. (1986) *J. Mol. Biol.* 189, 377–382.
285. Redfield, C., Smith, L.J., Boyd, J., Lawrence, G.M., Edwards, R.G., Smith, R.A., & Dobson, C.M. (1991) *Biochemistry* 30, 11029–11035.
286. Garrett, D.S., Powers, R., March, C.J., Frieden, E.A., Clore, G.M., & Gronenborn, A.M. (1992) *Biochemistry* 31, 4347–4353.
287. Powers, R., Garrett, D.S., March, C.J., Frieden, E.A., Gronenborn, A.M., & Clore, G.M. (1992) *Science* 256, 1673–1677.
288. Driscoll, P.C., Gronenborn, A.M., Beress, L., & Clore, G.M. (1989) *Biochemistry* 28, 2188–2198.
289. Hansen, P.E. (1991) *Biochemistry* 30, 10457–10466.
290. Delaglio, F., Kontaxis, G., & Bax, A. (2000) *J. Am. Chem. Soc.* 122, 2142–2143.
291. Huber, J.G., Moulis, J.M., & Gaillard, J. (1996) *Biochemistry* 35, 12705–12711.
292. Golden, B.L., Hoffman, D.W., Ramakrishnan, V., & White, S.W. (1993) *Biochemistry* 32, 12812–12820.
293. Habazettl, J., Myers, L.C., Yuan, F., Verdine, G.L., & Wagner, G. (1996) *Biochemistry* 35, 9335–9348.
294. Myers, L.C., Verdine, G.L., & Wagner, G. (1993) *Biochemistry* 32, 14089–14094.
295. Braun, W., Vasak, M., Robbins, A.H., Stout, C.D., Wagner, G., Kagi, J.H., & Wuthrich, K. (1992) *Proc. Natl. Acad. Sci. U.S.A.* 89, 10124–10128.
296. Smith, L.J., Sutcliffe, M.J., Redfield, C., & Dobson, C.M. (1993) *J. Mol. Biol.* 229, 930–944.
297. Otting, G., & Wuethrich, K. (1989) *J. Am. Chem. Soc.* 111, 1871–1875.
298. Xu, R.X., Meadows, R.P., & Fesik, S.W. (1993) *Biochemistry* 32, 2473–2480.
299. Gardner, K.H., Pan, T., Narula, S., Rivera, E., & Coleman, J.E. (1991) *Biochemistry* 30, 11292–11302.
300. Billeter, M., Kline, A.D., Braun, W., Huber, R., & Wuthrich, K. (1989) *J. Mol. Biol.* 206, 677–687.
301. Fraenkel, E., & Pabo, C.O. (1998) *Nat. Struct. Biol.* 5, 692–697.
302. Shaanan, B., Gronenborn, A.M., Cohen, G.H., Gilliland, G.L., Veerapandian, B., Davies, D.R., & Clore, G.M. (1992) *Science (Washington, D.C.)* 257, 961–964.
303. Zink, T., Ross, A., Luers, K., Cieslar, C., Rudolph, R., & Holak, T.A. (1994) *Biochemistry* 33, 8453–8463.
304. Gopal, B., Haire, L.F., Cox, R.A., Jo Colston, M., Major, S., Brannigan, J.A., Smerdon, S.J., & Dodson, G. (2000) *Nat. Struct. Biol.* 7, 475–478.

305. Chan, M.K., Gong, W., Rajagopalan, P.T., Hao, B., Tsai, C.M., & Pei, D. (1997) *Biochemistry* 36, 13904–13909.
306. Li, N., Zhang, W., White, S.W., & Kriwacki, R.W. (2001) *Biochemistry* 40, 4293–4302.
307. Chou, J.J., Li, S., Klee, C.B., & Bax, A. (2001) *Nat. Struct. Biol.* 8, 990–997.
308. Lukin, J.A., Kontaxis, G., Simplaceanu, V., Yuan, Y., Bax, A., & Ho, C. (2003) *Proc. Natl. Acad. Sci. U.S.A.* 100, 517–520.
309. Bertini, I., Dikiy, A., Kastrau, D.H., Luchinat, C., & Sompornpisut, P. (1995) *Biochemistry* 34, 9851–9858.
310. Schiffer, C.A., Huber, R., Wuthrich, K., & van Gunsteren, W.F. (1994) *J. Mol. Biol.* 241, 588–599.
311. Gallagher, T., Alexander, P., Bryan, P., & Gilliland, G.L. (1994) *Biochemistry* 33, 4721–4729.
312. Baldwin, E.T., Weber, I.T., St. Charles, R., Xuan, J.C., Appella, E., Yamada, M., Matsushima, K., Edwards, B.F., Clore, G.M., Gronenborn, A.M., et al. (1991) *Proc. Natl. Acad. Sci. U.S.A.* 88, 502–506.
313. Ulmer, T.S., Ramirez, B.E., Delaglio, F., & Bax, A. (2003) *J. Am. Chem. Soc.* 125, 9179–9191.
314. Baistrocchi, P., Banci, L., Bertini, I., Turano, P., Bren, K.L., & Gray, H.B. (1996) *Biochemistry* 35, 13788–13796.
315. Wolf-Watz, M., Thai, V., Henzler-Wildman, K., Hadjipavlou, G., Eisenmesser, E.Z., & Kern, D. (2004) *Nat. Struct. Mol. Biol.* 11, 945–949.
316. Wand, A.J., Ehrhardt, M.R., & Flynn, P.F. (1998) *Proc. Natl. Acad. Sci. U.S.A.* 95, 15299–15302.
317. Arrowsmith, C.H., Pachter, R., Altman, R.B., Iyer, S.B., & Jardetzky, O. (1990) *Biochemistry* 29, 6332–6341.
318. Kelly, M.J., Ball, L.J., Krieger, C., Yu, Y., Fischer, M., Schiffmann, S., Schmieder, P., Kuhne, R., Bermel, W., Bacher, A., Richter, G., & Oschkinat, H. (2001) *Proc. Natl. Acad. Sci. U.S.A.* 98, 13025–13030.
319. Tugarinov, V., Choy, W.Y., Orekhov, V.Y., & Kay, L.E. (2005) *Proc. Natl. Acad. Sci. U.S.A.* 102, 622–627.
320. Peters, A.R., Dekker, N., van den Berg, L., Boelens, R., Kaptein, R., Slotboom, A.J., & de Haas, G.H. (1992) *Biochemistry* 31, 10024–10030.
321. Meadows, D.H., Markley, J.L., Cohen, J.S., & Jardetzky, O. (1967) *Proc. Natl. Acad. Sci. U.S.A.* 58, 1307–1313.
322. Zhou, M.M., Davis, J.P., & Van Etten, R.L. (1993) *Biochemistry* 32, 8479–8486.
323. Meadows, D.H., Jardetzky, O., Epand, R.M., Ruterjans, H.H., & Scheraga, H.A. (1968) *Proc. Natl. Acad. Sci. U.S.A.* 60, 766–772.
324. Matthew, J.B., & Richards, F.M. (1982) *Biochemistry* 21, 4989–4999.
325. Botelho, L.H., & Gurd, F.R. (1978) *Biochemistry* 17, 5188–5196.
326. Bycroft, M., & Fersht, A.R. (1988) *Biochemistry* 27, 7390–7394.
327. Pesando, J.M. (1975) *Biochemistry* 14, 675–681.
328. Pesando, J.M. (1975) *Biochemistry* 14, 681–688.
329. Zhang, P.H., Graminski, G.F., & Armstrong, R.N. (1991) *J. Biol. Chem.* 266, 19475–19479.
330. Botelho, L.H., Friend, S.H., Matthew, J.B., Lehman, L.D., Hanania, G.I., & Gurd, F.R. (1978) *Biochemistry* 17, 5197–5205.
331. Glickson, J.D., Phillips, W.D., & Rupley, J.A. (1971) *J. Am. Chem. Soc.* 93, 4031–4038.
332. Mendz, G.L., Moore, W.J., & Martenson, R.E. (1983) *Biochim. Biophys. Acta* 742, 215–223.
333. McIntosh, L.P., Hand, G., Johnson, P.E., Joshi, M.D., Korner, M., Plesniak, L.A., Ziser, L., Wakarchuk, W.W., & Withers, S.G. (1996) *Biochemistry* 35, 9958–9966.
334. Joshi, M.D., Sidhu, G., Pot, I., Brayer, G.D., Withers, S.G., & McIntosh, L.P. (2000) *J. Mol. Biol.* 299, 255–279.
335. Kohda, D., Sawada, T., & Inagaki, F. (1991) *Biochemistry* 30, 4896–4900.
336. Oda, Y., Yamazaki, T., Nagayama, K., Kanaya, S., Kuroda, Y., & Nakamura, H. (1994) *Biochemistry* 33, 5275–5284.
337. Katayanagi, K., Miyagawa, M., Matsushima, M., Ishikawa, M., Kanaya, S., Nakamura, H., Ikehara, M., Matsuzaki, T., & Morikawa, K. (1992) *J. Mol. Biol.* 223, 1029–1052.
338. Shrager, R.I., Cohen, J.S., Heller, S.R., Sachs, D.H., & Schechter, A.N. (1972) *Biochemistry* 11, 541–547.
339. Szyperski, T., Antuch, W., Schick, M., Betz, A., Stone, S.R., & Wuthrich, K. (1994) *Biochemistry* 33, 9303–9310.
340. Chau, M.H., Cai, M.L., & Timkovich, R. (1990) *Biochemistry* 29, 5076–5087.
341. Dardel, F., Laue, E.D., & Perham, R.N. (1991) *Eur. J. Biochem.* 201, 203–209.
342. Osborne, M.J., Lian, L.Y., Wallis, R., Reilly, A., James, R., Kleanthous, C., & Moore, G.R. (1994) *Biochemistry* 33, 12347–12355.
343. Kawata, Y., Goto, Y., Hamaguchi, K., Hayashi, F., Kobayashi, Y., & Kyogoku, Y. (1988) *Biochemistry* 27, 346–350.
344. Endo, T., Ueda, T., Yamada, H., & Imoto, T. (1987) *Biochemistry* 26, 1838–1845.
345. Molday, R.S., Englander, S.W., & Kallen, R.G. (1972) *Biochemistry* 11, 150–158.
346. Perrin, C.L., Lollo, C.P., & Johnston, E.R. (1984) *J. Am. Chem. Soc.* 106, 2749–2753.
347. Bai, Y., Milne, J.S., Mayne, L., & Englander, S.W. (1993) *Proteins: Struct., Funct., Genet.* 17, 75–86.
348. Englander, S.W., & Staley, R. (1969) *J. Mol. Biol.* 45, 277–295.
349. Hvidt, A., & Linderstrom-Lang, K. (1954) *Biochim. Biophys. Acta* 14, 574–575.
350. Nakanishi, M., Tsuboi, M., & Ikegami, A. (1972) *J. Mol. Biol.* 70, 351–361.
351. Johnson, R.S., & Walsh, K.A. (1994) *Protein Sci.* 3, 2411–2418.
352. Katta, V., & Chait, B.T. (1993) *J. Am. Chem. Soc.* 115, 6317–6321.
353. Deng, Y., & Smith, D.L. (1999) *J. Mol. Biol.* 294, 247–258.
354. Zhang, Z., & Smith, D.L. (1993) *Protein Sci.* 2, 522–531.
355. Wang, F., Blanchard, J.S., & Tang, X.J. (1997) *Biochemistry* 36, 3755–3759.
356. Andersen, M.D., Shaffer, J., Jennings, P.A., & Adams, J.A. (2001) *J. Biol. Chem.* 276, 14204–14211.
357. Zhang, Z., Post, C.B., & Smith, D.L. (1996) *Biochemistry* 35, 779–791.
358. Resing, K.A., & Ahn, N.G. (1998) *Biochemistry* 37, 463–475.
359. Wagner, G., & Wuthrich, K. (1982) *J. Mol. Biol.* 160, 343–361.
360. Skelton, N.J., Kordel, J., Akke, M., & Chazin, W.J. (1992) *J. Mol. Biol.* 227, 1100–1117.

361. Linderstrom-Lang, K. (1955) *Chem. Soc. (London), Spec. Publ. No. 2*, 1–20, discussion 21–24.
362. Englander, S.W., & Kallenbach, N.R. (1983) *Q. Rev. Biophys.* 16, 521–655.
363. Ehrhardt, M.R., Urbauer, J.L., & Wand, A.J. (1995) *Biochemistry* 34, 2731–2738.
364. Wand, A.J., Roder, H., & Englander, S.W. (1986) *Biochemistry* 25, 1107–1114.
365. Haruyama, H., Qian, Y.Q., & Wuthrich, K. (1989) *Biochemistry* 28, 4312–4317.
366. Wang, Q.W., Kline, A.D., & Wuthrich, K. (1987) *Biochemistry* 26, 6488–6493.
367. Jandu, S.K., Ray, S., Brooks, L., & Leatherbarrow, R.J. (1990) *Biochemistry* 29, 6264–6269.
368. Perrin, C.L., Dwyer, T.J., Rebek, J., Jr., & Duff, R.J. (1990) *J. Am. Chem. Soc.* 112, 3122–3125.
369. Hvidt, A., & Nielsen, S.O. (1966) *Adv. Protein Chem.* 21, 287–386.
370. Roder, H., Wagner, G., & Wuthrich, K. (1985) *Biochemistry* 24, 7407–7411.
371. Liu, K., Cho, H.S., Hoyt, D.W., Nguyen, T.N., Olds, P., Kelly, J.W., & Wemmer, D.E. (2000) *J. Mol. Biol.* 303, 555–565.
372. Wagner, G. (1980) *Biochem. Biophys. Res. Commun.* 97, 614–620.
373. Roder, H., Wagner, G., & Wuthrich, K. (1985) *Biochemistry* 24, 7396–7407.
374. Arrington, C.B., & Robertson, A.D. (1997) *Biochemistry* 36, 8686–8691.
375. Neira, J.L., Sevilla, P., Menendez, M., Bruix, M., & Rico, M. (1999) *J. Mol. Biol.* 285, 627–643.
376. Wagner, G., Stassinopoulou, C.I., & Weuthrich, K. (1984) *Eur. J. Biochem.* 145, 431–436.
377. Kossiakoff, A.A. (1982) *Nature* 296, 713–721.
378. Wlodawer, A., & Sjeolin, L. (1982) *Proc. Natl. Acad. Sci. U.S.A.* 79, 1418–1422.
379. Paterson, Y., Englander, S.W., & Roder, H. (1990) *Science* 249, 755–759.
380. Schechter, A.N., Moravek, L., & Anfinsen, C.B. (1969) *J. Biol. Chem.* 244, 4981–4988.
381. Deligiannakis, Y., & Rutherford, A.W. (1996) *Biochemistry* 35, 11239–11246.
382. Voss, J., Salwinski, L., Kaback, H.R., & Hubbell, W.L. (1995) *Proc. Natl. Acad. Sci. U.S.A.* 92, 12295–12299.
383. Stone, T.J., Buckman, T., Nordio, P.L., & McConnell, H.M. (1965) *Proc. Natl. Acad. Sci. U.S.A.* 54, 1010–1017.
384. Johnson, D.A., Gassner, G.T., Bandarian, V., Ruzicka, F.J., Ballou, D.P., Reed, G.H., & Liu, H.W. (1996) *Biochemistry* 35, 15846–15856.
385. Rocklin, A.M., Tierney, D.L., Kofman, V., Brunhuber, N.M., Hoffman, B.M., Christoffersen, R.E., Reich, N.O., Lipscomb, J.D., & Que, L., Jr. (1999) *Proc. Natl. Acad. Sci. U. S. A.* 96, 7905–7909.
386. Lee, H.-I., Sorlie, M., Christiansen, J., Song, R., Dean, D.R., Hales, B.J., & Hoffman, B.M. (2000) *J. Am. Chem. Soc.* 122, 5582–5587.
387. Parast, C.V., Wong, K.K., Lewisch, S.A., Kozarich, J.W., Peisach, J., & Magliozzo, R.S. (1995) *Biochemistry* 34, 2393–2399.
388. Itoh, S., Ogino, M., Haranou, S., Terasaka, T., Ando, T., Komatsu, M., Ohshiro, Y., Fukuzumi, S., Kano, K., et al. (1995) *J. Am. Chem. Soc.* 117, 1485–1493.
389. Parast, C.V., Wong, K.K., Kozarich, J.W., Peisach, J., & Magliozzo, R.S. (1995) *J. Am. Chem. Soc.* 117, 10601–10602.
390. Yonetani, T., & Schleyer, H. (1967) *J. Biol. Chem.* 242, 3919–3925.
391. Libertini, L.J., Waggoner, A.S., Jost, P.C., & Griffith, O.H. (1969) *Proc. Natl. Acad. Sci. U.S.A.* 64, 13–19.
392. Altenbach, C., Oh, K.J., Trabanino, R.J., Hideg, K., & Hubbell, W.L. (2001) *Biochemistry* 40, 15471–15482.
393. McHaourab, H.S., Oh, K.J., Fang, C.J., & Hubbell, W.L. (1997) *Biochemistry* 36, 307–316.
394. Sjoberg, B.M., Reichard, P., Graslund, A., & Ehrenberg, A. (1978) *J. Biol. Chem.* 253, 6863–6865.
395. Feher, G. (1956) *Phys. Rev.* 103, 834–835.
396. Feher, G. (1959) *Phys. Rev.* 114, 1219–1244.
397. Tierney, D.L., Huang, H., Martasek, P., Masters, B.S., Silverman, R.B., & Hoffman, B.M. (1999) *Biochemistry* 38, 3704–3710.
398. Edmondson, D.E., & D'Ardenne, S.C. (1989) *Biochemistry* 28, 5924–5930.
399. Bender, C.J., Sahlin, M., Babcock, G.T., Barry, B.A., Chandrashekar, T.K., Salowe, S.P., Stubbe, J., Lindstroem, B., Petersson, L., Ehrenberg, A., & Sjoberg, B. (1989) *J. Am. Chem. Soc.* 111, 8076–8083.
400. Ivancich, A., Jouve, H.M., Sartor, B., & Gaillard, J. (1997) *Biochemistry* 36, 9356–9364.
401. McElroy, J.D., Feher, G., & Mauzerall, D.C. (1972) *Biochim. Biophys. Acta* 267, 363–374.
402. Deisenhofer, J., Epp, O., Miki, K., Huber, R., & Michel, H. (1984) *J. Mol. Biol.* 180, 385–398.
403. Toyoshima, C., Nakasako, M., Nomura, H., & Ogawa, H. (2000) *Nature* 405, 647–655.
404. Toyoshima, C., & Mizutani, T. (2004) *Nature* 430, 529–35.
405. Sorensen, T.L., Moller, J.V., & Nissen, P. (2004) *Science* 304, 1672–5.
406. Toyoshima, C., Nomura, H., & Tsuda, T. (2004) *Nature* 432, 361–8.
407. Olesen, C., Sorensen, T.L., Nielsen, R.C., Moller, J.V., & Nissen, P. (2004) *Science* 306, 2251–5.
408. Toyoshima, C., & Nomura, H. (2002) *Nature* 418, 605–11.

Chapter 13

Folding and Assembly

Each polypeptide begins its existence by emerging, amino terminus foremost, from a ribosome. Its initial amino acid sequence is the complete translation of the sequence in which the codons are arranged between the start codon and the stop codon on the messenger RNA. At some point in its early history, the polypeptide folds to assume its native state. The **native state** of a polypeptide is the limited set of equilibrating conformations in which it will spend the remainder of its lifetime and in which it is capable of performing its role within or on behalf of the living organism in which it was synthesized. The native state is the set of conformations of the polypeptide represented by the crystallographic molecular models of the protein. It is also referred to as the **folded state**. On the basis of the easily verified existence and identity of the native state, a denatured state of a polypeptide can be defined as its antonym. A **denatured state** of a polypeptide is any set of equilibrating conformations of that polypeptide that is not or does not contain the set of conformations of the native state. As it emerges from the ribosome, the nascent polypeptide is in a denatured state.

The initial folded state of the polypeptide can undergo posttranslational modification, it can combine with several other identically folded polypeptides of the same sequence or several other folded polypeptides of a different sequence and structure, or it can enter a helical polymeric protein as one of the protomers. The product of these steps is the mature native state of the protein encountered in the living tissue. The order in which these later processes occurs cannot be predicted, but all of them usually follow the folding of the unadorned polypeptide because it is usually only the folded polypeptide that contains the information controlling them.

Accordingly, the steps in the maturation of a protein can be divided into folding, posttranslational modification, and assembly. **Folding** is any process by which the polypeptide initially in a denatured state, for example, its set of conformations as it emerges from the ribosome, assumes the folded native state. **Assembly** is the process by which individual folded polypeptides associate to form their ultimate oligomeric or polymeric protein.

Thermodynamics of Folding

A polypeptide is a polymer of amino acids (2–15). It is known from studies of polymers in general that their

conformational behavior depends critically on the solvent in which they are dissolved.¹ If the functional groups of its repeating units are miscible with the solvent, the polymer is free to expand and expose all of those functional groups to that solvent without penalty. Such a solvent is a **good solvent**. When a polypeptide is dissolved in a good solvent, rotation about each bond between amide nitrogen and α carbon and between α carbon and acyl carbon is permitted, within the confines of the clashes represented in the Ramachandran plot (Figure 6–4) and within the requirement that no two atoms anywhere in the polypeptide can occupy the same space at the same time. As with any other unconfined organic molecule in solution, the conformation of a polypeptide in a good solvent is continuously changing as these rotations occur at random. Such a protean polymer in a good solvent is a **random coil**. This term incorporates unavoidably the uncontrolled and continuous motion of this process. A random coil is a special type of unfolded state. The **unfolded state** of a polypeptide is a state in which the polypeptide is significantly expanded relative to the native state so that most its structure is exposed to the solution even though it may not be a fully random coil.

Unfortunately, there are few good solvents for naturally occurring polypeptides. This is due to the fact that almost all^{2,3} natural polypeptides are created to fold. To fold, they must be composed of a mixture of hydrophobic and hydrophilic amino acids, placed in a particular sequence. There is almost no solvent in which the resulting mixture of side chains is miscible. In particular, water, although it is a good solvent for the hydrophilic side chains, is a bad solvent for the hydrophobic side chains.

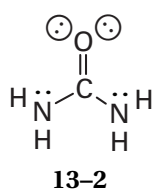
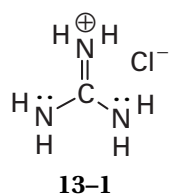
A **bad solvent** is a solvent in which functional groups of the repeating units of a polymer are only sparingly soluble. In a bad solvent, a polymer contracts to decrease the exposure of those sparingly soluble functional groups to the solution. The hydrophobic effect is a force that seeks to minimize the exposure of a hydrophobic solute to water. Because of the hydrophobic effect that is exerted on the hydrophobic side chains in a natural polypeptide, water is a bad solvent for such a polypeptide at neutral pH and at an ionic strength of 0.2 M, which are the conditions under which most proteins are found. If water were not a bad solvent, natural polypeptides would not fold. Natural polypeptides, because they have evolved in water, fold in

water. Within a cell, each polypeptide begins its life emerging from a ribosome into the cytoplasm. Even though water is a bad solvent for the emerging polypeptide, it probably remains in an unfolded state until it reaches a length at which there are a large enough number of hydrophobic side chains to accomplish its contraction. Consequently, beyond a certain length, the polypeptide has the potential to contract to form a state other than a random coil, yet there are experimental observations suggesting that at least some incomplete polypeptides adopt expanded unfolded states when they are dissolved in aqueous solution.

Only glycerol, another cohesive, hydroxylic solvent, is also able to promote folding.⁴ Other pure solvents, because they dissolve hydrophobic functional groups rather than exclude them (Figure 5-22) cannot promote the folding of a polypeptide, but they are nevertheless bad solvents because they cannot solvate the polar side chains adequately. Consequently, if a polypeptide is not in its native state in a particular solvent, it will usually also not be a random coil. When an organic solvent that is miscible with water, such as ethanol, is added to a solution of protein, it causes the protein to denature because it solvates its nonpolar groups, but it also diminishes the solvation of the polar groups, preventing the formation of a random coil. **Denaturants** are solutes that, when added to an aqueous solution of a protein, promote the formation of a denatured state of that protein. Most denaturants do not turn water into a good solvent.

If one is studying its folding, both the native state and the denatured state of a polypeptide must be well-defined. The only denatured state of a polypeptide that can be defined with sufficient accuracy is a random coil. Therefore, the folding of a polypeptide is most informatively studied if the process that is monitored is the **isomerization between the random coil and the native state**, even though this may not be what occurs in a cell. For this study to be accomplished, a good solvent is required. One of the few ways to create a good solvent is to add either guanidinium chloride or urea to an aqueous solution of the protein. Both of these solutes are denaturants, but they are denaturants that create a good solvent. Regardless of whether or not a polypeptide in a cell is a random coil at its birth, experimentally an examination of the thermodynamics of protein folding usually begins with the polypeptide as a random coil in a concentrated solution of guanidinium chloride or urea.

When almost all natural proteins, the cystines of which have been reduced to cysteines, are dissolved in solutions of **guanidinium chloride** (13-1) or **urea** (13-2)



at concentrations of 6 or 8 M, respectively, they become completely unfolded, and their constituent polypeptides become random coils.

There are several ways to demonstrate this fact.⁵ The **molar masses** of the proteins determined from the colligative properties of these solutions are those of the constitutive polypeptides rather than the oligomers. The **intrinsic viscosities** of proteins dissolved in these solutions range from 15 to 100 cm³ g⁻¹ even though the intrinsic viscosities of the native proteins are between 3 and 5 cm³ g⁻¹. Furthermore, within a set of proteins, the intrinsic viscosities of their polypeptides dissolved in those solutions are correlated to the length of the constituent polypeptides by a relationship that agrees with theoretical expectation for the behavior of random coils. The optical rotatory dispersion spectra and **circular dichroic spectra** of proteins in such solutions are those theoretically expected from a polypeptide lacking any regular secondary structure, even if the spectra of the native proteins indicate that they are predominantly α helix and β structure. The **acid-base titration curves** of proteins dissolved in these solutions lose the normally observed shifts in intrinsic pK_a brought about by the electrostatic features of the native state and become simple sums of the constituent intrinsic acid-base titrations of the constituent amino acids (Table 2-2). All of the tyrosines in the protein display ultraviolet spectrophotometric acid-base titrations with expected intrinsic values of pK_a. The rates of **amido proton exchange** become very rapid when proteins are dissolved in these solutions, and no evidence for a class of slowly exchanging amido protons is usually found. The **ultraviolet spectra** between 270 and 300 nm of proteins dissolved in these solutions are simple summations of the spectra of phenylalanine, tyrosine, and tryptophan and display none of the spectral shifts characteristic of the native states.⁶

Solutions of either guanidinium chloride or urea promote the unfolding of a polypeptide by **increasing the stability of the random coil**. This increase in stability is due to favorable changes in the solvation both of the side chains of the amino acids and of the polypeptide backbone brought about by these solutes. From measurements of the solubility of various amino acids, as well as diglycine and triglycine, in solutions of either urea⁷ or guanidinium chloride,⁸ the standard free energies of transfer of both the side chains of the amino acids and the peptide bond between water and solutions of urea or guanidinium chloride have been estimated (Table 13-1). These **standard free energies of transfer** were derived from the differences between the solubilities of each of the amino acids and peptides and the solubilities of glycine in water, 7 M urea, or 5 M guanidinium chloride. To arrive at these estimates, it was assumed that the differences between the standard free energies of solution of glycine and each of the other amino acids in the various solutions of denaturants would give the standard

free energies of transfer of each of the side chains or the peptide bond, respectively.

The values obtained for leucine, phenylalanine, and tryptophan agree quite closely with direct measurements of the standard free energies of transfer for isobutane, toluene, and skatole as models of the respective side chains (Table 13–1).⁹ The *N*-acetyl ethyl esters of leucine and phenylalanine, however, were found to have standard free energies of transfer, relative to ethyl *N*-acetylglucinate, that were much less negative (Table 13–1).¹⁰ It may be premature to attach any significance to the absolute numerical values of these various estimates.

It has been uniformly observed that the free energies of transfer of both hydrophobic solutes and neutral hydrophilic solutes such as peptides between water and either 7 M urea or 5 M guanidinium chloride have negative values. Unlike the hydrophobic effect, which is imposed only on hydrogen–carbon bonds, the increase in solvation performed by urea and guanidinium ion is linearly related to the **accessible surface area** of the side chain, regardless of its polarity,¹¹ so that both hydrophobic and hydrophilic functional groups that are exposed to the solution upon formation of the random coil are more favorably solvated in solutions of urea or guanidinium chloride than they would be in water. This stabilization of the random coil increases monotonically, but not linearly¹² with the concentration of denaturant.^{7,8,10} At some concentration of the denaturant, which differs for each protein, the unfolded polypeptide becomes more stable than the folded polypeptide. This point is reached not only because of the increase in favorable solvation but also because the unfolded polypeptide is more disordered.

The favorable solvation of both polar and nonpolar functional groups in a polypeptide by urea or guanidinium ion quantified in the free energies of transfer in

Table 13–1 has been explained as the result of preferential binding of the denaturant to the random coil.^{5,13,14}

There is no evidence, however, for the existence of particular binding sites for either of these denaturants, which if they existed would have to be distributed rather uniformly over the dramatically heterogeneous surface of the random coil. A more realistic explanation would be that these denaturants partition favorably into the peculiar layer of water solvating the random coil, relative to the water in the bulk of the solution.¹⁵ Regardless of the molecular explanation, the experimental observation that accounts for the increase in the stability of the random coil relative to that of the native state is that both urea and guanidinium ion preferentially solvate those portions of a polypeptide exposed to the solution,^{15,16} and a random coil simply exposes more of the polypeptide than does the native state. The **preferential solvation** (Equation 1–57) of bovine serum albumin in its native state¹⁵ by urea is +0.10 g mL^{–1}; and by guanidinium ion, +0.14 g mL^{–1}. The increase in preferential solvation that occurs during the unfolding of lysozyme from *Gallus gallus* by urea¹⁶ is +0.25 g mL^{–1}. These positive experimentally measured preferential solvations demonstrate that both of these denaturants are significant salting-in solutes. They do not exert their effects by decreasing the cohesion of the water and producing in turn a decrease in the hydrophobic effect because both urea and guanidinium ion increase the surface tension of an aqueous solution.¹⁷

The favorable solvation of the hydrophobic side chains by urea and guanidinium chloride does make a major contribution to the stabilization of the random coil (Table 13–1). This observation suggests that urea and guanidinium chloride cause the solution to become more like a usual organic solvent in its properties (Figure 5–22). This effect may result from the stable introduction

Table 13–1: Estimates of the Standard Free Energy of Transfer of Various Side Chains of the Amino Acids between Water and Solutions of Urea or Guanidinium Chloride

amino acid side chain	$\Delta G^{\circ}_{\text{transfer, H}_2\text{O} \rightarrow 7 \text{ M urea}}$ (kJ mol ^{–1})			$\Delta G^{\circ}_{\text{transfer, H}_2\text{O} \rightarrow 5 \text{ M GdmCl}}$ (kJ mol ^{–1})		
	amino acid ^a	alkane model ^b	<i>N</i> -acetyl ethyl ester ^c	amino acid ^a	alkane model ^b	<i>N</i> -acetyl ethyl ester ^c
leucine	–1.1	–1.0	+0.1	–1.8	–1.3	–0.1
phenylalanine	–2.2	–1.9	–0.7	–2.8	–2.5	–1.3
tryptophan	–3.2	–3.2		–4.6	–4.0	
methionine	–1.5			–2.0		
threonine	–0.4			–0.5		
tyrosine	–2.8			–2.9		
histidine	–1.0			–1.7		
asparagine	–1.6			–2.4		
glutamine	–0.8			–1.4		
peptide bond	–0.8		–0.5	–1.3		–0.8

^aCalculated from the difference between solubility of glycine and the appropriate amino acid.^{7,8} ^bFree energy of transfer of isobutane, toluene, or skatole.⁹ ^cDifference in free energy of transfer of ethyl *N*-acetylglucinate and *N*-acetyl ethyl ester of the amino acid.¹⁰

of the nonpolar π clouds of the denaturants (2–26) into the solution. *N*-Alkyl-, *N,N'*-dialkyl-, and *N,N,N',N'*-tetraalkylureas are even more effective at increasing the solubility of naphthalene, indole, and ethyl *N*-acetyltryptophanate in water than is urea itself.^{18,19} This observation also suggests that it is nonpolar noncovalent interactions between urea and the hydrophobic amino acids that explain its ability to solvate them favorably.¹⁸ The fact that methylurea, dimethylurea, and tetramethylurea are each in turn increasingly better denaturants of proteins than urea itself¹⁹ and the fact that some alkylureas are better denaturants than even guanidinium chloride²⁰ suggest that the favorable solvation by urea of the hydrophobic functionalities revealed during the formation of the random coil, in addition to its ability as a donor or acceptor of hydrogen bonds,¹⁴ is the major feature of its ability to promote unfolding.

A polypeptide will fold only if the free energy of the native state is less than the free energy of all accessible denatured states. Because of this requirement, for example, a nascent polypeptide cannot fold until it is long enough for the native state to contain a large enough collection of noncovalent interactions to overcome the significant unfavorable loss of standard entropy that must always accompany folding. It is also the case that a polypeptide which has undergone extensive covalent **posttranslational modification** after it originally folded may not be able to fold again after it has been returned to a denatured state. For example, proinsulin can be unfolded and its cystines reduced to cysteine. The protein will then refold spontaneously to its native state, and the proper cystines will reform under oxidizing conditions.²¹ Insulin, however, which is a posttranslationally modified fragment of proinsulin, missing 25 amino acids from the middle of the polypeptide, does not refold spontaneously after it has been unfolded and its cysteines reduced, and it can be refolded only with subterfuge. The only fact that seems to be inescapable is that, at some point in its lifetime, a polypeptide has a covalent structure capable of folding to produce either the mature native state directly or an initial native state, which is modified subsequently but retains its basic folded state.

A polypeptide that has not been modified so extensively as to cause the mature native state to be higher in free energy than the random coil or higher in free energy than any other accessible denatured state will, under the proper circumstances, **spontaneously refold** to its mature native state after it has been purposely turned into a random coil by dissolving it in 6 M guanidinium chloride or 8 M urea. Most of our understanding of the folding of polypeptides has been derived from the study of such conformational isomerizations. Their existence states that all of the information necessary to achieve the proper native state resides in the amino acid sequence of the polypeptide.

The conformational isomerization that encom-

passes the process of protein folding can be presented as the equilibrium



where F is the polypeptide folded in its native state and U is the unfolded state. The rate constants k_F and k_U are composite rate constants including any kinetic steps between the two extremes, and the **equilibrium constant for folding**, K_{Fd} , is defined by

$$K_{Fd} = \frac{[F]_{eq}}{[U]_{eq}} = \frac{k_F}{k_U} \quad (13-2)$$

Because polypeptides folded in their native state are by design reasonably stable in aqueous solution at physiological temperatures and ranges of pH, the molar concentration of the unfolded state under normal circumstances is immeasurably low, and in such a situation, neither the equilibrium constant nor the thermodynamic changes associated with folding can be measured directly by following the concentrations of the two forms of the protein. One solution to this problem is to shift the equilibrium by introducing an unnatural perturbation. Because the unfolded states produced by adding guanidinium ion or urea are random coils, the least controversial perturbation that can be used to shift the equilibrium is to add increasing concentrations of one or the other of these solutes to a series of solutions of the protein. As the concentrations of these perturbants are increased, the unfolded form of the protein becomes more and more stable until the equilibrium constant for Equation 13–1 is small enough for measurable amounts of the unfolded form of the protein to exist.

Raising the temperature or lowering the pH of the solution or a combination of these perturbations also stabilizes unfolded states of the protein relative to the native state. The decrease in the magnitude of the equilibrium constant for Equation 13–1 brought about by **raising the temperature** is presumably due to increases in thermal motion that always shift reactions in favor of more disordered states. The decrease in the magnitude of the equilibrium constant brought about by **lowering the pH** is due to the fact that, because an unfolded state is expanded relative to the folded state, it can support a greater net charge and thus can take up more protons than the folded state as the pH is lowered. The relationship between the equilibrium constant for folding and the pH of the solution is governed by the differential equation⁵

$$\frac{\partial \ln K_{Fd}}{\partial \ln a_{H^+}} = \bar{Z}_{H,F} - \bar{Z}_{H,U} \quad (13-3)$$

where a_{H}^+ is the activity of protons in the solution and $\bar{Z}_{\text{H,U}}$ and $\bar{Z}_{\text{H,F}}$ are the mean net proton charge numbers of the unfolded and the folded states of the protein, respectively.

The unfolded state of the protein can support a greater mean net proton charge number because the values of $\text{p}K_{\text{a}}$ for its functional groups are higher. Consequently, it takes up more protons as the pH is lowered than does the folded state. This fact is verified by measuring the **acid–base titration curves** of the protein (Figure 1–11) in the absence and the presence of 6 M guanidinium chloride.^{22–24} The two titration curves are then related to each other in absolute terms by measuring the moles of protons that must be added to a solution to maintain a constant pH as the protein is unfolded by adding guanidinium chloride.²⁴ In the acid region of the titration curves, the mean net proton charge number of the unfolded state is usually greater than that of the folded state, so the equilibrium constant for folding decreases as the pH is lowered (Equation 13–3). The decrease in K_{Fd} becomes more pronounced the more the pH is lowered so that a larger and larger fraction of the carboxylates in the protein become involved in the titration.²³ At low ionic strength, there also may be a small increase in the repulsion between the positively charged side chains in the compact native state, which destabilizes it relative to the denatured state.²⁵

Because a side chain that is an acid–base often titrates anomalously when it becomes incorporated into the native structure of a protein, its incorporation will affect the equilibrium constant for folding. If the side chain is a carboxylic acid, the effect of its incorporation is most readily understood by considering the folding–unfolding at a pH low enough that it is fully protonated in both the unfolded and the folded states. If it is buried in the folded state so that its $\text{p}K_{\text{a}}$ is elevated,²⁶ the carboxylic acid will lose its proton at a higher pH in the folded state than in the unfolded state. Consequently, as the pH is raised, $\bar{Z}_{\text{H,F}}$ will be greater than it would have been if the carboxylic acid had not been buried, and the equilibrium constant for folding will be smaller than it would have been.^{27,28} If the carboxylic acid has its $\text{p}K_{\text{a}}$ lowered by participating as the acceptor in one or more hydrogen bonds in the folded state,^{29,30} it will lose its proton at a lower pH in the folded state than in the unfolded state, and the equilibrium constant for folding will be larger than it would have been.³¹ For example, Aspartate 76 in ribonuclease T₁ from *Aspergillus oryzae* participates in several hydrogen bonds in the native state that lower its $\text{p}K_{\text{a}}$ to 0.5, coincident with an increase in the equilibrium constant for its folding of a factor of 400.³⁰ A buried lysine, the $\text{p}K_{\text{a}}$ of which is lowered because it remains unprotonated in the folded state of a protein,²⁶ also decreases the stability of the folded state relative to the unfolded state³² for the same reasons that a buried carboxylic acid does, but the argument begins with a pH high enough that the lysine is unprotonated in

the unfolded state and the pH is decreased. Similarly a histidine with an elevated $\text{p}K_{\text{a}}$ stabilizes the folded state.²⁹ All of these shifts can be explained quantitatively by considering the effects of the respective values of $\text{p}K_{\text{a}}$ for a particular side chain on the magnitudes of $\bar{Z}_{\text{H,F}}$ and $\bar{Z}_{\text{H,U}}$ ⁵ in the integrated form of Equation 13–3.^{31–33}

A solution of **guanidinium chloride at 6 M** seems to produce the most complete unfolding to the random coil.⁵ The values of the various physical parameters for the same protein dissolved in 8 M urea rather than 6 M guanidinium chloride are slightly but significantly displaced in the direction of a folded state. Proteins dissolved in solutions of low pH the temperatures of which have been raised until no further change in optical rotation occurs will still display further changes when guanidinium chloride is added,^{5,34} even when no intramolecular hydrogen bonds seem to remain in the denatured protein,³⁴ and this observation suggests that reversible thermal denaturation does not produce a random coil. Lowering the pH of a solution of a protein without applying heat often leads to its denaturation, but the denatured state produced by acid alone usually also retains residual structure.^{35,36} When either the temperature is increased or the pH is lowered, hydrophobic clusters (Figure 6–21) in otherwise unfolded polypeptides should remain associated. This would account for the incomplete unfolding observed in these situations.

In any meaningful measurement of the properties of folding, the conditions must be such that the reaction remains reversible. When a concentrated solution of ovalbumin and lysozyme, otherwise known as the white of an egg, is heated, the polypeptides unfold but then rapidly coagulate among themselves to form a white, intractable, gelatinous precipitate. In the unfolded state produced initially by raising the temperature, otherwise buried hydrophobic amino acids on these polypeptides all become simultaneously exposed to the solution and noncovalent intermolecular polymerization takes place. There is little doubt that, in this example, a significant portion if not the majority of the changes in standard enthalpy and standard entropy proceeding during this process are those of the coagulation, which is of only marginal interest.

In all studies of protein folding, the first result presented should demonstrate the complete **reversibility of the reaction**. Foldings perturbed by the addition of urea or guanidinium chloride usually are reversible. With larger proteins, the rates of renaturation from a concentrated solution of urea or guanidinium chloride, however, can be slow; and if the concentration of denaturant is abruptly decreased by dilution, an otherwise reversible folding can become irreversible.³⁷ Denaturation produced by acid is usually reversible because the denatured polypeptides are so positively charged that they will not coagulate.³⁶

Although a few foldings perturbed solely by increases in temperature are reversible at neutral pH,^{38,39}

most proceed irreversibly, usually with coagulation.⁴⁰ When thermal unfolding is performed in a **scanning calorimeter**, however, the solution is heated continuously while the absorption of excess heat is monitored. It is possible that, under these conditions, the transition from folded protein to unfolded protein takes place in a short enough period that little coagulated protein accumulates, and the reaction remains reversible during that interval and only becomes irreversible upon coagulation at the higher temperatures experienced beyond the range of temperatures encompassing the transition to the denatured state. It has been argued that the behavior of many foldings in a scanning calorimeter—namely, the shapes of the curves, the effects of ligands, and the molecularities of the apparent reactions—is that expected of a simple reversible isomerization.^{41,42} At low pH (pH 2–3), however, most thermally perturbed foldings, even though they would proceed with coagulation at higher pH, often become reversible.^{43,44} Presumably, this is due to the fact that coagulation is prevented by charge repulsion among the denatured polypeptides. It is usually observed that a polypeptide denatured thermally and reversibly at low pH will coagulate visibly and irreversibly as the pH is increased, and often the onset of this coagulation is found to occur abruptly within a very narrow range of pH.⁴⁵

When a physical property of a protein such as its intrinsic viscosity, its sedimentation velocity, its optical rotation, its molar ellipticity, its intrinsic fluorescence, its

absorption of ultraviolet light, its capacity to take up protons from the solution at constant pH,²⁴ its absorption of heat at constant temperature,⁴⁶ its elution volume on chromatography by molecular exclusion,⁴⁷ its electrophoretic mobility,¹¹ or its nuclear magnetic resonance absorptions^{48,49} is measured as a function of temperature, pH, or the concentration of urea or guanidinium chloride, changes indicative of a **shift in the value of the equilibrium constant** K_{Fd} for folding (Equation 13–1) are observed (Figure 13–1A).⁵⁰

Each pair of experimental points (a square and a circle) in Figure 13–1A represents a solution of cold shock-like protein from *Thermotoga maritima* at a particular temperature, pH, and concentration of guanidinium chloride. Two different initial solutions of protein were used. Either a solution of the native protein in the absence of denaturant was diluted by mixing into a solution of guanidinium chloride (open symbols) or a solution of the unfolded protein in 5.5 M guanidinium chloride was diluted by mixing into a solution of guanidinium chloride (solid symbols). In each case, the mixture was formulated to produce the noted final concentration of denaturant. One member of each pair of points is the initial fluorescence of the solution immediately after mixing (squares). For each final concentration of guanidinium chloride, the solution was allowed to reach equilibrium (circles), which was assumed to be the state after all changes in fluorescence had ceased.

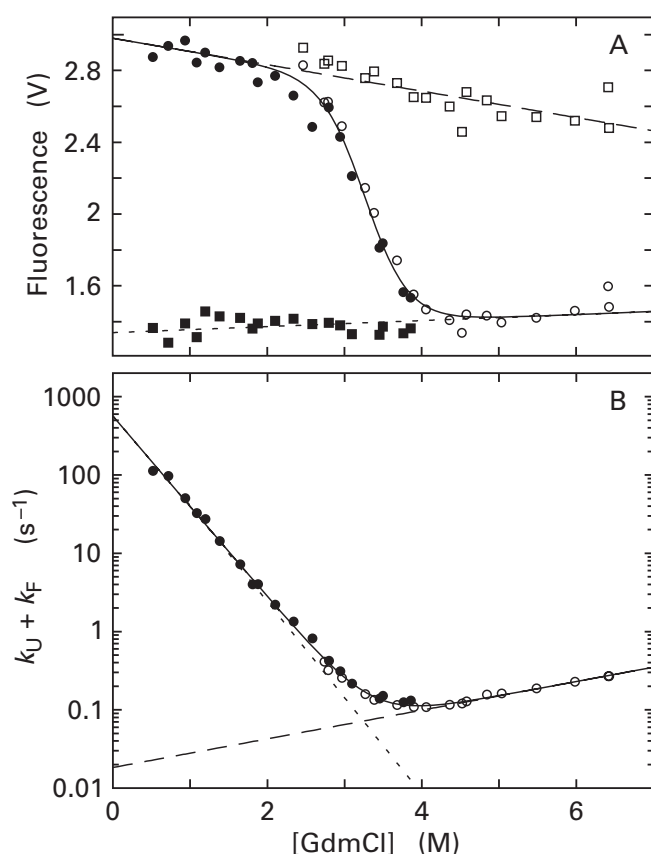


Figure 13–1: Shift of the equilibrium constant for the folding of the cold shock-like protein from *Thermotoga maritima* ($n_{aa} = 66$) in solutions of guanidinium chloride.⁵⁰ Solutions of cold shock-like protein and solutions of guanidinium chloride, both at 25 °C, were mixed in a rapid mixing chamber that then introduced the mixture immediately into the cuvette of a fluorometer so that the fluorescence of the solution (expressed as the voltage from the photomultiplier) could be monitored continuously. The emission at wavelengths greater than 300 nm upon excitation at 280 nm (intrinsic fluorescence of tryptophan) was monitored as a function of time. (A) Initial and equilibrium levels of fluorescence. A solution of protein (15 μ M) was diluted 11-fold by mixing into solutions of guanidinium chloride prepared so that the final concentration (molar) of guanidinium chloride (GdmCl) after mixing would be that noted on the horizontal axis. The initial solution of protein was prepared either in an aqueous buffer at pH 7.0 (open symbols) or in the same aqueous buffer 5.5 M in guanidinium chloride (solid symbols). The latter solution unfolded the protein completely. The initial fluorescence of each sample immediately after mixing was estimated by extrapolating the traces back to zero time (squares) to correct for the short interval between mixing and the start of the monitoring. Each isomerization was allowed to progress until the lack of further changes in fluorescence indicated that equilibrium had been reached. The fluorescence of the solution at equilibrium (circles) is plotted as a function of the final concentration of guanidinium chloride for protein that was refolded (\bullet) from 5.5 M guanidinium chloride or that was unfolded (\circ). (B) Rate constants for folding. For each sample, the fluorescence was monitored continuously as a function of time at 25 °C after mixing. The fluorescence in each sample either decreased or increased, respectively, as unfolding or folding progressed with first-order kinetics. Plots of these changes in fluorescence were fit by nonlinear least-squares to single-exponential functions to obtain first-order rate constants, $k_U + k_F$ (second⁻¹), which are plotted as a function of the final concentration (molar) of guanidinium chloride (GdmCl). Reprinted with permission from ref 50. Copyright 1998 Nature Publishing Group.

Below a certain concentration of guanidinium chloride (about 2 M), there is, at equilibrium, a linear increase in the intrinsic fluorescence with decreasing concentration of guanidinium chloride. Even in samples of native protein that eventually will unfold (\square), the immediate magnitude of the fluorescence upon addition of guanidinium chloride, before unfolding commences, falls upon this baseline. This **baseline** traces the perturbation in the intrinsic fluorescence of the fully folded state due only to addition of the denaturant in the absence of any unfolding or after complete folding has been achieved.

Above a certain concentration of guanidinium chloride (about 4 M) there is, at equilibrium, a linear increase in intrinsic fluorescence with increasing concentration of guanidinium chloride, presumed to reflect the effect of increasing the concentration of denaturant on the intrinsic fluorescence of the unfolded polypeptide. When fully unfolded protein is diluted into the range of concentrations of guanidinium chloride where it will fold (\blacksquare), the immediate fluorescence of the solution before folding commences also falls upon this other baseline.

At intermediate concentrations of guanidinium chloride, in the region of transition, the observed magnitudes of the intrinsic fluorescence at equilibrium fall between the extremes of fluorescence of the fully folded protein and the fluorescence of the fully unfolded protein. The **region of transition** is that range of denaturant concentration, pH, temperature, or pressure in which measurable concentrations of both denatured and native states are present in the solution in equilibrium with each other. It is flanked on its two sides by the ranges in which the polypeptide is almost completely folded and almost completely unfolded, respectively. It is in the region of transition that the equilibrium constant between the native state and the denatured state can be measured because sufficient concentrations of both states are present in the solution so that both are registered by the physical property being monitored.

Within the region of transition, the same equilibrium value is reached whether the folded protein (\square) or the unfolded protein (\blacksquare) is the initial state, and this demonstrates that a reversible process is being monitored. If an equilibrium between folded state and unfolded state has not been established over the interval of observation, such a coincidence between the curve for folding and the curve for unfolding is not observed.⁵¹ That the changes in the measured parameter within the region of transition cease at **intermediate values** between the extremes is also consistent with the conclusion that equilibrium has been achieved.

Assume for the moment that at all concentrations of guanidinium chloride the solution contains an equilibrium mixture of only the fully native protein and its random coil. This is the **two-state assumption**.⁵² At high enough concentrations of guanidinium chloride, the

concentration of random coil becomes sufficiently large that it contributes significantly to the intrinsic fluorescence. At this point, the isomerization of the folding (Equation 13-1) continuously interconverts measurable quantities of native state and measurable quantities of unfolded state in equilibrium with each other. As the concentration of guanidinium chloride is increased further, a greater fraction of the protein is in the unfolded state until, finally, immeasurably small amounts of the native state are present at equilibrium.

A similar monotonic transition between the native state and a denatured state, the two in equilibrium with each other, is observed when a series of solutions of a protein are each brought to a different temperature, as long as the thermal denaturation is reversible (Figure 13-2).⁵³ In the example presented in the figure, the increase in the stability of the denatured state relative to that of the native state with decreasing pH, as defined by Equation 13-3, is apparent in the shifts of the regions of transition to lower temperatures as the pH of the solution is lowered. Similar shifts of the region of transition caused by either pH⁵⁴ or temperature⁴⁶ are observed when the equilibrium between folded and unfolded states is being shifted with guanidinium chloride or urea. A monotonic transition reflecting the shift in equilibrium

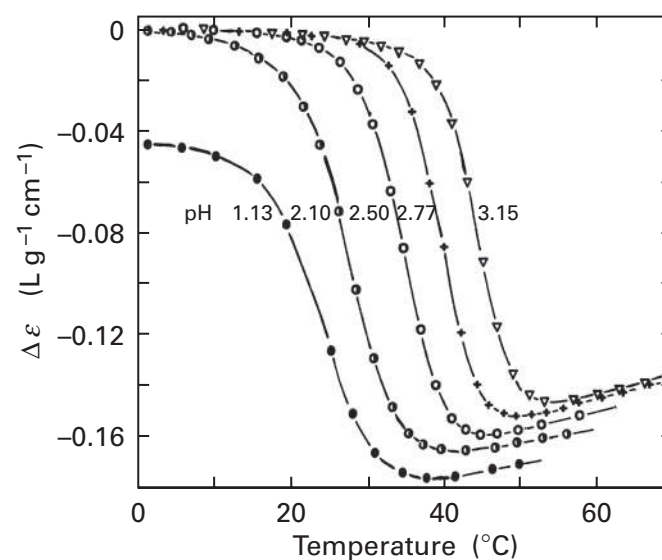


Figure 13-2: Shift of the equilibrium constant for the folding of bovine ribonuclease A produced by increases in temperature at different values of pH: 1.13 (\bullet), 2.10 (\bullet), 2.50 (\circ), 2.77 ($+$), and 3.15 (∇).⁵³ The difference in extinction coefficient ($\Delta\epsilon$; liters gram⁻¹ centimeter⁻¹) is the difference between the extinction coefficient at 287 nm for ribonuclease at pH 7, 25 °C, and the samples at the noted pH and temperature. The changes in extinction coefficient over the ranges monitored were fully reversible. The samples were buffered with 40 mM glycine (pH 2.77 and 3.15) or by the protein itself (pH < 2.7). For each point on the curves, the solution was brought to the noted temperature (degrees Celsius), and the absorbance was tabulated after it no longer changed. The change in extinction coefficient is plotted as a function of the final temperature. Reprinted with permission from ref 53. Copyright 1967 American Chemical Society.

between the native state and a denatured state can also be observed by differential scanning calorimetry,³⁹ provided the rate of temperature increase is slow enough that equilibrium is reached at each temperature and the process remains reversible over the interval in which the region of transition is traversed.^{55,56}

If the two-state assumption is made, the fluorescence of the solution (F_{obs}) observed at equilibrium at each concentration of guanidinium chloride in Figure 13-1A is

$$F_{\text{obs}} = f_F F_{0,F} + f_U F_{0,U} \quad (13-4)$$

where $F_{0,F}$ is the intrinsic fluorescence that would be observed if all of the protein were fully folded, $F_{0,U}$ is the intrinsic fluorescence that would be observed if all of the protein were fully unfolded, f_F is the fraction of the protein in the folded state, and f_U is the fraction of the protein in the unfolded state, all at the particular concentration of guanidinium chloride. Because $F_{0,U}$ and $F_{0,F}$ at each concentration of guanidinium chloride are known from the respective baselines and because $f_F + f_U = 1.0$ as a result of the two-state assumption, f_F and f_U at each concentration of guanidinium chloride can be calculated. From f_F and f_U , the equilibrium constant for folding ($K_{\text{Fd}} = f_F/f_U$) can be determined for that concentration of guanidinium chloride, temperature, and pH. The same analysis can be applied to the behavior of any other **physical property** that is directly proportional to the concentration of native protein and denatured protein, respectively, such as absorbance, optical rotation, circular dichroism, or specific viscosity.

If the two-state assumption is correct and significant concentrations of only the native state and the random coil are present at each concentration of guanidinium chloride, then the situation is dramatically simplified. It is, however, reasonable that this should be the case. When a polypeptide folds from a random coil to form the native state under physiological conditions, it must pass through intermediate states between the random coil and the native state. If, however, these intermediate states were as stable or more stable than the folded state, there would be significant, measurable concentrations of them at equilibrium, a possibility that has rarely been observed and that would be unfortunate for the protein in terms of both its function and its ability to avoid endopeptidolytic digestion. That these intermediate states remain less stable than the native state as guanidinium chloride is added to the solution is not surprising, so long as they are about as compact as the native state. There are several observations which suggest that, in many cases, the two-state assumption is valid.

In the region of transition, a point on the curve in Figure 13-1A should represent, if the two-state assumption is valid, an equilibrium mixture of only fully native protein and its random coil. The same equilibrium mixture forms when either the native state in the absence of

guanidinium chloride (□) or the random coil in a concentrated solution of guanidinium chloride (■) is transferred into a solution at that pH and final concentration of guanidinium chloride. Either of these reactions is a special case of a general kinetic category referred to as an **approach to equilibrium**.

The approach to equilibrium of either the unfolding or the folding polypeptide, respectively, should be governed only by the two composite first-order rate constants k_F and k_U (Equation 13-1) if the two-state assumption is valid. Either the unfolded state or the folded state, respectively, should be the exclusive product formed in the two reactions as the equilibrium is established. The rate at which the concentration of either species, U or F, in Equation 13-1 changes is

$$-\frac{d[U]}{dt} = \frac{d[F]}{dt} = k_F[U] - k_U[F] \quad (13-5)$$

Because the concentration of total protein, $[\text{protein}]_{\text{TOT}}$, remains constant, it follows that

$$[U] + [F] = [\text{protein}]_{\text{TOT}} = [U]_{\text{eq}} + [F]_{\text{eq}} \quad (13-6)$$

where $[U]_{\text{eq}}$ and $[F]_{\text{eq}}$ are the concentrations of native state and random coil at equilibrium. Combining Equations 13-5 and 13-6 and focusing on the concentration that is decreasing, arbitrarily chosen to be the unfolded form for the following derivation, then

$$-\frac{d[U]}{dt} = (k_F + k_U)([U] - [U]_{\text{eq}}) - k_U[F]_{\text{eq}} + k_F[U]_{\text{eq}} \quad (13-7)$$

At equilibrium no further changes occur in the concentrations of either the native state or the random coil so

$$k_U[F]_{\text{eq}} = k_F[U]_{\text{eq}} \quad (13-8)$$

at all times, and, because $[U]_{\text{eq}}$ is a constant

$$\begin{aligned} -\frac{d[U]}{dt} &= -\frac{d([U] - [U]_{\text{eq}})}{dt} = \\ &= (k_F + k_U)([U] - [U]_{\text{eq}}) = k_{\text{obs},F}([U] - [U]_{\text{eq}}) \end{aligned} \quad (13-9)$$

where $k_{\text{obs},F}$ is the observed rate constant for the approach to equilibrium during net folding.

Equation 13-9 is a simple first-order differential equation in the variable $([U] - [U]_{\text{eq}})$ and describes a first order-process in this variable. Upon integration

$$\frac{[U] - [U]_{\text{eq}}}{[U]_0 - [U]_{\text{eq}}} = \exp(-k_{\text{obs,F}} t) \quad (13-10)$$

where $[U]_0$ is the concentration of unfolded form at the beginning of the approach to equilibrium. Because the process is symmetric (Equation 13-5), if unfolding were being monitored rather than folding, the variable would have been $([F] - [F]_{\text{eq}})$ rather than $([U] - [U]_{\text{eq}})$ but the **observed rate constant for the approach to equilibrium**, $k_{\text{obs,U}}$, would still be $(k_{\text{F}} + k_{\text{U}})$.

Suppose one were to monitor any physical property Y , such as intrinsic fluorescence, absorbance, optical rotation, circular dichroism, or specific viscosity, that is directly proportional to the concentration of unfolded state and directly proportional to the concentration of folded state, respectively. The observed magnitude of that physical property for any solution containing a mixture of unfolded state and folded state at any time

$$Y_{\text{obs}} = \zeta_{\text{U}} [U] + \zeta_{\text{F}} [F] \quad (13-11)$$

where ζ_{U} and ζ_{F} are the constants of proportionality between the concentration of each species and its respective contribution to the overall magnitude of the physical property for the mixture. When Equation 13-11 is combined with Equation 13-6,

$$Y_{\text{obs}} - Y_{\text{obs,eq}} = (\zeta_{\text{U}} - \zeta_{\text{F}})[U] - (\zeta_{\text{U}} - \zeta_{\text{F}})[U]_{\text{eq}} \quad (13-12)$$

where $Y_{\text{obs,eq}}$ is the magnitude of that physical property at equilibrium. It follows from Equation 13-10 and 13-12 that

$$\frac{Y_{\text{obs}} - Y_{\text{obs,eq}}}{Y_{\text{obs,0}} - Y_{\text{obs,eq}}} = \exp(-k_{\text{obs,F}} t) \quad (13-13)$$

Equation 13-13 states that the fraction of the total change that occurs in the magnitude of a particular physical property monitoring the isomerization between unfolded state and folded state of a polypeptide should decrease exponentially as a function of time if during that isomerization only two states are significantly populated, namely, the unfolded state and the folded state. Again, because the process is **symmetric**, if the derivation just presented had been based on $[F]$ instead of $[U]$ because unfolding was being monitored rather than folding, the same outcome would have occurred, and because $k_{\text{obs,F}} = k_{\text{obs,U}} = k_{\text{U}} + k_{\text{F}}$, Equation 13-13 is the same whether net folding is progressing from an initially unfolded state or net unfolding is progressing from an initially folded state.

This derivation for the approach to equilibrium as monitored by a physical property is completely general and can be applied to any situation where the equilibrium can be described by two first-order rate constants, forward and reverse. In the case of cold shock-like protein (Figure 13-1A), clearly first-order, exponential approaches to equilibrium were observed whether unfolded protein (■) was folded (●) or folded protein (□) was unfolded (○).⁵⁰ An uncomplicated **first-order approach to equilibrium** is generally accepted as support for a two-state assumption.⁵⁴

The observed rate constants for these approaches to equilibrium for cold shock-like protein (Figure 13-1B)⁵⁰ have identical values at the same concentrations of guanidinium chloride whether the equilibrium was approached in the direction of folding (●) or in the direction of unfolding (○), another observation consistent with the two-state assumption. Furthermore, the observed rate constants for the approach to equilibrium decrease smoothly as the concentration of guanidinium chloride is increased until the region of transition is reached and increase smoothly with guanidinium chloride beyond the region of transition, as expected if only one rate constant, k_{F} , dominates in the former portion of the plot and another, k_{U} , dominates in the latter. Such uncomplicated behavior of these observed rate constants is also presented as evidence for two-state behavior. In the region of transition, both k_{F} and k_{U} contribute to the observed rate constant for the approach to equilibrium $k_{\text{F}} + k_{\text{U}}$.

It is the decrease in k_{F} and the increase in k_{U} (dashed lines in Figure 13-1B) with increasing concentrations of guanidinium chloride that together shift the equilibrium constant into the measurable range. As the guanidinium chloride has its greatest effect on the stability of the unfolded state, it is not surprising that k_{F} is affected more than k_{U} .

Often conditions are purposely chosen to ensure that the approaches to equilibrium of the folding reaction in the region of the transition between fully native and fully denatured protein are simple first-order processes. Under other conditions of temperature, pH, or concentration of denaturant, either folding or unfolding or both are not first-order processes and these conditions are avoided. For example, the equilibrium constant for folding of myoglobin at 25 °C is shifted into the measurable range at pH 4.2. At this pH both the folding and unfolding of the polypeptide are first-order processes.⁵⁷ They both proceed with clean isosbestic points in the Soret region of the visible spectrum, and this also indicates that both folding and unfolding at this pH are two-state processes. At higher or lower values of pH, however, the kinetics of both the folding and unfolding reactions become complex.

If a point in the region of transition in Figure 13-1A represents a mixture containing only the native state and the random coil, then Equation 13-4, with appropriate

substitutions, should describe the behavior of every physical parameter measured as long as it is directly proportional to the concentrations of folded state and unfolded state. In this equation, the values for the fraction of native state, f_F , and the fraction of random coil, f_U , must be the same at a particular pH and at a given concentration of guanidinium chloride regardless of the physical property used to follow folding.⁵⁴ A convincing demonstration of the **coincidence of physical measurements** was presented for the thermal denaturation of bovine ribonuclease A (Figure 13-2). When the fraction of denatured ribonuclease A, f_U , as a function of temperature at pH 2.1 was monitored by intrinsic viscosity, optical rotation, and ultraviolet absorption, all of these properties gave the same values for this quantity (Figure 13-3).⁵⁸ If even one intermediate were present, the values of all three of these physical properties for this intermediate would have to assume the same fractional deviation relative to the two extremes of each respective property. It seems unlikely that an intermediate could exist, for example, that had an intrinsic viscosity a third of the way between the intrinsic viscosities of the native protein and the random coil and an optical rotation also a third of the way between those of the native protein and the random coil and an ultraviolet absorption also a third of the way between those of the native protein and the random coil. In effect, this test is formally equivalent to the observation of an isosbestic point in a series of spectra, and this observation has always been accepted as evidence for a transition between only two states.

Intermediate states between the unfolded state and the native state of a polypeptide must be less stable than the native state at physiological temperature and pH and may remain so as the equilibrium is shifted towards the unfolded state. If, however, the perturbation does happen to destabilize these intermediate states less than it does the native state, they may become more stable than the native state before the denatured state dominates the equilibrium as the level of the perturbation is increased. If the criteria just discussed are routinely used to establish that such intermediate states are not present at significant concentrations and that the two-state assumption is valid, then these same criteria must also be able to detect the presence of such **stable intermediate states at equilibrium** when they are revealed by the perturbation. It has already been noted that, in the folding of myoglobin, the departure of the approach to equilibrium from simple first-order behavior indicated the presence of intermediate states.⁵⁷ In the case of phosphoribosylanthranilate isomerase from *Escherichia coli*, both the kinetics of folding and the kinetics of unfolding at all concentrations of guanidinium chloride proceeded in two well-resolved first-order exponential phases rather than a single phase, an observation indicating that an intermediate must be present.⁵⁹ Furthermore, plots of the transition between the folded state and the unfolded state of phosphoribosylanthranilate isomerase differed

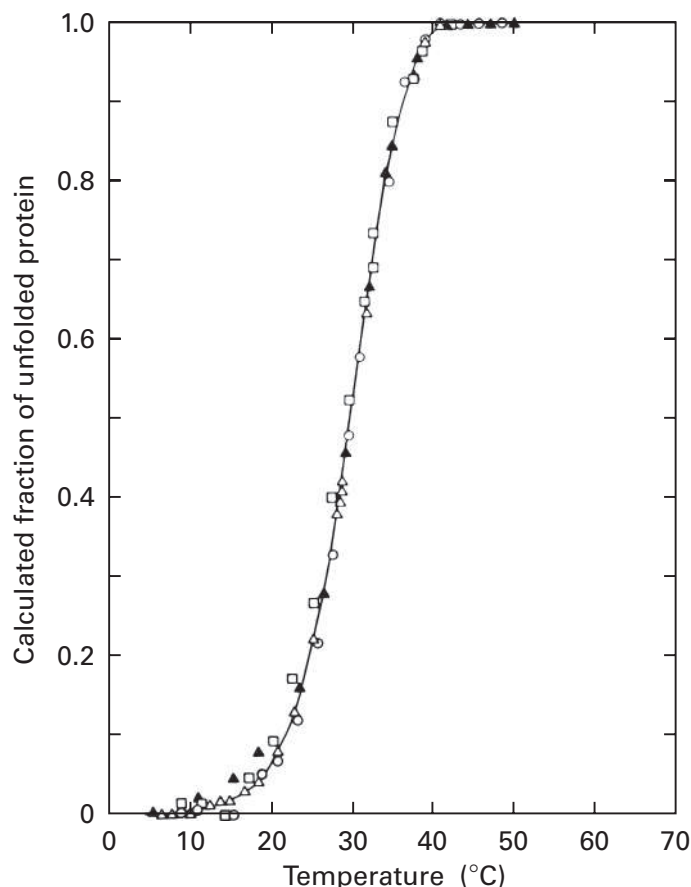


Figure 13-3: Change in the fraction of unfolded bovine ribonuclease A, f_U , calculated by Equation 13-4, for a solution of the protein at pH 2.10 (Figure 13-2) as a function of temperature.⁵⁸ The variation in three physical properties of a solution of ribonuclease A at pH 2.10 and ionic strength of 20 mM were measured as a function of temperature. These three properties were the absorbance at 287 nm (Δ), the intrinsic viscosity (\square), and the optical rotation, $[\alpha]_{365}$, at 365 nm (\circ). In each case, the direct observations were first plotted as a function of temperature as in Figure 13-2. The behavior of the physical property for the native state and the fully denatured state as a function of temperature was estimated by linear extrapolation as in Figure 13-1. The fraction of the denatured state f_U was then calculated from each separate curve by Equation 13-4. The values of the calculated fraction of unfolded protein determined by each physical property are presented together as a function of temperature (degrees Celsius). To demonstrate that the process being followed was reversible, the absorbance at 287 nm was followed a second time (\blacktriangle) after a sample was heated to 40.8 °C for 16 h and then cooled. Reprinted with permission from ref 58. Copyright 1965 American Chemical Society.

significantly in their shape and their dependence on the concentration of guanidinium chloride depending on whether the transition was monitored by circular dichroism at 278 nm, circular dichroism at 222 nm, or intrinsic fluorescence. It was concluded that at least one intermediate state of this protein was present in the region of transition.

A clear example of the existence of a stable intermediate state during the transition between the folded state and the unfolded state has been observed for bovine γ -crystallin B (Figure 13-4A).⁶⁰ When the transi-

tion was monitored by circular dichroism, sedimentation velocity, fluorescence emission at 320 nm, and fluorescence emission at 360 nm, the changes observed differed dramatically. Each curve, however, showed an obvious plateau at intermediate concentrations of urea, consistent with an almost completely populated intermediate state in this range. The kinetics of both folding and unfolding over the entire range of concentrations of urea displayed two distinct exponential phases with different rate constants during the approach to equilibrium, and these two sets of observed rate constants when plotted against the concentration of urea gave two separate, overlapping curves, each resembling the one in Figure 13-1B. The plot for the observed rate constants of the isomerization between the unfolded state and the intermediate was displaced to higher concentrations of urea than the one for the isomerization between intermediate and the native state. Consequently, it was concluded that this unfolding promoted by urea is a **three-state process** with a discrete intermediate state that is almost completely populated at concentrations of urea between 3 and 4 M.

In situations such as the one described, in which a discrete intermediate is thought to exist, the curves plotting the transition from native state to fully unfolded state as a function of the concentration of denaturant often show an obvious **plateau** or inflection,^{59,61} and these curves can be fit by equations similar to Equation 13-4 based on a three-state assumption to obtain the fraction of native state, intermediate state, and unfolded state at each concentration of denaturant. From these fractions, equilibrium constants for the isomerizations between these three states as a function of the concentration of denaturant can be calculated. In some situations, four states—the native state, the unfolded state, and two discrete intermediates—can be detected in the inflections within the plots of the magnitude of physical properties as a function of the concentration of a denaturant.^{62,63}

Often, however, the curve for a particular physical property is a smooth function with no inflections, and the only indication that an intermediate is present is the

lack of coincidence of the curves for different physical properties.^{64,65} For example, when the shift in the equilibrium between the native state and the unfolded state of bovine α -lactalbumin (Figure 13-4B)⁶⁶ was followed

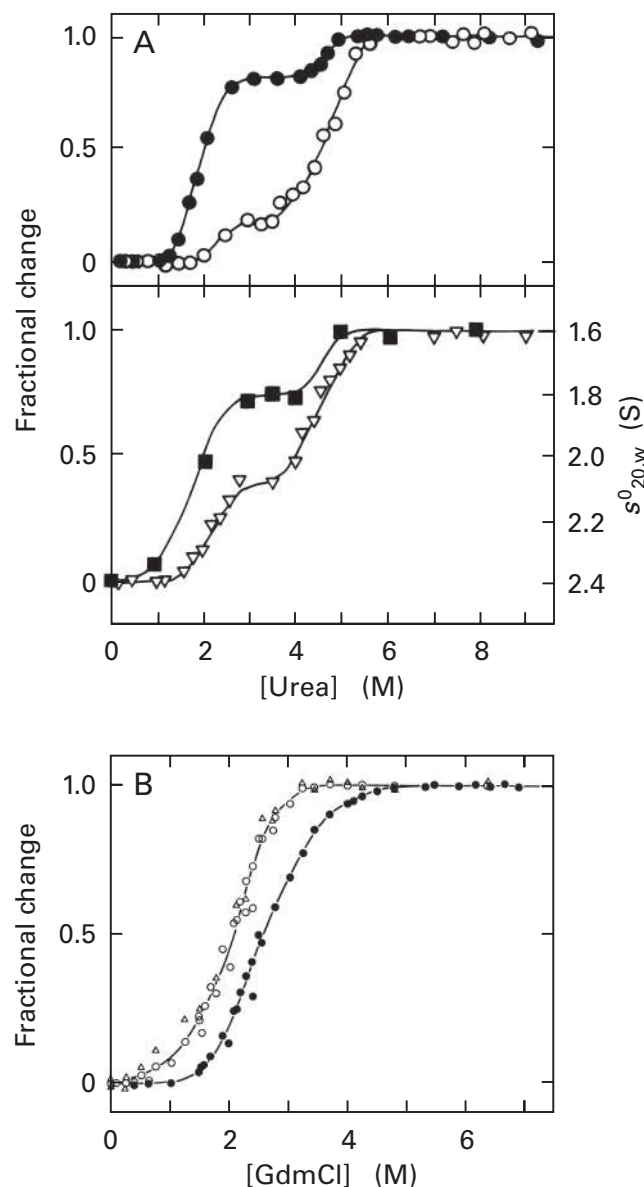


Figure 13-4: Evidence for the existence of stable intermediate states appearing during transitions between the native state and the random coil produced by increasing concentrations of urea or guanidinium chloride. (A) Shift of the equilibrium constant for the folding of bovine γ -crystallin B.⁶⁰ Solutions of the protein ($n_{aa} = 174$) in 0.1 M NaCl at pH 2.0 and 20 °C were diluted into solutions of urea at pH 2.0 and 20 °C so that the final concentration of urea (molar) would be as noted. Either the intrinsic fluorescence emission of the tryptophans of the protein at 360 nm (●; excitation at 280 nm, final concentration = 0.04 mg mL⁻¹), the intrinsic fluorescence of its tryptophans at 320 nm (○; excitation at 280 nm, final concentration = 0.04 mg mL⁻¹), its sedimentation velocity ($s_{20,w}^0$ ■; final concentration = 0.2 mg mL⁻¹), or its circular dichroism at 222 nm (▽; final concentration of protein = 0.1 mg mL⁻¹) were recorded as a function of the concentration of urea after each of the solutions was brought to equilibrium (24 h). Reprinted with permission from ref 60. Copyright 1990 held by authors. (B) Shift of the equilibrium constant for the folding of bovine α -lactalbumin.⁶⁶ Solutions of the protein ($n_{aa} = 123$) at pH 6.7 and 25 °C were diluted into solutions of guanidinium chloride at pH 6.7 and 25 °C so that the final concentration of guanidinium chloride (molar) would be as noted. The molar ellipticities of these solutions of protein were measured at 296 nm (△), 270 nm (○), and 222 nm (●) after each was brought to equilibrium. Reprinted with permission from ref 66. Copyright 1976 Academic Press. For the measurements presented in both panels, the direct results at equilibrium were plotted as in Figure 13-1 for each set of observations. Lines were drawn for the behavior of each physical property for fully native and fully unfolded protein as a function of the concentration of urea or guanidinium chloride, respectively, and the apparent fractional change of each measurement from the behavior of that physical property for the fully folded state were estimated from the positions of each data point relative to these lines. These estimated values of the fractional change are plotted in each panel as a function of the concentration (molar) of urea or guanidinium chloride (GdmCl).

by circular dichroism at 222 nm (●), the transition observed did not coincide with the one measured by circular dichroism at 270 and 296 nm (○, Δ).

The shift in the equilibrium constant for the folding of bovine carbonate dehydratase has been followed by changes in circular dichroism at 269 nm, ultraviolet absorption at 290 nm, and optical rotation at 400 nm as a function of the concentration of guanidinium chloride at pH 7.0.⁶⁷ The circular dichroism smoothly traced one transition and the optical rotation smoothly traced another transition proceeding at a higher concentration of guanidinium chloride. The change in absorbance traced a curve between the other two that displayed an inflection, suggesting that it was able to monitor both transitions. Furthermore, the kinetics of the refolding of the random coil was not a homogeneous, first-order process. It was concluded that one or more stable conformers of the polypeptide of carbonate dehydratase, other than the fully folded state and the random coil, are present in solutions of guanidinium chloride between 2 and 3 M in concentration. The properties of these other states are distinct from those of either the native state or the random coil. From observations of different transitions with different physical properties, the fractions of native, intermediate, and unfolded states as a function of the concentration of denaturant can also be estimated,⁶⁸ even in situations where very little intermediate forms,⁶⁹ and equilibrium constants among the states can be calculated.

If the protein being unfolded or refolded is an **oligomer** of two or more subunits, the respective dissociation or association of those subunits causes the transition between folded state and unfolded state to depend on the concentration of protein in the solution. Suppose that the native protein is an α_2 dimer. The equilibrium between the unfolded state α_U and the native state $(\alpha_F)_2$ is



and because

$$[\text{polypeptide}]_{\text{TOT}} = 2[(\alpha_F)_2] + [\alpha_U] \quad (13-15)$$

if there are only two states present, native dimer and unfolded monomer, at equilibrium

$$K_{\text{Fd}} = \frac{[(\alpha_F)_2]}{[\alpha_U]^2} = \frac{1 - f_U}{2f_U^2 [\text{polypeptide}]_{\text{TOT}}} \quad (13-16)$$

As the magnitude of the perturbation is increased, the equilibrium constant between folded and unfolded state is shifted in the direction of the unfolded state. At the midpoint of the transition between fully folded and

fully unfolded states that is being monitored by a particular physical property, $f_F = f_U$ (Equation 13-4); $f_U = 1/2$; and, instead of K_{Fd} being equal to 1 at this point, as with a monomer, K_{Fd} for a dimer is equal to $[\text{polypeptide}]_{\text{TOT}}^{-1}$. Consequently, as the total concentration of protein is increased, the equilibrium constant between unfolded and folded states must be shifted more and more before the midpoint is reached, which requires a greater and greater perturbation. If there are only the two states, as the total concentration of protein is increased, the midpoints of the curves describing the transition between folded dimer and unfolded monomer move systematically to higher and higher concentrations of guanidinium chloride^{63,70-72} or urea⁷³ or to higher temperatures.^{38,74} Because it is only the **molecularity of the reaction** that causes these shifts in the curves with the concentration of protein, they are no longer observed when the dimer is artificially converted to a monomer by joining the carboxy terminus of one of its subunits with the amino terminus of the other.^{74,75} Even greater shifts with the concentration of protein are observed in the curves following folding of higher oligomers⁷⁶ as a function of the perturbation.

Often when the equilibrium between folded oligomer and unfolded monomer is shifted by a perturbation, stable intermediate states are formed. For example, during the increase in the perturbation, a dimer may dissociate to monomers before the polypeptides unfold, and if the physical property detects only unfolding, the curves following the transition between the folded state and the unfolded state will not shift as the concentration of protein is increased³⁸ because the formation of the monomeric intermediate goes undetected. Similarly, a tetramer can dissociate into dimers before the dimers unfold to monomers.⁷² In some cases, the intermediate state is detected. In one such instance, the curves showed a plateau as in Figure 13-4A, but because both the native protein and the intermediate were dimers, it was only the portion of the curve monitoring the transition from the dimeric intermediate to the unfolded monomer that shifted with concentration of protein.⁷³

When the protein binds a **ligand**, the addition of the ligand also causes the curves following the transition between the native state and the unfolded state to shift to higher levels of perturbation. Because only the folded protein can bind the ligand, L, if the ligand is present at saturation so that only liganded native protein is present at all concentrations of denaturant



and

$$K'_{\text{Fd}} = \frac{[F \cdot L]}{[U][L]} = \frac{f_F}{f_U [L]} \quad (13-18)$$

Consequently, as the concentration of ligand is increased beyond its level of saturation, a greater perturbation is required to shift the equilibrium to a point at which $f_F = f_U$, and the curves move toward greater perturbation, for example to higher concentrations of guanidinium chloride, as the concentration of ligand is increased. It is also the case that, at the same concentrations, a ligand with a smaller dissociation constant will shift the curve a greater distance than one with a larger dissociation constant.⁷⁷

The change in **standard enthalpy of folding**, ΔH°_{Fd} , can be measured directly in a differential scanning calorimeter^{39,41} or it can be calculated from the dependence of the equilibrium constant of folding, K_{Fd} , on temperature. From the van't Hoff relationship

$$\left[\frac{\partial \ln K_{Fd}}{\partial \left(\frac{1}{T} \right)} \right]_P = - \frac{\Delta H^\circ_{Fd}}{R} \quad (13-19)$$

If a folding is followed as the temperature is varied and the value of the logarithm of the equilibrium constant for folding is plotted as a function of T^{-1} , the slope of the plot will be directly proportional to the change in standard enthalpy. When the folding of β -lactoglobulin was made reversible and kinetically first-order in both directions by adding appropriate concentrations of urea and adjusting the pH to 3, the equilibrium constant for folding, K_{Fd} , could be measured at each concentration of urea for temperatures between 10 and 50 °C. The behavior of $\log K_{Fd}$ as a function of T^{-1} (Figure 13-5A)⁷⁸ demonstrates that the change in standard enthalpy for the reaction, ΔH°_{Fd} , is not constant but varies considerably with temperature.

In fact, the values of the change in standard enthalpy for folding, ΔH°_{Fd} (Figure 13-5B),⁷⁸ calculated from the slopes of this first plot, vary from **exothermic to endothermic** over the range of temperatures sampled, a fact suggesting that the change in standard enthalpy for folding is by itself uninformative. Furthermore, the change in standard enthalpy for folding of a series of mutants of the same protein is usually linearly related to the change in standard entropy (Equation 5-63) with a slope T_c of about 350 K.^{79,80} Although the slope is somewhat greater than most other noncovalent processes, the compensation observed suggests that as with the hydrophobic effect both standard enthalpy and standard entropy are registering mainly compensatory changes in the water.

When the change in standard enthalpy, ΔH°_{Fd} , is plotted against the temperature (Figure 13-5B), the slope of the relationship observed at each point is the **standard heat capacity change of folding**, $\Delta C^\circ_{p,Fd}$, for that temperature

$$\left(\frac{\partial \Delta H^\circ_{Fd}}{\partial T} \right)_P = \Delta C^\circ_{p,Fd} \quad (13-20)$$

The change in standard heat capacity can also be measured directly in a calorimeter⁸¹ or by combining measurements of unfolding in solutions of urea and with temperature in a different manner.⁸²

From the observations presented in Figure 13-5B, it could be calculated that the change in standard heat capacity⁷⁸ for the folding of the polypeptide of β -lactoglobulin ($n_{aa} = 162$) between 5.5 and 4.4 M urea, pH 2.5 and 3.2, and 15 and 50 °C is $-8700 \pm 700 \text{ J K}^{-1} \text{ mol}^{-1}$, or $-54 \text{ J K}^{-1} (\text{mol of amino acid})^{-1}$. The measured values for the changes in standard heat capacity for the folding of proteins composed of a single polypeptide and lacking cystines are $-60 \pm 10 \text{ J K}^{-1} (\text{mol of amino acid})^{-1}$, regardless of the perturbation used to shift the equilibrium.^{78,79,82-87}

Unlike the changes in standard entropy and standard enthalpy that vary considerably from situation to situation, this uniform decrease in standard heat capacity seems to be a fundamental property of folding. It must arise from a combination of the decrease in heat capacity that occurs when **hydrophobic amino acids** are transferred from the aqueous phase into the interior of the molecule of protein,⁸⁸ the increase in heat capacity that arises from the **desolvation of polar amino acids** when they are transferred into the interior,⁸⁹ and the decrease in **conformational heat capacity** that occurs when vibrations and rotations along the polypeptide become more hindered after it is folded.⁸³ Of the three contributors, however, the difference in conformational heat capacity between the native state and the random coil may not be very significant because the observed heat capacity of an unfolded polypeptide is quite close to the heat capacity calculated only from the individual side chains and the individual peptide bonds composing that polypeptide.^{85,90,91}

The value for the change in standard heat capacity of folding is consistent with the decrease in standard heat capacity (-200 to $-400 \text{ J K}^{-1} \text{ mol}^{-1}$) observed for the transfer of alkanes and arenes from water into an organic phase (Table 5-8) if it is recalled that hydrophobic amino acids make up only a fraction (30%) of the amino acids in a polypeptide and that many of them remain accessible to water in the native state after the protein has folded. These considerations suggest that the uniform decrease in standard heat capacity [$-60 \text{ J K}^{-1} (\text{mol of amino acid})^{-1}$] associated with the folding of a polypeptide is one of the few signatures of the hydrophobic effect arising from the removal of hydrophobic amino acids from the solvent during their burial in the interior of the native state upon folding. The **hydrophobic effect** is the only noncovalent force that can provide a significant favorable contribution to the standard free energy of folding.

Although not always the case,⁷⁹ it has been pointed

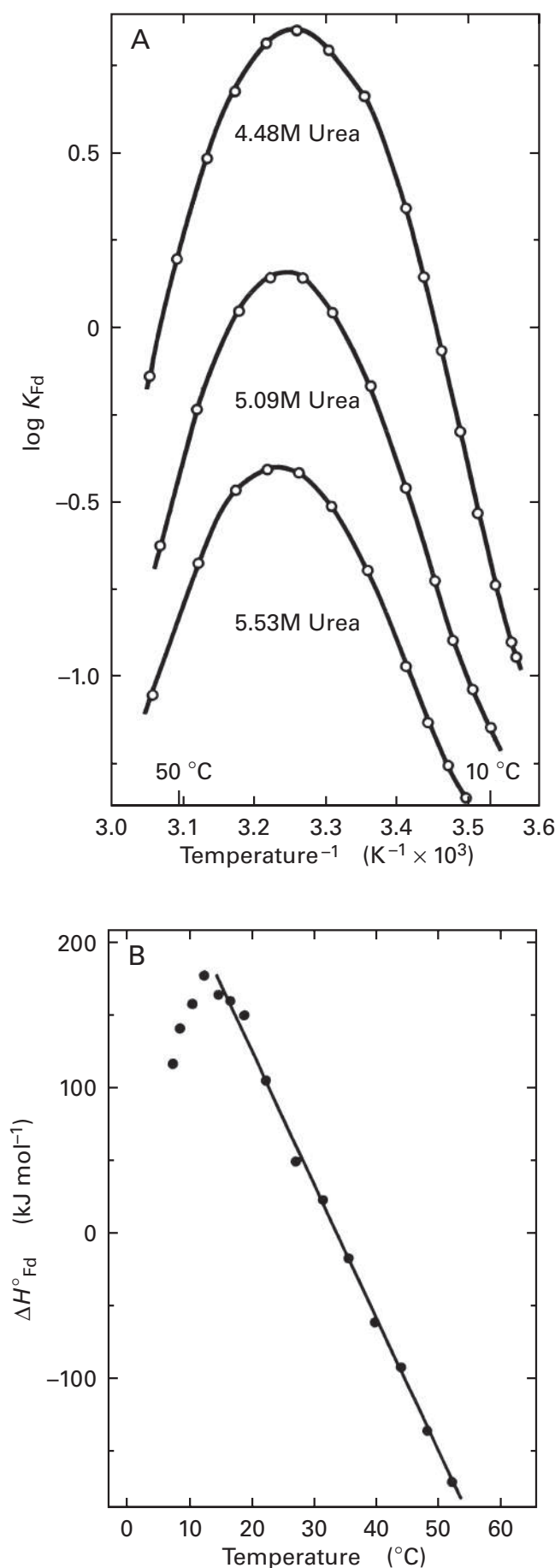


Figure 13-5: Determination of ΔH°_{Fd} and $\Delta C^\circ_{p,Fd}$ for the folding of β -lactoglobulin in solutions of urea between pH 2.5 and 3.5.⁷⁸ A series of measurements of the optical rotation at 365 nm of solutions of β -lactoglobulin as a function of the concentration of urea were made at various temperatures. From the lines extrapolated from the pre-transition and posttransition regions of these smooth curves, the optical rotation of the fully native state and the fully unfolded state, respectively, of β -lactoglobulin at any particular temperature and at any concentration of urea were estimated. Solutions were then prepared at 4.48, 5.09, and 5.53 M urea, concentrations at which equilibrium constants could be measured over the temperature range of 10–50 °C. For each of these solutions, from the optical rotation at a given temperature and the estimated optical rotations of the native state and the unfolded state at that temperature and concentration of urea, the equilibrium constant for folding, K_{Fd} , could be calculated (Equation 13-2 and 13-4). (A) Logarithms to the base 10 of K_{Fd} ($\log K_{Fd}$) plotted as a function of the inverse of the temperatures [$1/T$ (Kelvin⁻¹ $\times 10^3$)]. The plots are curves, from the slopes of which the standard enthalpies of folding, ΔH°_{Fd} , can be calculated (Equation 13-19) at any particular temperature. (B) Standard enthalpy of folding, ΔH°_{Fd} (kilojoules mole⁻¹), plotted as a function of the temperature (degrees Celsius). The slope of the line is the change in standard heat capacity, $\Delta C^\circ_{p,Fd}$, for folding. Reprinted with permission from ref 78. Copyright 1968 American Chemical Society.

out that there is a tendency for the characteristic change in standard heat capacity of folding to decrease in magnitude as the frequency of disulfides in the polypeptide increases.^{84,87} This effect is thought to arise from a reduction in the otherwise complete exposure of hydrophobic side chains to the water upon unfolding because of the inability of the cross-linked unfolded state to expand fully,⁹² but the presence of cystines in the unfolded state must also decrease its conformational heat capacity. This decrease in the magnitude of the change in standard heat capacity may also result from an evolutionary compensation because the unfolded state, losing standard configurational entropy by the introduction of the disulfides, requires less of a contribution from the hydrophobic effect to achieve the proper standard free energy of folding.

The fact that the changes in standard heat capacity of folding for all proteins are significant negative numbers dictates that the change in standard enthalpy of folding, ΔH°_{Fd} , must decrease significantly as the temperature is raised (Figure 13-5B) and must pass through a value of zero at some temperature. This causes the equilibrium constant for folding to pass through a maximum at that same temperature (Figure 13-5A). From these considerations it follows that if the negative change in standard heat capacity is an intrinsic property of folding, each protein must have a **characteristic temperature of maximum stability**.^{78,93,94} These temperatures of maximum stability vary from less than 0 °C to more than 35 °C. In the presence of moderate concentrations of urea⁹⁵ or guanidinium chloride⁹⁴ or at high pressure,⁹⁶ a protein that is fully folded at room temperature will often unfold as the temperature is decreased to 0 °C or below.

It is also possible to shift the equilibrium constant for folding in the direction of denaturation by applying pressure to a solution of protein.⁹⁶⁻⁹⁹ This observation

requires that the solvated denatured state have a smaller volume than the solvated native state because

$$\left(\frac{\partial \Delta G_{\text{Fd}}^{\circ}}{\partial P} \right)_T = \Delta V_{\text{Fd}}^{\circ} = V_{\text{F}} - V_{\text{U}} \quad (13-21)$$

where $\Delta G_{\text{Fd}}^{\circ}$ is the standard free energy of folding ($-RT \ln K_{\text{Fd}}$), $\Delta V_{\text{Fd}}^{\circ}$ is the **standard volume change of folding**, V_{F} is the molar volume of the folded state, and V_{U} is the molar volume of the unfolded state. The volume changes for folding calculated from these results are positive as expected. At pH 2.0 and 0 °C the volume change for folding of bovine ribonuclease A⁹⁷ is +48 cm³ mol⁻¹ and that for bovine chymotrypsinogen⁹⁸ is +14 cm³ mol⁻¹, while that for myoglobin⁹⁹ from pH 4 to 6 at 20 °C is +92 ± 5 cm³ mol⁻¹.

The changes in **isoentropic compressibility of folding**

$$\Delta \kappa_{\text{S,Fd}} = - \left[\frac{1}{V_{\text{F}}} \left(\frac{\partial V_{\text{F}}}{\partial P} \right)_S - \frac{1}{V_{\text{U}}} \left(\frac{\partial V_{\text{U}}}{\partial P} \right)_S \right] \quad (13-22)$$

are more informative. The isoentropic changes in compressibility of folding for ribonuclease A and chymotrypsinogen^{97,98} are both about -0.015 GPa⁻¹. The negative values for these isoentropic compressibilities indicate that the solvated denatured state is more compressible than the native state. This is not a surprising result because the isoentropic compressibilities of native proteins are very small, 10-fold smaller than those of organic liquids and 2-fold smaller than those of amorphous organic solids.¹⁰⁰ The greater compressibility of the denatured state is probably due in part to its more fluid structure, but it is also possible that the hydrophobic functional groups revealed in the denatured state increase the structure of the water surrounding them and thereby increase its compressibility. This increase in the structure of water, if it is significant, would resemble the increase in its structure caused by decreasing the temperature, and decreasing the temperature of liquid water increases its compressibility (Figure 5-5).

High pressures also are able to dissociate multimeric proteins into monomers, reversibly, without causing denaturation, even at neutral pH. The volume change is small; in the case of enolase at 10 °C and pH 7.4, $\Delta V^{\circ} = 0.025 \text{ cm}^3 (\text{mol of amino acid})^{-1}$. Presumably the individual volume changes occur only at the faces of the subunits that are exposed during the dissociation.¹⁰¹

To be able to calculate the equilibrium constant K_{Fd} for folding from the measured concentrations of the two states of the protein, it must be decreased significantly by one or a combination of rather unphysiological perturbations such as increasing the temperature or pressure, lowering the pH, or adding guanidinium chloride or urea to the solution. It would be of interest to be able to esti-

mate the value at pH 7 and 25 °C of this **equilibrium constant in the absence of any perturbation**. This is generally accomplished by extrapolation¹⁰² from realms of pH, temperature, pressure, and concentrations of urea and guanidinium chloride where measurements can be made.

Various equations have been derived for extrapolating values of the equilibrium constant K_{Fd} to small or zero concentrations of urea and guanidinium chloride^{103,104} and from acidic to neutral pH.¹⁰⁴ Most of these equations plot the observed **standard free energies of folding**, $\Delta G_{\text{Fd}}^{\circ}$, as functions of the magnitude of the perturbation to perform the extrapolations. It is also possible to perform nonlinear least-squares fits of empirical equations to plots of the directly observed changes of a physical property as a function of denaturant.¹⁰⁵ Unfortunately, each theoretical curve, although it is successful at reproducing the behavior in the measurable regions, deviates from the other theoretical curves beyond the measurable regions. The values for the standard free energy of folding measured both at elevated temperatures and in the presence of urea can be extrapolated simultaneously to obtain an estimate of the value for standard free energy of folding at 25 °C in the absence of urea.⁸⁶ Extrapolation both from high concentrations of guanidinium chloride and from low pH can also be performed simultaneously.¹⁰⁶ It is also possible to measure the thermal unfolding of a protein in a differential scanning calorimeter at a series of concentrations of urea below the range of concentrations at which the transition is observed at 25 °C.

The **extrapolation** that has become most widely accepted is one for values of standard free energies of folding, $\Delta G_{\text{Fd}}^{\circ}$, observed in solutions of guanidinium chloride or urea, and the equation for performing this extrapolation that has emerged as the most popular is¹⁰⁷

$$\Delta G_{\text{Fd,[D]}}^{\circ} = \Delta G_{\text{Fd,H}_2\text{O}}^{\circ} - m [\text{D}] \quad (13-23)$$

where $\Delta G_{\text{Fd,[D]}}^{\circ}$ is the standard free energy of folding calculated (Equation 5-14) from the observed equilibrium constant at a given concentration of the denaturant; $\Delta G_{\text{Fd,H}_2\text{O}}^{\circ}$ is the standard free energy of folding for the protein in aqueous solution at the same pH, ionic strength, and temperature; and m is the slope of a line that is fit to the observations. This equation states that the standard free energy of folding is a simple linear function of the concentration of denaturant, which seems to ignore the observation that the changes in solvation brought about by guanidinium chloride and urea (Table 13-1) are not directly proportional to their molar concentrations.^{8,12,25,103,108}

Nevertheless, there are many observations supporting the **validity** of this relationship. In situations where there are **wide ranges of the concentration** of denaturant over which the equilibrium constant for folding can be

measured accurately (Figure 13–6),^{109,110} the standard free energies of folding do in fact vary linearly with the concentration of denaturant over the entire range of measurements.* Most of the time, however, the range of concentrations of denaturant over which measurements of the equilibrium constants for folding can be made is much narrower (Figure 13–7).¹⁰⁵ Extrapolations of standard free energies of folding perturbed by **different denaturants** give the same value for $\Delta G^\circ_{\text{Fd,H}_2\text{O}}$, in spite of the long distances over which those extrapolations must be made.^{105,107,111} Standard free energies of folding beyond the range of denaturant concentrations in which they can be directly measured can be estimated from measurements of unfolding induced by raising the temperature in a differential scanning calorimeter, and these estimates usually fall close to the line of extrapolation.¹¹¹ Measurements of the first-order **rate constants** for the approach to equilibrium can be made for the entire range of concentrations of denaturant, and from a plot of these observed rate con-

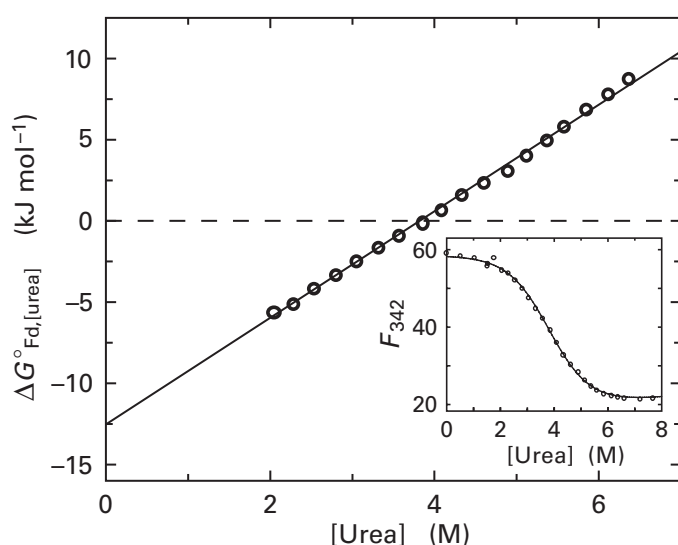


Figure 13–6: Estimation of the standard free energy of folding in the absence of denaturant by extrapolation. Cold shock protein CspB ($n_{\text{aa}} = 67$) from *Bacillus subtilis*¹⁰⁹ was dissolved in a series of solutions of increasing molar concentrations of urea at pH 7 and 25 °C. Two final concentrations of protein, 1.35 μM and 13.5 μM , were used. The emission of fluorescence of each solution, F_{342} , was monitored at 342 nm upon excitation at 280 nm (inset). After equilibrium was reached, the equilibrium constant for folding was estimated for each concentration of urea, and from these equilibrium constants, the respective standard free energies of folding (kilojoules mole⁻¹) were calculated. These standard free energies of folding (kilojoules mole⁻¹) are plotted as a function of the concentration of urea (molar). A line was fit to the data by linear least-squares analysis. The dashed line at zero (equilibrium constant of 1) emphasizes that direct measurements of the equilibrium constant can usually be made only over a limited range. Reprinted with permission from ref 109. Copyright 1995 Nature Publishing Group.

* The ionic strength of the solution must be maintained as the concentration of guanidinium chloride is decreased to retain linear behavior of $\Delta G^\circ_{\text{Fd,[GdmCl]}}$.¹¹¹

stants, values for k_{F} and k_{U} in the absence of denaturant can be estimated by extrapolation (dashed lines in Figure 13–1B). If the folding is a two-state process, the equilibrium constant for folding calculated from the estimates of these two rate constants (Equation 13–2) gives a value for the standard free energy of folding that agrees^{50,112} with that obtained by use of Equation 13–23.

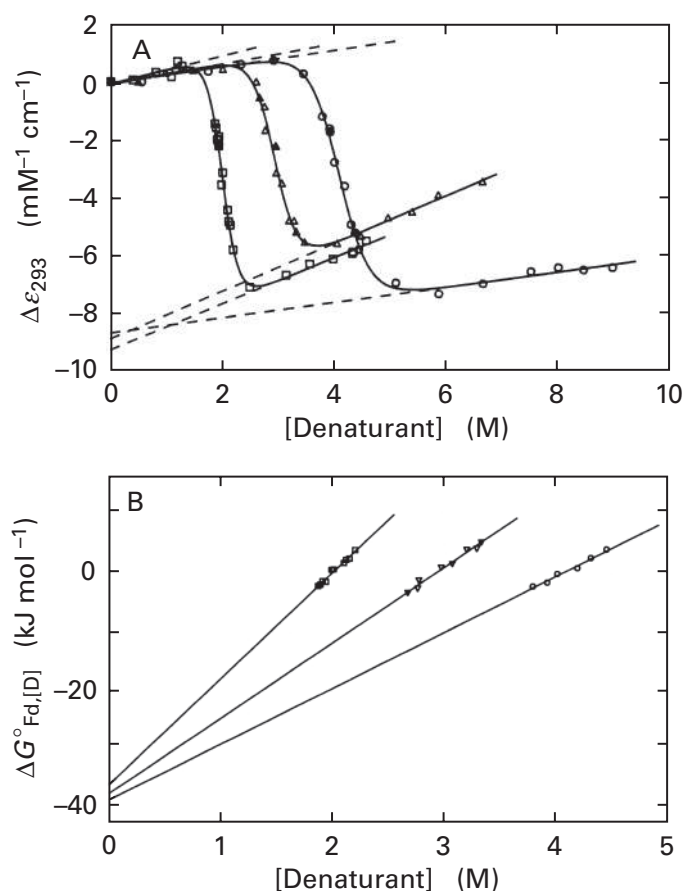


Figure 13–7: Estimation of the standard free energy of folding by extrapolating standard free energies of folding observed in solutions of different denaturants.¹⁰⁵ (A) Shifts of the equilibrium constants of folding. Bovine chymotrypsin ($n_{\text{aa}} = 241$), which had been sulfonated with phenylmethanesulfonyl fluoride to inactivate the endopeptidase, was dissolved in solutions of different concentrations of urea (\circ), 1,3-dimethylurea (Δ), and guanidinium chloride (\square) at pH 4.0 and 25 °C. The several folding isomerizations were monitored by the change in extinction coefficient at 293 nm ($\Delta \epsilon_{293}$; millimolar⁻¹ centimeter⁻¹), which is plotted as a function of the concentration (molar) of urea. In the respective regions of transition, the fraction of the protein in the folded state and the fraction in the unfolded state were calculated from the distance of each data point from the values of the change in absorbance for the fully folded (upper dashed lines) and the fully unfolded (lower dashed lines) states. Equilibrium constants for folding were calculated from these fractions for each concentration of denaturant, and from each of these equilibrium constants, standard free energies of folding $\Delta G^\circ_{\text{Fd,[D]}}$ were calculated at the respective concentrations of denaturant. (B) Standard free energies of folding (kilojoules mole⁻¹) plotted against the respective concentrations (molar) of each denaturant. Each of the lines was fit to the respective set of data by linear least-squares analysis. Reprinted with permission from ref 105. Copyright 1988 American Chemical Society.

That the numerical value for $\Delta G_{\text{Fd,H}_2\text{O}}^\circ$, obtained by use of Equation 13–23 is a reasonable estimate of its actual value can also be demonstrated by evaluating the **effect of pH** on its magnitude. Acid–base titration curves for the folded state of the protein and its unfolded state in 8 M urea or 6 M guanidinium chloride can be measured directly^{23,24} or calculated from its composition of amino acids,²⁵ and an integrated form of Equation 13–3³³ can be used to calculate the variation expected in $\Delta G_{\text{Fd,H}_2\text{O}}^\circ$ caused by changes in pH. These calculated variations reproduced the observed variations with pH of estimates of $\Delta G_{\text{Fd,H}_2\text{O}}^\circ$ obtained by the extrapolation defined by Equation 13–23 for bovine ribonuclease A^{23,25} and bovine chymotrypsin²⁴ when they were unfolded in solutions of urea and guanidinium chloride.

The observed changes in the dependence of $\Delta G_{\text{Fd,H}_2\text{O}}^\circ$ on pH caused by **site-directed mutation** of a particular amino acid in the protein also agree quantitatively with those calculated with integrated forms of Equation 13–3. The observed values of $\Delta G_{\text{Fd,H}_2\text{O}}^\circ$ between pH 5 and 8 for the carboxy-terminal domain of protein L9 from the 50S subunit of the ribosome of *E. coli* fell on the curve calculated with an integrated form of Equation 13–3 by use of the values of $\text{p}K_a$ for its four histidines in the native state as determined directly by nuclear magnetic resonance.¹¹³ When each of these histidines was mutated in turn, the observed changes in the behavior of $\Delta G_{\text{Fd,H}_2\text{O}}^\circ$ were again those predicted from their individual values of $\text{p}K_a$. In particular, the mutation of Histidine 134, which is buried in the interior and has the lowest $\text{p}K_a$ of the histidines in the native protein, caused the greatest change in the observed pH dependence of $\Delta G_{\text{Fd,H}_2\text{O}}^\circ$. Aspartate 26 is buried in the native state of thioredoxin from *E. coli*, which causes its $\text{p}K_a$ to be 7.5, a fact that destabilizes the folded state relative to the unfolded (Equation 13–3). The destabilization calculated from the difference in the values of $\text{p}K_a$ for just Aspartate 26 is equal to the difference in the values of $\Delta G_{\text{Fd,H}_2\text{O}}^\circ$ estimated by Equation 13–23 for the wild-type protein and a mutant in which Aspartate 26 is replaced by alanine.²⁷ Differences in $\Delta G_{\text{Fd,H}_2\text{O}}^\circ$ calculated from observed shifts in the values of $\text{p}K_a$ for histidines in ribonuclease T₁ from *A. oryzae* caused by mutation of Glutamate 58 to alanine also agreed with differences in values of $\Delta G_{\text{Fd,H}_2\text{O}}^\circ$ estimated with the extrapolation of Equation 13–23.¹¹⁴

The constant fragment C_L of the light chain of immunoglobulin G (Figure 11–1) is a small protein, the folding of which as a function of the concentration of guanidinium chloride has been measured.¹¹⁵ The protein contains a deeply buried cystine that is readily reduced by dithiothreitol when it is unfolded in solutions of guanidinium chloride. The rate of its reduction in the random coil in the absence of guanidinium chloride can be estimated by extrapolation of its rate of reduction at higher concentrations of guanidinium chloride. The actual rate of its reduction in the absence of guanidinium chloride when the protein is folded is much slower than

the extrapolated value. If it is assumed that the reduction of this cystine in the absence of guanidinium chloride occurs only when the native protein is briefly and reversibly a random coil, the difference between the rate of reduction for the native state and that estimated for the random coil is consistent with a value for the standard free energy of folding of -26 kJ mol^{-1} . This is close to the value (-30 kJ mol^{-1}) obtained by extrapolation from ranges of guanidinium chloride concentrations in which the equilibrium constant can be measured.

Measurements of **proton exchange** also support an extrapolation of standard free energy of folding that is linear in the concentration of denaturant. The amide protons of peptide bonds buried deeply in the interior of a protein, when they exchange at the EX₂ limit (Equation 12–63), often register a conformational change that is the global unfolding and folding of the protein.¹¹⁶ Consequently, in these situations, K_{conf} (Equation 12–63)* is actually K_{Fd}^{-1} . When the standard free energy of this conformational change revealed by proton exchange, $\Delta G_{\text{HX}}^\circ$, is monitored, it is found to be a linear function of the concentration of the denaturant (Figure 13–8).^{116,117} In the case of cysteineless type I ribonuclease H from *E. coli*, only Methionine 47 is buried deeply enough to respond only to the global unfolding and folding over the range of rates that could be measured, but the change in standard free energy of the conformational change that it monitors remains a linear function of the concentration of guanidinium chloride to concentrations well below those at which global folding can be monitored directly (inset in Figure 13–8). The range of values for proton exchange that can be measured for deeply buried amide hydrogens can be extended by raising the temperature, and at higher temperature, they remain linear functions of the concentration of guanidinium chloride until none is left in the solution.¹¹⁶ Furthermore, the standard free energies of folding in water, $\Delta G_{\text{Fd,H}_2\text{O}}^\circ$, estimated from these plots of $\Delta G_{\text{HX}}^\circ$ as a function of the concentration of denaturant, agree satisfactorily with values of standard free energies of folding in water estimated from linear extrapolations of standard free energies of folding calculated from direct measurements of the equilibrium constant for folding at higher concentrations of denaturants (inset to Figure 13–8).¹¹⁸

There are several observations, however, suggesting that the standard free energy of folding of at least some proteins that fold in a two-state process may not be a linear function of the concentration of denaturant all

* By convention, the equilibrium constant for the conformational change producing exchange is defined for the opening of the structure, while the equilibrium constant for folding is defined reciprocally, namely, for the closing of the structure. Consequently, $K_{\text{Fd}}^{-1} = K_{\text{conf}}$ and $\Delta G_{\text{Fd}}^\circ = -\Delta G_{\text{HX}}^\circ$ when the conformational change being monitored by the exchange is the global unfolding and folding of the protein.

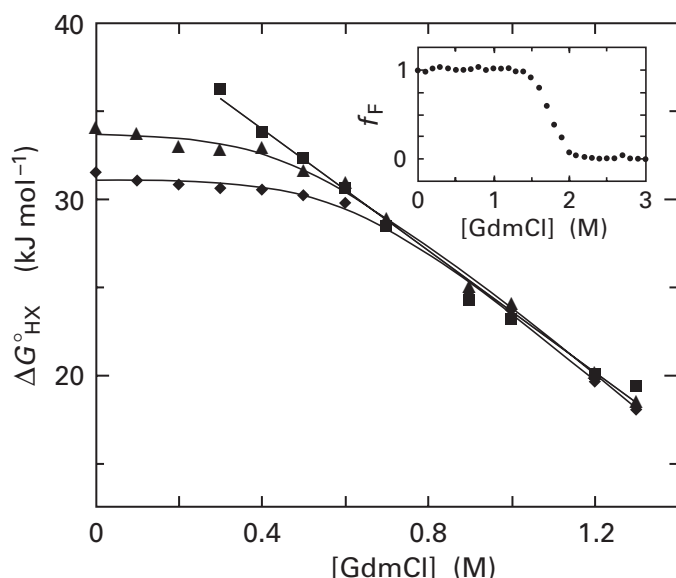


Figure 13-8: Standard free energies of unfolding estimated from the rates of exchange for particular amido protons in cysteineless type 1 ribonuclease H from *Escherichia coli*.¹¹⁷ All three of the cysteines in ribonuclease H were mutated to alanines. The resulting cysteineless protein was dissolved at a concentration of 1 mM in a series of solutions of deuterated guanidinium chloride prepared in deuterium oxide at p²H 5.1 and 25 °C. The rates at which 53 of the amido protons in the polypeptide backbone of the protein exchanged with deuterons from the solvent were slow enough to be monitored by two-dimensional nuclear magnetic resonance spectroscopy (Figure 12-33). Observed rate constants were calculated by fitting the amplitudes of the peaks as a function of time to single exponential decays by nonlinear least-squares analysis. Each proton exchanged in the EX₂ limit, and from the observed rate constant of its exchange, an equilibrium constant K_{conf} for the conformational change that exposed it to solvent was calculated (Equation 12-63). It was assumed that for the most deeply buried positions showing the smallest values of K_{conf} , the value of K_{conf} would actually be the equilibrium constant for global unfolding (the reciprocal of the equilibrium constant for folding). It was concluded that the exchange rates of Methionine 47 (■), Glutamine 105 (▲), and Alanine 110 (◆) were monitoring this global unfolding equilibrium of the protein. From the values of K_{conf} for each of these three exchanges, the respective free energies for the conformational change exposing the protons for exchange, $\Delta G^{\circ}_{\text{HX}}$, were calculated. These standard free energies of exposure (kilojoules mole⁻¹) are plotted as a function of the concentration (molar) of guanidinium chloride (GdmCl). The inset presents the fraction of the protein that is folded (f_F) as a function of the concentration of guanidinium chloride (molar), as monitored by circular dichroism at 220 nm. Reprinted with permission from ref 117. Copyright 1996 Nature Publishing Group.

the way to the point at which none remains. Values for the changes in standard heat capacity of folding ($\Delta C^{\circ}_{p,\text{Fd}}$), the standard enthalpies of folding ($\Delta H^{\circ}_{\text{Fd}}$), and the temperatures at which the concentrations of folded and unfolded states are equal (the melting temperatures, T_m) were measured for ribonuclease from *Bacillus amyloliquifaciens*⁸⁶ at concentrations of urea below those at which the equilibrium constant for folding could be measured at 25 °C. From these thermodynamic parameters, the standard free energy of folding at 25 °C and at

each of these concentrations of urea could be calculated with the relationship

$$\Delta G^{\circ}_{\text{Fd}} = \Delta H^{\circ}_{\text{Fd},m} + \Delta C^{\circ}_{p,\text{Fd}} (T - T_m) - \Delta H^{\circ}_{\text{Fd},m} \left(\frac{T}{T_m} \right) - T \Delta C^{\circ}_{p,\text{Fd}} \ln \left[\left(\frac{T}{T_m} \right) \right] \quad (13-24)$$

where T is the temperature (kelvins) for which these calculations are to be made (in this case, 298 K) and $\Delta H^{\circ}_{\text{Fd},m}$ is the standard enthalpy of folding at T_m , which can be measured in the calorimeter. The change in standard heat capacity of folding $\Delta C^{\circ}_{p,\text{Fd}}$ was estimated from the temperature dependence of $\Delta H^{\circ}_{\text{Fd}}$ (Equation 13-20). When these calculated values of $\Delta G^{\circ}_{\text{Fd,[urea]}}$ for concentrations of urea below those for the region of transition at 25 °C were plotted, they deviated from the line fit to the values of $\Delta G^{\circ}_{\text{Fd,[urea]}}$ measured directly within the region of transition at 25 °C.

The deviation was a gradual curvature sending the plot of the actual standard free energy of folding below the linear extrapolation. A similar downward curvature has been directly observed for the standard free energy of folding for helical peptides as a function of the molar concentration of urea in the range from 4 M down to 0 M.¹¹⁹ If these thermodynamic calculations and direct observations are correct and applicable to folding in general, then the linear extrapolations normally performed consistently **underestimate the magnitudes of the standard free energy** of folding at 25 °C in the absence of denaturant. In the case of ribonuclease from *B. amyloliquifaciens*, the actual value for the standard free energy of folding would be about 15% more negative than that obtained by extrapolation,⁸⁶ but a somewhat larger underestimate has been reported for the magnitude of the standard free energy of folding for lysozyme from *G. gallus* by a similar approach.¹²⁰ A drawback of the heavy reliance on measurements of thermal unfolding in these experiments, however, is that the thermally denatured state of a protein is not a random coil.

Another approach, which is somewhat simpler to accomplish than extrapolating a set of measurements at different concentrations of denaturant, is widely used when the standard free energies of folding for mutants are compared to the standard free energy of folding of the unmutated protein.^{121,122} The equilibrium constant for folding is measured over the region of transition at higher temperatures in which the concentration of denatured state becomes significant (Figure 13-2). With the van't Hoff relationship (Equation 13-19), the standard enthalpies of folding are estimated for both wild type and the mutants within this range of temperatures. The melting temperatures T_m for wild type and mutants are the temperatures at the midpoint of the thermal transition; the change in standard heat capacity of folding $\Delta C^{\circ}_{p,\text{Fd}}$ is

assumed to be the same for wild type and mutants, a value that has been either directly measured or estimated from $60 \text{ J K}^{-1} (\text{mol aa})^{-1}$; and the differences in standard free energy of folding, $\Delta\Delta G^\circ_{\text{Fd}}$, between mutants and wild type at a particular temperature within the range of the measurements are calculated with Equation 13–24.

One might assume that the relative stabilities of a series of mutants of a protein could be estimated from the concentrations of guanidinium chloride required to shift the equilibrium constants for their folding to a value of 1. This is not the case, however, because mutants requiring greater concentrations of guanidinium chloride to shift their equilibrium constants can have less negative standard free energies of folding in the absence of guanidinium chloride.¹²³

Representative values for extrapolated standard free energy changes of folding* under physiologically relevant conditions in the absence of denaturants at 25 °C have been assembled in Table 13–2. Each of the several values for a particular protein is the result of a different extrapolation, often from the same experimental data. The remarkable feature of this tabulation is that the standard free energies of folding for at least these 12 proteins are similar and fall between -20 and -60 kJ mol^{-1} . These are not large changes of standard free energy when the magnitude and the number of the individual noncovalent interactions involved in the process are considered. They are clearly the **sums of a large number of positive and negative terms** that cancel each other to produce small negative numbers. That they are all negative is merely the result of evolution by natural selection and consequently uninformative. The small magnitudes of these values may also be a consequence of evolution by natural selection. It is possible, if one is lucky, to increase the stability of proteins by site-directed mutation,¹²⁹ and proteins from hyperthermophilic organisms are generally more stable than those from organisms adapted to normally encountered temperatures¹³⁰ so the opportunity to evolve more stable proteins must exist, but it is usually not exploited.

With these values for the standard free energy changes (Table 13–2), the values for the equilibrium constants for the folding of these native proteins at 25 °C should be between 10^4 and 10^{10} . If this is the case, 10^{-4} to 10^{-10} of the lifetime of each of these proteins is spent in the fully unfolded state under normal circumstances.

* The standard state for free energy of folding has not been well defined. Because most foldings that have been studied are intramolecular isomerizations, their equilibrium constants should be independent of their concentration; and either infinite dilution or a corrected volume fraction of 1 (Equations 5–13 and 5–14) could be chosen as the standard state with little effect on the values for the standard free energies. The foldings of oligomers, however, are concentration-dependent and require that a finite concentration be chosen for standard state.

Table 13–2: Standard Free Energies of Folding in the Absence of Guanidinium Chloride or Urea, $\Delta G^\circ_{\text{Fd,H}_2\text{O}}$, at 25 °C

protein	pH	perturbation extrapolated	$\Delta G^\circ_{\text{Fd,H}_2\text{O}}$ (kJ mol ⁻¹)
ribonuclease A	6.0	GdmCl ¹²⁴	-50
<i>Bos taurus</i>	6.0	GdmCl ¹²⁴	-40
	6.6	urea ^{103,107}	-30
	6.6	GdmCl ^{103,107}	-40
lysozyme	6.0	GdmCl ¹²⁴	-60
<i>G. gallus</i>	6.0	GdmCl ¹²⁴	-50
	7.0	GdmCl, pH ¹²⁵	-50
	7.0	GdmCl, pH ¹²⁵	-80
α -chymotrypsin	4.3	urea ^{103,107}	-30
<i>B. taurus</i>	4.3	GdmCl ^{103,107}	-30
	4.3	GdmCl ¹⁰³	-50
	4.3	GdmCl ¹⁰⁷	-40
[(phenylmethyl)sulfonyl]- α -chymotrypsin	4.0	GdmCl ¹⁰⁵	-40
<i>B. taurus</i>	4.0	urea ¹⁰⁵	-40
	4.0	1,3-dimethylurea ¹⁰⁵	-40
	6.0	urea, GdmCl ²⁴	-50
myoglobin	7.0	GdmCl, pH ¹⁰⁶	-50
<i>Equus caballus</i>	6.0	GdmCl ¹²⁶	-40
	6.0	GdmCl ¹²⁶	-50
cytochrome <i>c</i>	6.5	GdmCl ¹²⁷	-50
<i>E. caballus</i>			
ribonuclease T ₁	7.0	urea ¹¹⁴	-30
<i>A. oryzae</i>			
ribonuclease	6.3	urea ⁸⁶	-40
<i>B. amyloliquifaciens</i>			
chymotrypsin inhibitor 2	6.0	GdmCl ¹²⁸	-30
<i>Hordeum vulgare</i>			
phosphocarrier protein HPr	6.0	urea ¹¹⁸	-20
<i>E. coli</i>			
ribonuclease H	6.0	GdmCl ¹¹⁸	-60
<i>Thermus thermophilus</i>			
thioredoxin	7.0	GdmCl ¹²³	-25
<i>E. coli</i>			

The protection factors for the proton exchange of deeply buried amide hydrogens are usually 10^6 or greater, and these most deeply buried amide protons are probably exchanged only during the brief periods when the native state has become reversibly and completely unfolded. Natural selection has settled on an equilibrium constant large enough to limit the time the protein spends in the unfolded state in part to protect it from degradation by endopeptidolytic enzymes.¹²⁹

The requirement that a protein must **unfold and refold during its lifetime** may be viewed as a consequence of the need to fold in the first place (Equation 13–1) and the inescapable dictates of microscopic reversibility. If the observed rate constant k_F for the spontaneous refolding of a recently unfolded polypep-

time¹³¹ is on the order of 10 s^{-1} at $25\text{ }^{\circ}\text{C}$ and the equilibrium constant K_{Fd} for folding is on the order of 10^8 , then the observed rate constant for the unfolding of a native protein to the random coil ($k_{\text{U}} = k_{\text{F}}/K_{\text{Fd}}$) must be on the order of 10^{-7} s^{-1} . This would state that a protein has a 50% chance of unfolding to the random coil every 100 days at $25\text{ }^{\circ}\text{C}$. This is not a major problem in the life of a protein.

Measurements of the equilibrium constants K_{conf} for the conformational changes permitting the exchange of amide protons at the EX₂ limit (Equation 12-63) are consistent with the proposal that the most deeply buried positions in the polypeptide backbone exchange only upon its complete unfolding (Figure 13-8). The conformational equilibrium constants governing the exchange rates for the less deeply buried amide protons, however, are larger than the one for the most deeply buried positions and are spread over a range of values.¹³²⁻¹³⁵ The larger values for these other conformational equilibrium constants, which produce faster rates of exchange, are the result of conformational changes confined only to portions of structure of the protein, for example to individual α helices or loops of random meander,^{116,132,136,137} rather than the result of the fundamental unfolding encompassing the entire structure.

These **local conformational changes** appear to be of two types: those involving considerable exposure of nonpolar functional groups to the solvent, similar to the exposure experienced during global unfolding, and those involving exposure of the amide protons to be exchanged without any significant expansion of the local structure into the solvent.¹³⁸ The former are recognized by the increase in their equilibrium constants produced by adding guanidinium chloride or urea; the latter, by the insensitivity of their equilibrium constants to the addition of these denaturants.^{116,132,137} For example, above 0.8 M guanidinium chloride, the α amide protons on Glutamine 105 and Alanine 110 of cysteineless type I ribonuclease H (Figure 13-8) must exchange during a major unfolding of the protein because the standard free energy for the conformational change permitting their exchange decreases significantly as the concentration of guanidinium chloride increases. At lower concentrations of guanidinium chloride, however, their respective rates of exchange are governed by other conformational changes the equilibrium constants for which are unaffected by the concentration of guanidinium chloride. These other conformational changes, therefore, must not involve significant increases in the exposure of the polypeptide to the solvent. Because they are unaffected by the concentration of guanidinium chloride, the equilibrium constants for these other conformational changes become larger than the equilibrium constant for the major unfolding at low concentrations of the denaturant.

These observations indicate that in solution, in its native state, the structure of a protein is **constantly fluctuating** as a result of conformational changes of various

extents, occurring in different locations, some involving isomerizations retaining compact globular structure, others involving large, rapid expansions into the solvent followed by a collapse back into the native state.¹³⁹

Because a polypeptide can fold in the first place and because it must refold in part or in its entirety during the span of its life, the **information** dictating the final native state of the protein must be contained within its amino acid sequence. Because the standard free energy of folding of most proteins is not a large negative number (Table 13-2), perhaps for cause, if some of the information is lost or misinformation is added, the protein will not fold. For example, whenever the sequence of a protein is changed by site-directed mutation, the possibility exists that the mutant will not fold, for reasons that will never be learned. Many site-directed mutations, however, have little effect on the ability of the protein to fold, and in a few instances, a site-directed mutation has been found to increase the stability of a protein. For example, when the amino acids at positions 40-49 of lysozyme from bacteriophage T4 were all replaced with alanines,¹⁴⁰ its standard free energy of folding increased by only 10 kJ mol^{-1} , while the appropriate replacement of five of the amino acids in type I ribonuclease H from *E. coli*¹²⁹ decreased its standard free energy of folding by 20 kJ mol^{-1} .

Incomplete polypeptides often lack sufficient information to fold properly. A form of the polypeptide of bovine ribonuclease A ($n_{\text{aa}} = 124$) that is missing the last six amino acids is unable to produce a folded protein with enzymatic activity, and what structure it does have at $20\text{ }^{\circ}\text{C}$ is eliminated by heating to only $40\text{ }^{\circ}\text{C}$ at pH 7.5 in the absence of denaturants.¹⁴¹ This truncated polypeptide is also susceptible to endopeptidolytic degradation, unlike the intact native protein. When the last 23 amino acids, which form only a small number of contacts with the bulk of the folded polypeptide in its crystallographic molecular model, are removed from the polypeptide of micrococcal nuclease ($n_{\text{aa}} = 149$), the polypeptide produced is a random coil by the criteria of circular dichroism, optical rotation, and ultraviolet absorption.¹⁴² It is also readily digested by trypsin, unlike the native enzyme. Its residual enzymatic activity of 0.1%, which is an intrinsic property of the shortened polypeptide,¹⁴³ suggests that it can still fold properly to form an active enzyme but that the equilibrium constant for folding is displaced heavily ($K_{\text{Fd}} \leq 10^{-3}$) in the direction of the random coil. When the first 12 amino acids and the last 9 amino acids are removed from the protein, it folds partially to form a state in which some of its normal secondary structures are formed but in low yield.^{144,145}

Another set of examples of the fact that a polypeptide can fold only when all the necessary information is present is proteins that are posttranslationally modified during their natural maturation. In many instances, the polypeptide that folds to produce the native state is

longer than the final product because the initial folded form is clipped, and the smaller piece or pieces resulting from the **posttranslational clipping of the polypeptide** dissociate.¹⁴⁶ For example, subtilisin E from *Bacillus subtilis* folds naturally when it is a polypeptide 352 amino acids in length. After it folds, it is posttranslationally modified. During this process, the peptide bond after Tyrosine 77 is cleaved, and the first 77 amino acids of the polypeptide, the prosequence, are lost. If the mature, enzymatically active form of the protein ($n_{aa} = 275$) is unfolded, it will not refold; but if the full-length polypeptide ($n_{aa} = 352$) is unfolded in 6 M guanidinium chloride, it readily refolds to produce the native state.¹⁴⁷ If the prosequence ($n_{aa} = 77$) is included in the solution when the mature protein is being refolded, considerable native state is recovered.¹⁴⁸ The yield of enzymatic activity is low but increases as the concentration of prosequence is increased up to a molar excess of 4-fold.¹⁴⁹ Only when the complete amino acid sequence of the longer polypeptide is intact, however, is there sufficient information to produce a high yield of the mature form. Once folded and posttranslationally modified, the mature protein is stable and biologically competent, as long as it is not unfolded. In the case of carboxypeptidase C from *Saccharomyces cerevisiae*, however, the mature posttranslationally modified form of the protein ($n_{aa} = 421$) is considerably more resistant to the effects of guanidinium chloride than is its intact precursor ($n_{aa} = 512$), a result suggesting that the prosequence provides information rather than standard free energy for folding.¹⁵⁰

There are many other examples of proteins that lose portions of their polypeptide, usually from the amino terminus, after they have folded. This is so common that the term **proprotein** is used to designate the longer polypeptide that folds, with the implication that the cleaved, mature native state is designated as the protein. Familiar examples of this designation are proinsulin, proalbumin, and prothrombin.

Fragments of a polypeptide, each lacking sufficient information to fold separately, can sometimes cooperate to produce the proper native state. The first example of this was the ability of the amino-terminal fragment of ribonuclease (Lysine 1–Alanine 20), which is almost structureless in isolation,¹⁵¹ to reassume its native structure as an α helix when combined with the remainder of the polypeptide (Serine 21–Valine 124).¹⁵² Both the fragment Alanine 1–Arginine 126 and the fragment Glycine 49–Glutamine 149 of micrococcal nuclease ($n_{aa} = 149$) are structureless in isolation.^{142,153} When they are mixed together, however, they combine with each other to form two different forms of the native state that both appear to be properly folded but together have only 10% of the nuclease activity of the native enzyme.¹⁵³

Higher yields of enzymatic activity have been observed upon **combination of fragments** of ribonuclease from *B. amyloliquifaciens* (fragments of 36 and 74 aa; yield 30%),¹⁵⁴ fragments of phosphoribosylanthranilate

isomerase from *S. cerevisiae* (fragments of 174 and 59 aa; yield 50%),¹⁵⁵ fragments of penicillin amidase from *E. coli* (fragments of 209 and 557 aa; yield 60%),¹⁵⁶ and fragments of porcine 3-oxoacid CoA-transferase (fragments of 250 and 270 aa; yield 85%).¹⁵⁷ In the case of the ribonuclease, each fragment was a random coil in the absence of the other, but in the cases of the isomerase and the amidase, one of the two fragments refolded on its own to form a compact structure. The two fragments of the transferase that were chosen for expression are structural domains in its crystallographic molecular model, and both formed compact structures in the absence of the other, but neither formed the structure it has in the intact protein. None of the fragments from any of these proteins had enzymatic activity on its own, and it was only upon mixing the two respective fragments that activity was regained.

When a protein is split into two fragments and the separated, incompetent fragments are mixed together in the hope of regenerating the native state of the protein, the situation is complicated by the fact that the fragments must associate with each other. For example, the complex of the fragments of ribonuclease from *B. amyloliquifaciens* and the complex of the fragments of phosphoribosylanthranilate isomerase from *S. cerevisiae* had dissociation constants of 0.4 μ M and 0.2 μ M, respectively, so the fragments had to be present at concentrations in excess of these dissociation constants for the full yield of the native state to be regained.^{154,155}

One solution to this problem of the bimolecular association of the fragments is to perform a **circular permutation**¹⁵⁸ of the protein. By genetic manipulation, the coding sequence for the protein in the DNA is severed at a particular position, and the portion to the 5' side of the break is moved to the 3' end of the remainder of the coding sequence. The 3' end of the 3' fragment is joined in phase to the 5' end of the 5' fragment with a linking sequence of DNA encoding a segment of polypeptide long enough to connect comfortably the carboxy terminus of the original unpermuted protein to the amino terminus of the original unpermuted protein.¹⁵⁹ Consequently, a protein is eligible for circular permutation only if its amino terminus and its carboxy terminus are near each other in its crystallographic molecular model so that after the circular permutant has been expressed and has folded properly, its former carboxy terminus and former amino terminus can be joined by a continuous stretch of polypeptide.

Following circular permutation, there is a break in the polypeptide elsewhere in the native structure of the protein, formally equivalent to the break that would otherwise produce two fragments of the protein, but the polypeptide is continuous from the former carboxy terminus to the former amino terminus. If the break is placed at a position in its amino acid sequence known to be a **disordered loop**, the circularly permuted protein will usually fold, display almost normal enzymatic activ-

ity or biological function, assemble into the native oligomer, and have similar standard free energies of folding to that of the wild type.^{159–163} In such a situation, the disordered loop is broken by the new amino and carboxy termini, and the former carboxy and amino termini, which are usually disordered anyway, are now joined together, to prevent the two fragments of the original protein from dissociating when it is unfolded.

Circular permutation can be used to examine the **information necessary to fold**. The position at which the amino acid sequence of the wild-type protein is broken to produce the new amino terminus or carboxy terminus can be varied at random, and circular permutants that are still enzymatically active can be selected genetically. In almost all of the enzymatically active circular permutants of aspartate carbamoyltransferase from *E. coli*, the new carboxy and amino termini were found in segments of the polypeptide between α helices and β structure in the crystallographic molecular model of the wild-type protein.¹⁶⁴ This result seemed reasonable at the time because such elements of secondary structure probably could not form if there were a discontinuity within them. When a similar analysis, however, was made of random circular permutants of thiol:disulfide interchange protein dsbA from *E. coli*, the majority of the new amino and carboxy termini of the enzymatically active permutants were located at positions in the sequence of amino acids that in the wild-type protein are α helices or β structure. Four of the nine α helices and three of the five β strands could be interrupted, and the resulting circular permutants folded and were enzymatically active.¹⁶⁵

Another approach is to place systematically the new carboxy and amino termini at each position in the amino acid sequence of the protein and measure the enzymatic activity and standard free energy of folding for each of the resulting circular permutants. When such an analysis¹⁶⁶ was performed on dihydrofolate reductase from *E. coli* ($n_{aa} = 159$), a set of 10 segments varying in length from 2 to 14 aa could be identified, the interruption of which by introducing new amino and carboxy termini at any position led to a protein incapable of folding and enzymatically inactive. Placing the interruption at almost any one of the 87 positions outside these 10 forbidden regions gave a circular permutant that could fold to produce an enzymatically active protein. As with thiol:disulfide interchange protein dsbA, many of the permissive positions were within segments that are α helices or β strands in the crystallographic molecular model of the wild-type protein. The **forbidden regions**, however, also failed to correlate with elements of secondary structure. These results suggest that the information necessary to fold a polypeptide may be distributed over its sequence of amino acids by rules that are not immediately obvious.

The fact that many, if not most, of the circular permutants of a protein can fold to produce enzymatically and biologically active proteins and even the proper

oligomers clearly states that one piece of information that has nothing to do with the folding of a protein is the order in which its amino acids emerge from the ribosome. If there are portions of the protein that do fold before the complete protein emerges, those portions are not required to fold before the complete protein emerges. It has been suggested, however, that if a protein has domains, each domain might be required to fold as it emerged from the ribosome during biosynthesis before the next emerged. There is no evidence in favor of this conjecture, and proteins containing two or more **domains** undergo reversible folding as readily as proteins with only one domain.^{44,167–169}

If the reaction producing the unique native state of a folded polypeptide is an isomerization between the random coil and that native state, the individual contributions to the overall standard free energy change for this isomerization determine its outcome. Neither the formation of a hydrogen bond between a donor and an acceptor in the random coil nor the formation of an ionic interaction between a positively charged side chain and a negatively charged side chain in the random coil can provide any net favorable standard free energy for the folding of a protein in aqueous solution. In fact, their formation would be unfavorable. Nor can van der Waals forces make any contribution because the isomerization occurs in a condensed phase. Therefore, by exclusion and perhaps for the lack of a better candidate, the hydrophobic effect has attracted the most attention in discussions of the folding of a polypeptide.¹⁷⁰ The hydrophobic effect provides favorable standard free energy for the formation of the native state because hydrophobic side chains, which are exposed to water in the random coil, are removed to the interior of the protein during the folding.¹⁷¹

One of the major **deficits of standard free energy in the folding** of a protein results from the requirement to unsolvate those hydrophilic functional groups destined for the interior. This loss is due to the fact that water participates in strong interactions with donors and acceptors of hydrogen bonds and charged functional groups and to the fact that when **charged side chains** are withdrawn from water they are usually neutralized first. The removal of even neutral **hydrogen-bond donors** from water, even though they may always find an acceptor in the interior of the protein, is a significantly endothermic transfer.¹⁷² It has already been noted, however, that the formation of a hydrogen bond between an acceptor and a donor on a side chain, in the context of a folded polypeptide, is usually favorable with a standard free energy of formation of around -5 kJ mol^{-1} (Table 6–6). For example, in 52 instances in which a tyrosine was mutated to a phenylalanine, the standard free energy of folding increased by $6 \pm 4 \text{ kJ mol}^{-1}$ when the tyrosine was involved in a hydrogen bond in the crystallographic molecular model of the protein but showed no change when it was not. In 40 instances in which a threonine was

mutated to a valine, the standard free energy of folding increased by $4 \pm 4 \text{ kJ mol}^{-1}$ when that threonine was engaged in a hydrogen bond in the crystallographic molecular model but showed no change when it was not.¹⁷³

The reason that these hydrogen bonds between side chains in the native state have the modestly favorable free energies of formation that they do is **approximation**. The hydrophobic effect drives the condensation of the random coil that unavoidably withdraws donors and acceptors in the backbone of the polypeptide from contact with water. These donors and acceptors then combine to form the hydrogen bonds that define the **secondary structure**. These hydrogen bonds form because these donors and acceptors can no longer participate in hydrogen bonds with water and must do so among themselves. α Helices and β structure appear not because they are beautiful (Figures 6–6 and 6–9) but because they are an efficient way to provide an acceptor to most if not all of the donors pulled out of the water by the condensation driven by the hydrophobic effect. The proper **packing** of the secondary structure then can juxtapose the donor on a side chain with an acceptor. It is this approximation, brought about by the complete, cooperative process of folding, that is the only reason the resulting hydrogen bond has a favorable standard free energy of formation relative to the separated donor and acceptor in the random coil.

Because the realization of this favorable standard free energy of formation results from approximation, there are significant geometric requirements for its favorability. In addition, if too many of the hydrophobic groups on the side chains in the interior were replaced with properly aligned donors and acceptors to exploit these favorable increments in standard free energy of formation, the polypeptide could not fold in the first place.³ These are among the reasons that there are few such hydrogen bonds involving a donor on a side chain in the interiors of proteins.¹⁷¹ Those few hydrogen bonds between side chains that are found are the result of evolution by natural selection so it is not surprising that they have favorable free energies of formation.

In addition to the unfavorable standard free energy of transfer associated with the dehydration of hydrophilic functional groups as they are pulled into the interior of the folded polypeptide,¹⁷⁴ the other major deficit that must be overcome during the folding process is the **configurational entropy of the random coil**. This is the positive, intrinsic standard entropy that arises from the fact that the random coil can assume a large number of different configurations. It represents a deficit during the folding of the polypeptide because the native state, to a first approximation, assumes only a few conformations. Therefore, when the random coil becomes the native state its configurational entropy almost disappears.

At first glance, it seems that the configurational

entropy of the random coil, dictated by the sum over all of its states, should be very large because each amino acid has at least the two dihedral angles, ψ and ϕ (Figure 6–2), each of which can assume a number of values as dictated by the Ramachandran plot (Figure 6–4). This initial intuition, however, neglects excluded volume.¹⁷⁵

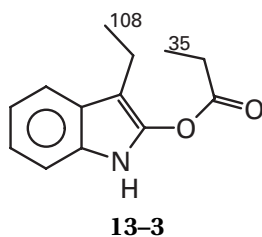
Excluded volume designates the qualification that every configuration of the random coil in which two or more atoms would occupy the same space at the same time is impossible and thus cannot contribute to the configurational entropy. This is the consequence of the steric effects that produce the Ramachandran plot itself operating over the whole polymer rather than just between neighboring amino acids. Excluded volume makes a large contribution to diminishing the configurational entropy of the random coil. For a polypeptide 100 amino acids in length, a set of configurations could be generated by randomly assigning values to the dihedral angles ψ and ϕ within their allowed ranges. The number of these configurations that do not superpose two or more atoms in the polypeptide has been estimated to be only 10^{-44} of the total number of randomly generated configurations.¹⁷⁶

Even though a consideration of excluded volume remarkably decreases the number of configurations available to the random coil, there are still a large number of configurations that are accessible. Only a small number of these configurations constitute the compact native state of the folded polypeptide. For the native state to be stable relative to the random coil, the configurational entropy resulting from the sum over all of the allowed unfolded configurations must be overcome by the hydrophobic effect realized upon the formation of the native state.

The presence of a **cystine** in a folded polypeptide makes a contribution to the change in configurational entropy for the isomerization between random coil and native protein. The polypeptide must first fold before the cysteines juxtaposed by the folding can be oxidized to cystines.¹⁷⁷ Because the folded native state is a prerequisite for the formation of a proper cystine and because the formation of a naturally occurring cystine usually has little effect on the structure or conformational freedom of the native protein,^{177–179} it necessarily follows that the cystine itself cannot significantly change the intrinsic configurational entropy of the properly folded protein and can change its intrinsic standard enthalpy only by the standard enthalpy of formation of the cystine. It has been demonstrated, however, that the standard enthalpy of formation for cystine in a random coil is about the same as that estimated for the standard enthalpy of formation of the same cystine in the native state.¹⁸⁰ Consequently, the incorporation of a cystine cannot affect the change in standard enthalpy of folding either. Rather, a cystine between two cysteines that are adjacent in the native structure increases the value of the equilibrium constant of the folding and decreases the change in

standard free energy of folding of a protein because it **decreases the configurational entropy of the random coil**.

The decrease in standard free energy of folding can be demonstrated experimentally by introducing a specific cross-link between two adjacent amino acids into the native structure of a protein and determining its effect on its folding. Glutamate 35 and the enol tautomer of the oxindole produced by the oxidation (Reaction 10–37) of Tryptophan 108 in lysozyme from *G. gallus* form an ester:



This ester introduces an intramolecular cross-link between these two amino acids¹⁸¹ that are adjacent to each other in the crystallographic molecular model of the protein. The cross-linked lysozyme has a standard free energy of folding at pH 2 in 2 M guanidinium chloride at 62 °C that is 22 kJ mol⁻¹ less than that of un-cross-linked lysozyme.¹⁸⁰ Lysine 7 and Lysine 41 in bovine ribonuclease A can be cross-linked specifically by 2-(*p*-nitrophenyl)-3-(3-carboxy-4-nitrophenyl)thio-1-propene (Figure 10–8).¹⁸² The difference in the standard free energy change of folding between the cross-linked and un-cross-linked ribonuclease at pH 2.0 and 40 °C¹⁸³ is –21 kJ mol⁻¹.

Many studies have incorporated single cystines by site-directed mutation between positions in the amino acid sequence of a protein that are adjacent to each other in its crystallographic molecular model. For example, cystines have been introduced into lysozyme from bacteriophage T4 ($n_{aa} = 164$). This protein has no cystines to begin with, so each mutant contained only one cross-link in its polypeptide.^{184,185} In one study, four different mutants were made containing cystines cross-linking positions 29, 34, 121, and 155 aa apart. In another, the same two site-directed mutants containing cystines cross-linking positions 121 and 155 aa apart were used and a third, cross-linking positions 94 aa apart, was made; and each of these mutants was submitted to the same circular permutation to produce permutants of T4 lysozyme with cystines cross-linking positions 49, 15, and 76 aa apart, respectively. Together, these manipulations gave eight mutants of the same protein, each with a cross-link producing a covalent loop in the denatured polypeptide of a different length. For each mutant, the difference in the standard free energy of folding ($\Delta\Delta G_{Fd,SS}^\circ$) between the protein with the cystine intact and the protein with the cystine cleaved by disulfide interchange (Figure 3–20) was estimated from its melting

temperature. The differences in standard free energies of folding varied from –3 to –14 kJ mol⁻¹, and the difference increased monotonically as the distance between the cysteines, and hence the **length of the loop**, increased.

A theoretical treatment of the expected decrease in configurational entropy caused by **cross-linking a random coil**, which accounts for excluded volume, predicts that the configurational entropy should decrease linearly with the natural logarithm of the distance between the cross-linked positions with a slope of 2.4.¹⁸⁶ When the experimental values of $(\Delta\Delta G_{Fd,SS}^\circ)/T$ are plotted against the natural logarithm of the distance, in number of amino acids, between the cystines in each of the mutants of T4 lysozyme, the expected relationship is observed. Furthermore, the difference in standard free energy of folding between native α -lactalbumin and α -lactalbumin in which the cystine between Cysteine 6 and Cysteine 120 has been reduced falls on the same line.¹⁸⁷

The effects of introducing cystines into other proteins, however, differ significantly from those measured in these observations. In some instances the magnitude of the difference in standard free energy of folding is much less than expected;¹⁸⁸ in others, much more.¹⁸⁹ The magnitudes of the differences in standard free energy of folding for lysozyme from *G. gallus* cross-linked through the oxindole ester (**13-3**) and bovine ribonuclease A cross-linked by 2-(*p*-nitrophenyl)-3-(3-carboxy-4-nitrophenyl)thio-1-propene are also greater than those observed with cystines introduced into lysozyme from bacteriophage T4.

The equilibrium constant for the folding of the constant fragment of the light chain of an immunoglobulin G, C_L (Figure 11–1), at pH 7.5 and 25 °C in solutions of guanidinium chloride is decreased when its single cystine is reduced. All of the change could be accounted for by the fact that the observed **rate constant for folding** (k_F in Equation 13–1) of the random coil with the cystine was 100-fold greater than the observed rate constant for the random coil without the cystine.¹⁹⁰ This is consistent with the conclusion that the intact, correct cystine decreases the configurational entropy of only the random coil, while retaining access to the properly folded structure, and permits the random coil to fold more rapidly. Whether or not the proper cystine was present had no effect on the observed rate constant of unfolding (k_U in Equation 13–1).

It has been shown that if the favorable noncovalent standard free energies of association between a subpopulation of the monomers along a polymer are significantly more negative than their individual standard free energies of solvation, the polymer should spontaneously collapse to a globular form.¹⁹¹ Because the constraints of excluded volume are even more extreme in this compact, globular form, the number of accessible configurations and hence its configurational entropy should be much smaller. Because the hydrophobic effect is the only inter-

action capable of producing significant net favorable standard free energies of association among the side chains of the amino acids in a random coil, it is generally assumed that the noncovalent force that would perform the condensation leading to a globular state of a polypeptide is the hydrophobic effect exerted upon the hydrogen-carbon bonds in those side chains in the polypeptide. This view of the folding of a polypeptide could be called the **condensation model**. Its central proposal is that the collapse of the random coil to a condensed state decreases the configurational entropy of the polypeptide dramatically and narrows the search for the native state to a much smaller number of accessible conformations.

On the basis of this model, folding can be treated theoretically as a process in which the unfavorable loss of the configurational entropy of the random coil is balanced only by the favorable removal of hydrophobic side chains from contact with the aqueous phase.^{176,192} The statistical treatment of the random coil developed by Flory,¹⁹³⁻¹⁹⁵ which takes account of excluded volume and the solvation of the monomeric units, can be expanded¹⁷⁶ to include the hydrophobic effect exerted during the sequestration of the monomers in the condensed state¹⁹⁵ and the fortuitous sequestration of the monomers in the random coil,¹⁷⁶ as well as the much smaller, but still significant, configurational entropy of the condensed polypeptide before it assumes the native state.

The process of folding is divided into two imaginary steps,¹⁷⁶ not necessarily related to the actual steps. These imaginary steps are the **condensation** of the random coil to a globular structure excluding water and the **reconfiguration** of the polymer in this condensed state to maximize the exposure of hydrophilic groups and minimize the exposure of the hydrophobic groups to the water. It is during this reconfiguration following the condensation that the donors and acceptors for hydrogen bonds that have been withdrawn away from the acceptors and donors of the water form hydrogen bonds among themselves to produce the α helices and β structure observed in the final native state of the protein. Before the condensation, water formed hydrogen bonds with those donors and acceptors.

With reasonable values both for the hydrophobic effect on the average hydrophobic amino acid (-8 kJ mol^{-1}) and for the fraction of the amino acids in the polypeptide that are hydrophobic (0.50), the formation of a unique globular state from a random coil should proceed with net negative standard free energy change for polypeptides greater than about 70 amino acids in length.¹⁹² Polypeptides less than about 70 amino acids in length should not fold because they should not be able to bury a large enough number of hydrophobic amino acids to overcome the configurational entropy of their random coils. It is the case that small, folded, cystineless, monomeric proteins of less than 70 amino acids are quite rare. Proteins composed of polypeptides shorter than 70

amino acids usually contain several cystines, are oligomeric,^{196,197} or have a relatively large hydrophobic core.¹⁹⁸ There are, however, a few small domains, for example the WW domains (35 aa)¹⁹⁹ or the peripheral subunit-binding domain from dihydrolipoyllysine-residue acetyltransferase (41 aa),²⁰⁰ that fold to form stable monomeric, well-defined native structures lacking cystines.

Although it was only for the sake of the computations that the folding of the polypeptide described in this condensation model was divided into the two steps of condensation and reconfiguration, there are stable, condensed but fluidly unstructured states of a polypeptide that seem to have the properties required of a condensed state on its way to the native state. These are the molten globules. A **molten globule** is a state of a polypeptide in which it has collapsed to a globular particle from the expanded random coil but remains fluid with a constantly changing conformation rather than achieving the limited set of conformations that is the native state. In such a fluid condensed state, the configurational entropy of the polypeptide should be significantly reduced relative to that of the random coil, and only a much smaller number of conformations that avoid the problems of excluded volume should be accessible.¹⁹¹ Many of these conformations should display α helices and β structure that form spontaneously.²⁰¹

Under conditions that differ significantly from those in the living system in which a particular polypeptide has evolved to fold, its native state may no longer be the most stable of the condensed conformations accessible to that polypeptide, and a number of other condensed, structured conformations may be as stable. Peculiar conditions such as low pH or the presence of denaturants, however, are necessary to prevent the polypeptide from assuming its native state, as it would do normally. It is argued that the intermediates detected in the folding of proteins under several such circumstances are examples of molten globular states and that all of these various intermediates represent a single configurational state assumed by a polypeptide that is at least as distinct as that of the random coil. This may be an overstatement. For example, two different molten globular states of apomyoglobin have been distinguished,²⁰² and there are intermediate states that do not have the properties assigned to a molten globule.²⁰³

Stable intermediates believed to be molten globules have been detected under many different circumstances. They have been observed for α -lactalbumin²⁰⁴ below pH 4.5 at concentrations of guanidinium chloride below 2.5 M; for α -lactalbumin,²⁰⁵ stripped of bound Ca^{2+} , at pH 8 and guanidinium chloride concentrations between 0.5 and 2.0 M; for cytochrome *c*^{206,207} below pH 3 either at chloride concentrations greater than 0.1 M or at concentrations of *O*- α -D-glucopyranosyl(1-3)- β -D-fructofuranosyl- α -D-glucopyranoside greater than 0.5 M; and for carbonate dehydratase⁶⁴ at temperatures

below 60 °C and values of pH less than 3.5. The mutation of Phenylalanine 173 to an alanine in murine interleukin 6 converts the protein into a molten globule.²⁰⁸ These are all unphysiological conditions, but proteins have evolved to be entirely in their native states under physiological conditions. Consequently, it would not be surprising that, to isolate intermediates in the normal process of folding, such peculiar conditions would be required. Molten globules often become stable relative to the native state at low pH. Presumably, their fluidity permits carboxy groups that are rigidly buried in the native state to reach the surface of the globule and be exposed to the solvent, thereby lowering the standard free energy of the molten globule relative to that of the native structure (Equation 13–3).

Various physical measurements have been made of these stable intermediates identified as molten globules. The **circular dichroic absorptions** of the native protein between 260 and 290 nm are largely lost in the molten globule, and this loss must result from the disappearance of the unique asymmetric environments around tryptophans, tyrosines, and phenylalanines.^{66,209} The complex **nuclear magnetic resonance spectrum** of the native state becomes much simpler and much more like that of the random coil upon formation of these molten globules,^{210,211} as would be expected if the unique environments around each amino acid had been lost and each side chain now sampled continuously a broad range of changing environments. When the internal dynamics of one of these molten globules are examined by **quasielastic neutron scattering**,²¹² it is observed that the potential barriers to bond rotations in the side chains are lower than those in the native state, while diffusive motions of side chains are greater, and significantly smaller units of structure diffuse cooperatively than those diffusing in the native state. Measurements of the **absorption of ultrasound** also indicate that such molten globules are more fluid than the native state,²¹³ and conformational relaxations in the interior that occur in the 2 MHz range are significantly enhanced.

The majority of the circular dichroic absorption between 200 and 240 nm seen in the native states is retained in the respective molten globules, and this suggests that they contain α helices and β structure.^{66,209} The slow **proton exchange** observed by nuclear magnetic resonance for buried peptide bonds in the native state increases by factors of 1000–100,000 upon the transition to one of these molten globules, even though the α amido protons of many of the same amino acids remain relatively less accessible.^{65,214} These observations suggest that some of the same elements of secondary structure but not all of them²¹⁵ remain at the same locations in the amino acid sequence of the polypeptide but open up 1000–100,000-fold more often. These accelerated rates of exchange are increased much more by adding denaturants¹³⁸ so the conformational changes within the molten globule leading to exchange of amido

protons do involve some expansion of the structure into the solvent, but the effect, and hence the expansion, is much less than when the native state unfolds to the random coil. Three of the eight α helices of the native state of apomyoglobin²¹⁶ are present in its molten globule,⁶⁵ but site-directed mutations at the interfaces in the native state between these α helices have little effect on the stability of the molten globule.²¹⁷ It was concluded that although these α helices had formed, they were not packed against each other in any stable arrangement.

The accessibility of tryptophans to the solvent, as judged by **quenching of fluorescence** (Equation 12–41), differs significantly in these molten globules, and tryptophans that are relatively more exposed in the native state become less exposed.²¹⁸ The fluorescence intensities of the tryptophans in cytochrome *c*, which are quenched by the nearby heme in the native state but are fully expressed in the random coil, remain quenched in its molten globule.²¹⁰ The **intrinsic viscosity**, **rotational relaxation times**, and **diffusion coefficients** of the molten globules are indistinguishable from those of the corresponding native states but are different from those of the random coil.^{209,210} All of these observations demonstrate that they are condensed, globular structures like the native state.

It is thought that these intermediates identified as molten globules represent the random coil that has collapsed to a **globular state** because of the hydrophobic effect, even though it is **fluid** and cannot assume the unique set of conformations that is the native state. In this regard, it is interesting that the majority (85%) of the change in standard heat capacity between the random coil and the native state of α -lactalbumin, which is a signature of the hydrophobic effect, is experienced in the transition between the random coil and the intermediate that has been characterized as a molten globule.²⁰⁵ The equilibrium constant for the formation from the random coil of an intermediate thought to be a molten globular state of apomyoglobin displays a significant temperature dependence, passing through a maximum between 0 and 20 °C. This observation is also consistent with a process accompanied by a large change in standard heat capacity,²⁰² but in this case, only about 50% of the overall change in standard heat capacity is realized in the transition from random coil to molten globule.

If these intermediate, molten globular states resemble intermediates on the normal kinetic pathway between the random coil and the native state, then the condensation model for the folding of a polypeptide may be an accurate rendition of the process. In this description of folding, the random coil spontaneously collapses under the influence of the hydrophobic force to form a condensed state that would be a molten globule. This molten globule would fluidly sample the limited number of conformations available to the condensed polymer until the native state, the set of conformations of lowest standard free energy, was encountered.

The alternative to the condensation model for the

folding of a polypeptide could be referred to as the **nucleation model**. In this view of the process, a short segment of the polypeptide or several short segments would spontaneously assume a metastable conformation similar to the conformation of that short segment or those short segments in the complete native state. This nucleus for folding would resemble the conformation of the native state in both its secondary and tertiary interactions in this restricted region, and it would represent the most independently stable region of the native state. From this nucleus, folding would rapidly spread to produce the entire native structure. Evidence for this proposal comes from the study of short segments of polypeptide that can assume structured states other than the random coil and from stable expanded states of some polypeptides.

Although almost all short segments of polypeptide have proven to be structureless, a few have been found that assume a structured state. For example, two peptides from bovine pancreatic trypsin inhibitor ($n_{aa} = 58$), Arginine 20–Phenylalanine 33 and Asparagine 43–Alanine 58, were chemically synthesized and joined by forming the cystine between Cysteine 30 and Cysteine 51 that occurs naturally in the native protein. This covalent complex, containing only half of the covalent structure of the full-length protein, nevertheless formed a structure²¹⁹ that had some of the structural features assumed by this region in the crystallographic molecular model of the protein. The antiparallel β sheet could be discerned in the nuclear magnetic resonance spectrum but not the α helix. The short, stable α -helical segments of polypeptide discussed earlier have also been proposed as models of nucleation points in protein folding.²²⁰

There are also stable conformations of a few polypeptides, observed under circumstances promoting denaturation, in which condensation has not occurred but elements of structure resembling those in the native state have formed. For example, there is an expanded form of the polypeptide of cytochrome *c* observed at low ionic strength and low pH in which α helices found in the native state of the protein are formed²²¹ and an expanded form of the α subunit of tryptophan synthase from *E. coli* in which a hydrophobic cluster has formed.²²² In the denatured form of ribonuclease from *B. amyloliquifaciens* observed at pH 2 and 25 °C, which is unfolded, two of the α helices found in the native state of the protein are formed in low yield.³⁵ All of these results suggest that portions of the polypeptide, when it is still in an almost fully expanded state, might assume their native structures initially to produce a point of nucleation for overall folding.

What is more likely, however, is that both condensation and nucleation occur during the folding of a polypeptide. The most obvious evidence for such a scenario is that some, but not all, of the secondary structure that is found in the native state of a protein is usually present in its molten globule. Such elements of secondary structure could nucleate the formation of the remainder of the native structure. A fragment of α -lactalbumin

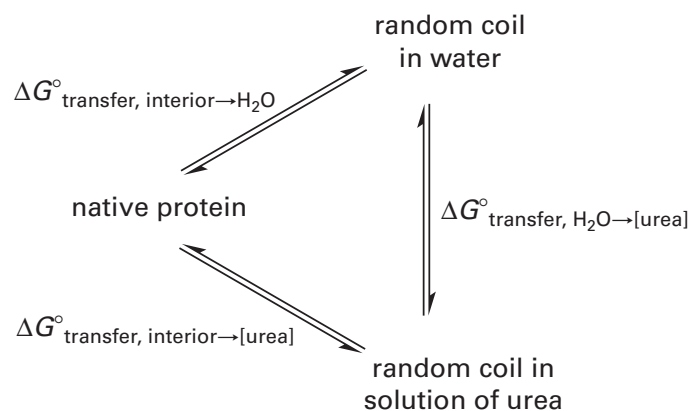
that assumes at equilibrium a structure containing the α helices that are present in its portion of the native structure of the protein has been proposed to represent a point of nucleation for the native state even though it is by itself a molten globule.²²³ These observations suggest that protein folding involves both condensation and nucleation, but not necessarily in that order. The question of the sequence of events in the folding of a polypeptide requires kinetic observations of the process.

Suggested Reading

- Salahuddin, A., & Tanford, C. (1970) Thermodynamics of the denaturation of ribonuclease by guanidine hydrochloride, *Biochemistry* 9, 1342–1347.
- Taniuchi, H., & Anfinsen, C.B. (1971) Simultaneous formation of two alternative enzymatically active structures by complementation of two overlapping fragments of staphylococcal nuclease, *J. Biol. Chem.* 246, 2291–2301.
- Perl, D., Welker, C., Schindler, T., Schroder, K., Marahiel, M.A., Jaenicke, R., & Schmid, F.X. (1998) Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins, *Nat. Struct. Biol.* 5, 229–235.

Problem 13–1: As urea is added to a solution containing a protein in its native state, the protein usually begins to unfold when the concentration of urea rises above 4–5 M. This unfolding is due to the ability of urea to stabilize the unfolded state.

Consider the side chain of an amino acid that is located in the interior of a protein and cannot see the solvent when the protein is folded. From the point of view of this interior side chain, the following series of equilibria govern the unfolding process:



The standard free energy changes $\Delta G^\circ_{\text{transfer, interior} \rightarrow \text{H}_2\text{O}}$ and $\Delta G^\circ_{\text{transfer, interior} \rightarrow [\text{urea}]}$ are standard free energy changes that occur as the amino acid is transferred from the interior of the protein either into pure water or into a solution of urea, respectively, as the protein unfolds to a random coil, and $\Delta G^\circ_{\text{transfer, H}_2\text{O} \rightarrow [\text{urea}]}$ is the standard free energy change involved in transferring the side chain of an amino acid, which is exposed during unfolding, from water into a solution of urea.

- (A) How are these three values of ΔG° related? What sign must each carry to explain the unfolding caused by urea?

The following is a table⁷ of the solubilities of a series of amino acids in solutions of several concentrations of urea at 25 °C.

amino acid	solubilities [g (100 g of solvent) ⁻¹] at noted concentration of urea				
	0 M	2 M	4 M	6 M	8 M
Gly	25.1	22.7	20.4	17.5	15.00
Ala	16.7	15.3	13.7	12.1	10.60
Leu	2.16	2.37	2.34	2.29	2.25
Phe	2.80	3.42	3.94	4.33	4.67
Trp	1.38	1.98	2.65	3.31	3.95
Met	5.59	6.19	6.74	7.00	6.99
Thr	9.80	9.56	9.07	8.31	7.41
Tyr	0.0451	0.0600	0.0732	0.0870	0.0986
His	4.33	4.66	4.70	4.46	4.23
Gln	4.30	4.49	4.49	4.30	4.02
Asn	2.51	2.89	3.08	3.22	3.32

- (B) Calculate $\Delta G^\circ_{\text{transfer, H}_2\text{O} \rightarrow [\text{urea}]}$ for each model compound in the units of joules mole⁻¹. Subtract $\Delta G^\circ_{\text{transfer, H}_2\text{O} \rightarrow [\text{urea}]}$ for glycine to estimate $\Delta G^\circ_{\text{transfer, H}_2\text{O} \rightarrow [\text{urea}]}$ for each side chain.^{7,8}

These values you have just calculated are tabulated below.

side chain	$\Delta G^\circ_{\text{transfer, H}_2\text{O} \rightarrow \text{denaturant}}$ (cal mol ⁻¹)							
	urea				GdmCl			
	2 M	4 M	6 M	8 M	1 M	2 M	4 M	6 M
Ala	0	+15	+10	+10	-10	-20	-30	-45
Val ^a	-60	-85	-125	-160	-85	-115	-195	-265
Leu	-110	-155	-225	-295	-150	-210	-355	-480
Ile ^a	-100	-140	-205	-265	-135	-190	-320	-430
Met	-115	-225	-325	-415	-150	-245	-400	-535
Phe	-180	-330	-470	-600	-215	-355	-580	-775
Tyr	-225	-395	-580	-735	-235	-385	-605	-770
Trp	-270	-505	-730	-920	-400	-630	-980	-1,235
Pro ^a	-75	-105	-155	-200	-100	-140	-240	-320
Thr	-40	-60	-90	-115	-65	-90	-120	-125
His	-100	-160	-205	-255	-180	-285	-385	-420
Asn	-135	-225	-330	-430	-200	-320	-490	-645
Gln	-80	-130	-190	-230	-135	-215	-315	-360

^aThe values for these side chains are estimates based on results for the other side chains and on results at a single concentration of denaturant.

- (C) Plot $\Delta G^\circ_{\text{transfer, H}_2\text{O} \rightarrow [\text{urea}]}$ against [urea] for each of these side chains. Determine the slopes of these lines that give values for $(\partial \Delta G^\circ / \partial [\text{urea}])$ in joules (mole of side chain)⁻¹ [liter (mole of urea)⁻¹].
- (D) How do these numbers correlate with your expectations in part A? Explain why the protein unfolds when [urea] rises above a certain critical level.

- (E) Plot $(\partial \Delta G^\circ_{\text{transfer, H}_2\text{O} \rightarrow [\text{urea}]} / \partial [\text{urea}])$ against the number of hydrogen-carbon bonds in each side chain. Is the major effect of urea to counteract the hydrophobic effect? Why?

The accessible surface area of each of these side chains has been calculated by a computer from molecular models.

model	surface area of side chain (nm ²)
Ala	0.21
Val	0.48
Thr	0.51
Leu	0.67
Met	0.90
Phe	0.93
Tyr	1.10
Trp	1.34
Asn	0.60
Gln	0.89
His	0.83

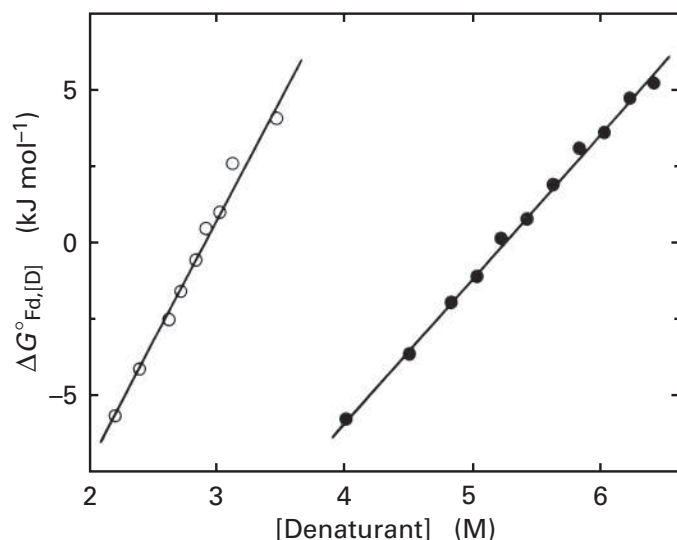
- (F) Plot $(\partial \Delta G^\circ_{\text{transfer, H}_2\text{O} \rightarrow [\text{urea}]} / \partial [\text{urea}])$ against accessible surface area, labeling each point on your curve to keep track of the side chain it represents. What is the value of $\partial(\partial \Delta G^\circ_{\text{transfer, H}_2\text{O} \rightarrow [\text{urea}]} / \partial [\text{urea}]) / \partial (\text{surface area})$?

Measurements of a physical property displayed by a protein can be used to obtain a value of the equilibrium constant for the transformation between the native state and the random coil at different concentrations of a denaturant. From each of these equilibrium constants, the standard free energy of folding, $\Delta G^\circ_{\text{Fd, [denaturant]}}$, for the reaction at that concentration of denaturant can be calculated. The figure on the next page is an example of the relationship between $\Delta G^\circ_{\text{Fd, [denaturant]}}$ and the concentration of denaturant for the unfolding of lysozyme promoted by urea and guanidinium chloride.

The slopes of these two lines $(\partial \Delta G^\circ / \partial [\text{denaturant}])$ are relative measures of the effectiveness of the two denaturants. By fitting a straight line to these data it is possible to obtain, by extrapolation, the standard free energy of folding in the absence of denaturant, $\Delta G^\circ_{\text{Fd, H}_2\text{O}}$. The following table gathers the results from four separate proteins, where $[\text{GdmCl}]_{1/2}$ or $[\text{urea}]_{1/2}$ is the concentration of denaturant when $[F] = [U]$ and $\Delta G^\circ_{\text{Fd}} = 0$.

protein	guanidinium chloride		urea	
	$[\text{GdmCl}]_{1/2}$ (M)	$\Delta G^\circ_{\text{Fd, H}_2\text{O}}$ (kJ mol ⁻¹)	$[\text{urea}]_{1/2}$ (M)	$\Delta G^\circ_{\text{Fd, H}_2\text{O}}$ (kJ mol ⁻¹)
bovine ribonuclease A	3.01	-39	6.96	-32
lysozyme <i>G. gallus</i>	3.07	-24	5.21	-24
bovine chymotrypsin	1.90	-32	4.04	-35
ovine β -lactoglobulin	3.23	-52	5.01	-44

- (G) From an examination of the figure and an understanding of where the two points tabulated fall



Apparent standard free energy of folding, $\Delta G^{\circ}_{\text{Fd},[D]}$, of lysozyme as a function of the molar concentrations of urea (●) or guanidinium chloride (○) at pH 2.9.¹⁰⁷ The apparent standard free energy of folding is zero at the concentration of denaturant at which $[U] = [F]$. Adapted with permission from ref 107. Copyright 1974 *Journal of Biological Chemistry*.

upon each line, calculate (Equation 13–23) the slope m of the line ($m = \partial\Delta G^{\circ}/\partial[\text{denaturant}]$) for each combination of protein and denaturant.

(H) Calculate the quantity

$$R_{\text{prot}} = \frac{(\partial\Delta G^{\circ}_{\text{Fd}}/\partial[\text{guanidinium}])_T}{(\partial\Delta G^{\circ}_{\text{Fd}}/\partial[\text{urea}])_T}$$

for each protein. This number will serve as a quantitative estimate of the relative effectiveness of the two denaturants.

(I) In part F you calculated a quantity $\partial(\partial\Delta G^{\circ}_{\text{transfer},\text{H}_2\text{O} \rightarrow [\text{urea}]} / \partial[\text{urea}]) / \partial(\text{surface area})$. Using the same methods, calculate a value for $\partial(\partial\Delta G^{\circ}_{\text{transfer},\text{H}_2\text{O} \rightarrow [\text{GdmCl}]} / \partial[\text{GdmCl}]) / \partial(\text{surface area})$ from the data in the table preceding part C.

(J) Calculate the quantity

$$R_{\text{transfer}} = \frac{(\partial\Delta G^{\circ}_{\text{transfer},\text{H}_2\text{O} \rightarrow [\text{GdmCl}]} / \partial[\text{guanidinium}])_T}{(\partial\Delta G^{\circ}_{\text{transfer},\text{H}_2\text{O} \rightarrow [\text{urea}]} / \partial[\text{urea}])_T}$$

and compare it to R_{prot} .

Problem 13–2: Bovine ribonuclease A is a protein containing 124 amino acids and four cysteines. Ribonuclease was added to a series of solutions containing different concentrations of guanidinium chloride, and the change in its extinction coefficient ($\Delta\epsilon_{287}$) was measured at 287 nm when each of the solutions had come to equilibrium.¹²⁴

[GdmCl] (M)	$\Delta\epsilon_{287}$ (M ⁻¹)	[GdmCl] (M)	$\Delta\epsilon_{287}$ (M ⁻¹)
0.00	0	3.01	–68
0.02	4	3.20	–96
0.34	4	3.30	–102
0.99	4	3.54	–119
1.61	7	4.03	–127
2.20	8	4.37	–126
2.67	–18	4.84	–121
2.81	–44	5.57	–118
2.95	–63	6.04	–114
		6.82	–110

The change in absorbance at 287 nm is a spectral indicator that reflects changes in the environments around the tryptophans in a protein. Make a plot of these data.

(A) What is $\Delta\epsilon_{287}$ of native ribonuclease at 4 M guanidinium chloride?

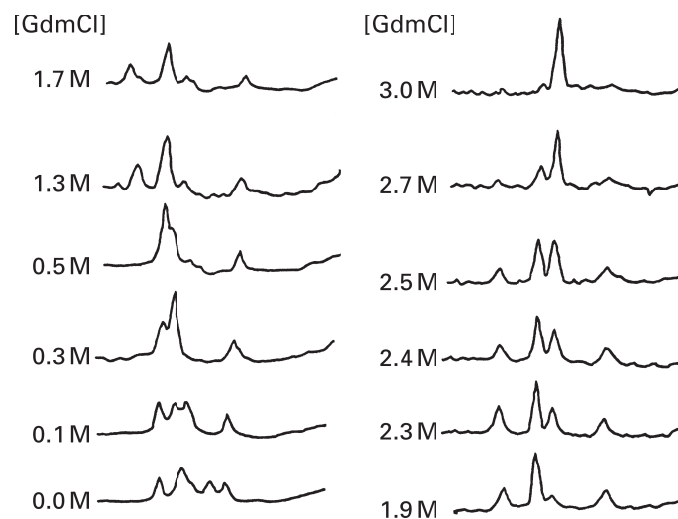
(B) What is $\Delta\epsilon_{287}$ of unfolded ribonuclease at 1 M guanidinium chloride?

For ribonuclease it has been proven that only the native state and the random coil are present at any concentration of guanidinium chloride.

(C) Calculate K_{Fd} for the equilibrium of Equation 13–1 for each concentration of guanidinium chloride and enter your values into a table.

(D) Plot $\ln K_{\text{Fd}}$ against [GdmCl] and determine, by extrapolation, the standard free energy of folding for ribonuclease in water, $\Delta G^{\circ}_{\text{Fd},\text{H}_2\text{O}}$.

Problem 13–3: In the region of the nuclear magnetic resonance spectrum of bovine ribonuclease A between 8.0 and 9.0 ppm, the only absorptions present are those from the carbons 2 of the imidazole rings of the histidines. The traces⁴⁸ are from this region of the nuclear magnetic resonance spectrum of ribonuclease. Changes in the spectrum occur when guanidinium chloride is added to the sample at the noted concentration.



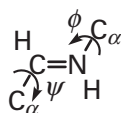
In the native protein (0.0 M guanidinium chloride), four absorptions from the protons on carbons 2 are observed. They have been assigned to Histidines 48, 119, 12, and 105, the four histidines of ribonuclease.

- (A) Why does each absorption have a unique position in the spectrum of native ribonuclease?
- (B) Why is there only one absorption from the protons on the carbons 2, which integrates as four protons from the protein, when it is dissolved in 3.0 M guanidinium chloride?
- (C) Between 0.0 and 1.7 M guanidinium chloride the resonances shift around, but above 1.7 M the four absorptions coalesce into the one absorption. What process is the spectrometer monitoring between 1.7 and 3.0 M guanidinium chloride?
- (D) What would be the position of the absorption from the protons on the carbons 2 of *N*^α-acetylhistidine ethyl ester in 3.0 M guanidinium chloride?

Problem 13-4: Below are listed several thermodynamic parameters that are involved in the process of protein folding.

- (a) Change in standard free energy for the hydrophobic effect
 - (b) Standard free energy of formation for hydrogen bonds
 - (c) Change in standard electrostatic free energy
 - (d) Configurational entropy of the random coil
 - (e) Configurational entropy of the native state
- (A) Which one is most affected by the steric constraints described in the Ramachandran plot?

Suppose proteins were held together by imine linkages rather than peptide bonds:



- (B) What effect would this have on the parameter you have chosen above?
- (C) How would the value of the standard free energy of folding $\Delta G_{\text{Fd}}^\circ$ be affected by this change?

Kinetics of Folding

The most straightforward way to initiate the folding of the random coil of a polypeptide that has been unfolded to a random coil in a concentrated solution of guanidinium chloride or urea is to dilute that solution. The **dilution** is performed so that the final concentration of denaturant is well below the region of transition so that the equilibrium between the folded state and the unfolded state is shifted from one heavily in favor of the

random coil to one heavily in favor of the folded state (Figure 13-1).*

Often the folding of the polypeptide is complete within a few seconds so the dilution must be performed rapidly. Usually, the solution of the random coil at a high concentration of denaturant is mixed with around 10 volumes of aqueous buffer of the appropriate ionic strength and pH in a **rapid mixing chamber**. The chamber is designed to mix the two solutions completely in less than a millisecond as they are forced through it at high velocity under considerable pressure. There are several different ways in which the solution emerging from the mixing chamber can then be monitored. Usually, a cuvette† is attached to the mixing chamber, and the mixture from the chamber is passed through the cuvette at the high velocity developed in the mixing chamber until it fills the cuvette uniformly. Once a steady state is reached, the flow is abruptly stopped, and changes that occur in the solution in the cuvette after cessation of the flow are monitored. The mean time the solution in the cuvette has spent between being mixed and the cessation of flow coincident with the initiation of the monitoring is the **dead time** of the apparatus. No measurements can be made of events that occur during the dead time. In most cases, the dead time of such a **stopped-flow apparatus** is 1–50 ms. Changes in the absorbance, molar ellipticity, or fluorescence of the solution can be monitored continuously from the dead time onward.

Often, during the folding of a protein, significant changes in absorbance, fluorescence, or molar ellipticity or two or three of these properties occur within the dead time of the apparatus. Type I ribonuclease H from *E. coli* displays such behavior (Figure 13-9).^{225,226} When its molar ellipticity at either 220 nm (Figure 13-9A) or 292 nm (Figure 13-9B) is monitored, the signal observed after flow has stopped decays to the value for the native state in a single, apparently first-order relaxation (Equation 13-13) with a rate constant of 0.6 s^{-1} . When these time courses are extrapolated through the dead time back to the instant of mixing, however, it can be seen that 83% of the change in molar ellipticity at 220 nm and 44% of the change in molar ellipticity at 292 nm did not occur during this apparently homogeneous transformation but in one or more kinetic steps that were much more rapid than the final isomerization. These steps occurred within the dead time of the apparatus, and, consequently, they could not be resolved.

A reaction that is unresolved in a stopped-flow experiment because it is complete during the dead time

* It is also possible to begin the experiment with a solution of the complex between the polypeptide and dodecyl sulfate and then rapidly strip the dodecyl sulfate from the protein.²²⁴ The difficulty with this approach is that the polypeptide in the complex with dodecyl sulfate is completely α -helical rather than a random coil.

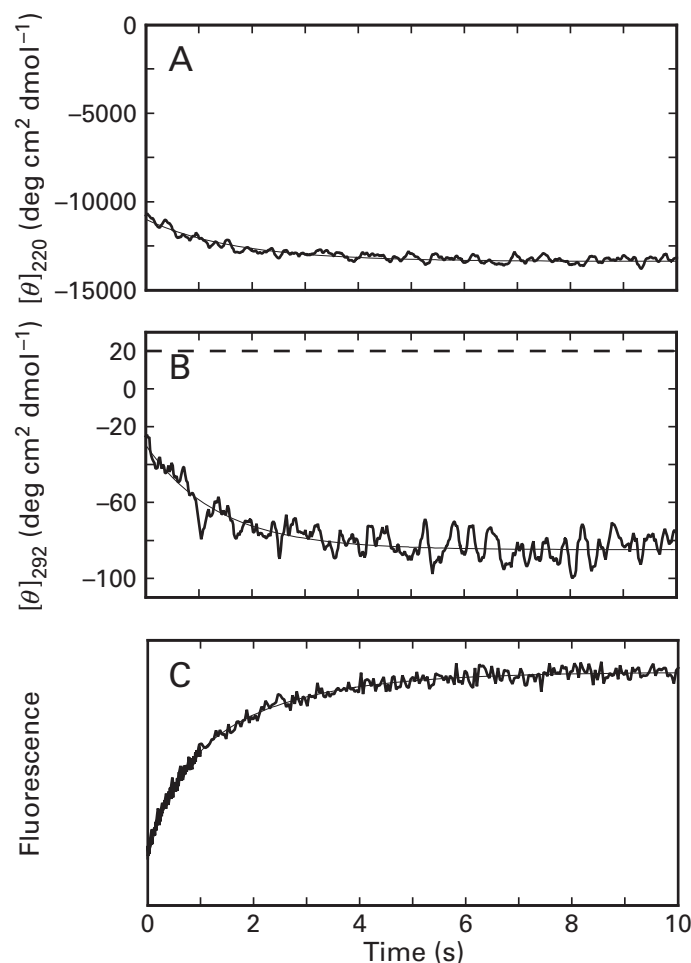
† A cuvette is a chamber with transparent walls through which spectrophotometric measurements can be made.

Figure 13-9: Kinetic burst during the refolding of type I ribonuclease H from *E. coli*.²²⁵ Type I ribonuclease H ($n_{aa} = 155$) was dissolved in 3.3 M guanidinium chloride and 10 mM sodium acetate, pH 5.5 at 25 °C. After it had unfolded completely, it was mixed in a stopped-flow apparatus with 10 volumes of 0.65 M guanidinium chloride and 10 mM sodium acetate, pH 5.5 (final concentration = 0.9 M guanidinium chloride), and the effluent from the mixing chamber was monitored following cessation of flow (dead time = 50 ms). (A) Molar ellipticity at a wavelength of 220 nm ($[\theta]_{220}$) in units of degrees centimeter² (decimole of peptide bond)⁻¹ as a function of time (seconds). The molar ellipticity of unfolded ribonuclease H at 220 nm in 0.9 M guanidinium chloride, as determined by extrapolation at equilibrium as in Figures 13-1A and 13-7A, should be 0. (B) Molar ellipticity at a wavelength of 292 nm ($[\theta]_{292}$) in units of degree centimeter² (decimole of peptide bonds)⁻¹ as a function of time (seconds). The molar ellipticity of unfolded ribonuclease H at 292 nm in 0.9 M guanidinium chloride, again as determined by extrapolation at equilibrium, should be 20 (dashed line). (C) Fluorescence of tryptophans in the protein monitored at emission wavelengths of greater than 300 nm with excitation at 280 nm. The scale for fluorescence was not quantified, so the relative fluorescence of the unfolded protein in 0.9 M guanidinium chloride was not reported. The three sets of data were fit with first-order relaxations (smooth curves) with rate constants of (A) 0.59 s⁻¹, (B) 0.74 s⁻¹, and (C) 0.51 s⁻¹ and 1.95 s⁻¹. Reprinted with permission from ref 225. Copyright 1995 American Chemical Society.

is a **kinetic burst**. The observation of a kinetic burst is interpreted to mean that one or more transformations of the random coil have occurred during the dead time and that they have produced an intermediate state. This intermediate state then turns into the native state as the reaction is monitored over time. In the case of type I ribonuclease H, this intermediate state becomes the native state in a reaction that appears by measurements of circular dichroism to display simple first-order kinetics with a rate constant of about 0.6 s⁻¹. This latter transformation is also revealed in the change in fluorescence of the solution (Figure 13-9C).

The kinetics of the folding of apomyoglobin from *Physeter catodon*,²²⁷ of micrococcal nuclease from *Staphylococcus aureus*,^{228,229} of equine cytochrome *c*,²³⁰ of dihydrofolate reductase from *E. coli*,²³¹ of equine β -lactoglobulin,²³² and of equine lysozyme^{233,234} all display similar **kinetic bursts producing intermediate states** that then apparently decay through one or several first-order steps to their native states. The appearance of these intermediates during a kinetic burst and their decay during the period of measurement can be detected by their absorbance, by their molar ellipticities in the range of 220–230 nm (the far ultraviolet), by their molar ellipticities in the range of 270–290 nm (the near ultraviolet), by their fluorescence, or by the transfer of energy between donors and acceptors placed at particular positions in their amino acid sequences. They are observed upon rapid dilution to 0.4–0.8 M urea or to 0.3–0.9 M guanidinium chloride. They are formed usually within less than 10 ms at temperatures between 10 and 25 °C, and they then decay at various rates.

A certain fraction of the changes in molar ellipticity or fluorescence that occurs in each of these kinetic bursts



results from the instantaneous changes in the molar ellipticity or fluorescence that occur in the random coil upon the abrupt decrease in the concentration of denaturant and that may or may not be able to be corrected for by linear extrapolation of the values at equilibrium for these physical properties from beyond the region of transition, as was done in Figure 13-1A.²³⁵ Their magnitude, however, is sufficiently large in most cases that they must reflect significant conformational changes in the unfolded state that produce real intermediates in the process of folding.

Within the range of denaturant concentration in the region of transition where the respective equilibrium constants for folding have been shifted into measurable ranges, most of these proteins display two-state behavior in the isomerization of their folding without evidence for intermediates. Why do intermediates in folding appear at lower concentrations of denaturant? The answer lies in the behavior of the observed rate constant*

* The progress of the folding of a protein monitored spectrophotometrically can usually be fit by a rate equation for one or more sequential first-order steps. Even though this fit is probably an oversimplification of the actual events, the observed apparently uncomplicated first-order rate constants obtained by such numerical analysis will be referred to as observed rate constants.

of folding, k_F , as a function of the concentration of denaturant.

When the logarithm of the observed rate constant for the approach to equilibrium for the folding of type I ribonuclease H from *E. coli* is plotted as a function of the concentration of guanidinium chloride (Figure 13–10),^{225,236} two-state behavior is observed in the region of transition, where the observed rate constant for unfolding dominates at high concentrations of denaturant and the observed rate constant for folding dominates at low concentrations. The logarithm of the observed rate constant for folding, however, does not display a continuous linear decrease below the region of transition as is observed in Figure 13–1B. Instead, its behavior is resolved further into two components, one dominant at intermediate concentrations of denaturant and the other at low concentrations. These two steps are distinguished by the two different slopes of the two distinct linear segments below the region of transition in

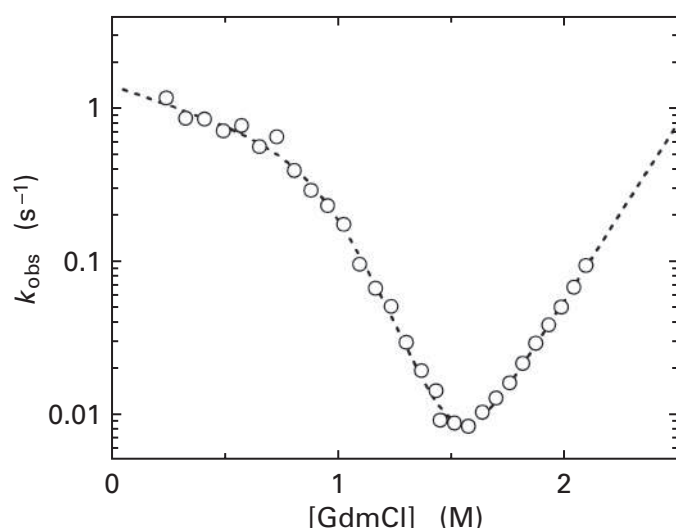


Figure 13–10: Observed rate constants for the approach to equilibrium for the folding isomerization of the cysteineless mutant of type I ribonuclease H from *E. coli*.²³⁶ A solution of 110 mM protein, 50 mM KCl, 20 mM sodium acetate, pH 5.5, and either no denaturant (unfolding) or 3 M guanidinium chloride (folding) was mixed in a stopped-flow apparatus with 11 volumes of 50 mM KCl, 20 mM sodium acetate, pH 5.5, and the appropriate concentration of guanidinium chloride to achieve the noted concentration following the mix. The fluorescence emission of the solution at wavelengths greater than 320 nm from an excitation at 295 nm was monitored as a function of time as in Figure 13–9C. The traces of fluorescence as a function of time could each be fit (Equation 13–13) by single-exponential increases (unfolding) or decreases (folding) of fluorescence. The observed rate constants k_{obs} (second⁻¹) for these exponential relaxations are plotted as a function of the concentration (molar) of guanidinium chloride (GdmCl). The dashed line is a fit of the data to an equation derived from a mechanism in which there are three states interconverting: the random coil, the native state, and a kinetic intermediate. It is assumed that the logarithm of each of the four first-order rate constants interconverting those three states is a linear function of the concentration of guanidinium ion as in Figure 13–1B. Reprinted with permission from ref 236. Copyright 1999 Elsevier B.V.

Figure 13–10.* This behavior is characteristic of a **change in rate-limiting step**† and consequently is consistent with a kinetic mechanism for folding in which there are two or more steps and one or more intermediates.²³⁸ At concentrations of guanidinium chloride between 1.5 and 1 M, the folding of type I ribonuclease H (Figure 13–10) has an observed rate constant that is strongly dependent on the concentration of guanidinium chloride. At concentrations of guanidinium chloride between 0.2 and 0.6 M, however, the folding of the protein has an observed rate constant that is only weakly dependent on the concentration of guanidinium chloride. At concentrations of guanidinium chloride between 0.6 and 1.0 M, the change in rate-limiting step occurs from a step strongly dependent on concentration of denaturant to a later step in the process of folding that is weakly dependent.

The two observed rate constants for the two respective steps between which the rate limitation shifts are the rate constant for the production of the intermediate present following the kinetic burst and the rate constant for its decay to the native state, respectively. The observed rate constant for the formation of this intermediate, which is defined by the linear segment of greater slope, is strongly **dependent on the concentration of denaturant**. Because of this strong dependence, above a certain concentration of denaturant the rate at which the intermediate is formed becomes slower than the rate at which it isomerizes to the native state. Consequently, above that concentration of denaturant, the intermediate, although it is still formed every time a polypeptide folds, cannot accumulate because it turns into the native state faster than it is formed, and the reaction, which is at least a three-state process at low concentrations of denaturant or in its absence, appears to become a two-state process above 1 M guanidinium chloride.

Such behavior indicative of a change in the rate-limiting step of folding is also observed for the folding of type I ribonuclease H from *E. coli* in solutions of urea²²⁶ as well as lysozyme from bacteriophage T4 in solutions of guanidinium chloride,²³⁶ the carboxy-terminal domain of the cell surface receptor CD2,²³⁹ the inhibitor barstar of the ribonuclease of *B. amyloliquifaciens*,²⁴⁰ the amino-terminal domain of phosphoglycerate kinase from

* In the simplest cases, where only one continuously varying apparent rate constant seems to control the folding at concentrations of denaturant below those of the region of transition, as in Figure 13–1B, it is observed that the logarithm of that apparent rate constant is a linear function of the concentration of denaturant, much as is the standard free energy of folding. It is customary to designate the slope of such a line with the symbol m , just as the slope is designated in Equation 13–23.

† The **rate-limiting step** in the mechanism of a reaction is the last step in the sequence that exerts any influence on the overall rate.²³⁷ By this definition, all of the steps that follow the rate-limiting step must be so fast that they occur immediately relative to the passage through the rate-limiting step.

Bacillus stearothermophilus,¹² cytochrome c_2 from *Rhodobacter capsulatus*,²⁴¹ and human lysozyme.²⁴²

In addition to explaining the appearance of these kinetic intermediates as the concentrations of denaturant are decreased and providing further evidence for the existence of one or more intermediates in the folding of each of these polypeptides in the absence of denaturant, these observations of a change in the rate-limiting step provide a clue about the structures of the intermediates formed during a kinetic burst. Because the observed rate constants for their formation decrease significantly as the concentration of denaturant is increased while the observed rate constants for their conversion to the respective native states decrease much less significantly, each of these intermediates must be a more **compact, condensed state of the polypeptide** than the random coil. This follows from the proposal that the slope m of the line relating the logarithm of an observed rate constant for the folding of a polypeptide to the concentration of guanidinium chloride or urea (for example, the slopes of the linear segments in Figures 13-1B and 13-10) is a measure of the change in exposure of that polypeptide to the solvent between its initial state and the transition state of either the rate-limiting step in the transformation being monitored^{104,243} or of one or more of the rate-determining steps* that together establish the value of the composite rate constant²⁴⁴ for that transformation or the change in exposure of that polypeptide experienced during an unfavorable preequilibrium that precedes the rate-limiting step for that transformation. Consequently, either during the rate-limiting step or prior to the rate-limiting step of the transformation occurring during the kinetic burst in which the intermediate is formed from the random coil, a significant decrease in the exposure of the polypeptide to the solvent must occur.

Further evidence that these intermediates formed during a kinetic burst are compact, condensed forms of the polypeptide is provided by studies of their **scattering of X-radiation at small angles**. It is possible to measure small-angle scattering of X-radiation from a sample in the cuvette of a stopped-flow apparatus. When the intermediate formed from the polypeptide of apomyoglobin during the kinetic burst²²⁷ was examined in this way, it was found that the angular dependence of its scattering (Figure 12-2) was indistinguishable from that of the native state of the protein but clearly different from that of its unfolded random coil.²⁴⁵ This result indicates that most if not all of the condensation required to occur between the random coil and the native state must be accomplished within the kinetic burst. Similar results were observed for the folding of bovine β -lactoglobu-

lin.²⁴⁶ The rotational relaxation time of 1-anilidonaphthalene-8-sulfonate tightly bound to the intermediate formed in the kinetic burst during the folding of dihydrofolate reductase from *E. coli* is almost identical to that of the same probe bound to the native state, a result also suggesting that most if not all of the condensation of the polypeptide has already occurred in this isomerization.²⁴⁷ Increases in energy transfer by resonance between donors and acceptors covalently attached to particular positions in a polypeptide also indicate that it condenses during one of these isomerizations occurring in a kinetic burst.²²⁸

In addition to being compact, the intermediates formed during a kinetic burst contain **secondary structure**. This conclusion follows from the fact that large changes in molar ellipticity in the far ultraviolet, similar to those accompanying the formation of β structure and α helices (Figure 12-10), usually accompany the burst (Figure 13-9).^{226,227,231,233,248,249} A random coil has a slight positive **molar ellipticity** in the range from 210 to 230 nm while both β structure and α helices have significant, negative molar ellipticities (Figure 12-10), so the changes observed are decreases in molar ellipticity in this range. In fact, the molar ellipticity of the polypeptide of bovine β -lactoglobulin at 222 nm actually decreases to a level 2-fold lower than that of the native state during the kinetic burst before increasing to the proper value during the formation of the native state.²³² This result suggests that extra α -helical secondary structure is transiently forming in the intermediate and then disappearing as the native state forms.

The positions in the amino acid sequence that participate in the secondary structure formed in these intermediates can be defined by measurements of the **exchange of specific amido protons** from the peptide backbone. To perform such measurements, the unfolded polypeptide as a random coil in a concentrated solution of denaturant is passed in turn through a series of mixing chambers (Figure 13-11).²²⁷ In a typical experiment, the unfolded polypeptide in $^1\text{H}_2\text{O}$ is rapidly diluted into aqueous buffer prepared in $^2\text{H}_2\text{O}$ at a pH low enough to suppress proton exchange for the time being (Figure 12-31), and folding commences. After various millisecond intervals, during which the folding of the protein has progressed normally, the pH of the solution is increased, usually to a level greater than 9, in a second rapid mixing chamber to initiate the rapid and complete exchange of all amido protons still exposed to the deuterated solvent (Figure 12-31). The pH and duration of this period of rapid exchange are set so that it is long enough for exposed amido protons to exchange completely but not long enough for buried amido protons to exchange significantly. Finally, in a third rapid mixing chamber the pH is dropped again to slow the exchange and permit folding to be completed in the absence of further exchange.

In the final folded protein, most of the amido protons are well protected from further exchange by its sec-

* A **rate-determining step** in a reaction is any step the rate of which affects the rate of the overall reaction. In other words, if a step is rate-determining, an increase or decrease in its individual rate will cause a change in the overall rate of the reaction, but not necessarily of the same magnitude.

ondary and tertiary structure. Consequently it can be submitted to two-dimensional nuclear magnetic resonance spectroscopy (Figure 12–33) for the periods of time necessary to obtain two-dimensional spectra and determine which amido protons had become protected from exchange during the time spent folding before rapid exchange was initiated. Those positions in the amino acid sequence of the protein that have lost their protons during the experiment are those that were accessible when the jump in pH occurred; those that have not are those that had become protected. The times spent in the various steps and the levels of pH established in each step of these triple mixing experiments vary,^{233,250–252} but the intentions of initiating folding, of performing rapid exchange after folding has progressed for a certain period, and of then locking the information in the native state of the protein remain the same.*

Although most of the amido protons in an intermediate formed during a kinetic burst exchange rapidly during the respective jumps in pH, many are already protected from exchange by the structures of these intermediates.^{226,233,239,249,250} Because amido protons protected from exchange during the kinetic burst are found in segments within which each of a string of consecutive positions in the amino acid sequence is protected, it is assumed that these **segments of continuous protection** represent either α helices or β structure that have already formed in the intermediate.

Proteins during the folding of which intermediates do not accumulate in a kinetic burst nevertheless will often have similar intermediate states that form more slowly, in the milliseconds following the dead time. For example, an intermediate forms during the refolding of cytochrome *c* with a rate constant of 50 s^{-1} at 10°C that is compact and contains several elements of secondary structure but lacks the complete secondary and tertiary structure of the native protein.²⁵¹ These intermediates also have strings of consecutive amido protons protected from exchange.^{251,252}

It seems that some of the same **secondary structures found in the final native state** of the protein are already assembled in their entirety in these early kinetic intermediates formed either during a kinetic burst (Figure 13–11) or in the period immediately following the dead time. To a certain extent, this impression is illusory. Because the secondary structures in the native state are used to store the information about the protection that occurred upon the formation of the intermediate, this information is automatically divided into segments bounded by those elements of native secondary structure. Any information about the regions of the polypeptide outside of these segments of native secondary structure has been automatically erased. Amido protons

in these erased regions may have been protected in the intermediate but not in the native state. It is also possible to identify amido protons on particular amino acids that participate in hydrogen bonds in an intermediate formed in a kinetic burst by examining the effect of the concentration of denaturant on the exchange of amido protons within the native structure of a protein.²⁵³ Again, however, the fact that amido protons formed in the intermediate identified in this way coincide with elements of secondary structure in the crystallographic molecular model may be only a consequence of the fact that exchange from the native protein is being monitored.

The fact that all of the positions within a particular element of native secondary structure register similar levels of protection suggests but cannot prove that many of the same elements of secondary structure found in the native state are formed as discrete units early in the process of folding. There are, however, exceptions to this correspondence. For example, only a portion of the amino acids in helix B of apomyoglobin is protected from exchange in the intermediate formed during the kinetic burst (Figure 13–11).

Many if not most of the intermediates observed during kinetic bursts in stopped-flow experiments are similar if not **identical to stable molten globules** of the same polypeptide that are observed at equilibrium under unphysiological concentrations of denaturant, at unphysiological temperatures, at low pH, in the absence of salt, or with some combination of these perturbations. It has already been noted that, as is a molten globule, these kinetic intermediates are condensed conformations of the polypeptide. In addition, by varying the wavelength at which the kinetic measurements are made, it has been possible to demonstrate that the molar ellipticities at several wavelengths, both in the near ultraviolet and in the far ultraviolet, match the values in the **circular dichroic spectrum** of a stable molten globular state of the same polypeptide formed at equilibrium, usually under acidic conditions (Figure 13–12).^{226,232} Furthermore, the protection factors for the amido protons in the peptide backbone buried during the formation of an intermediate in a kinetic burst²³⁹ are often in the range of those observed for a molten globule at equilibrium rather than in the range of the much larger **protection factors** observed for amido protons locked within the secondary and tertiary structure of the native state. The effects of site-directed mutations of particular isoleucines to leucines and valines or particular leucines to isoleucines and valines in the hydrophobic core of the crystallographic molecular model of dihydrofolate reductase from *E. coli* were dramatically different on the yield of the intermediate observed in the kinetic burst than on the stability of the native state, a result suggesting that the packing of the native structure had not yet been established in the intermediate, as is the case in a molten globule.²⁵⁴

The most explicit evidence that these intermediates

* In most of these experiments the protein is unfolded in $^2\text{H}_2\text{O}$ and then folded in $^1\text{H}_2\text{O}$, and the gain of protons rather than their loss is monitored.

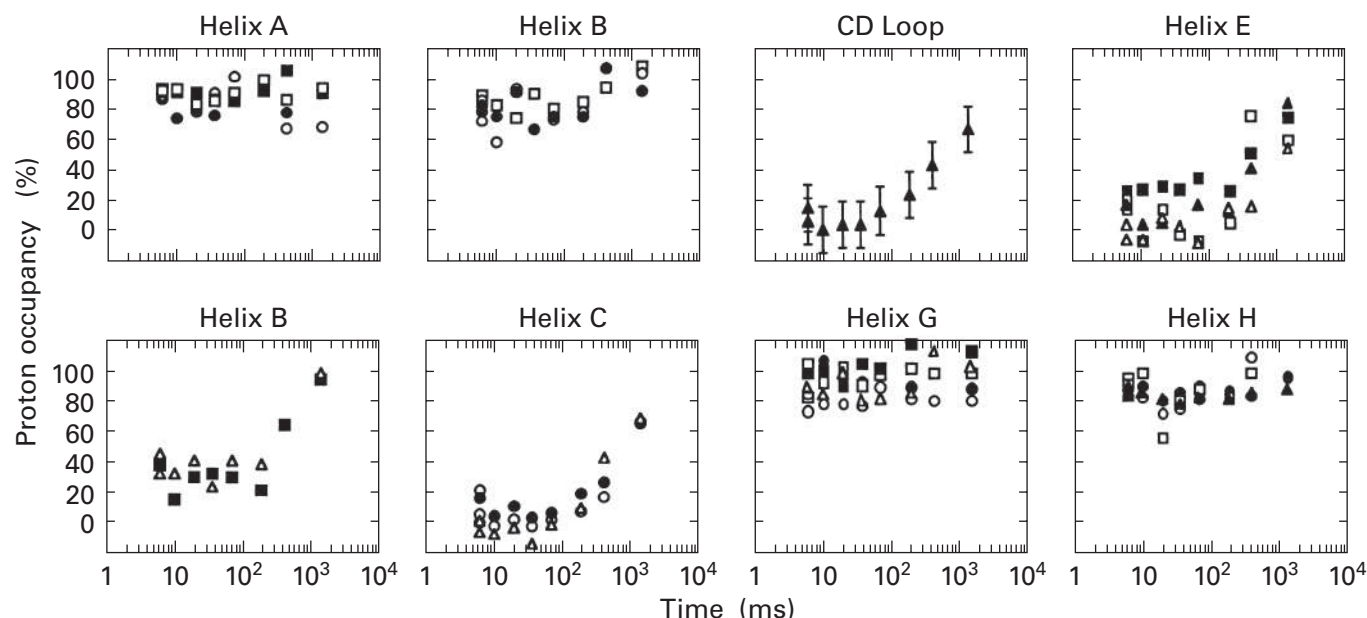


Figure 13-11: Sequestration of amido protons during the folding of apomyoglobin as followed by rapid mixing.²²⁷ A series of three rapid mixing chambers fed by four syringes were assembled in a cold room at 5 °C. Apomyoglobin from *P. catodon* was dissolved in 6 M urea and 10 mM sodium acetate, pH 6.1, in $^1\text{H}_2\text{O}$ and stood until it was fully unfolded. Flow through the mixing chambers was then initiated. The solution was diluted in the first mixing chamber 8.5-fold with 10 mM sodium acetate, pH 6.1, in $^2\text{H}_2\text{O}$. This mixture then travelled in the tubing connecting the first mixing chamber to the second for various periods of time (milliseconds). In the second rapid mixing chamber, the solution was mixed with an equal volume of a solution containing the buffers tris(hydroxymethyl)methylammonium ion, *N*-(ethylsulfonato)morpholinium ion, and acetate ion at a final ionic strength of 0.2 M, pH 10.2 in $^2\text{H}_2\text{O}$, which immediately brought the pH of the mixture to pH 10.2. This second mixture passed through a piece of tubing in which it spent 20 ms in transit to the third mixing chamber. In the third rapid mixing chamber it was diluted by mixing with a solution of the same ionic buffers at an ionic strength of 0.25 M, pH 1.9 in $^2\text{H}_2\text{O}$, that adjusted the final pH to 5.6. The effluent from the last mixing chamber was directed into a solution of hemin to turn the apomyoglobin to myoglobin and lock the protein in its native state. The amplitudes of the absorptions from the amido protons in two-dimensional nuclear magnetic resonance spectra (Figure 12-33) of the final solutions were determined for each sample. The different absorptions had been previously assigned to specific amido protons in the amino acid sequence of the protein.^{216,588} The percentage that each position in a set of representative positions was occupied (percentage occupancy) by a proton is plotted as a function of the time (milliseconds) the polypeptide was allowed to fold before the protons were exchanged with deuterons at pH 10.2. The amido protons are grouped according to the secondary structure they occupy in the crystallographic molecular model of myoglobin.^{216,588} The eight α helices in the crystallographic molecular model are designated in alphabetical order from the amino terminus and the CD loop is the segment of random meander between helix C and helix D. Occupancies of the amides of Leucine 29, Isoleucine 30, and Phenylalanine 33 from helix B are plotted in the upper panel; those of Isoleucine 28 and Arginine 31 from helix B are shown in the lower panel. Reprinted with permission from ref 227. Copyright 1993 American Association for the Advancement of Science.

formed during the kinetic bursts are similar if not identical to well-characterized molten globules of the same polypeptide observed at equilibrium is the **correspondence in the specific amido protons protected** during their formation with the respective amido protons protected in the respective molten globule. For example, in the intermediate formed during the kinetic burst in the folding of apomyoglobin, amido protons in positions in the amino acid sequence that form the first (A), a portion of the second (B), the seventh (G), and the eighth (H) α helices in the native state of the protein²¹⁶ are protected, but those that form the rest of the second, the third (C), and the fifth (E) α helices are not protected (Figure 13-11). This is the same pattern of protection observed in the molten globule of this polypeptide that is the dominant state at equilibrium between pH 4 and 5.⁶⁵ Likewise, in the intermediate formed in the kinetic burst in the folding of the cysteineless version of type I

ribonuclease H from *E. coli*, the pattern in which protected amido protons are distributed over its sequence of amino acids²²⁶ closely matches the pattern in which those amido protons are protected in the molten globule that predominates at equilibrium at levels of pH less than 2.²⁵⁵ When amino acids in regions that have been observed to be protected in the intermediate formed during the kinetic burst are mutated, the yield of the intermediate decreases; but when amino acids that have not been observed to be protected are mutated, the yield of the intermediate is unaffected.²⁵⁶

Apomyoglobin and type I ribonuclease H both seem to form an intermediate that already contains some but not all of the specific secondary structures that will eventually end up in their native states. The kinetic intermediates of β -lactoglobulin, however, which is also a molten globule, displays a circular dichroic spectrum indicating that it contains significant amounts of α helix

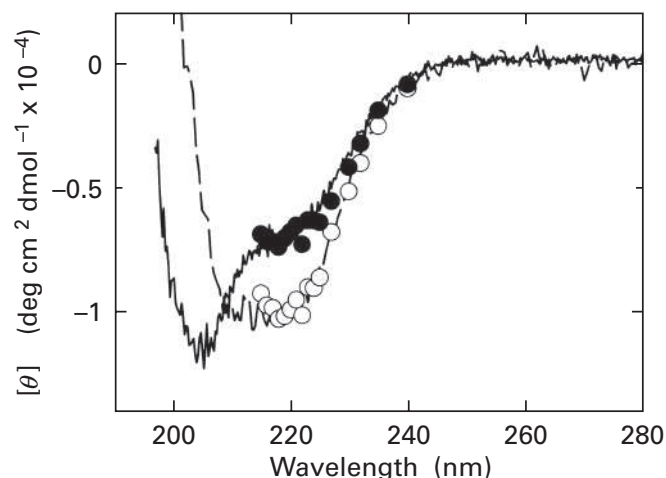


Figure 13-12: Circular dichroic spectrum of the kinetic intermediate observed during the kinetic burst in the folding of a cysteineless version of type I ribonuclease H from *E. coli*.²²⁶ Folding was initiated by an 11-fold dilution of the unfolded polypeptide in 7 M urea, and the kinetics of refolding (Figure 13-9) were monitored by following molar ellipticity at 16 different wavelengths, each in a separate run. The molar ellipticity [degrees centimeter² (decimole of peptide bonds)⁻¹] observed at the dead time (burst amplitude; ●) in each run is plotted as a function of the wavelength (nanometers) set for that run. The molar ellipticity observed at the end of each run (final folded state; ○) is also plotted as well as the circular dichroic spectrum of a molten globule of the protein that forms at equilibrium at pH 1.0 (solid line) and the circular dichroic spectrum of the native protein (dashed line). Reprinted with permission from ref 226. Copyright 1997 Nature Publishing Group.

even though only 12% of its native state is α helical.^{232,246} Consequently, in this instance, the intermediate contains at least some secondary structure that will not be present in the final native state.

That stable molten globules of the same polypeptides at equilibrium are similar if not identical to the respective kinetic intermediates formed during kinetic bursts means that the former should be valid models for the latter. From physical characterizations of these equilibrium states at rest, a much better understanding of the structure and dynamics of the kinetic intermediates on the move can be gained. The details of the physical properties of molten globules at equilibrium have already been discussed.

With some proteins, there are intermediates formed during a kinetic burst that are not molten globules. For example, the polypeptide of lysozyme from *G. gallus* isomerizes during the kinetic burst to an intermediate that has a radius of gyration intermediate between that of the random coil and that of the native state.²⁵⁷ A molten globule of this polypeptide would have a radius of gyration much closer to that of the native state. In the folding of dihydrofolate reductase from *E. coli*, little of the accessible surface that is eventually buried in the native state seems to be buried in the kinetic intermediate formed during kinetic burst.²⁵⁸ This

observation, however, seems to be contradicted by the fact that the rotational correlation time of this intermediate is indistinguishable from that of the native state, and therefore it should be as compact as the native state.²⁴⁷

In an observation of folding by stopped-flow, the formation of one of these kinetic molten globular intermediates usually takes place within the dead time of the apparatus. The dead time of such an observation, however, can be shortened by monitoring **continuous flow** through a rapid mixing chamber rather than stopped-flow. A transparent tube is attached directly to the mixing chamber, and the two fluids being mixed flow continuously at a high velocity through both the chamber and the tube. The fluorescence of the sample passing through the tube is monitored as a function of the distance along the tube, and hence the time spent in the tube, following mixing. In this way, events can be followed from about 100 μ s to 2 ms. The drawback is that, because of turbulence,²⁵⁹ only emission of light from the sample rather than absorption of light can be measured at times less than 500 μ s. The purpose of most of these experiments, however, is to monitor the formation of intermediates that have already been characterized extensively by stopped-flow observations and triple mixing experiments and to determine whether the step during which they are formed is preceded by yet an earlier step.

The rate of formation of a molten globular intermediate that appears during a kinetic burst in stopped-flow can often be resolved in continuous flow. For example, the observed rate constants for the formation of molten globular intermediates in the folding of bovine β -lactoglobulin,²⁶⁰ the B1 domain of immunoglobulin G binding protein G from *Streptococcus*,²⁶¹ intestinal fatty acid-binding protein from *Rattus norvegicus*,²⁶² and colicin E7 immunity protein from *E. coli*²⁶³ are 7000 s⁻¹ (20 °C), 2300 s⁻¹ (20 °C), 1500 s⁻¹, and 3000 s⁻¹ (10 °C), respectively. The similarity in the values of all of these observed rate constants may have more to do with the range over which rate constants can be measured by continuous flow than a similarity in the process being measured.

In the case of β -lactoglobulin, the B1 domain, and immunity protein, the isomerizations of the random coil producing the molten globular intermediates at these respective rates appear to be single first-order relaxations that are not preceded by any other kinetic burst, so the transformation of random coil to molten globular intermediate appears to proceed in one step with no prior kinetic intermediates. This conclusion follows from the fact that the kinetic traces of fluorescence extrapolate through the dead times (100–150 μ s) to the value for the fluorescence of the random coil at the final concentration of denaturant. Consequently, in these cases, the random coil appears to isomerize directly to the molten globule. In the case of fatty acid-binding protein, however, the extrapolation did not coincide with the fluores-

cence of the random coil, and the extrapolated values at different wavelengths of emission produced a spectrum for an additional intermediate formed in the kinetic burst that was distinct from the spectrum of the unfolded state. It follows that, in this case, at least one other intermediate precedes the molten globular intermediate.

Events that occur within even shorter intervals can be monitored by **temperature jump**. This approach exploits the fact that at a low temperature, the stability of a protein increases as the temperature is raised (Figure 13–5). A solution of protein in a concentration of denaturant within the region of transition is brought to a low temperature, which increases the concentration of the denatured state at the expense of the folded state. The temperature of the solution is then jumped by the rapid application of heat, and the solution stabilizes at a higher temperature within a hundred nanoseconds. The approach to the new equilibrium now favoring the folded state is then monitored. When the folding of equine cytochrome *c*²⁶⁴ and barstar from *B. amyloliquifaciens*²⁶⁵ are examined after a temperature jump of 10 °C, relaxations with rate constants of 11,000 s^{−1} and 3000 s^{−1} were observed, similar to those observed for other proteins by continuous flow. A much faster relaxation with a rate constant of 200,000 s^{−1} at 10 °C is observed for apomyoglobin, and this rate constant is significantly affected by the viscosity of the solvent, an observation suggesting that it represents the collapse of the denatured state,²⁶⁶ and a relaxation with a similar rate constant has been observed for the folding of bovine ribonuclease A.²⁶⁷ That these very rapid relaxations, however, monitor the initial global collapse of the random coil has been questioned,²⁶⁸ and it is possible that the initial collapse of these proteins usually occurs much more slowly, with rate constants in the range below 10,000 s^{−1}.

When it is dissolved in 0.01 M HCl, equine cytochrome *c* is in a denatured state that is not a random coil but is at least as expanded.^{269,270} Upon jumping of the pH to 4.0, the protein refolds from this expanded state. The refolding can be followed from 45 μs to 1 ms by continuous flow (Figure 13–13).²⁷⁰ It appears to pass through two clearly resolved steps with observed rate constants of 17,000 s^{−1} and 2300 s^{−1}. No kinetic burst is observed. During the first step 60% of the fluorescence from Tryptophan 59, the only tryptophan in the protein, is quenched by the covalently attached heme as the expanded conformation collapses, but no secondary structure forms beyond the residual α helix found in the expanded denatured state.²⁵⁹ During the second step, about 70% of the α-helical content of the native state is regained, to produce a molten globule.

What seems to be the same second step, involving the formation of the same level of α-helical content, occurs more slowly (50 s^{−1} at 10 °C) when the random coil in 4.4 M guanidinium chloride is diluted to 0.7 M guanidinium chloride, so it is possible to determine by rapid

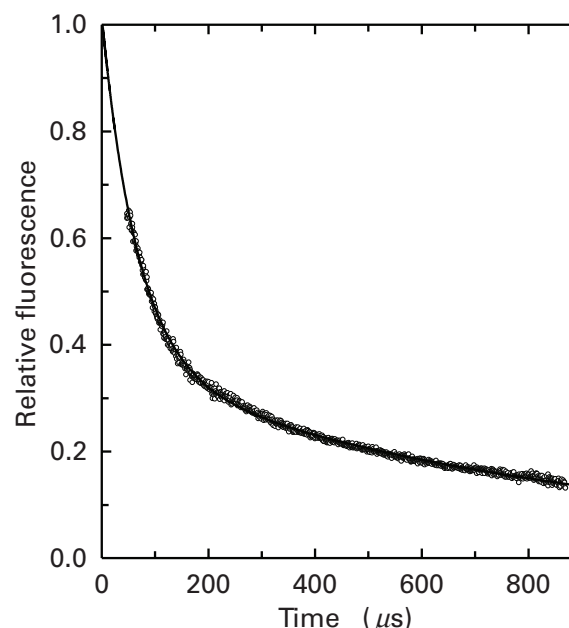


Figure 13–13: Folding of cytochrome *c* monitored by continuous flow.²⁷⁰ Equine cytochrome *c* was dissolved in 0.01 M HCl and deionized by molecular exclusion chromatography in 0.01 M HCl. It was then mixed in a rapid mixing chamber with 10 volumes of 50 mM sodium acetate and 50 mM sodium phosphate, pH 5.1. The final pH of the mixture was 4.5. The effluent from the mixing chamber was passed through a 0.25 mm × 0.25 mm channel in a quartz block at 0.62 mL s^{−1} (0.99 μm μs^{−1}). The block was illuminated with light at 280 nm wavelength, and fluorescence emission at a wavelength greater than 324 nm was measured as a function of the distance along the channel. The fluorescence relative to that of the denatured polypeptide in 0.01 M HCl (1.0) and the folded native state at pH 4.5 (0.0) is presented as a function of time (microseconds). The data are fit with the solid curve, which is the sum of two first-order exponentials with rate constants of 17,000 s^{−1} and 2300 s^{−1} and amplitudes of 0.60 and 0.29. The dead time of the apparatus was measured directly and found to be 45 μs. Reprinted with permission from ref 270. Copyright 1998 Nature Publishing Group.

proton exchange that both the amino- and carboxy-terminal α helices of the native structure form during this step, but not the α helices between positions 60 and 80 in the amino acid sequence (Figure 7–9).²⁵¹ The molten globule formed during this second step, however, displays none of the molar ellipticity at 420 nm indicative of the asymmetric environment of the native structure surrounding the heme.²⁴⁸ Finally the native state arises in a biphasic process with rate constants of 2.5 s^{−1} and 0.25 s^{−1} at 10 °C²⁵¹ or 8 s^{−1} and 0.8 s^{−1} at 25 °C.²⁴⁸

The expanded denatured state of equine cytochrome *c* in 0.01 M HCl has an α-helical content that is 20% that of the native state,²⁵⁹ and although the α-helical content does not increase during its collapse, the earliest observed collapsed intermediate does contain this amount of α helix. Moreover, when the folding is performed by diluting the random coil of cytochrome *c*, which contains no α helix, from 4.4 to 0.4 M guanidinium chloride, there is still no kinetic burst and all of the

change in fluorescence at times less than 1 ms can be resolved into two steps with rate constants of $21,000\text{ s}^{-1}$ and 730 s^{-1} at $22\text{ }^{\circ}\text{C}$.²⁷⁰ The initial collapsed state, however, again has 20–30% of the α -helical content of the native state.^{230,259}

These observations raise the question of whether or not the polypeptide of a protein is able to collapse hydrophobically in an isomerization that involves no formation of secondary structure. Is there a purely **hydrophobic collapse**? Certainly, all of the condensed kinetic intermediates observed during the folding of polypeptides, when they are assayed for secondary structure, do contain it. Furthermore, the fastest events in protein folding involving condensation of the random coil to form a globular state usually have rate constants of less than $20,000\text{ s}^{-1}$, often much less. For example, in the case of the immunoglobulin binding domain of protein L from *Peptococcus magnus*, the condensation of the random coil to a globular state²⁷¹ occurs in a first-order reaction with a rate constant of only 0.12 s^{-1} . Yet there are measurements indicating that the purely hydrophobic collapse of a polypeptide should have a rate constant of at least 10^6 s^{-1} at $20\text{ }^{\circ}\text{C}$,²⁷² and theoretical treatments²⁶⁴ suggest that the rate constant should be 10^7 s^{-1} . Furthermore, a purely hydrophobic collapse should have no energy of activation, but the observed relaxations assigned to the collapses of denatured states do.

Explanations are required for both the slower than expected rate constants observed for these condensations and the fact that most if not all of the initial condensed states contain significant **secondary structure**. Even the extremely rapid relaxation of apomyoglobin with a rate constant of $200,000\text{ s}^{-1}$ observed by temperature jump nevertheless seems to involve an intermediate with significant secondary structure.^{266,268}

Suppose that kinetically, a purely hydrophobic collapse occurs before the formation of any of the secondary structure characteristic of a molten globule, and that the mechanism for folding is



where C is the hydrophobically collapsed state and MG is the subsequent molten globule. If both steps were intrinsically fast reactions, if k_2 were greater than k_1 , and if k_{-1} were greater than k_2 , the reactions would be coupled, and little purely hydrophobically collapsed state would be observed. It has just been noted, however, that k_1 , the hydrophobic collapse, should be much faster than the observed rate for the overall formation of the molten globule.

If $k_{-1} \gg k_2$, then the unfolded state and the hydrophobically collapsed state are in a **rapid preequilibrium** that precedes the step in which the secondary structure, which distinguishes the molten globule from

the hydrophobically collapsed state, is formed. The rate equation for the formation of the molten globule from the unfolded state by this mechanism is

$$[\text{MG}] = [\text{protein}]_{\text{TOT}} \left[1 - \exp \left(- \frac{k_2 K_{\text{cpse}}}{1 + K_{\text{cpse}}} t \right) \right] \quad (13-26)$$

where $[\text{protein}]_{\text{TOT}}$ is the total concentration of protein and K_{cpse} is the equilibrium constant for the hydrophobic collapse:

$$K_{\text{cpse}} = \frac{[\text{C}]}{[\text{U}]} = \frac{k_1}{k_{-1}} \quad (13-27)$$

Equation 13-26 defines a first-order formation of the molten globule with an observed rate constant

$$k_{\text{obs}} = \frac{k_2 K_{\text{cpse}}}{1 + K_{\text{cpse}}} \quad (13-28)$$

Upon initiation of the reaction, the equilibrium between the unfolded state and the hydrophobically collapsed state would be established immediately, and the molten globule would appear in a kinetically first-order reaction. If the equilibrium constant between the unfolded state and the hydrophobically collapsed state is less than 1, no purely hydrophobically collapsed state should be observed, and none is.

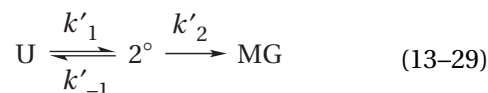
It is reasonable that this equilibrium constant should be less than 1. Few if any polypeptides should be able to bury enough hydrogen-carbon bonds upon their hydrophobic collapse to overcome both the unfavorable loss of the configurational entropy of the random coil and the unfavorable loss of solvation arising from the inescapable transfer of donors and acceptors of hydrogen bonds from the water into the interior of the collapsed state. Certainly the small values for standard free energy of folding (Table 13-2) suggest that this must be the case. Many of these unbonded donors and acceptors buried during the collapse, however, do become occupied within the secondary structure of the molten globule when it forms. These buried hydrogen bonds within α helices and β structure of the molten globule, because they are formed in the absence of water, stabilize it relative to a hydrophobically collapsed state lacking any internal hydrogen bonds. Consequently, the molten globule with its characteristic secondary structure can be the first intermediate observed even though the mechanism of folding passes obligatorily through a hydrophobically collapsed state lacking any secondary structure.

This explanation, however, is inconsistent with the observation that the reciprocals of the observed rate con-

stants for the formation of the molten globular intermediates seem to be linearly related to the **viscosity of the solvent**.²⁷³ Because it is a state function, the equilibrium constant for collapse cannot be affected by the viscosity of the solvent. Because neither K_{cpse} nor k_2 should be affected by viscosity, the rate-limiting step in the formation of these intermediates cannot be the formation of secondary structure within an already collapsed state. It could, however, be the collapse of an expanded state already containing sufficient secondary structure to stabilize the collapsed state, much as the expanded acid-denatured form of cytochrome *c* with 20% α helix collapses upon a jump in pH²⁵⁹ or as the expanded cold-denatured form of barstar, which also contains significant residual secondary structure,⁹⁵ collapses upon a jump in temperature. If this is the mechanism for the collapse of an expanded denatured state, the secondary structure that will stabilize the molten globule relative to the hydrophobically collapsed state forms first within the random coil and is then trapped by the collapse of this structured denatured state to form the molten globule directly.

There are indications that, in the unfolded state of a protein in solutions of guanidinium ion or urea, **metastable segments of secondary structure** form and dissolve continuously even though they are not present at significant concentrations.⁹⁵ During the folding of bovine acyl-CoA-binding protein at 0.5 M guanidinium chloride,²⁷⁴ an intermediate forms during a kinetic burst in which considerable protection is afforded to the exchange of amide protons, but this protection seems to result from the formation of a set of relatively stable conformations of the uncollapsed polypeptide that contain elements of secondary structure. It is difficult, however, to determine just how uncollapsed these conformations are. Clusters of secondary structure in either evanescent or more stable conformations of the uncollapsed random coil could be trapped during its collapse to form a molten globule directly. A random coil is an ensemble of conformations in rapid equilibrium with each other, and subsets of those conformations may contain elements of secondary structure waiting to be enclosed.

In this case, the preequilibrium would be not hydrophobic collapse (Equation 13–25) but the formation of these fleeting unstable elements of secondary structure:



where 2° is any subset of the conformations of the random coil containing secondary structure extensive enough and appropriately located to support the stable formation of a molten globular intermediate. The formation of this molten globule would be a first-order reaction (Equation 13–26). If $k_{-1}' > k_2'$, then the observed rate constant for the mechanism of Equation 13–29 is

$$k_{\text{obs}} = \frac{k'_2 K_{2^\circ}}{1 + K_{2^\circ}} \quad (13-30)$$

where K_{2° is the equilibrium constant for the formation of the secondary structure within the random coil the presence of which is required before collapse can occur. No secondary structure would be observed in the random coil because the equilibrium constant for its formation, K_{2° , would be significantly less than 1. If $K_{2^\circ} \ll 1$, then the observed rate constant is $K_{2^\circ} k_2'$. The formation of the molten globular intermediate would exhibit an apparent energy of activation that is the sum of the actual standard free energy of activation for the reaction governed by rate constant k_2' and the change in standard free energy for the reaction governed by the equilibrium constant K_{2° for the preequilibrium in which the secondary structure is formed.

The decision between a mechanism in which hydrophobic collapse precedes any formation of secondary structure (Equation 13–25) and a mechanism in which the formation of sufficient secondary structure to stabilize the collapsed state precedes hydrophobic collapse (Equation 13–29) depends on the interpretation of the effects of solutes that increase the viscosity of the solvent on the observed rate constants for the formation of the molten globular intermediates. If the effects of these solutes are on the stability of intermediates in the process of folding²⁷⁵ rather than exclusively on the viscosity,²⁷³ this decision is ambiguous. It is often impossible to distinguish these two possibilities experimentally.

The transition between one of the molten globular intermediates formed during a kinetic burst and the final native state of a protein appears to occur in one or several consecutive kinetic steps. For example, the molten globular kinetic intermediates of apomyoglobin,²²⁷ intestinal fatty acid binding protein,²⁶² and the B1 domain of protein G²⁶¹ become the respective native state in apparently single first-order steps with rate constants of 1 s^{-1} (5°C), 5 s^{-1} , and 600 s^{-1} (20°C). The molten globular intermediate of type 1 ribonuclease H becomes the native state in what appears to be a single first-order step with a rate constant of 0.6 s^{-1} at 25°C when monitored by molar ellipticity at 220 nm and 292 nm but an additional faster step of 2 s^{-1} is detected by fluorescence (Figure 13–9). This latter result illustrates the fact that some steps in the folding of a protein go unregistered by certain physical measurements.

The apparently single steps that occur following the rapid formation of an intermediate during a kinetic burst and that in turn produce the native state often seem to involve only a portion of the entire protein. The **folding of the rest of the protein** must occur either during the kinetic burst or during rapid unregistered steps following the rate-determining steps and the rate-limiting step that produce this slow observed rate constant. For example, some site-directed mutations of ribonuclease from

B. amyloquifaciens affect the observed rate constant for the production of the native state from the intermediate formed during the kinetic burst, while others affect the stability of that intermediate.²⁷⁶ The observed rate constant (20 s^{-1} at 20°C) for the principal relaxation observed in both molar ellipticity and fluorescence following the kinetic burst during the folding of lysozyme from bacteriophage T4 is affected significantly by site-directed mutations in the carboxy-terminal half of the polypeptides but not by mutations in the amino-terminal half even though the mutations in both halves affect the stability of native state.²⁷⁷ Presumably the process being registered as an apparent single step by both circular dichroism and fluorescence is the folding of only the one half of the protein and not the other.

It is more common to observe two or more steps rather than just one during the transition between a kinetic molten globular intermediate and the native state. For example, four additional steps can be discerned following the formation of the first kinetic intermediate during the folding of β -lactoglobulin;²⁶⁰ two, during the folding of human lysozyme;²⁴² two, during the folding of cytochrome *c*;^{230,251} and two, during the folding of micrococcal nuclease from *S. aureus*.²⁷⁸ Again, the number of phases detected often depends on how many spectral properties have been monitored. For example, five additional steps in the refolding of dihydrofolate reductase from *E. coli* were discerned following the formation of a molten globular intermediate if molar ellipticity at 220 nm, molar ellipticity at 235 nm, absorbance, and intrinsic fluorescence were all monitored.²³¹

Some if not most of these **multiple steps** may each involve the assembly of a particular portion of the final secondary structure of the native state.^{278,279} Intermediates formed after the formation of the initial molten globule, however, often have all of the secondary structure of the native state, but some of that secondary structure is in a less stable state than it will eventually assume upon complete folding,^{230,251} as if the rigid conformation of the native structure maintained by the proper packing of the secondary structures locks in place only in the final step or one of the final steps in the process. The last or almost the last property to appear in the folding of a protein is the spatial arrangement of the constellation of side chains responsible for its function.^{258,280}

The order in which different elements of secondary structure form and lock into place during the transition from the initial molten globular intermediate to the final native state can be inferred from studies of **native-state proton exchange**. The amido protons of the peptide bonds buried in the native state of a protein and defining its secondary structure exchange with deuterons in the solvent at different rates. When standard free energies for the conformational equilibria leading to their exposures are followed as a function of the concentration of a denaturant, it is observed that they generally coalesce

into groups as the concentration of denaturant is increased (Figure 13–14).^{116,117} These groups are groups of amido protons involved in particular secondary structures in the native protein. For example, in the crystallographic molecular model of cytochrome *c* (Figure 7–9), Arginine 91 through Lysine 100 (Figure 13–14A) are within one α helix; Methionine 65 through Asparagine 70 (Figure 13–14B) are within another α helix; and the ϵ amido proton of Tryptophan 59 and the α amido protons of Leucine 64, Lysine 60, Phenylalanine 36, and Glycine 37 are within a cluster of adjacent hydrogen bonds (Figure 13–14C). The coalescence of the respective sets of standard free energies of exposure indicate that each of these elements of secondary structure opens to exchange its amido protons cooperatively.

There are indirect observations suggesting that these openings of the structure, which are occurring all the time in the native protein, actually occur sequentially.²⁸¹ For example, the cluster of hydrogen bonds involving the α amido proton of Lysine 60 must open before the α helix containing Leucine 68, which must open before the α helix containing Leucine 98 during the exchange of the amido protons in the latter α helix. Furthermore, as was the case with the exchange of the amido proton at Methionine 47 in cysteineless type I ribonuclease H from *E. coli* (Figure 13–8), the exchange of the amido protons of the secondary structure that is the last to open, namely the α helix containing Leucine 98 in cytochrome *c*, tracks the global unfolding of the protein.

All of these observations suggest that the exchange of amido protons in the native state of a protein reveal, in reverse, the **steps in the formation of secondary structure** and the **locking in of that secondary structure** as the elements pack next to each other during the normal folding of the protein. In other words, during the folding of cytochrome *c*, the α helix containing Leucine 98 forms before that containing Leucine 68, which forms before the cluster of hydrogen bonds involving the α amido proton of Lysine 60.

The **slowest steps** in the folding of a protein from its random coil are often the **isomerizations of peptide bonds to the amino-terminal side of prolines**. In the crystallographic molecular models of proteins, about 6% of the peptide bonds on the amino-terminal sides of proline are *cis* peptide bonds^{282,283} and the rest are *trans* peptide bonds (Equation 6–1). These are geometric isomers of each other. The peptide bond on the amino-terminal side of a proline at a particular position in the amino acid sequence of a particular protein usually will be either *cis* in every molecule of the native state or *trans* in every molecule of the native state. In the random coil, however, every proline is free to adopt either geometric isomer, and the *cis* and *trans* isomers slowly come to equilibrium. In dipeptides, the equilibrium constants between *cis* and *trans* isomers of proline vary with pH, but they fall between 10 and 1.5 in favor of the *trans* isomer. The more

prolines that must be one or the other isomer before the native state can be achieved, the more random coils with at least one incorrect isomer of proline will be present in the solution.

When bovine ribonuclease A is added to 5 M guanidinium chloride at pH 2.3, the unfolding of the polypeptide is very rapid (<10 s), and the unfolding produces a random coil, cross-linked by its four cysteines.¹³¹ If the solution containing this random coil is diluted within 15 s to 1.3 M guanidinium chloride, pH 6.4, at 25 °C, all of the polypeptide (> 95%)¹³¹ refolds to the native state, capable of full enzymatic activity,²⁸⁴ in an uncomplicated first-order relaxation²⁸⁵ with a rate constant of about 10 s^{-1} . This result demonstrates that a random coil that a moment ago was native ribonuclease can refold rapidly, and with no obvious complications, back into native ribonuclease.

If rapidly unfolded ribonuclease, however, is allowed to sit as a random coil in a solution of guanidinium chloride over a period of 10 min, the kinetics of refolding are split into several phases, one with the same observed rate constant as that of the initially produced random coil and several others that are much slower. Consequently, a portion of the initially produced random coil that all refolds rapidly has isomerized to random coils that refold slowly. At equilibrium, the rapidly folding isomer of the random coil accounts for 20% of the protein; the **slowly folding isomers of the random coil**, for 80%,²⁸⁴ and the rate constant for the approach to this equilibrium between the rapidly folding isomers and the slowly folding isomers is 0.005 s^{-1} at 25 °C.¹³¹

Because the π system of the amide (Figure 2–3) must be broken for conversion to occur between the *cis* and *trans* geometric isomers of a peptide bond, this conversion is a slow process. The rate constants for the approach to the equilibrium between the *cis* and *trans* isomers of a set of dipeptides containing a carboxy-terminal proline are slow, between 0.002 s^{-1} and 0.005 s^{-1} at 25 °C above pH 3,^{286,287} rates that are similar to observed rate constant for the approach to equilibrium of the rapidly and slowly folding forms of ribonuclease.

In addition to this similarity of rates, there are several properties of the transitions between the rapidly

folding form and the slowly folding forms of the random coil of ribonuclease consistent with them being due entirely to isomerization of peptide bonds on the amino-terminal sides of prolines in the sequence. The observed rate constant for the approach to the equilibrium between the rapidly folding form and slowly folding

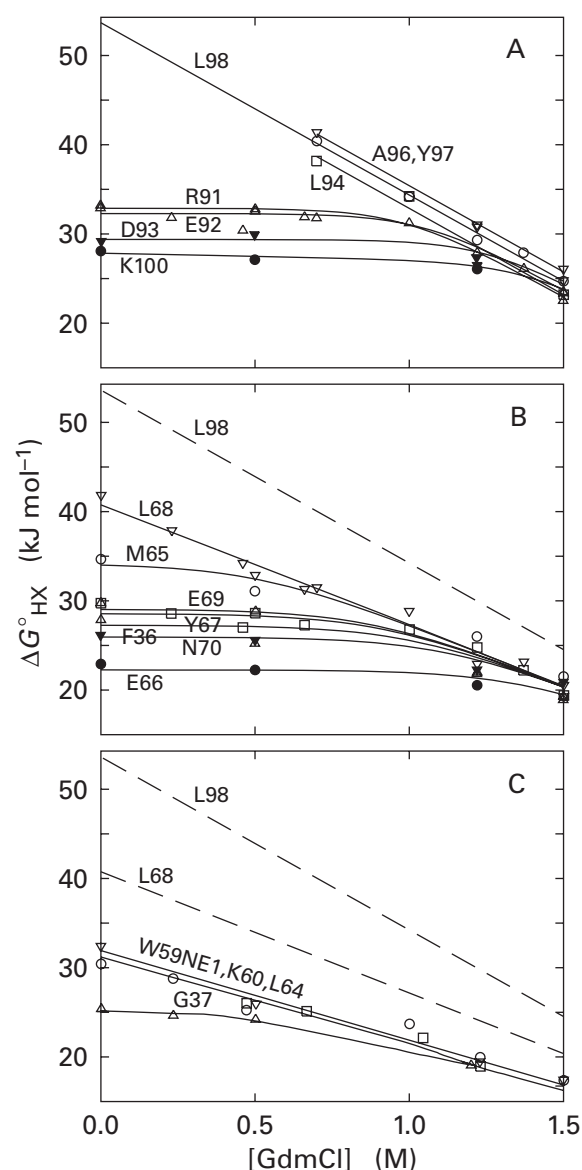


Figure 13–14: Rates of exchange of particular amido protons along the polypeptide backbone of native equine cytochrome *c* as a function of the concentration of guanidinium chloride.¹¹⁶ Equine cytochrome *c* was dissolved at p²H 7 in ²H₂O at the noted concentrations of guanidinium chloride. After different intervals, samples were removed, the pH was adjusted to 5 to slow the rates of exchange, and a two-dimensional nuclear magnetic resonance spectrum was gathered. The amplitudes of each of the peaks in each of the respective spectra were tabulated as a function of the time spent at p²H 7 in ²H₂O before the pH was lowered, and rates of exchange were calculated from the decreases in these amplitudes as a function of time. Each exchange was at the EX₂ limit, and the equilibrium constant for the formation of the conformation exposing each amido proton at each concentration of guanidinium chloride was calculated (Equation 12–63). From these equilibrium constants, standard free energies for the exposures of each proton, $\Delta G^\circ_{\text{HX}}$, were calculated. These standard free energies of exposure (kilojoules mole⁻¹) are plotted as a function of the concentration (molar) of guanidinium chloride (GdmCl). The standard free energies of exposure coalesced into specific groups as the concentration of guanidinium chloride was raised. Separate plots for three of these groups are presented: (A) the amido protons in the α helix containing Arginine 91 (R91) to Lysine 100 (K100); (B) the α helix containing Methionine 65 (M65) to Asparagine 70 (N70); and (C) a cluster of hydrogen bonds containing the ϵ amido proton of Tryptophan 59 and the α amido protons of Leucine 64 (L64), Lysine 60 (K60), Phenylalanine 36 (F36), and Glycine 37 (G37). In each successive panel, the lines for the most extensively protected amido protons from the previous plots are drawn as a dashed line to identify the discrete groups. Reprinted with permission from ref 116. Copyright 1995 American Association for the Advancement of Science.

forms of the random coil²⁸⁸ and the observed rate constant for the formation of the native state from the slowly folding forms of the random coil of ribonuclease²⁸⁹ are both **increased by strong acid**, as are the rate constants for the *cis*–*trans* isomerization in dipeptides of proline.²⁸⁶

The **standard enthalpy of activation** for the approach to the equilibrium between the rapidly folding form and the slowly folding forms of the random coil is between 75 and 90 kJ mol^{−1}, either at low pH or in 5 M guanidinium chloride,^{288,290} which compares favorably to the values for the standard enthalpy of activation (80–90 kJ mol^{−1}) for the *cis*–*trans* isomerization of dipeptides of proline.²⁸⁶ Neither the rate of the approach to *cis*–*trans* equilibrium of dipeptides of proline nor the approach to the equilibrium between the rapidly folding form and the slowly folding forms of the random coil is affected by the concentration of guanidinium chloride.²⁹¹

In the crystallographic molecular model of bovine ribonuclease A, the peptide bonds on the amino-terminal sides of Proline 93 and Proline 114 are *cis* peptide bonds. The amount of the peptide bond amino-terminal to Proline 93 that is in the *cis* form in the random coil can be monitored by its insensitivity to endopeptidolytic cleavage by Xaa-Pro dipeptidase.²⁸⁷ In 8.5 M urea at 10 °C, 70% of this peptide bond is *cis*. When the urea is diluted to 0.3 M, the 30% of this peptide bond that is *trans* slowly and completely reverts to the *cis* isomer with a rate constant of 0.01 s^{−1}, as the polypeptide folds. Under these conditions, 30% of the random coil refolds in the slowest phase and 30% of the activity of the enzyme is regained in this slowest phase, both with a rate constant of 0.01 s^{−1}. Proline 93 is preceded by Tyrosine 92, and the fluorescence from this tyrosine tracks the slow isomerizations that produce fully native enzyme during the refolding of the equilibrated random coil after a decrease in the concentration of guanidinium chloride.²⁸⁹ It was presumed that the slow process monitored by the fluorescence of Tyrosine 92 is the state of isomerization of the peptide bond between Tyrosine 92 and Proline 93.

If both Proline 93 and Proline 114 are mutated, the former to an alanine and the latter to a glycine, the stability of the native protein is decreased significantly, but its random coil is still able to fold to produce a protein that is enzymatically active.²⁹² The folding of this double mutant, when monitored by molar ellipticity, is a single first-order reaction with a rate constant of 0.07 s^{−1} in 0.4 M guanidinium chloride at 10 °C with no evidence for any slower phase.

It has been concluded from all of these observations that all of the slow isomerizations of the random coil of bovine ribonuclease A that in turn produce the slowly folding forms from the rapidly folding form are isomerizations of peptide bonds on the amino-terminal sides of prolines from the isomer found in the native state and that the most disruptive isomerization of the random coil is to a *trans* proline at position 93. When the equilibrium mixture of rapidly folding and slowly folding

random coils of ribonuclease is diluted into conditions favorable to folding, some of the slowly folding random coils assume **nativelike conformations** in which the critical peptide bonds on the amino-terminal sides of prolines are nevertheless the incorrect isomer. One intermediate, I₁, is sufficiently folded to trap almost 20 amido protons in stable hydrogen bonds,²⁹³ and another, I_N, is compactly folded by several criteria,²⁹⁴ including its insensitivity to digestion by pepsin.²⁹⁵ These compact intermediates, however, differ from the native state and can be distinguished from it by having the incorrect isomers at particular prolines.²⁸⁹

A similar but more dramatic effect of the slow isomerizations of prolines is observed in the folding of ribonuclease T₁ from *A. oryzae*. In the crystallographic molecular model of this even shorter protein (104 aa), the peptide bonds preceding both Proline 39 and Proline 55 are *cis*.²⁹⁶ If the native protein is unfolded in 6.0 M guanidinium chloride, pH 1.6, and the resulting random coil is diluted after 5 s to 1.0 M guanidinium chloride, pH 5.0, 80% refolds in a single first-order relaxation with a rate constant of 6 s^{−1}.²⁹⁷ As it sits in 6.0 M guanidinium chloride, however, the percentage of the fast-folding isomer decreases to 3% and the approach to this equilibrium has a rate constant of around 0.05 s^{−1}. From an analysis of the kinetics of this loss of the rapidly folding state of the random coil in both wild-type protein and protein in which Proline 55 had been mutated to asparagine, it could be concluded that the random coil with both prolines in the *cis* isomer folds to the native state with full enzymatic activity at 6 s^{−1}, that the isomerization of *cis*-Proline 55 to *trans*-Proline 55 has a rate constant of 0.05 s^{−1} and a *cis*–*trans* equilibrium constant of 0.16, and that the isomerization of *cis*-Proline 39 to *trans*-Proline 39 has a rate constant of 0.02 s^{−1} and a *cis*–*trans* equilibrium constant of 0.1. At equilibrium, 78% of the random coils have both prolines in the *trans* conformation.

Nevertheless, when this mixture of geometric isomers of the random coil is diluted to 1 M guanidinium chloride, pH 5.0, at least 70% of the random coils collapse at a rate of 50 s^{−1} to molten globules in which the central β sheet has formed²⁹⁸ and the α helix of the native state then forms within these molten globules with a rate constant of 20 s^{−1}. When the same equilibrium mixture is diluted to 0.15 M guanidinium chloride, the entire far-ultraviolet circular dichroic spectrum of the native protein is regained in a few seconds.²⁹⁹ The resulting condensed, molten globular states, however, do not become native protein until both Proline 55 and Proline 39 have become *cis*. The isomerizations that produce the proper *cis* isomers, and hence the native state, proceed in these molten globular states with rate constants between 0.01 s^{−1} and 0.0003 s^{−1} at 10 °C (Figure 13–15A).²⁹⁹ The isomerization of Proline 39 is retarded significantly by the formation of these partially folded molten globules, and this retardation is in part responsible for the slowest rate constant of 0.0003 s^{−1} for 66% of the protein.³⁰⁰

In the laboratory, the foldings of many polypeptides have slow phases with rate constants between 0.1 s^{-1} and 0.002 s^{-1} at $25\text{ }^{\circ}\text{C}$ that are attributed to proline isomerization.^{228,240,286,301–306} Many of these attributions have been validated by demonstrating that when particular prolines in the polypeptide are eliminated by **site-directed mutation**, the slow phases disappear.^{228,240,301–303} A problem with this approach is that if the proline that must be mutated is critical to the structure of the protein, in particular if it is *cis* in the native state, its removal often destabilizes the protein significantly and alters the kinetics of folding.³⁰⁷ For example, the double mutant of bovine ribonuclease A folds about 100-fold more slowly than the wild-type polypeptide with both Proline 93 and Proline 114 in the *cis* isomer.²⁹²

In the folding of small proteins that contain only one or two domains, most if not all of the steps that have rate constants less than 0.1 s^{-1} result from required isomerizations of one or more prolines. The **rate constants for the isomerization of proline** between *cis* and *trans* isomers in a random coil are in the range from 0.1 s^{-1} to 0.002 s^{-1} at $25\text{ }^{\circ}\text{C}$.²⁹⁷ The isomerization of *trans* to *cis* is always slower by a factor of 2–10 because of the equilibrium constants, so if the proline is *cis* in the native state, which because of its peculiar geometry is usually the more critical for proper folding, any step during folding that involves the formation of that *cis* isomer will be quite slow (Figure 13–15).

A related set of isomerizations proceed at a more rapid rate than those for *cis*-prolines. In the fully equilibrated random coil, about 0.0015 of the peptide bonds amino-terminal to amino acids other than proline are in the *cis* conformation.³⁰⁸ Although this is a small fraction, in a protein with 100 amino acids only 86% of the random coils will be all *trans* at equilibrium. If the native state has a *cis* peptide bond amino-terminal to an amino acid other than proline or if a significant percentage of its positions cannot tolerate *cis* peptide bonds during its folding, a fraction of the random coils will be in geometric isomers incapable of folding to the native state. The monocationic or monoanionic forms of dipeptides approach the equilibrium between their *cis* and *trans* isomers at a rate constant of about 1 s^{-1} at $25\text{ }^{\circ}\text{C}$,³⁰⁹ and a slow phase with a rate constant of 2.5 s^{-1} at $25\text{ }^{\circ}\text{C}$ involving 5% of the random coils during the folding of α -amylase inhibitor HOE-467A ($n_{\text{aa}} = 74$) from *Streptomyces tendae* has been attributed to random coils in which critical peptide bonds are in the incompatible *cis* isomer.³⁰⁸

Most of the time, the isomerizations of only a few prolines, usually to the *cis* isomer, are required steps in the complete folding of a protein. The isomerizations of many of the peptide bonds amino-terminal to prolines in a protein have no effect on its kinetics of folding,³¹⁰ and the folding of a number of proteins show no slow phases that result from proline isomerization.²⁴² In fact, the native states of some proteins have two conformations in slow equilibrium with each other, one in which a proline

is in the *cis* isomer and the other in which it is in the *trans* isomer.^{304,311} Because the slow isomerizations of the prolines significantly complicate the kinetics of folding, most of the proteins chosen for detailed studies of folding are those that do not have prolines that have to isomerize.

This choice is appropriate because in the cytoplasm and the extracytoplasmic spaces in which all polypeptides normally fold, as opposed to the laboratory, the isomerizations between the *cis* and *trans* isomers of the peptide bonds to the amino-terminal sides of prolines are catalyzed by **peptidylprolyl isomerases**. The first enzyme with this catalytic activity that was purified to homogeneity^{312,313} was assayed by its ability to catalyze the *cis*–*trans* isomerization in *N*-glutaryl-Ala-Ala-Pro-Phe 4-nitrophenylanilide.³¹⁴ This particular enzyme is able to increase the rates of the slowest phases in the refolding of, among other proteins, ribonuclease A,^{315,316} the light chain of immunoglobulin,³¹⁷ human acylphosphatase,³⁰⁶ and type III collagen.³¹⁸

In most of these instances, the increases observed in the rate constants for these slowest phases in folding were relatively unremarkable (less than a factor of 10) even at high concentrations of the peptidylprolyl isomerase. It has been found, however, that there are many **different isoforms** of peptidylprolyl isomerase in a given organism; for example, there are at least 25 different peptidylprolyl isomerases encoded by the human genome, and several are often found within the same cell.^{319,320} It is possible that if the folding protein were matched with its proper peptidylprolyl isomerase under conditions resembling those encountered by the folding protein in the cytoplasm, much more effective rates of catalysis would be observed, but peptidylprolyl isomerases from bacteria, fungi, and animals are about as effective when catalyzing the folding of the same protein.³²¹ Within an intact mitochondrion, the increase in the rate of folding catalyzed by endogenous peptidylprolyl isomerase was observed to be a factor of only 2–6.³²² It is possible that the effect of peptidylprolyl isomerases on the rate of folding in the cytoplasm and the various extracytoplasmic spaces is usually modest.

Peptidylprolyl isomerases have been isolated from bacteria,³²³ fungi,³²⁴ and plants³²⁵ as well as from animals. The peptidylprolyl isomerase associated with the ribosome in *E. coli* appears to be one of the most effective.³²⁶

The **extent of exposure** of the peptide bonds amino-terminal to proline is an important factor in the efficiency with which peptidylprolyl isomerases can function^{299,327} because they must be able to find the bond before they can isomerize it. For example, peptidylprolyl isomerase is able to catalyze the faster of the isomerizations of proline that must occur during the folding of ribonuclease T₁ to its native state, but not the slowest (Figure 13–15). This slowest step involves the isomerization of Proline 39, which has already been slowed considerably from its rate in the random coil²⁹⁷ by being

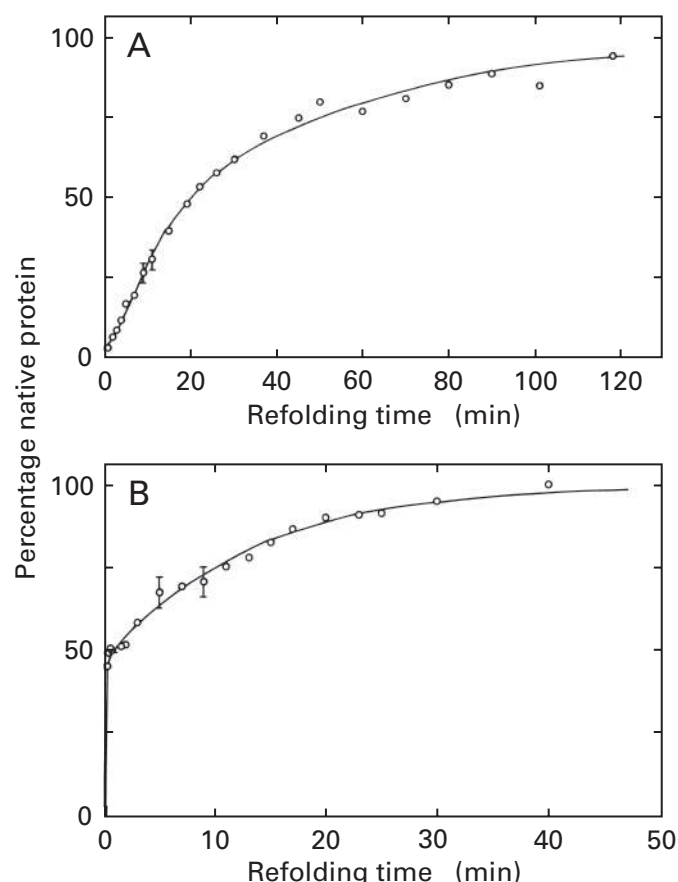


Figure 13-15: Slow steps in the folding of ribonuclease T_1 from *A. oryzae* due to isomerization of the peptide bonds amino-terminal to prolines.²⁹⁹ Unfolded ribonuclease T_1 at 0.3 mM in (A) 6.0 M guanidinium chloride, pH 1.9, or (B) 8 M urea, pH 8.0, was diluted 40-fold with 0.1 M tris(hydroxymethyl)ammonium chloride, pH 8.0, to initiate folding. At various times, the percentage of the protein in its native state was determined by monitoring the decrease in fluorescence (emission at wavelengths greater than 320 nm; excitation at 268 nm) that occurred upon dilution of the sample to 5.6 M guanidinium chloride, pH 2.0. (A) Refolding in the absence of prolyl isomerase. (B) Refolding in the presence of 0.7 μ M prolyl isomerase. The percentage of the molecules in their native state is plotted as a function of time (minutes). The curves are multiexponential fits to the data with first-order rate constants of (A) 0.01 s^{-1} , 0.005 s^{-1} , 0.002 s^{-1} (31%), and 0.0003 s^{-1} (66%) and (B) 0.1 s^{-1} , 0.04 s^{-1} , 0.5 s^{-1} (46%), and 0.0013 s^{-1} (51%). For both fits an unresolved kinetic burst of 3% was assumed. Reprinted with permission from ref 299. Copyright 1990 American Chemical Society.

buried in a molten globular intermediate.³⁰⁰ In the molten globules formed during the folding of isoform 2 of cytochrome *c* from *S. cerevisiae*, peptidylprolyl isomerase is unable to hasten any of the slow isomerizations at prolines in the final steps, even though it can if the equilibrium between molten globule and unfolded forms is shifted significantly by adding guanidinium chloride.³²⁸ How this problem of accessibility is solved in the cytoplasm of a cell is unclear.

There are a number of small proteins or small domains removed from larger proteins that fold in what

kinetically appear to be single steps with the formation of no intermediates. One example of such a protein is the cold shock-like protein from *T. maritima* (Figure 13-1).^{109,329} Even at 0.5 M guanidinium chloride, this small protein ($n_{aa} = 66$) folds in what kinetically appears to be a simple first-order isomerization, and there is no obvious change in rate-limiting step observed in the plot of the observed first-order rate constant for folding as a function of the concentration of guanidinium chloride (Figure 13-1B). **Folding with no kinetic intermediates** has been observed for the competent proline isomer of the random coil of chymotrypsin inhibitor 2A ($n_{aa} = 83$) from *Hordeum vulgare*,³⁰⁵ for which the observed first-order rate constant for folding shows no evidence of a change in rate-limiting step even in the absence of any denaturant, as well as for, among others, human acylphosphatase ($n_{aa} = 98$),³³⁰ protein S6 from the 30S subunit of ribosomes from *Thermus thermophilus* ($n_{aa} = 101$),³³¹ the engrailed homeodomain from *Drosophila melanogaster* ($n_{aa} = 61$),³³² and human ubiquitin ($n_{aa} = 76$).³³³

It is also possible to follow even faster folding reactions that exhibit apparently two-state behavior in the region of transition down to low concentrations of denaturant by analysis of the line widths of the absorptions or analysis of relaxation rates in nuclear magnetic resonance. In this way, it can be demonstrated for some proteins and the detached domains of other proteins that there is no evidence of a change in rate-limiting step and that their folding remains kinetically a one-step reaction even in the absence of denaturant. Such demonstrations have been made for the amino-terminal domain of the repressor from bacteriophage λ ($n_{aa} = 80$),^{334,335} the B-domain of protein A from *S. aureus* ($n_{aa} = 58$),¹¹⁰ and the peripheral subunit binding domain of dihydrolipoyl-sine-residue acetyltransferase of *B. stearothermophilus* ($n_{aa} = 41$).³³⁶

All of these foldings appear to occur in a kinetically single step because there is **no change in the rate-limiting step** as denaturant is decreased to zero. In the foldings in which there is a change in the rate-limiting step (Figure 13-10), it is slower steps unaffected or less affected by the concentration of denaturant, such as the progression of the formation of secondary structure within the molten globular intermediate, the final locking together of the secondary structures into their native packing, and isomerizations of peptide bonds amino-terminal to prolines, that become slower than the initial condensation of the random coil to the molten globule as the concentration of denaturant is lowered. In the foldings that appear to proceed in a single kinetic step, these later steps simply remain faster than the initial condensation; they still must occur, but they are kinetically silent because they occur after the rate-limiting step.

The observed rate constant for the first-order relaxation of one of these foldings that appear to proceed in a single kinetic step is determined by the rate constant of

the rate-limiting step, the rate constants of any rate-determining steps preceding the rate-limiting step, and the equilibrium constants of any unfavorable preequilibria that precede the rate-limiting step (Equations 13–28 and 13–30). Favorable preequilibria preceding the rate-limiting step would require the formation of observable intermediates in the reaction.

The following facts indicate that, immediately upon completion of the rate-limiting step in one of these foldings that appears to proceed through a single kinetic step, a **molten globular intermediate** containing significant secondary structure has formed from the random coil. The observed rate constants of these foldings are significantly affected by the concentration of denaturant (Figure 13–1B); they increase by factors as large as 10,000 between a solution containing a concentration of denaturant within the region of transition and one containing no denaturant.³³¹ This fact requires that considerable accessible surface area be lost either in the transition state of the rate-limiting step, during rate-determining steps preceding the rate-limiting step, or during preequilibria preceding the rate-limiting step. The **apparent molar activation volume** for one of these foldings is positive and even larger than the positive change in molar volume between random coil and native state,³³⁷ an observation indicating that the transition state for the rate-limiting step or for rate-determining steps that precede it is globular or that an intermediate formed in an unfavorable preequilibrium is globular, as is the native state, but is somewhat less compact than the native state. All³³⁸ or most³³⁹ of the **change in standard heat capacity** between random coil and native state has occurred by the time the transition state in the rate-limiting step appears, an observation indicating that the hydrophobic functional groups buried in the native state are buried during the rate-limiting step or before it. The **effect of pH** on the observed first-order rate constant for one of these foldings indicates that some carboxylates have been sequestered from the solvent, but far fewer than those sequestered in the native state, by the time the transition state of the rate-limiting step has formed.³³⁹ In the case of the carboxy-terminal domain of protein L9 from the 50S subunit of the ribosome from *E. coli*, the mutation of Histidine 134, the histidine in the protein with the lowest pK_a in the native state, caused the most dramatic change in the dependence of the rate constant k_f on pH, a result suggesting that this same histidine is also the most buried by the time the transition state of the rate-limiting step has formed and that at this point the protein resembles the native state.¹¹³

It is the effects of **site-directed mutations** on the observed rate constants of these foldings appearing to proceed in a single kinetic step which indicate that secondary structure has formed before or during the rate-limiting step. For example, when one or the other of the two α helices in acylphosphatase is stabilized by mutating one of its amino acids to alanine, the observed rate

constant for its folding increases if the mutation is in the second α helix but not if it is in the first.³⁴⁰ When the two α helices of transcriptional repressor *arc* from bacteriophage P22 were destabilized by mutations of alanines to glycines, the observed rate constants of folding decreased significantly,³⁴¹ while alanine to glycine mutations in only one of the α helices in the amino-terminal domain of the repressor from bacteriophage λ affect the observed rate constant of its folding.³⁴² The magnitudes of the changes in the observed rate constants of folding relative to the change in the standard free energy of folding for mutations at 37 positions in chymotrypsin inhibitor 2A suggest that some elements of secondary structure have formed by the time the transition state of the rate-limiting step has been reached but that they are less stable than they are in the native state.³⁴³ The effects of site-directed mutation on the folding of the WW domain from peptidylprolyl isomerase Pin 1, however, indicate that its folding is most sensitive to changes in what is an external loop in its crystallographic molecular model,³⁴⁴ an observation suggesting that secondary structure in the native state may not correspond to secondary structure in a molten globular intermediate.

All of the observations discussed so far can be explained as effects either on the rate-limiting step or on one or more unfavorable preequilibria preceding the rate-limiting step. It is the case, however, that the observed rate constant for at least one of these foldings that appears to proceed in a single kinetic step is linearly related to the inverse of the **viscosity of the solvent**.²⁷³ This fact requires that the condensation of a conformation or set of conformations of the random coil occur during the rate-limiting step, not before it. This conclusion follows from the fact that the viscosity of the solvent must affect both contraction and expansion of a polypeptide equally and hence an equilibrium constant for contraction not at all.

The observed rate constants for these foldings that appear to proceed in a single kinetic step span a wide range from 0.2 s^{-1} (28 °C) to $120,000 \text{ s}^{-1}$ (37 °C). The fact that these rate constants span such a wide range and the fact that nevertheless they all seem to register condensations producing molten globular intermediates suggest that their rate-limiting steps are not hydrophobic collapse, for which the observed rate constant would be related solely to the length of the polypeptide and would not vary so dramatically,²⁷² but hydrophobic collapse of a conformation or set of conformations of the random coil that are present at low occupancy; the lower that occupancy, the slower the observed rate constant (Equation 13–30). Presumably, the subset of conformations of the random coil that are present at such low levels of occupancy are those that contain sufficient secondary structure to condense to a stable molten globule. Consistent with this presumption is the fact that the observed rate constant for at least one of these foldings that appears to occur in a single step is significantly

increased by the addition of low concentrations (0.006 mole fraction) of trifluoroethanol or 1,1,1,3,3,3-hexafluoro-2-propanol.³³⁰ These cosolvents are known to stabilize secondary structure in unfolded polypeptides.

The apparent single step registered in each of these foldings is equivalent to the step producing the molten globular intermediates formed during the kinetic burst in the foldings of proteins that proceed through multiple steps. In the case of the proteins folding in an apparent single step, the steps following the molten globular intermediate are kinetically silent. The advantage of the foldings in a single step is that attention can be directed exclusively on this one step, uncomplicated by the later ones; the disadvantage is that the later steps cannot be studied at all.

What seems to occur during folding of a polypeptide can be summarized. The random coil has conformations within its ensemble that are evanescently present at low concentrations and that contain sufficient secondary structure to stabilize a molten globular intermediate. One or more of these minor conformations of the random coil collapses hydrophobically to form one or more molten globules.* Within these molten globules, some of the secondary structure of the native state is present and the rest forms in a series of steps before or as it locks into its proper packing to produce the native state. Superimposed on this progression are the slow isomerizations of the peptide bonds amino-terminal to prolines, each of which has the potential to decrease the observed rate constant of any one of these steps dramatically but only for those isomers of the polypeptide that happen to contain an incompatible isomer of proline. If isomerizations of peptide bonds amino-terminal to prolines were not or are not involved, the folding of a polypeptide would be or is usually complete within less than 10 s at 25 °C, which is remarkably fast for such a complicated process.

Up to this point, for the sake of clarity, the transformations between each of the major intermediate states encountered during the folding of a protein have been presented as if they were uncomplicated first-order reactions proceeding in single steps just as the bimolecular nucleophilic displacement of an iodide in the alkylation of a thiolate anion (Equation 3–17) is an uncomplicated second-order reaction proceeding in a single step through a single transition state. It is certainly the case, however, that in a process as complicated as protein folding none of these transformations between the denatured state, major intermediate states, and the final native state proceeds in a single step. Consequently, the transformations **only appear to be single steps**, and the

observed rate constants are only apparent rate constants. For example, the kinetic progress of the folding of cytochrome *c'* from *Rhodopseudomonas palustris*, although apparently proceeding through four steps when the absorbance of the heme is followed at 440 nm, can be fit more exactly by a set of 80 rate constants spanning a range from 10^5 to 10^{-2} s⁻¹.³⁴⁵ This observation by itself is not surprising. Were the kinetic measurements of a simple chemical reaction that definitely had only four steps to be fit with an equation derived for 80 steps, the fit for 80 steps would necessarily be better than the fit for four steps. Nevertheless, this exercise suggests that there are more than just four steps involved in the folding of this cytochrome *c'*. Careful examination of Figure 13–13 also suggests that a mechanism involving more than two steps in the range monitored would yield a rate equation more successful at fitting the actual data.

It has been shown that the apparently first-order rates of the transformations between apparently discrete intermediate states in the folding of a protein and the behavior of those rates as a function of the concentration of denaturant are more successfully fit by kinetic models involving a large number of intermediate states in sequence.²³⁶ It has already been noted that the number of transitions observed during the folding of a polypeptide usually increases in number as more and more physical properties are monitored during the same process, so observations based on only one physical property usually fail to register all of the discrete steps in the process of folding. It has also been proposed that the differences observed among the effects of denaturant on the rates of exchange of different amide protons along the backbone of a polypeptide (Figures 13–8 and 13–14) are evidence for the existence of a **continuum of intermediate states** in the process of folding.³⁴⁶ These various intermediate states are in preequilibrium with each other prior to a rate-limiting step, represent a set of rate-determining steps, are in a steady state with each other so that the observed first-order rate constant is actually a composite rate constant,²⁴⁴ or are related to each other in some combination of these three possibilities. Because it is so unlikely that any one of these apparently first-order transformations is actually a simple single step, a discussion of the transition state associated with the observed first-order rate constant for any one of these transformations is meaningless.

Another complication is the existence of multiple, **parallel pathways** in the folding of a protein. The most obvious example of such a situation results from the isomerization of prolines. Each geometric isomer of the random coil at each proline that is required to have a particular configuration for proper folding folds along its own distinct pathway (Figure 13–15). In the case of lysozyme from *G. gallus*^{347–349} and dihydrofolate reductase from *E. coli*,^{350,351} however, there are two and four parallel pathways, respectively, through which these foldings progress, none of which are distinguished by

* If the effect of solutes increasing the viscosity of the solution have been misinterpreted, it is also possible that hydrophobic collapse is the unfavorable preequilibrium preceding the formation of sufficient secondary structure to stabilize the molten globule or molten globules.

differences in the isomerization of prolines. At one or more points in the process of folding, some of the molecules assume one state and the others assume another state, and from there on, each of the two populations proceeds through a different sequence of steps to form the same final native state of the protein. In the case of dihydrofolate reductase, the existence of parallel pathways may be related to the fact that there are two different conformations of the native state in equilibrium with each other that are both substantially occupied under most circumstances.^{352,353}

There are also **kinetic dead-ends** that can complicate the process of folding. For example, during the folding of equine cytochrome *c* from its random coil, the wrong side chains can form the fifth and sixth ligands to the covalently bound heme,³⁵⁴ and they must dissociate and the proper side chains must associate before folding can proceed successfully.³⁵⁵

One of the most consequential kinetic dead-ends is aggregation of the folding polypeptides. When, in the laboratory, the concentration of urea or guanidinium chloride is rapidly lowered by dilution, the hydrophobic side chains on a random coil are no longer favorably solvated. They can either promote the desired intramolecular hydrophobic collapse, or they can associate with hydrophobic side chains on other unsolvated polypeptides to form an undesired intermolecular aggregate, which just as successfully removes them from contact with the water. When the **intermolecular aggregation** is reversible so that the aggregate is in equilibrium with an unaggregated, denatured intermediate produced during some step in the process of proper folding, the aggregation has the effect of lowering the concentration of that intermediate on the pathway to the native state and decreasing the rate of folding. In this situation, the amount of intermolecular aggregate decreases as the total concentration of protein is decreased, and the rate of folding increases³⁵⁶ as the concentration of protein is decreased. The aggregate at the dead end can also appear to be an intermediate in the folding process, forming rapidly and then disappearing as the folding depletes the solution of the intermediate in equilibrium with the aggregate.³⁵⁷

When the intermolecular aggregation of the folding polypeptide is **irreversible**, however, it competes with proper folding, and the final yield of the native state, rather than its rate of formation, is decreased accordingly. Often such irreversible aggregation can be minimized by lowering the total concentration of protein,^{358,359} but in many instances the folding protein passes through an intermediate so prone to aggregation that very little if any native state forms at any concentration of protein. This catastrophic problem is solved within the cell by the chaperones.

A **chaperone** is a protein that intercepts and suppresses the unproductive, nonspecific, irreversible intermolecular aggregation of a folding polypeptide so that it

can fold intramolecularly to achieve its native state.^{360,361}

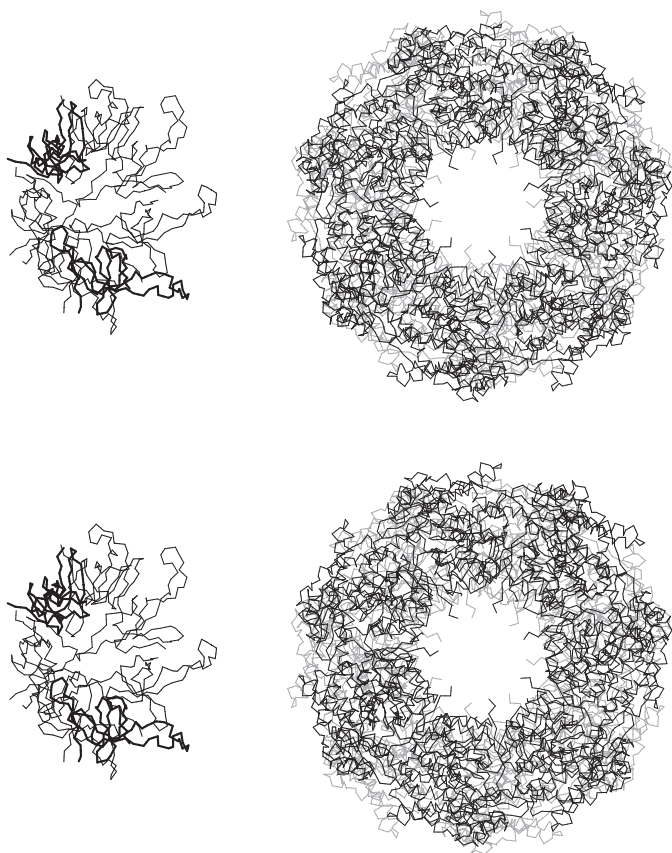
There are two species of chaperones that are responsible for most of this suppression of aggregation and protection of proper folding. These two species are represented by **chaperonin 60** (the product of the *groEL* gene) and **heat shock protein 70** (the product of the *dnaK* gene), respectively, from *E. coli*. Chaperonins 60 have been purified from fungi,³⁶² chloroplasts,³⁶³ mitochondria,³⁶⁴ and eukaryotic cytoplasm,³⁶⁵ and heat shock proteins 70 are also universally distributed. The most obvious distinction between these two different species of chaperones, other than their unrelated amino acid sequences, is that chaperonins 60 are all large oligomers of 14–16 subunits arranged with the respective symmetries of the point group 722(*D*₇) or the point group 822(*D*₈),^{366–368} while heat shock proteins 70 are monomers or dimers.

The chaperone about which the most is known is chaperonin 60 from *E. coli*. It is a homotetradecamer with symmetry of the point group 722(*D*₇) enclosing a large central cavity (Figure 13–16)^{366,369–371} that is divided into two halves (upper and lower heptamers in Figure 13–16) at its middle by a thick septum formed from the irregular coalescence of the 14 carboxy-terminal segments, each 23 amino acids long, which are disordered in the crystallographic molecular model.³⁷² Each of the 14 subunits in turn is divided into three domains (arranged one on top of the other and consequently difficult to distinguish in the view presented in Figure 13–16). The apical domains are symmetrically arrayed around the central 7-fold rotational axis of symmetry to form the entrances to the upper and lower cavities, the equatorial domains surround the central septum, and each of the intermediate domains connects its apical domain to its respective equatorial domain.

Chaperonin 60 prevents proteins that aggregate irreversibly during their folding in its absence from doing so in its presence.^{373,374} It accomplishes this task by recognizing and associating with **intermediates in the pathway of folding that are prone to aggregation**.³⁷⁵ By itself, chaperonin 60 forms a tight complex with such an intermediate.^{376,377} If it were the only capability of chaperonin 60, the formation of this **tight complex** would interrupt the folding of a protein and prevent its completion. Consequently, the bound intermediate must dissociate so that folding can proceed. If, however, the bound intermediate were to dissociate from chaperonin 60 unchanged, it would then proceed to aggregate as it was about to do just before it was bound. Consequently, the structure of the bound intermediate must be altered so that it dissociates in a form that is not prone to aggregation. The standard free energy necessary to promote the dissociation of the bound intermediate and to change its structure before it dissociates is provided by the binding and hydrolysis of MgATP.

The structure of the form of the folding protein recognized and bound by chaperonin 60 has not been clearly defined. In theory, chaperonin 60 should recog-

Figure 13-16: Chaperonin 60 and chaperonin 10 from *E. coli*. The α -carbon skeleton of the tetradecamer of chaperonin 60 (upper structure) is drawn from the crystallographic molecular model of the protein in a complex with 14 molecules of MgKATP.³⁶⁹ The tetradecamer is viewed down the 7-fold rotational axis of symmetry. The seven identical subunits ($n_{aa} = 547$) in the upper heptamer are drawn with line segments of different widths. Those in the lower heptamer are drawn with gray lines. In the crystallographic molecular model, the carboxy-terminal 22 amino acids of each subunit are disordered, unresolved, and, consequently, absent from the model. They extend in the plane of the page from the 14 carboxy termini of the crystallographic molecular model protruding into the center of the central cavity and are thought to coalesce³⁷² to form the thick equatorial septum observed in image reconstructions from electron micrographs³⁷⁰ that divides the central cavity into upper and lower halves. The seven 2-fold rotational axes of symmetry of the point group 722 (D_7) are in the plane of the page and run through the center of the drawing so the structures of the upper and lower cavities are identical to each other. The α -carbon skeleton of chaperonin 10 is drawn below that of chaperonin 60. It is drawn from the crystallographic molecular model of the complex between chaperonin 60 and chaperonin 10.³⁷¹ Chaperonin 10, a symmetrical heptamer of seven identical subunits ($n_{aa} = 97$), is a cap that can sit upon the top of the chaperonin 60 with its 7-fold rotational axis of symmetry coincident with the 7-fold rotational axis of symmetry of the tetradecamer and seal the respective cavity. It is drawn so that a -135° flip around the x axis puts the cap on the upper cavity of the chaperonin 60.



nize only intermediates in the process of folding that are prone to aggregation and ignore intermediates that are not, but when it is added in the absence of MgATP to solutions of some proteins that are in the process of folding, their folding slows dramatically^{378,379} or ceases,³⁸⁰ as if almost all or all of the states of these proteins other than the native state are recognized and bound. With other proteins, the rate of their folding is unaffected or slightly accelerated upon the addition of chaperonin 60,³⁸¹ but these are usually proteins that are not susceptible to aggregation during their folding. Chaperonin 60 also forms complexes when added to solutions of certain native proteins,^{382,383} presumably by recognizing denatured forms in equilibrium with the native state and shifting that equilibrium by sequestering those denatured forms.³⁸⁴ Site-directed mutants in which hydrophobic amino acids have been replaced with alanine or glycine are recognized and bound by chaperonin 60 less successfully during their folding,³⁷⁸ suggesting that it is exposed hydrophobic side chains that elicit the attention of the chaperone.

After it has been bound by chaperonin 60, a folding protein is usually in a **molten globular state in the complex**.³⁸⁵ This conclusion follows from the fact that the rates of the exchange of its amido protons are rapid.^{380,382,383,386,387} With some proteins, unlike what is observed in the usual molten globular state, the exposure of all amido protons is increased to the same extent,^{383,388} a result suggesting that the association of the folding protein with the chaperone has disrupted all secondary structure. With other proteins exposure of amido protons varies along the sequence,³⁸⁰ a result suggesting that some secondary structure is preferentially preserved in the bound form of the protein. With yet other proteins almost no difference in exchange rates between the bound form of the protein and the native state is observed,³⁸⁹ a result suggesting that the bound protein is similar in its structure to the native state. It has also been observed that a tridecapeptide that is structureless in solution is an α helix when bound to chaperonin 60.³⁹⁰

Because it is not known which conformation or set of conformations of a folding protein is recognized and bound by a chaperonin 60, it is possible that the structure of that folding protein once it has been bound is identical to the structure that it had when it was recognized and bound, but it is also possible that the structure of the protein has been altered significantly during its binding to the chaperone to produce the conformation or set of conformations that are observed in the complex. This alteration in structure caused by the binding itself would be at least a portion of the alteration required so that a form of the protein that is not prone to aggregation can dissociate from the chaperone.

The folding protein is bound by one or more of the **apical domains** that surround the entrances to each of the two central cavities in chaperonin 60 (Figure 13-16).^{370,391} In some cases, the bound protein protrudes

outward from the cavity;³⁹¹ in other cases, it protrudes into the cavity and fills it.^{392,393} Whether the bound protein **ends up within the cavity** or protruding away from it is in part a function of its size;^{369,394} the larger proteins protrude outward because they do not fit within. A folding polypeptide larger than 600 aa is too large to fit in the cavity.³⁹⁵

When the apical domains of chaperonin 60 are detached genetically and expressed separately, they may³⁹⁶ retain the ability to recognize and bind forms of a folding protein that are prone to aggregation.³⁹⁷ A site in the crystallographic molecular model of a detached apical domain to which an extended hydrophobic segment of polypeptide has bound has been tentatively assigned as the site for binding of the folding protein to the apical domain.^{369,398}

The next step in the rescue of a folding protein from aggregation performed by chaperonin 60 is its **dissociation from this bound state**. This dissociation requires the addition of MgATP. When MgATP is added to a complex of chaperonin 60 and denatured protein, it promotes the dissociation of that protein, permitting its folding to recommence.^{399–405} The dissociation of the denatured protein from the complex is more rapid than the hydrolysis of the MgATP^{400,404} and also occurs when analogues of ATP that cannot be hydrolyzed are used instead of ATP,^{376,402,403,406} so it is the **binding of MgATP** and not its hydrolysis that promotes the dissociation. The binding of MgATP to chaperonin 60 occurs at a site located in each equatorial domain,⁴⁰⁷ while the folding protein is bound to the apical domain. These two sites are 3.2 nm apart at their closest approach, so no direct interaction between them is possible. Consequently, there must be two global conformations of chaperonin 60 in equilibrium with each other, one of which binds denatured protein at the apical domain strongly and MgATP at the equatorial zone weakly and the other of which binds denatured protein weakly and MgATP strongly, and the binding of MgATP must shift the equilibrium between these two conformations in favor of the one that binds denatured protein weakly.^{244,403,404,407}

In the conformation of chaperonin 60 that is stabilized by the binding of MgATP, an α helix in the apical domain that forms part of the site at which the folding protein is thought to bind³⁹⁸ is rotated by 102° relative to its orientation in the conformation of the protein that is the more stable in the absence of MgATP.³⁶⁹ This rotation sequesters a set of hydrophobic side chains and makes the site considerably less hydrophobic, perhaps promoting the dissociation of the folding protein.

The subsequent **slow hydrolysis of the MgATP** (0.06 s^{-1} at 25 °C),⁴⁰⁸ at all seven of the sites in one half of the protein,⁴⁰⁹ serves to regenerate the conformation of the chaperone that can again recognize intermediates prone to aggregation. During the folding of a protein that requires significant assistance to avoid aggregation, several cycles of binding to chaperonin 60 and release are

needed before the native state is achieved.^{404,410} Each cycle of binding and release requires that the MgATP be hydrolyzed to regenerate the competent conformation of the chaperone, so the presence of a folding protein in need of assistance elicits an ATPase activity from chaperonin 60.⁴⁰⁴

The dissociation of the folding protein from chaperonin 60 that is promoted by the binding of MgATP is accelerated by the addition of chaperonin 10.⁴⁰² Chaperonin 10 is a homoheptamer of subunits 97 aa in length that can sit as a cap upon the seven apical domains of chaperonin 60 and seal off the respective cavity from the solution (Figure 13–16).^{371,393,411} Its addition is required for the maximum yield of the native state of some folding proteins prone to aggregation; with other proteins that are prone to aggregation, its addition increases the rate of their folding.

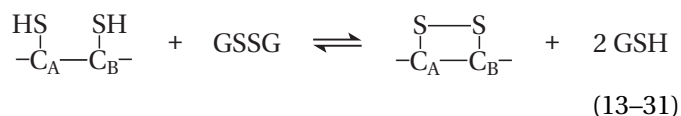
When the folding protein is too large to fit in the cavity, chaperonin 10 nevertheless still increases the rate of folding or the yield of properly folded protein or both of these outcomes but does so by binding to the opposite end of chaperonin 60 from the end at which the folding protein was bound.³⁹⁴ When the folding protein is small enough, chaperonin 10 binds at the end to which it is associated and traps it within the respective cavity of chaperonin 60,^{408,412} where it is definitely protected from intramolecular aggregation because it is alone. In the presence of chaperonin 10, a **trapped folding protein**, although it has been dissociated from its binding site on chaperonin 60 by the binding of MgATP, remains associated with the complex of chaperonin 60 and chaperonin 10. Only after the MgATP has been hydrolyzed is the folding protein released from the complex along with the chaperonin 10.^{408,409,413,414} Because the hydrolysis of the MgATP is so slow, however, this means that the complex remains intact for, on the average, about 15 s, which is a long period of time in the folding of a protein and during which considerable progress towards the native state can be achieved.

The requirement that chaperonin 60 alter the structure of the folding protein and release it in a form not prone to aggregation seems to be accomplished both before and after MgATP is bound. The alterations in the structure of the folding protein before MgATP is bound have already been described, but it has been noted that the addition of MgATP and chaperonin 10 to the complex between a folding protein and chaperonin 60 further increases the rates of amide proton exchange of the folding protein,⁴¹⁵ a result suggesting that significant structural alterations are made to the folding protein after the binding of MgATP but before it dissociates from its binding site on chaperonin 60.

The chaperones of the species of heat shock proteins 70, for which heat shock protein 70 from *E. coli* is the paradigm, also rescue folding proteins from aggregation by binding them and releasing them in associations and dissociations that are coupled to the binding and

hydrolysis of MgATP,⁴¹⁶ but the details of the process are much less clear. It is the complex between heat shock protein 70 and MgATP that binds the folding protein in a rapidly reversible equilibrium, and upon hydrolysis of the MgATP, the folding protein is locked onto the chaperone.^{417–419} The binding of the next molecule of MgATP rapidly releases the bound protein into the solution, presumably in an altered conformation. In a crystallographic molecular model of the complex between heat shock protein 70 and a peptide thought to mimic the bound folding protein, the peptide is bound in an unstructured extended conformation.⁴²⁰ Heat shock protein 40 increases the rate at which the ATPase recycles heat shock protein 70 among its various conformations,⁴¹⁹ as does the chaperonin 10 with chaperonin 60. There is, however, no cavity in any of these proteins similar to that in the tetradecamer of chaperonin 60.

As a polypeptide folds to its native state in the cytoplasm of a cell, the high concentration of reduced glutathione (3–5), or some other mercaptan with the same function,⁴²¹ prevents its cysteines from forming adventitious cystines, and for the same reason, the cysteines of the native protein remain reduced throughout its lifetime. Most of the proteins, however, that are excreted from a cell into the extracellular spaces contain **cystines**. In a eukaryotic cell, these cystines are formed as these soon to be excreted proteins fold within the lumen of the endoplasmic reticulum. In the lumen of the endoplasmic reticulum, the ratio of oxidized glutathione to reduced glutathione is much higher (0.5)⁴²² than it is in the cytoplasm (0.02). The native conformation of the protein juxtaposes the two cysteines that will form a correctly paired cystine.⁴²³ This juxtaposition increases the equilibrium constant for the formation of that cystine dramatically,⁴²⁴ so the ambient ratio of oxidized to reduced **glutathione** in the endoplasmic reticulum is sufficient to form a cystine from the two cysteines once they have been juxtaposed.^{423,425}



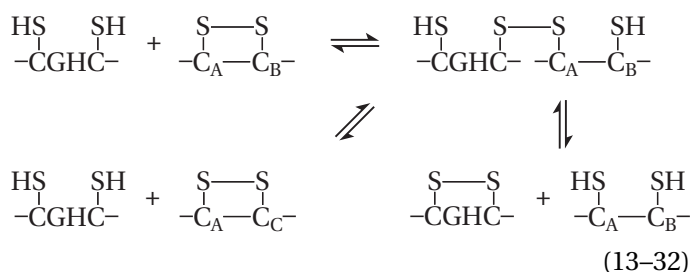
where C_A and C_B are the two cysteines juxtaposed, GSSG is oxidized glutathione, and GSH is reduced glutathione. The fact that one mole of oxidized glutathione yields two moles of reduced glutathione pulls the reaction to the right.

The problem, however, is that the ratio of oxidized to reduced glutathione in the endoplasmic reticulum is also high enough to produce adventitious cystines within intermediates in the pathway of folding that happen to juxtapose incorrect cysteines. This problem, which would lead to the accumulation of stable, improperly folded forms of the protein, is solved by the enzyme **protein disulfide-isomerase**,⁴²⁶ which is present at high con-

centration in the lumen of the endoplasmic reticulum.⁴²⁵ This enzyme fulfills two roles in the formation of correct cystines during the folding of a protein. It breaks cystines to reverse their incorrect formation, and it oxidatively couples pairs of adjacent cysteines to form cystines.⁴²⁷ Consequently, it catalyzes the rapid rearrangement of cystines in a folding protein until the correct partners are joined to produce the native state of the protein. In fact, there are some proteins, for example mammalian pancreatic human insulin-like growth factor⁴²⁸ and bovine ribonuclease A,⁶¹ that cannot fold stably unless all of their cystines have formed correctly, so their folding is required to proceed in tandem with the rearrangement of their cystines by protein disulfide-isomerase until the combination of the correct tertiary structure and the correct pairing of cysteines as cystines is reached. It is both the packing of the tertiary structure and the properly paired cystines that create their native states.

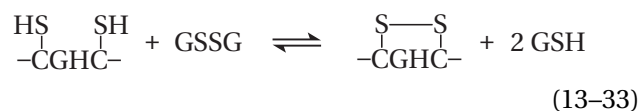
Protein disulfide-isomerase (490 aa) contains two domains (amino acids 5–100 and 350–440) that are homologous to each other and to thioredoxin (110 aa).⁴²⁹ As is the case with thioredoxin, each domain contains a **pair of cysteines** in the sequence –VEFYAPWCGHCK–. It is these pairs of cysteines found in protein disulfide-isomerase that are responsible for the catalysis of the rearrangement and formation of cystines in a folding protein. Two shorter proteins, thiol:disulfide interchange protein DsbA (218 aa) and thiol:disulfide interchange protein DsbC (216 aa), each with only one pair of cysteines in the sequences –LEFFSFFCPHCY– and –TVFTDITCGYCH–, respectively, fulfill the roles of protein disulfide-isomerase for proteins excreted from bacteria. The former can form cystines, but the latter is the enzyme responsible for the shuffling of incorrectly formed cystines.^{430–433}

The two domains in each molecule of protein disulfide-isomerase, which each contain an identical pair of cysteines, are similar in their catalytic abilities and act independently.⁴²⁵ The second cysteine in each of these pairs of cysteines is responsible for **breaking and rearranging cystines** by disulfide interchange during the folding of a protein:



where C_C is a cysteine elsewhere in the folding protein that contains cysteines C_A and C_B . The central intermediate in this shuffling is the **mixed disulfide** between protein disulfide-isomerase and the folding protein (upper right complex in Equation 13–32). The cystine that forms between a pair of cysteines in protein disulfide-iso-

merase by disulfide interchange with oxidized glutathione



is able to convert a pair of juxtaposed cysteines in a folding protein into a cystine (reverse of the reactions on the lower right and top of Equation 13-32).⁴²⁵ Protein disulfide-isomerase also catalyzes the formation of a mixed disulfide between glutathione and a cysteine on a folding protein,⁴³⁴ which is also a central intermediate⁴³⁵ in the formation of cystines by oxidized glutathione (Equation 13-31).⁴²⁵

Because protein disulfide-isomerase contains one pair of cysteines in each of its two domains, and because most folding proteins have enough cysteines to give rise to two or more cystines while they are folding, at the proper molar ratios protein disulfide-isomerase and a folding protein form a precipitate,⁴³⁶ just as an antigen and a set of polyclonal immunoglobulins form a precipitate at equivalence. The existence of this precipitate serves to demonstrate the importance of the mixed disulfide in Equation 13-32 in the reactions catalyzed by protein disulfide-isomerase.

In order to participate in any of these reactions, in particular the formation of the mixed disulfide between it and the folding protein, protein disulfide-isomerase must be able to find the cystine or the cysteine on the folding protein. This necessity requires both that the folding protein be molten enough to expose unpaired cysteines or mispaired cystines on its surface and that the reaction between protein disulfide-isomerase and those exposed cysteines and cystines be rapid and efficient.^{61,423} In a role that may be connected to this **search for incorrect cystines**, protein disulfide-isomerase also acts as a chaperone.⁴³⁷ Once the proper cystines are formed, however, they can be buried without penalty. It is probably the burial of the correctly paired cystines within the proper native structure that terminates the rearrangements of the cystines catalyzed by protein disulfide-isomerase. The same problem of accessibility is faced by peptidylprolyl isomerase, and it is interesting that protein disulfide-isomerase and peptidylprolyl isomerase function synergistically to catalyze the folding of a protein.⁴³⁸

The histidine between the two cysteines in each domain of protein disulfide-isomerase, as opposed to a proline at the homologous position in thioredoxin, causes the cystine in the former protein to have a **reduction potential** 30–40 mV more positive than that in the latter.^{439,440} This higher reduction potential causes a cystine in protein disulfide-isomerase to be more effective at forming cystines in folding proteins by disulfide interchange than a cystine in thioredoxin would be. The homologous cystine in thiol:disulfide interchange protein DsbA from *E. coli* is also exceptionally reactive.⁴³¹

In order to perform both the rearrangement of cystines and their formation most efficiently during its catalysis of the folding of a protein (Equation 13-32), protein disulfide-isomerase must be poised at a level of reduction potential where only a portion of its cysteines are cystines, and in the laboratory, the optimal potential for the catalysis of the folding of a protein is reached at a ratio of oxidized glutathione to reduced glutathione of 0.2–0.5.⁴²⁷ This ratio is not significantly different from the ratio found in the endoplasmic reticulum.

The proteins the foldings of which have been discussed so far have been fairly small and in each case the entire protein has folded as a unit to produce the native state. The folding of a larger protein is usually complicated not only by the larger number of prolines it contains but also by the fact that larger proteins usually contain **domains**, which often fold independently of each other. For example, lysozyme, itself not a large protein, nevertheless contains two structural domains. On the basis of measurements of amide proton exchange, it was demonstrated that one of these domains folds more rapidly than the other,⁴⁴¹ and the completion of the foldings of the two domains is followed by a step in which they become properly oriented and associate correctly with each other to form the native state.³⁴⁹ The rates of the final slow steps in the foldings of both aspartate kinase-homoserine dehydrogenase from *E. coli*³⁵⁹ and D-octopine dehydrogenase from *Pecten jacobaeus*⁴⁴² are inversely proportional to the viscosity of the solvent and are thought to represent the association of independently folding domains. The transfer of energy by resonance between donors and acceptors positioned at several locations on phosphoglycerate kinase from *S. cerevisiae* has been used to follow the changes in the distances between these locations during the unfolding of the protein.⁴⁴³ To the extent that unfolding is the reverse of folding, the fact that the first step in the unfolding of the protein is the dissociation of its two domains suggests that the last step in its folding is the association of these two domains.

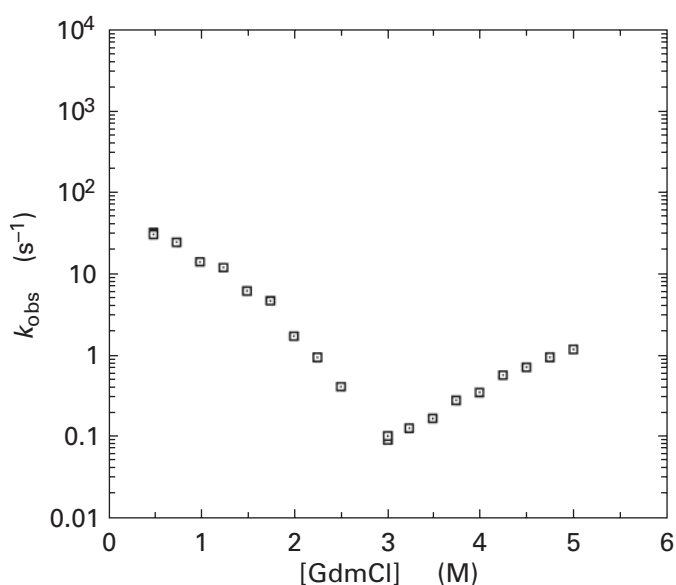
The results of these experiments suggest that, in larger proteins, the individual domains fold independently as if they were the small unitary proteins that have been discussed in detail until there is sufficient structure developed for them to recognize each other and associate. Following their association, the information developed during this association may or may not dictate further folding to reach the final native state depending on the intimacy and interdependence of their interaction.

Suggested Reading

- Jennings, P.A., & Wright, P.E. (1993) Formation of a molten globule intermediate early in the kinetic folding pathway of apomyoglobin, *Science* 262, 892–896.
- Raschke, T.M., & Marqusee, S. (1997) The kinetic folding intermediate of ribonuclease H resembles the acid molten globule and partially unfolded molecules detected under native conditions, *Nat. Struct. Biol.* 4, 298–304.

Mayr, L.M., Odefey, C., Schutkowski, M., & Schmid, F.X. (1996) Kinetic analysis of the unfolding and refolding of ribonuclease T1 by a stopped-flow double-mixing technique, *Biochemistry* 35, 5550–5561.

Problem 13–5: The figure displays the behavior of the observed rate constants, k_{obs} , in units of seconds⁻¹ for folding and unfolding of human lysozyme. The rate constants for folding were obtained by rapidly diluting the unfolded polypeptide from a solution of 4.5 M guanidinium chloride to the noted final concentration; and those for refolding, by rapidly mixing the native protein with a solution of guanidinium chloride to produce the noted final concentration.



- The rate at which the unfolded protein refolds is governed by Equation 13–10. Derive a similar equation describing the change in the concentration of the folded form ($[F]$) as a function of time.
- Why is the curve for the observed rate constant of folding continuous with the curve for the observed rate constant of unfolding?
- Which intrinsic rate constant dominates the observed rate constant in the unfolding region?
- Which dominates in the folding region?
- Why does the behavior of k_F as a function of the concentration of guanidinium chloride indicate that there is at least one kinetic intermediate in the folding reaction?
- Estimate the rate constant for the formation of this intermediate in the absence of guanidinium chloride.
- Estimate the rate constant for the formation of the native state from the intermediate state or the

intermediate states in the absence of guanidinium chloride.

- Is the isomerization of peptide bonds amino-terminal to prolines involved in the folding of human lysozyme? How did you decide?
- How does the value of k_U change as the concentration of guanidinium is increased up to 5 M?
- In the transition region in which the equilibrium constant for folding can be measured, what are the relative values of the rate constants k_F and k_U ?

Problem 13–6: Using the notation of Equations 13–31 through 13–33, write a set of equations that describes the possible ways that protein disulfide-isomerase could catalyze the formation of the mixed disulfide between glutathione and a cysteine on a folding protein.

Assembly of Oligomeric Proteins

When oligomeric proteins are dissolved in solutions of guanidinium chloride, they dissociate into their constitutive polypeptides that unfold to random coils. When the guanidinium chloride is removed from the solution, the random coils refold, and the **refolding monomers reassociate to form the native state**. For example, the inorganic diphosphatase from *E. coli* is an $(\alpha_3)_2$ hexamer in its native state. When dissolved in 5 M guanidinium chloride at pH 7, it dissociates into single α polypeptides ($n_{\text{aa}} = 175$), as judged by sedimentation equilibrium, that are random coils, as judged by their sedimentation coefficient ($s_{20,w}^0 = 0.59$ S) and intrinsic viscosity ($[\eta] = 22$ cm³ g⁻¹). When the guanidinium chloride is removed by dialysis, 80–90% of the enzymatic activity slowly returns, and the protein that results is an $(\alpha_3)_2$ hexamer indistinguishable in sedimentation coefficient, optical rotatory dispersion, or ultraviolet absorption spectrum from the original native enzyme.⁴⁴⁴ The native protein contains no cystines, but it does contain cysteines. Renaturation is successful only if an external mercaptan, which mimics the reduced glutathione normally present in the cytoplasm, is added to prevent adventitious intermolecular and intramolecular formation of cystine, a reaction that interferes with proper refolding.

The steps in the assembly of an oligomeric protein can be followed by **quantitative cross-linking**. Phosphoglycerate mutase from *S. cerevisiae* is an $(\alpha_2)_2$ tetramer. When it is dissolved in 4 M guanidinium chloride, it dissociates into random coils of the α polypeptide as judged by circular dichroism (Figure 13–17A).⁴⁴⁵ When the solution is diluted 40-fold to 0.1 M guanidinium chloride, greater than 80% of the molar ellipticity of the native state is regained in less than 30 s (Figure 13–17B). At this point greater than 80% of the protein is still monomeric. The appearance of dimers and tetramers as a function of time could be followed by quantitative cross-linking to cata-

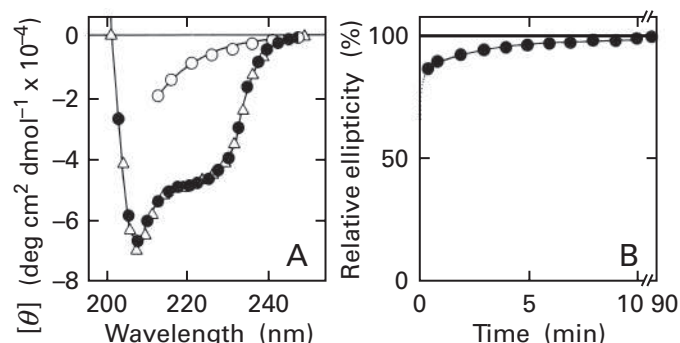
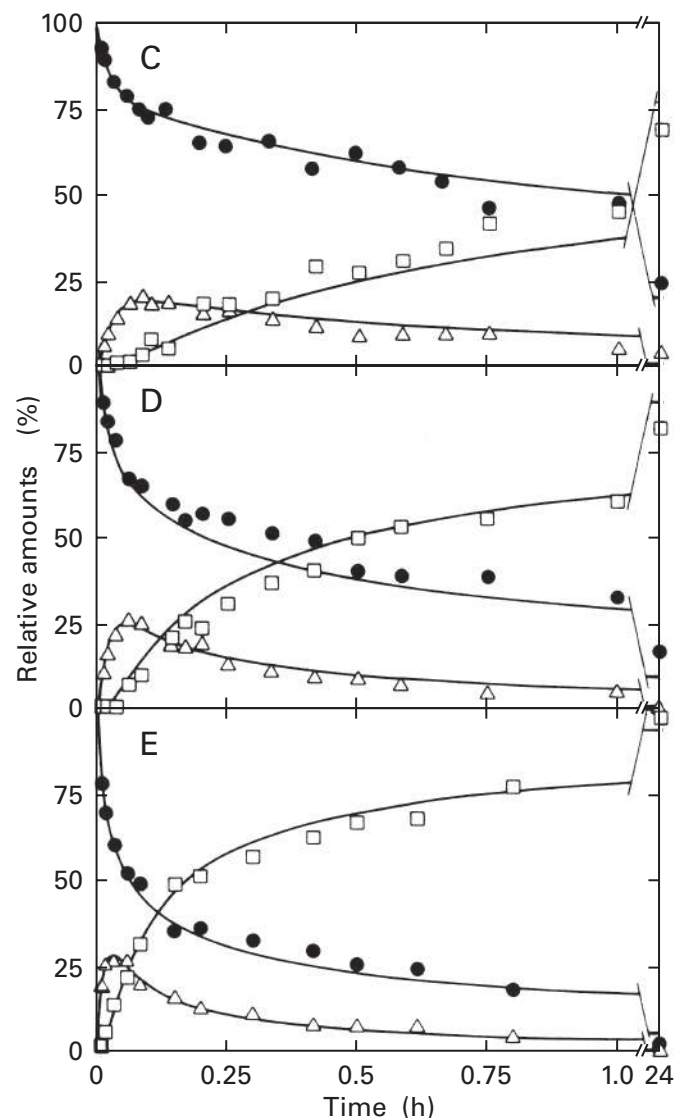
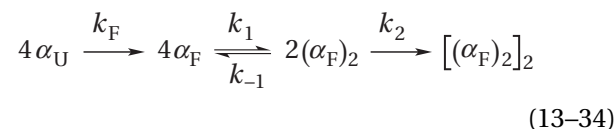


Figure 13-17: Assembly of yeast phosphoglycerate mutase following dilution from 4 to 0.1 M guanidinium chloride.⁴⁴⁵ (A) Far-ultraviolet circular dichroic spectra of native enzyme (●), native enzyme in 0.1 M guanidinium chloride (△), and enzyme in 4 M guanidinium chloride (○); all measurements were made at a concentration of protein of 1.7 mg mL⁻¹. Molar ellipticity ($[\theta]$) in units of degree centimeter² (decimole of peptide bonds)⁻¹ is presented as a function of wavelength (nanometers). (B) Regain of molar ellipticity at 225 nm upon dilution from 4 to 0.1 M guanidinium chloride. Ellipticity is presented as a function of time (minutes) in relative units where 0% is the molar ellipticity of fully unfolded protein and 100% is the molar ellipticity of the native protein (see panel A). (C-E) Assembly of the oligomer. At the noted times after initiation of folding by dilution from 4 to 0.1 M guanidinium chloride, samples were removed and cross-linked quantitatively with 1% glutaraldehyde for 2 min, and the complexes between the resulting covalent oligomers and dodecyl sulfate were submitted to electrophoresis on gels of polyacrylamide in the presence of dodecyl sulfate. The amounts of monomer (●), dimer (△), and tetramer (□) were assessed by scanning the stained gels for absorbance. The relative amounts of monomer, dimer, and tetramer (as a percentage of the sum of the three amounts) are plotted as a function of the time (hours) between dilution of the guanidinium chloride and the addition of the glutaraldehyde. The final concentrations of protein were (C) 11, (D) 21, and (E) 37 $\mu\text{g mL}^{-1}$. The solid curves were drawn in all three panels with integrated rate equations based on Equation 13-34 with $k_1 = 6.25 \times 10^3 \text{ M}^{-1} \text{ s}^{-1}$, $k_{-1} = 6.0 \times 10^{-3} \text{ s}^{-1}$, and $k_2 = 2.75 \times 10^4 \text{ M}^{-1} \text{ s}^{-1}$. The temperature for all of these experiments was 20 °C, and they were run at pH 7.5. Reprinted with permission from ref 445. Copyright 1983 *Journal of Biological Chemistry*.



logue the species present at each point (Figure 13-17C-E).⁴⁴⁵ No trimers were observed, as would be expected. From an examination of the progress of the reaction, it could be concluded that the dimer was the initial oligomer, which, as it built up in concentration, dimerized to produce tetramer. Although the circular dichroism of the sample changed insignificantly as the reaction progressed (Figure 13-17B), the intrinsic fluorescence of the protein increased in concert with the oligomerization. Both the rate of the oligomerization (Figure 13-17C-E) and the rate of the increase in fluorescence (except for a small immediate increase of 20% that was invariant) were dependent on the absolute concentration of the protein.

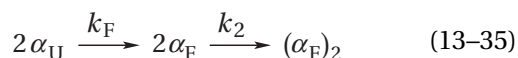
The **kinetics** of both the oligomerization of phosphoglycerate mutase and the increase in fluorescence could be accounted for quantitatively⁴⁴⁵ by the mechanism



wherein all unfolded polypeptides, α_U , have folded in the first 30 s; the folded monomer, α_F , regains 20% of the fluorescence of the native state; and both the dimer and the tetramer have the full fluorescence of the native state. That the folded monomer and reassembled dimer can be digested with trypsin while the reassembled tetramer cannot⁴⁴⁶ suggests that the polypeptides in the monomer and the dimer are loosely folded and regain their fully compact native state only following tetramerization. That both the monomer and the dimer possess some enzymatic activity⁴⁴⁶ suggests that they are properly folded. The tetramer is produced in quantitative yield

with full enzymatic activity at these concentrations of protein ($<50 \mu\text{g mL}^{-1}$).⁴⁴⁶ In this oligomerization, the association of two dimers to form the tetramer is the rate-limiting step in the reaction (Equation 13–34), but in the oligomerization of the tetramerization domain ($n_{\text{aa}} = 30$) of human cellular tumor antigen p53, it is the association of two monomers to form a dimer that is the rate-limiting step, and the association of the two dimers to form the tetramer is so rapid that it is kinetically silent.⁴⁴⁷

The assembly of a dimer from its dissociated random coils is an even simpler reaction. Porcine mitochondrial malate dehydrogenase is an α_2 dimer that can be reversibly unfolded in several different ways. After random coils, α_{U} , of the α polypeptide are transferred to a solution at neutral pH, coincident with the dilution of the denaturant, the reappearance of enzymatic activity shows the same time course regardless of the mode of the original denaturation (Figure 13–18).⁴⁴⁸ The time course displays two phases, a lag followed by an increase. The increase in activity has a second-order dependence on the concentration of protein. The lag is unaffected by the concentration of protein and is a first-order process. The results can be explained quantitatively with the following mechanism:



if only the dimer, $(\alpha_{\text{F}})_2$, and not the folded monomer, α_{F} , is enzymatically active. At pH 7.6 and 20 °C,

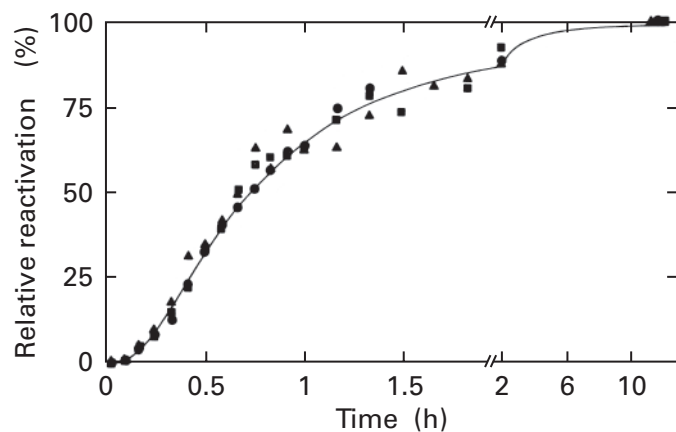


Figure 13–18: Reassembly and reactivation of porcine mitochondrial malate dehydrogenase at a concentration of 60 nM and at pH 7.6 and 20 °C after dilution from various denaturants.⁴⁴⁸ The protein was unfolded at pH 2.3 alone (●), in 6 M guanidinium chloride (■), or in 6 M urea (▲). After it was fully unfolded and dissociated into its separate polypeptides in each instance, it was diluted to initiate refolding and reassembly. Samples were removed at the noted times and assayed for enzymatic activity; the enzymatic activity is presented in relative units with 0% being the immediately observed enzymatic activity ($<4\%$ of the final) and 100% being the enzymatic activity after full reactivation (>24 h). The curve is for the integrated rate equation for the mechanism of Equation 13–35 with $k_1 = 6.5 \times 10^{-4} \text{ s}^{-1}$ and $k_2 = 3 \times 10^4 \text{ M}^{-1} \text{ s}^{-1}$. Reprinted with permission from ref 448. Copyright 1979 American Chemical Society.

$k_1 = 0.0006 \text{ s}^{-1}$ and $k_2 = 30,000 \text{ M}^{-1} \text{ s}^{-1}$. The value of k_{F} is too small to be the rate constant for the rapid refolding of a polypeptide with the correct proline isomers to form a structure with no domains. The rate is probably slow because isomerizations of peptide bonds amino-terminal to prolines are required or because the two structural domains of the monomer observed in the crystallographic molecular model of the protein⁴⁴⁹ associate slowly or because both of these problems must be overcome before the monomer has regained **sufficient native structure to recognize another monomer** and dimerize with it.

In the case of the reassembly of the dimer of aspartate transaminase from *E. coli*, there are two consecutive, slow, first-order, unimolecular steps which produce a molten globular monomer that is enzymatically inactive but that has sufficient native structure to dimerize. This monomer displays a circular dichroic spectrum in the far ultraviolet similar to that of the native protein. Its dimerization and the formation of the final enzymatically active native state are rapid, and because these later steps follow the slow rate-limiting formation of the molten globular monomer, they are kinetically silent.⁴⁵⁰ In the folding and assembly of the dimer of steroid Δ -isomerase from *Pseudomonas testosteroni*, however, all three steps, the unimolecular formation of an enzymatically inactive monomeric intermediate (60 s^{-1} at 25 °C), its bimolecular association ($60,000 \text{ M}^{-1} \text{ s}^{-1}$ at 25 °C), and the formation of the final enzymatically active native structure (0.017 s^{-1} at 25 °C) could be resolved kinetically.⁴⁵¹ These experiments demonstrate that a monomeric intermediate does not have to assume its fully native state before it oligomerizes and that, in such a situation, **further isomerizations** can then occur within each subunit **following oligomerization** to produce the final native state. The final step in the reactivation has a rate constant suggesting that the isomerization of a peptide bond amino-terminal to a proline is involved.

Transcriptional repressor arc from bacteriophage P22 is also an α_2 dimer but of much smaller subunits ($n_{\text{aa}} = 53$). At low concentrations, its rate of folding and assembly is second-order in the concentration of protein, can be fit by a bimolecular rate equation, shows no evidence of any intermediates, and is complete in less than a second. Because the reaction is second-order, the rate-limiting step is dimerization (Equation 13–35) under these conditions. As the concentration of protein is increased above $20 \mu\text{M}$, however, there is a change in the rate-limiting step as the rate of dimerization becomes so fast that a preceding unimolecular step (Equation 13–35), the rate constant of which does not depend on the concentration of protein, becomes rate-limiting.⁴⁵² Because the change in fluorescence with time shows no evidence for the formation of any intermediate when the dimerization is the rate-limiting step,⁴⁵³ the unimolecular step is probably an **unfavorable preequilibrium** producing

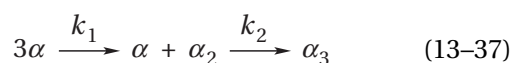
an unstable monomeric state that is competent to dimerize. This unstable monomer is then captured and stabilized by the favorable dimerization.

The assembly of a trimer (Figure 9–11) is somewhat more complicated than that of either a tetramer or a dimer because the addition of the third subunit to the dimer is quite different from the initial combination of two monomers to form the dimer. The catalytic subunit of aspartate carbamoyltransferase (Figure 9–37) is an α_3 trimer. The assembly of trimers of the catalytic subunit from random coils of the α polypeptide is a first-order process with no evident intermediates and a rate constant of $2 \times 10^{-4} \text{ s}^{-1}$ at 0°C .⁴⁵⁴ It seems that again a slow isomerization of the partially folded, monomeric α polypeptide is the rate-limiting step in the assembly from random coils. To circumvent the barrier presented by this isomerization to the kinetic observation of intermediates in the process, native α_3 trimer was dissociated into globular rather than unfolded α monomers with thiocyanate ion ($\text{S}=\text{C}=\text{N}^-$),⁴⁵⁵ which is a milder denaturant than either urea or guanidinium ion. It is an anion that salts in protein as does urea or guanidinium but not so vigorously. The enzymatically inactive α monomers that result retain most of the circular dichroic ellipticity and ultraviolet absorption of the native α_3 trimers and have a frictional ratio f/f_0 of 1.27.⁴⁵⁵ These **globular α monomers** assemble readily to form α_3 trimers after the dilution of the thiocyanate.

When the assembly was followed by quantitative cross-linking, α monomers turned directly into α_3 trimers with no evidence for the formation of any α_2 dimers (<3%). The appearance of α_3 trimers was coincident with the return of enzymatic activity. Both of these processes, however, were strictly second-order in the concentration of α monomer:⁴⁵⁵

$$\frac{d[\alpha_3]}{dt} = k_{\text{obs}}[\alpha]^2 \quad (13-36)$$

A mechanism consistent with both of these results is



where $k_2 \gg k_1$. When the third monomer adds to the dimer, two interfaces form simultaneously, and this reaction could have a much lower standard free energy of activation than the formation of the dimer itself. Because the second step in Equation 13–37 is so much faster than the first, no α_2 dimer accumulates. The first step in Equation 13–37, however, a bimolecular reaction, is the rate-limiting step.

When homooligomeric proteins are assembled from random coils, the observations are consistent with the first step in the process being the folding of the random coil to a globular structure. In many instances,

this structure is loosely folded. For example, it may be sensitive to endopeptidolytic cleavage. This globular monomer either combines directly with other globular monomers to form the oligomer or it undergoes one or more isomerizations before it is competent to assemble. These isomerizations may be isomerizations of peptide bonds amino-terminal to critical prolines or rearrangements of domains, but these possibilities have not been validated. The competent monomers then assemble in simple, reasonable bimolecular steps to form the enzymatically active oligomer. When the reactions can be observed, the rate constants measured for these bimolecular steps are between 10^4 and $10^6 \text{ M}^{-1} \text{ s}^{-1}$ at 25°C ,^{445,448,455} several orders of magnitude below diffusion-controlled rates for the collision of molecules of this size. Therefore, they proceed with significant standard free energies of activation.

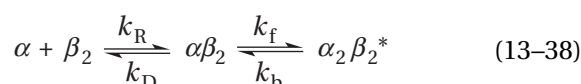
Whether or not **enzymatic activity** is displayed by the various compact intermediates in this process seems to be a property of the individual protein. Both α monomer and α_2 dimer of phosphoglycerate mutase have enzymatic activity.⁴⁴⁶ Fumarate hydratase, an $(\alpha_2)_2$ tetramer, can be denatured to random coils and reassembled to an α_2 dimer. The α_2 dimer is enzymatically inactive until it assembles to the $(\alpha_2)_2$ tetramer.⁴⁵⁶ Porphobilinogen synthase from *Pisum sativum*, an $[(\alpha_2)_2]_2$ octamer, can be disassembled to α_2 dimers by dilution. Only the octamer and the $(\alpha_2)_2$ tetramer are enzymatically active.⁴⁵⁷ The single active site of HIV-1 retropepsin from human immunodeficiency virus type 1 is formed from both subunits of the α_2 dimer, so it is not surprising that the α monomer has no enzymatic activity.⁴⁵⁸ When fructose-bisphosphate aldolase, an $(\alpha_2)_2$ tetramer, is denatured to random coils that are then transferred to a solution at pH 5.5, the random coils fold to form α monomers that have the sedimentation coefficient of a globular protein of their length and the circular dichroic spectra and ultraviolet spectra of the native protein. Their enzymatic activity cannot be determined because these α monomers oligomerize too rapidly to $(\alpha_2)_2$ tetramers when mixed with substrates,⁴⁵⁹ but they must bind those substrates for their assembly to be affected by them.

The **assembly of heterooligomers** constructed from several copies of each of two or more different polypeptides is somewhat more complex than that of homooligomers. When the assembly of a heterooligomer is studied, the reactants employed are the globular, homooligomeric subunits, such as catalytic subunits and regulatory subunits of aspartate carbamoyltransferase (Figure 9–37). It is generally assumed, in the absence of any evidence, that under physiological circumstances the homooligomeric subunits assemble first and then combine to form the heterooligomer. Only those proteins formed from two or more polypeptides translated from different messenger RNAs are of interest. Proteins containing different polypeptides arising from the post-

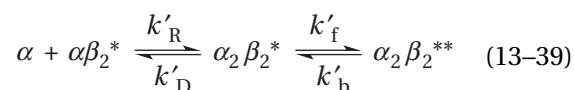
translational cleavage of identical larger polypeptides fold and assemble as simple homooligomers before the posttranslational modification occurs.

Alkanal monooxygenase (FMN-linked) from *Vibrio harveyi* is an $\alpha\beta$ heterodimer in which the α subunit ($n_{aa} = 355$) and the β subunit ($n_{aa} = 324$) are homologous and superposable, but neither of these subunits will form a homodimer. When they are expressed separately, each is a globular monomer containing about 50–60% of the α helix of a monomer in the native heterodimer but no discernible tertiary structure as judged from their nuclear magnetic resonance spectra. These properties are consistent with these two monomers being molten globules. Nevertheless, when they are mixed together, these α monomers and β monomers heterodimerize and form native alkanal monooxygenase.⁴⁶⁰ Either the two molten globular forms dimerize and then assume their native states while in the dimer, or the fully native states of the α monomer and the β monomer are present in the separate solutions of each at undetectably low equilibrium concentrations, and it is these forms that dimerize while the dimerization pulls the equilibria in the direction of the native states.

Tryptophan synthase from *E. coli* is another simple example of the assembly of a heterooligomer.⁴⁶¹ This protein is an $\alpha\beta_2\alpha$ heterotetramer.⁴⁶² When it is dissociated into its components, the products are α monomers and β_2 dimers, and both can be obtained in globular, folded states. When α monomer is mixed with excess β_2 dimer, the major product is the complex $\alpha\beta_2^*$. It forms in a reaction the kinetics of which are consistent with the mechanism



where the complex $\alpha\beta_2^*$ is an isomerized form of the initial intermediate $\alpha\beta_2$. The rate constants k_R , k_D , k_f , and k_b for this process at 25 °C are $1 \times 10^6 \text{ M}^{-1} \text{ s}^{-1}$, 3 s^{-1} , 6 s^{-1} , and 0.001 s^{-1} , respectively. When excess α monomer is then added to the $\alpha\beta_2^*$ complex, the next step in the assembly has kinetic behavior consistent with the mechanism



and the rate constants k'_R , k'_D , k'_f , and k'_b for this process at 25 °C are $1.6 \times 10^6 \text{ M}^{-1} \text{ s}^{-1}$, 26 s^{-1} , 16 s^{-1} , and 0.002 s^{-1} , respectively.

Each time an interface forms between an α monomer and one of the two β monomers in the β_2 dimer of tryptophan synthase, an isomerization of the structure of either the participating β monomer or the conjoined α monomer, or both, occurs, producing the asterisked conformer. The isomerizations producing

the conformers $\alpha\beta_2^*$ or $\alpha\beta_2^{**}$ are too rapid to be isomerizations of prolines. They presumably represent **rearrangements of the structures after the association** of the α and β subunits and are similar to the changes that permit the tetramer of phosphoglycerate mutase to resist endopeptidolytic degradation or that permit enzymatically inactive subunits to regain full enzymatic activity after oligomers such as malate dehydrogenase, aspartate transaminase, and fumarate hydratase reach the native stoichiometry. The equilibrium constant ($K_{is} = k_f/k_b$) for the isomerization following the addition of the first α monomer ($K_{is} = 6000$) is the same as that following the addition of the second α monomer ($K'_{is} = 8000$), indicating that the same local adjustments are occurring after each α monomer adds in turn.

Aspartate carbamoyltransferase is constructed from two catalytic C subunits that are α_3 trimers and three regulatory R subunits that are β_2 dimers. From its crystallographic molecular model (Figure 9–37), it is clear that only certain steps are possible in its assembly from separated C subunits and R subunits (Figure 13–19).⁴⁶³ The intermediates that appear during the assembly of the intact $(\alpha_3)_2(\beta_2)_3$ heterododecamer (C_2R_3 heteropentamer) have been followed by quenching the assembly of radioactive catalytic or regulatory subunits and their unradioactive complements with large excesses of unradioactive catalytic or succinylated catalytic subunits, respectively.⁴⁶³ The specific radioactivity of the various mosaic oligomers, which, because of the excess negative

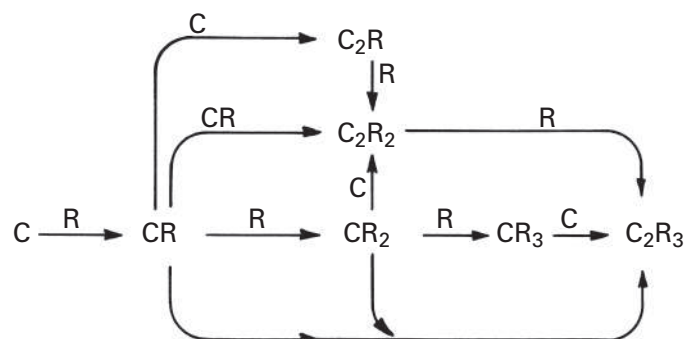


Figure 13-19: Intermediates in the assembly of aspartate carbamoyltransferase.⁴⁶³ Catalytic (C) or regulatory (R) subunits that had been made radioactive by iodinating their tyrosines with ¹²⁵I (Reaction 10–33) were mixed with excess unradioactive R or C subunits, respectively, to initiate assembly. At various times, the assembly reaction was quenched with either excess C subunit or excess succinylated C subunit to cap off partially formed complexes and scavenge all unreacted R subunit. From examining the specific radioactivity of complexes separated by electrophoresis, an estimate of the relative concentrations of all of the intermediates in the process of assembly at each time point could be made. The changes in these relative concentrations with time were used to formulate the assembly diagram displayed in the figure. Four of the steps in this process are rapidly reversible: $C + R \rightleftharpoons CR$, $CR + R \rightleftharpoons CR_2$, $CR_2 + R \rightleftharpoons CR_3$, and $CR + C \rightleftharpoons C_2R$. In contrast, processes forming the complexes C_2R_2 and C_2R_3 are essentially irreversible because these complexes are so stable. Reprinted with permission from ref 463. Copyright 1980 *Journal of Biological Chemistry*.

charge (Equation 10–27) on the succinylated C subunits, can be separated from each other by electrophoresis, permits the concentrations of the various intermediates at the time the reaction was quenched to be calculated.

When a limiting concentration of C subunit is mixed with various excesses of R subunit, equilibrium mixtures of the intermediates CR, CR₂, and CR₃ are formed. Subsequent addition of excess C subunit causes CR₃ to be trapped as CR₃C, the intact native protein, and CR₂ to be trapped as CR₂C.⁴⁶³ When excess C subunit is mixed with a limiting concentration of R subunit, the only two products other than unreacted C subunit are CR₃C and CR₂C, with the former in the majority.^{464,465} The complex CR₂C can be isolated as a stable protein. When it is combined with R subunit, it produces CR₃C in a clean bimolecular reaction.⁴⁶⁴ In these experiments, most of the intermediates in the general scheme (Figure 13–19) have been directly observed, and rate constants and equilibrium constants for their interconversion have been established.⁴⁶⁶ Most of the steps in the scheme seem to occur simultaneously, and **different pathways** become more or less important as concentrations of the subunits are changed.

The pyruvate dehydrogenase complex of *E. coli* is composed of three different polypeptide chains, α , β , and γ . The protein can be resolved into these three independent components. These are the dihydrolipoyllysine-residue acetyltransferase core, the pyruvate dehydrogenase (acetyl-transferring) subunits, and the dihydrolipoyl dehydrogenase subunits. The dihydrolipoyllysine-residue acetyltransferase core is an octahedral α_{24} oligomer (Figure 9–23), pyruvate dehydrogenase (acetyl-transferring) is a β_2 dimer, and dihydrolipoyl dehydrogenase is a γ_2 dimer. No association can be detected between the β_2 dimers of pyruvate dehydrogenase (acetyl-transferring) and the γ_2 dimers of dihydrolipoyl dehydrogenase.⁴⁶⁷ Therefore, the α_{24} oligomer of the dihydrolipoyllysine-residue acetyltransferase serves as the **point of attachment** of the other components.

Unlike the closely related dihydrolipoyllysine-residue succinyltransferase from the 2-oxoglutarate dehydrogenase complex, which can associate with only six β_2 dimers of oxoglutarate dehydrogenase (succinyl-transferring) at saturation because of a steric effect,⁴⁶⁸ the empty α_{24} oligomer of the dihydrolipoyllysine-residue acetyltransferase from *E. coli* can associate with up to 24 β_2 dimers of pyruvate dehydrogenase (acetyl-transferring).^{467,469} Presumably, in the saturated complex, one of the two faces on each of the 24 β_2 dimers occupies one of the 24 equivalent faces of the octahedral α_{24} oligomer with no steric hindrance. The empty α_{24} oligomer of the dihydrolipoyllysine-residue acetyltransferase can also associate with as many as 20 γ_2 dimers of dihydrolipoyl dehydrogenase in the absence of pyruvate dehydrogenase (acetyl-transferring).⁴⁶⁹

When both β_2 dimers of pyruvate dehydrogenase

(acetyl-transferring) and γ_2 dimers of dihydrolipoyl dehydrogenase are added together to the dihydrolipoyllysine-residue acetyltransferase core, substoichiometric amounts of each are bound,⁴⁶⁹ presumably because of **steric crowding**. Certainly the native protein, which has an average of about 12 γ_2 dimers of dihydrolipoyl dehydrogenase and an average of somewhat less than 24 β_2 dimers of pyruvate dehydrogenase (acetyl-transferring),⁴⁶⁷ appears to be a crowded structure (Figure 13–20).⁴⁷⁰ When a preformed complex containing an average of 12 γ_2 dimers of dihydrolipoyl dehydrogenase for each α_{24} oligomer of dihydrolipoyllysine-residue acetyltransferase is mixed with increasing amounts of pyruvate dehydrogenase (acetyl-transferring), about 22 β_2 dimers of pyruvate dehydrogenase (acetyl-transferring) bind to the α_{24} oligomers at saturation, and the overall enzymatic activity increases in direct proportion to the number bound.⁴⁶⁷ All of these results suggest that β_2 dimers of pyruvate dehydrogenase (acetyl-transferring) and γ_2 dimers of dihydrolipoyl dehydrogenase add at random to the respective faces on the α_{24} oligomer of dihydrolipoyllysine-residue acetyltransferase, at least under the circumstances of these experiments, until there is no more room left around the core. What is not clear is whether the dimers of dihydrolipoyl dehydrogenase and pyruvate dehydrogenase (acetyl-transferring) add at random to the core during normal assembly within the cell until no more can fit or there is some ordered sequence that determines the final stoichiometry.

The 30S subunit of a ribosome from *E. coli* (Figure 11–5) is composed of a single strand of 16S ribosomal RNA ($n_{\text{nuc}} = 1541$)⁴⁷¹ and 21 polypeptides that, when

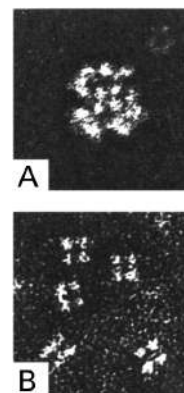


Figure 13–20: Electron micrographs of (A) the pyruvate dehydrogenase complex from *E. coli* and (B) the core of dihydrolipoyllysine-residue acetyltransferase from the same protein.⁴⁷⁰ Both specimens were adsorbed onto a thin, supported layer of amorphous carbon on an electron microscopic grid and negatively stained with sodium methylphosphotungstate. Magnification is 300000 \times . The complete complex was purified directly from a homogenate of the bacteria; the acetyltransferase core was prepared from the complete complex by stripping away dihydrolipoyl dehydrogenase and pyruvate dehydrogenase (acetyl-transferring). Reprinted with permission from ref 470. Copyright 1971 Cold Spring Harbor Laboratory.

folded and assembled, constitute its 21 subunits. When the ribosomal RNA and the separated individual polypeptides are mixed together, they spontaneously reassemble in high yield to form 30S subunits that are fully competent to participate in protein biosynthesis.⁴⁷² The assembly of the intact 30S subunit from its components (Figure 13–21)⁴⁷³ proceeds through an **explicit sequence of steps** beginning with the binding of a few of the subunits to the 16S ribosomal RNA itself. As the assembly progresses, the binding to the 16S ribosomal RNA of the subunits earlier in the sequence of events or the binding of polypeptides to complexes between the 16S ribosomal RNA and other polypeptides creates sites to which subunits later in the sequence of events can attach (Figure 13–21). If a polypeptide is added to the mixture before all of the subunits that must precede it have been incorporated, it will not bind to the partially assembled 30S subunit. An **assembly map**, necessarily of greater complexity but describing a similar hierarchically ordered process, has been drawn for the assembly of the

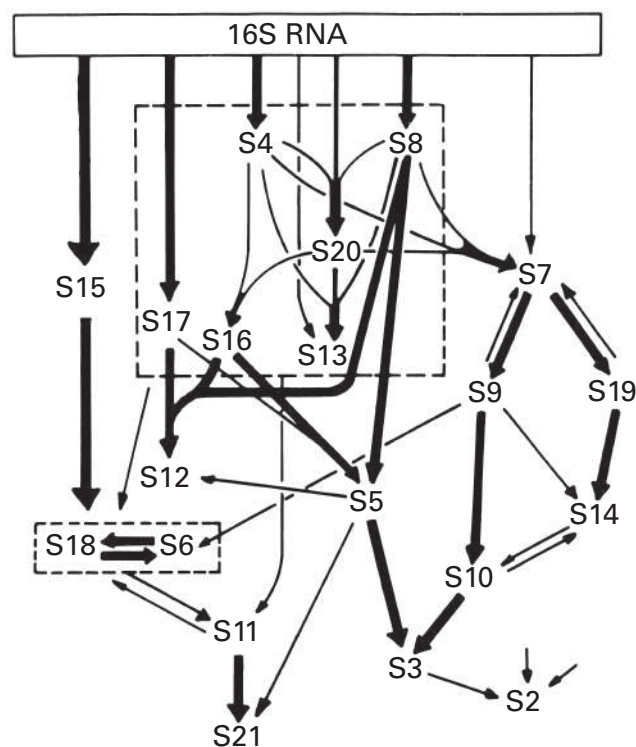


Figure 13–21: Assembly diagram for the 30S subunit of the ribosome from *E. coli*.⁴⁷³ The sequence of events was determined by mixing, in various combinations, the 21 purified polypeptides with the purified 16S RNA and assaying for formation of a complex or complexes among the components. For example, only polypeptides S15, S17, S4, and S8 would bind alone to 16S ribosomal RNA. Polypeptide S20 will form a complex with 16S RNA only when subunits S4 and S8 have been incorporated. Polypeptide S13 binds to 16S RNA only when subunits S4, S8, and S20 have been incorporated, and so forth. Upon binding each polypeptide becomes a subunit of the assembling 30S ribosomal subunit. Reprinted with permission from ref 473. Copyright 1974 *Journal of Biological Chemistry*.

50S ribosomal subunit of *E. coli* from 23S ribosomal RNA, 5S ribosomal RNA, and 31 polypeptides.⁴⁷⁴

From the crystallographic molecular model of the 30S subunit,⁴⁷⁵ some inferences can be drawn to explain the order in which its subunits are incorporated into the assembling particle (Figure 13–21). None of the polypeptides except S4, S8, S17, and S15 can add until other subunits have been incorporated. Subunits S6 and S18 form an intimate complex in one corner of the complete 30S subunit, and subunits S10, S14, and S3 form an intimate complex in another corner. These close associations explain the interdependences between the additions of the polypeptides of these subunits during assembly. Most of the subunits of the intact 30S subunit, however, have little if any contact with each other in the final particle.

The last polypeptides to add to the assembling particle, S3, S10, S14, S11, S18, S5, S12, and S9 (Figure 13–21), all form contacts with at least one and as many as five double helices of RNA also contacted by the subunit or subunits that must precede them onto the particle. These relationships suggest that the preceding subunit or subunits control the **orientations of these double helices** so that the site for the binding of the following subunit among these double helices is either created or stabilized.

The earlier subunits to add to the assembling particle, however, subunits S19, S13, S7, and S20, do not share either direct contacts or contacts with double helices in common with the subunits that must precede them. This fact suggests that **global conformational changes** of the 16S ribosomal RNA are effected by the earliest polypeptides to add, namely those of subunits S4, S8, S20, and S7, to create the distant sites for the polypeptides of subunits S19, S13, S7, and S20.

Not only does the structure of the 16S ribosomal RNA seem to adjust upon the association of the individual subunits but also the conformations of the separated subunits seem to adjust, sometimes dramatically, upon their association. One polypeptide that seems to be almost structureless before it associates with the 16S ribosomal RNA is polypeptide S4 ($n_{aa} = 203$). The nuclear magnetic resonance spectrum of polypeptide S4 under the conditions in which assembly takes place is almost indistinguishable from its spectrum in 8 M urea, which is the spectrum of the sum of the amino acids from which it is composed.⁴⁷⁶ This result indicates that, when alone in solution, polypeptide S4 cannot assume a unique native state. The circular dichroic spectrum⁴⁷⁶ and frictional ratio ($f/f_0 = 1.7$), however, are not those of a fully random coil ($f/f_0 = 2.4$ for $n_{aa} = 245$),⁴⁷⁷ and an explanation of these results and those from nuclear magnetic resonance spectroscopy would be that the polypeptide in solution is rapidly passing through an array of loosely folded conformations, none of which is unique. When it is bound to the 16S ribosomal RNA, the subunit S4 assumes a defined structure with seven α helices and

four strands of β structure, but it is flattened and spread over the surface of the 30S subunit.⁴⁷⁵ As it associates with the 16S ribosomal RNA, its final structure could easily be dictated solely by the noncovalent interactions in which it participates as it spreads over the surface of the folded polynucleotide.

Some ribosomal polypeptides seem to enter the assembling 30S subunit as folded globular proteins. For example, polypeptide S17 ($n_{aa} = 83$) under the conditions of assembly has both the frictional ratio ($f/f_0 = 1.24$; calculated from its sedimentation coefficient) and the intrinsic viscosity ($[\eta] = 4.2 \text{ cm}^3 \text{ g}^{-1}$) of a globular protein and the molar mass, as determined by sedimentation equilibrium, of a monomer.⁴⁷⁸ Other polypeptides, however, seem to be elongated but compact proteins under the conditions of assembly. For example, polypeptides S3, S5, S6, and S7 have frictional ratios $f/f_0 = 1.4\text{--}1.6$ that are too large to be those of globular proteins.^{477,478} Like subunit S4, subunit S3 is flattened against the surface of the assembled, intact 30S subunit. Subunit S5 also assumes a flattened, elongated conformation, but subunits S6 and S7 are both globular in the final structure and must become so as they associate with the assembling particle.

Suggested Reading

Hermann, R., Rudolf, R., Jaenicke, R., Price, N.C., & Scobbie, A. (1983) The reconstitution of denatured phosphoglycerate mutase, *J. Biol. Chem.* 258, 11014–11019.

Kim, D.H., Jang, D.S., Nam, G.H., Yun, S., Cho, J.H., Choi, G., Lee, H.C., & Choi, K.Y. (2000) Equilibrium and kinetic analysis of folding of ketosteroid isomerase from *Comamonas testosteroni*, *Biochemistry* 39, 13084–13092.

Problem 13-7: The dissociation constant for the first step in the assembly of tryptophan synthase (Equation 13-38) is

$$K_{d1} = \frac{2[\alpha][\beta_2]}{[\alpha\beta_2^*]}$$

(A) Why is the 2 present in the numerator?

The dissociation constant for the second step (Equation 13-39) is

$$K_{d2} = \frac{[\alpha][\alpha\beta_2^*]}{2[\alpha_2\beta_2^{**}]}$$

(B) Why is the 2 present in the denominator?

(C) Calculate the equilibrium constants K_{d1} and K_{d2} at 25 °C. Recall that at equilibrium

$$k_R[\alpha][\beta_2] = k_D[\alpha\beta_2]$$

and

$$k_f[\alpha\beta_2] = k_b[\alpha\beta_2^*]$$

Assembly of Helical Polymeric Proteins

There are two classes into which helical polymeric proteins can be divided when the question of assembly is considered: those that assemble irreversibly and those that assemble reversibly. Those that assemble irreversibly polymerize initially by the noncovalent assembly of monomeric subunits, but the polymers are often strengthened secondarily by the formation of naturally occurring covalent cross-links between adjacent subunits (Figure 3-19). Those that assemble reversibly require that reversibility for their biological role, and their assembly is not an approach to equilibrium but the result of a steady state.

Examples of helical polymeric proteins that assemble irreversibly are collagen (Figure 9-34), intermediate filaments (Figure 9-35), thick filaments, and fibrin.

The thick fibers of **fibrin** that form clots of blood are readily observed in scanning electron micrographs. They are lateral aggregates of thinner protofibrils of fibrin; these **protofibrils** of fibrin are assembled irreversibly from a protomeric unit known as **fibrinogen** (Figure 13-22A),⁴⁷⁹⁻⁴⁸¹ which is a freely soluble $(\alpha\beta\gamma)_2$ heterohexamer.⁴⁷⁹ Each molecule of fibrinogen is constructed from two $\alpha\beta\gamma$ heterotrimers arrayed about a 2-fold rotational axis of symmetry. Each of the two $\alpha\beta\gamma$ heterotrimers contains a long segment of rope (111 amino acids) in which the three strands, one from each of the α , β , and γ polypeptides, are wound around each other in a triple α -helical coiled coil (Figure 6-28). At the two ends of each rope are globular domains known as the terminal domain and the central domain, respectively. A terminal domain is composed of the carboxy-terminal domains of the β subunit and the γ subunit; the central domain is composed of the folded amino-terminal regions of all three polypeptides. Two $\alpha\beta\gamma$ heterotrimers are associated at their central domains around the 2-fold rotational axis of symmetry^{481,482} to produce the molecule of fibrinogen with two terminal domains.

Fibrinogen does not assemble into a protofibril until four short peptides, two **fibrinopeptides A** and two **fibrinopeptides B**, are removed by the endopeptidase thrombin from the amino-terminal ends of the two α polypeptides and the two β polypeptides to produce **fibrin monomers**. The fibrinopeptides have sequences that vary extensively among species (Problem 7-5) and seem to satisfy only the requirement that they be polar and structureless. Carboxy-terminal to the arginines at which the endopeptidolytic cleavages occur that remove the four fibrinopeptides, the amino acid sequences of the α and β polypeptides become highly conserved among

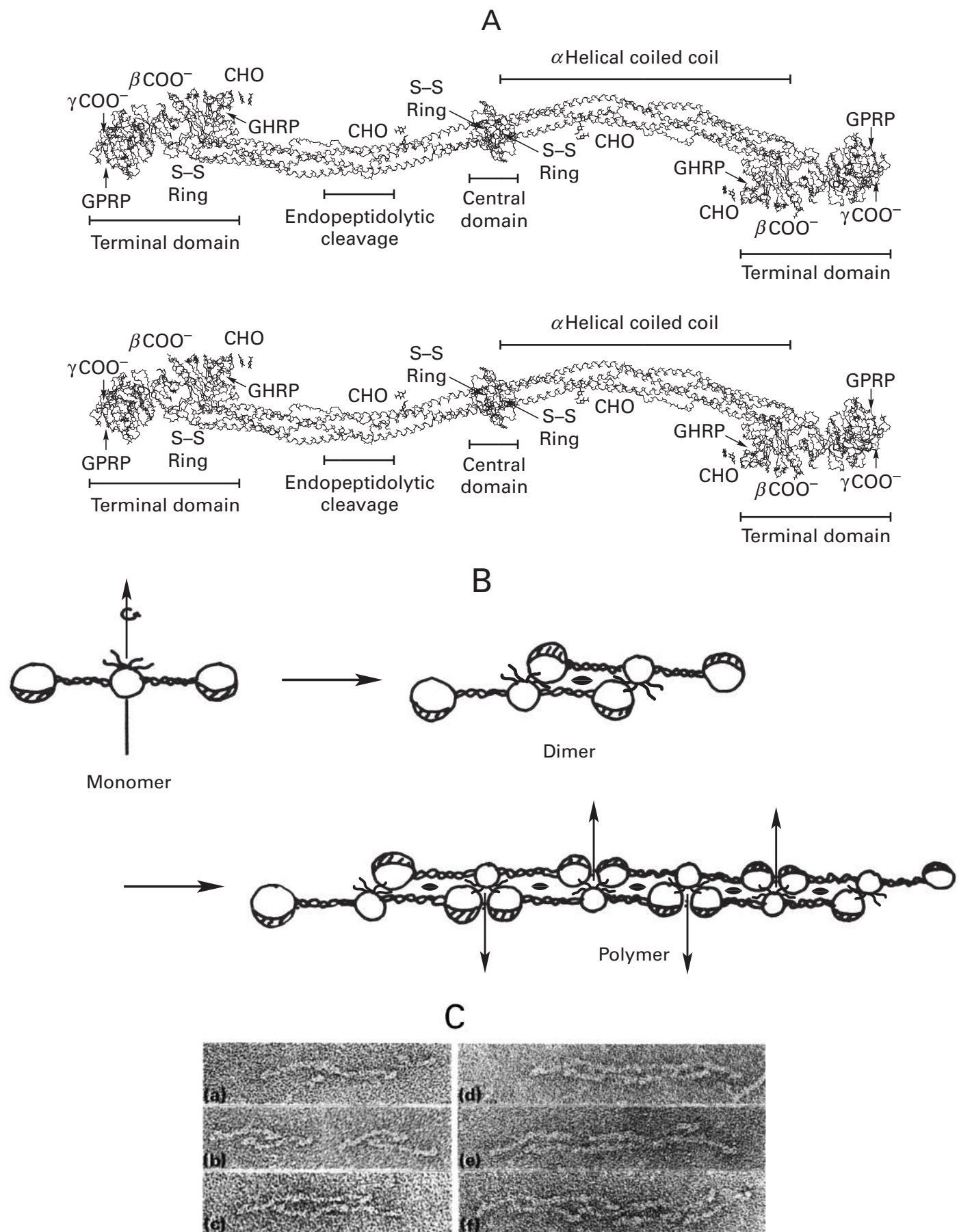


Figure 13-22: Assembly of fibrin from fibrinogen. (A) Skeletal drawing in stereo of the polypeptide backbones of the subunits in the crystallographic molecular model of fibrinogen from *G. gallus*.⁴⁸¹ The molecule is an $(\alpha\beta\gamma)_2$ heterohexamer with a rotational axis of symmetry normal to the plane of the page passing through the center of the central domain. Features noted on only one side of the molecule are reproduced symmetrically on the other. The amino termini of the six polypeptides α , α' , β , β' , γ , and γ' are all in the central domain, but the amino-terminal 26, 62, and 3 aa are missing from the maps of electron density of the α , β , and γ polypeptides, respectively, because these segments are disordered in the crystals. Two symmetrically displayed, identical rings of cystines (S-S rings) each connect Cysteine $\alpha 45$ to Cysteine $\gamma 23$, Cysteine $\gamma 19$ to Cysteine $\beta 80$, and Cysteine $\beta 76$ to Cysteine $\alpha 49$. One of these rings is on each side of the central domain, and together they mark its boundaries. At these rings, the two symmetrically displayed, three-stranded α -helical coiled coils commence in opposite directions, and each proceeds for 16 nm (111 aa) until terminating in another ring of cystines. In the middle of each coiled coil, there are intentional disruptions that make each susceptible to the endopeptidolytic cleavages that dissolve the fibrin clot. At each of the peripheral rings of cystines, each of the two terminal domains commences. The globularly and homologously folded carboxy-terminal 262 aa of a β subunit and 270 aa of a γ subunit together, side by side, form each terminal domain. The structures of these two halves of each of the peripheral domains are superposable.⁴⁸⁵ The carboxy-terminal 16 amino acids of the γ polypeptide, which contain the sites of intermolecular cross-linking, are disordered in the crystals. The carboxy-termini of each of these polypeptides in the crystallographic molecular model (βCOO^- and γCOO^- , respectively) are indicated. The carboxy-terminal 557 aa of each α polypeptide emerges from the respective peripheral ring of cystines, proceeds towards the center of the molecule along the α -helical coiled coil, and then disappears in disorder beyond Glutamate 218. There are oligosaccharides (CHO) at Asparagines $\gamma 52$ and Asparagines $\beta 363$. The binding sites for the tetrapeptides GPRP and GHRP that were cocrystallized with the fibrinogen are indicated. These are mimics of the amino termini from within an α polypeptide (GPRIL-) and a β polypeptide (AHRPL-) produced by thrombin cleavage. (B) Schematic drawing of the initial events in the polymerization of fibrin monomers to form a protofibril. Removal of the fibrinopeptides exposes new amino termini on the central domain (the four tails) that can then interact with complementary sites on terminal domains of other molecules. An additional set of contacts (end-to-end) comes into play upon addition of the third molecule. There are orthogonal 2-fold axes of symmetry (designated by arrow and sharpened ellipse) that alternate along the protofibril (polymer). Reprinted with permission from ref 479. Copyright 1984 Annual Reviews Inc. (C) Electron micrographs of intermediates in the polymerization of fibrin monomers. Thrombin was added to a solution of bovine fibrinogen (0.3 mg mL⁻¹) to initiate polymerization. At short times after nucleation of polymerization (about 10 min), the macroscopic clot was removed and samples of the clear solution that remained were placed on hydrophilic films of carbon supported by networks of formvar. The adsorbed complexes of fibrin monomers were negatively stained with 1.0% uranyl acetate.⁴⁸⁰ Magnification 290000 \times . The complexes of fibrin monomers presented in the gallery are (a) a dimer, (b) two dimers, (c) a tetramer, (d) a pentamer, (e) a hexamer, and (f) a heptamer. Reprinted with permission from ref 480. Copyright 1981 Academic Press.

species.⁴⁸³ The α polypeptide of a mammalian fibrin monomer has the amino-terminal sequence GPRAlk-, where Alk is the alkyl group of valine, leucine, or isoleucine; and the β polypeptide of a fibrin monomer

has the amino-terminal sequence GHRP-. A synthetic peptide of the sequence GPRP can inhibit completely the polymerization of fibrin monomers to fibrin polymer.⁴⁸⁴ It does so by competing with the amino termini of the α polypeptides of those fibrin monomers, which are in the central domain, for a binding site on the globular, carboxy-terminal domain of a γ subunit,^{485,486} which together with the homologous globular, carboxy-terminal domain of a β subunit⁴⁸⁵ forms one of the two terminal domains of the fibrin monomer (Figure 13-22A).

These facts suggest that the protofibril is assembled by the noncovalent binding of the central domain and terminal domain of one molecule of the fibrin monomer to the terminal domain and central domain, respectively, of another fibrin monomer to form a rotationally symmetric, doubly bonded dimer (Figure 13-22B).⁴⁷⁹ The dimer would be elongated to the polymer by adding other fibrin monomers, dimers, or oligomers through steps each creating identical interfaces (Figure 13-22B). Each of the consecutive, individual noncovalent interactions holding the helical polymer together is between the amino terminus of an α polypeptide on a central domain exposed by the cleavage produced by thrombin and the **binding site for its amino-terminal sequence, GPR-, on one of the terminal domains** of another fibrin monomer. Electron micrographs of intermediates in the polymerization of fibrin monomers are consistent with this structural proposal (Figure 13-22C).⁴⁸⁰

The binding site for the GPR- has been located in a crystallographic molecular model of the terminal domain.⁴⁸⁵ It is a **hole on the surface of the globular carboxy-terminal domain of the γ subunit** into which the amino-terminal sequence GPR- inserts with the glycine at the bottom of the hole. The amino-terminal 26 amino acids of the α polypeptide, amino-terminal to the symmetric interchain cystine between Cysteine $\alpha 28$ and Cysteine $\alpha' 28$, are missing from the crystallographic molecular models of both a detached central domain⁴⁸⁷ and a complete molecule of fibrinogen.⁴⁸¹ Consequently, the amino-terminal 11 aa of an α polypeptide in a fibrin monomer are probably structureless, and as a result, they are able to reach out to the hole on the peripheral domain. The cystine between Cysteine $\alpha 28$ and Cysteine $\alpha' 28$ itself forms a loop that juts out from the central domain. Both of these structural characteristics make it easier for the GPR- to find its site on the terminal domain of another fibrin monomer and suggest that the connection it forms between a central domain of the one fibrin monomer and the terminal domain of the other is a flexible tether.

A fibrin monomer is a knotted segment of rope with a 2-fold rotational axis of symmetry centered on the knot. It assembles into a protofibril, which is a helical cable. An infinite helix defined by a smooth curve is a geometric structure with 2-fold **rotational axes of symmetry** intersecting every one of its points, so the creation of an infinite helical polymer from a monomer with a molecular

2-fold axis of symmetry produces a structure with a 2-fold rotational axis of symmetry at each molecular 2-fold rotational axis of symmetry. When this polymer is finite, the two ends are necessarily identical in structure.

The protofibril that forms as the assembly proceeds has a thickness, as determined by light scattering,⁴⁸⁸ of about two fibrin monomers (Figure 13-22B,C). When fibrin monomers are created instantly by adding a large excess of thrombin to a solution of fibrinogen, the initial rate of formation of these protofibrils as monitored by light scattering is bimolecular in the concentration of fibrin monomers (Figure 13-23A)⁴⁸⁸ and shows no evidence of a lag. The time required for half of the final light scattering to be established is inversely proportional to the initial concentration of fibrin monomers (Figure 13-23B). Because light scattering is not proportional to bulk concentration of polymer, this is simply the time required for a particular fraction α of the polymer to form.

It can be shown¹⁹³ that such behavior is that expected from a polymerization in which end-to-end connections among monomers, dimers, and oligomers

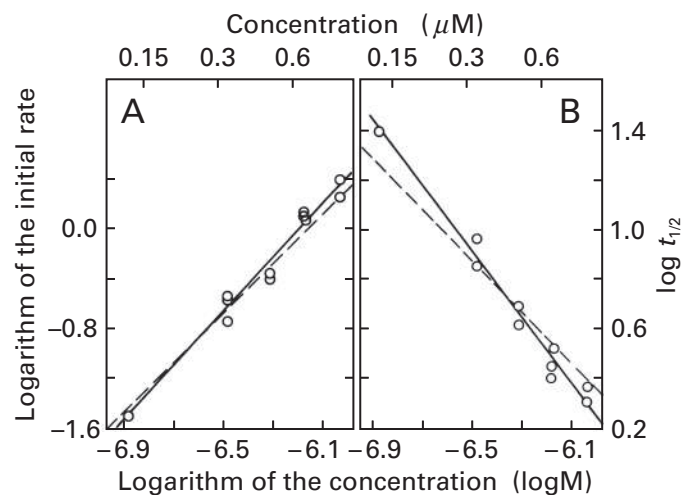


Figure 13-23: Dependence of the initial rate for the polymerization of fibrin (A) and the time required to reach half of the total increase in light scattering (B) on the initial concentration of fibrin monomer.⁴⁸⁸ Solutions of fibrinogen (0.05–0.35 mg mL⁻¹ final concentration) were mixed rapidly (<0.5 s) with solutions of thrombin at concentrations high enough to remove the fibrinopeptides from the molecules of fibrinogen immediately (<0.5 s). The polymerization of the fibrin was then followed by monitoring the increase in light scattering of the solution at 633 nm as a function of time (seconds). The initial rate of increase in light scattering was determined for each trace as well as the time (2–25 s) required for the light scattering to reach half its final value. The logarithm of the initial rate and the logarithm of the time required to reach half of the final value of light scattering, $\log t_{1/2}$, are presented, respectively, as functions of the logarithm of the concentration (molar) of fibrinogen. The concentration of protein (micromolar) is also noted at the top of each graph. The polymerization was performed at 23 °C in 0.5 M NaCl. The high ionic strength prevented side-to-side aggregation of the polymers so that only formation and elongation of protofibrils were occurring during the measurements. The solid lines are linear least-squares fits to the data; the dashed lines are lines of slope 2 and -1, respectively. Adapted with permission from ref 488. Copyright 1979 *Journal of Biological Chemistry*.

form at random with no initiation required and in which each connection has the same rate constant of formation regardless of the lengths of the two participants, including unconnected monomers. It is possible that, during the polymerization of fibrin monomers, an interface forms between the two adjacent terminal domains of two different fibrinogen molecules in a protofibril, which cannot form in the initial dimer (Figure 13-22B). Either this interface is much slower to form than the interface formed by the insertion of the amino terminus of an α subunit into the hole on a γ subunit or the difference in rate between the formation of a dimer and formation of a longer protofibril is irrelevant because almost all interfaces form between oligomers and between oligomers and monomers.

If the assumption is made that the degree of polymerization is equal to the molar concentration of connected interfaces between monomers, [interfaces], and it is realized that the molar concentration of amino termini of α polypeptides on central domains is always equal to the molar concentration of holes on terminal domains, then

$$\frac{d[\text{interfaces}]}{dt} = k[\text{faces}]^2 \quad (13-40)$$

where [faces] is the total molar concentration at any time, t , of **open faces**, each of which is an amino terminus and a hole on the same monomer and each of which is located at the **open end** of a fibrin monomer, dimer, or oligomer (Figure 13-22B), and each open end has both an amino terminus and a hole. The initial rate of polymer formation will be second-order in the concentration of fibrin monomer because $[\text{faces}]_0 = 2[\text{monomer}]_0$, where $[\text{faces}]_0$ and $[\text{monomer}]_0$ are the initial concentrations of faces and fibrin monomers, respectively.

Because one interface is formed from two faces, one from each monomer (Figure 13-22B)

$$2 \frac{d[\text{interfaces}]}{dt} = - \frac{d[\text{faces}]}{dt} \quad (13-41)$$

Upon combination of Equations 13-40 and 13-41 and rearrangement

$$- \frac{d[\text{faces}]}{[\text{faces}]^2} = 2k dt \quad (13-42)$$

Upon integration between $t = 0$ and t

$$\frac{1}{[\text{faces}]} - \frac{1}{[\text{faces}]_0} = 2kt \quad (13-43)$$

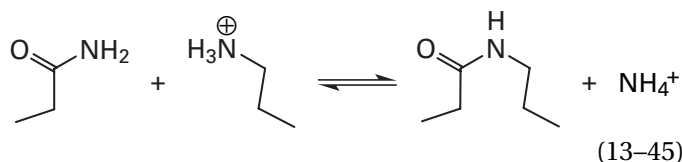
Choose any time, t_α , at which a particular fraction α of the polymer has formed. Because that fraction of the

faces has disappeared, $[\text{faces}] = (1-\alpha)[\text{faces}]_0$. When this equality is inserted into Equation 13-43

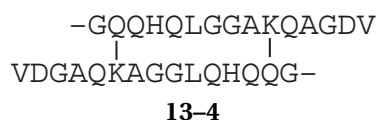
$$\frac{1}{[\text{faces}]_0} = 2 \left(\frac{1-\alpha}{\alpha} \right) k t_\alpha \quad (13-44)$$

It follows that the time at which any particular fraction, α , of the polymer has formed will be inversely proportional to the initial concentration of monomer, as was observed. The kinetic mechanism just described, based on the assumption that the combination of two faces is independent of whether they are faces on fibrin monomers, dimers, or oligomers, is completely consistent with the proposed molecular mechanism (Figure 13-22B).

The final step in the irreversible polymerization of fibrin is the **covalent cross-linking of the fibrin** monomers among themselves by the enzyme protein-glutamine γ -glutamyltransferase. This enzyme catalyzes replacement of the ammonia in a glutamine on one monomer with the ϵ amine of a lysine on another monomer:



Two symmetrically disposed pairs of lysines and glutamines near the carboxy termini of two γ subunits are cross-linked in this way to produce a covalent dimer of γ polypeptides with the following structure:⁴⁸⁹



where each valine is the carboxy terminus of a γ polypeptide. The covalently symmetric juxtaposition of these two sequences from the γ polypeptides of different fibrin monomers in the protofibril produced by the protein-glutamine γ -glutamyltransferase is consistent with a structural model in which there is a 2-fold rotational axis of symmetry at this location in the protofibril. Amide cross-links are also formed among α polypeptides juxtaposed during the lateral association of the protofibrils.⁴⁹⁰ The pairing in these cross-links reflects the side-by-side orientation of the α polypeptides in different protofibrils in the final fiber of fibrin.

After it has polymerized and been covalently cross-linked, fibrin cannot depolymerize. The fibrin clot is eliminated when necessary by the endopeptidolytic cleavage of the three polypeptides within the α helical coiled coils (Figure 13-22A) by plasmin. Collagen, also

because it is covalently cross-linked, cannot depolymerize. It too is eliminated, when necessary, by endopeptidolytic digestion. These two proteins are examples of helical polymeric proteins that are **assembled irreversibly**. Examples of helical polymeric proteins that **assemble reversibly** are actin (Figure 9-1B) and microtubules. These helical polymeric proteins are continuously assembled and disassembled during the life of a cell.

Microtubules are hollow cylinders of indefinite length. The overall radius of a microtubule is about 12 nm, and it has a hollow center with a radius of about 6 nm. When viewed in the electron microscope, by negative staining, microtubules are tubular bundles of 10-16 indistinguishable rows of protein (Figure 13-24A).^{491,492} Each row is parallel to all of the others and parallel to the axis of the microtubule for microtubules of 13 rows but skewed somewhat relative to the axis for microtubules with 10, 11, 12, 14, 15, or 16 rows.⁴⁹² At the end of a microtubule these rows can be frayed into individual threads.⁴⁹³ From this observation it can be concluded that the interfaces forming the rows are stronger than the interfaces between them.⁴⁹¹ Each thread of protein forming one of these rows is a **protofilament**.⁴⁹²

Each of the protofilaments in an intact microtubule is a string of globular protein subunits. Each of the subunits is pointed in the same direction. The top of one of these subunits is joined to the bottom of the subunit above it in the same protofilament by an interface. Each subunit is related to the one above it by a screw axis of symmetry coincident with the axis of the microtubule. Because each of the protofilaments in a microtubule of 13 rows is parallel to the axis of the tubule, the angle relating two consecutive subunits by the screw axis of symmetry is zero in this type of microtubule. In the other types of microtubules in which the protofilaments are skewed relative to the central axis, the angles of the screw axes are somewhat less or somewhat more than zero. In each case, however, the translation along the screw axis of symmetry that superposes a subunit in a protofilament upon the one above it is 4.0 nm.

The primordial microtubule was a polymer composed of identical monomeric subunits each related to its neighbor above by the screw axis of symmetry defined by the strong interfaces creating a protofilament. At some point, the gene encoding the subunit duplicated and the two resulting isoforms of the common ancestral subunit began to evolve separately. The results of this evolution are that along a protofilament, the two isoforms, the α subunit and the β subunit, alternate; that the interface between a β subunit and the α subunit above it is stronger than the interface between a β subunit and the α subunit below it; and that as a result, when a microtubule dissociates, it dissociates into $\alpha\beta$ heterodimers.^{494,495} Because the individual identical subunits in the primordial microtubule were related to each other by a screw axis of symmetry with a rise of

4.0 nm and an angle of rotation equal to about zero, the two subunits in the $\alpha\beta$ heterodimer are now related to each other by a screw axis of pseudosymmetry with the same rise and a rotation of almost zero.

The protomer from which a microtubule is now formed is this **$\alpha\beta$ heterodimer**. It is composed of an α subunit ($n_{aa} = 450$)⁴⁹⁶ and a β subunit ($n_{aa} = 445$)⁴⁹⁷ the amino acid sequences of which are homologous to each other [41% identity with 1.1 gaps (100 aa)⁻¹].⁴⁹⁷ Their native structures are also homologous and superposable and indistinguishable at low resolution. In the $\alpha\beta$ heterodimer, they sit one on top of the other, each

pointed in the same direction.^{498–500} This $\alpha\beta$ heterodimer is **tubulin**, the monomer from which a microtubule is formed by its polymerization. The $\alpha\beta$ heterodimer of tubulin in free solution will be referred to as **monomeric tubulin**; $\alpha\beta$ heterodimers of tubulin within a microtubule will be referred to as **protomeric tubulin**.

Upon the microtubule there is a **helical surface lattice**.⁴⁹¹ This lattice was originally defined by image reconstruction of electron micrographic images of intact microtubules formed from 13 rows (Figure 13–24A).^{491,492} The reciprocal lattice of the Fourier transform of digitized images from electron micrographs was assigned to

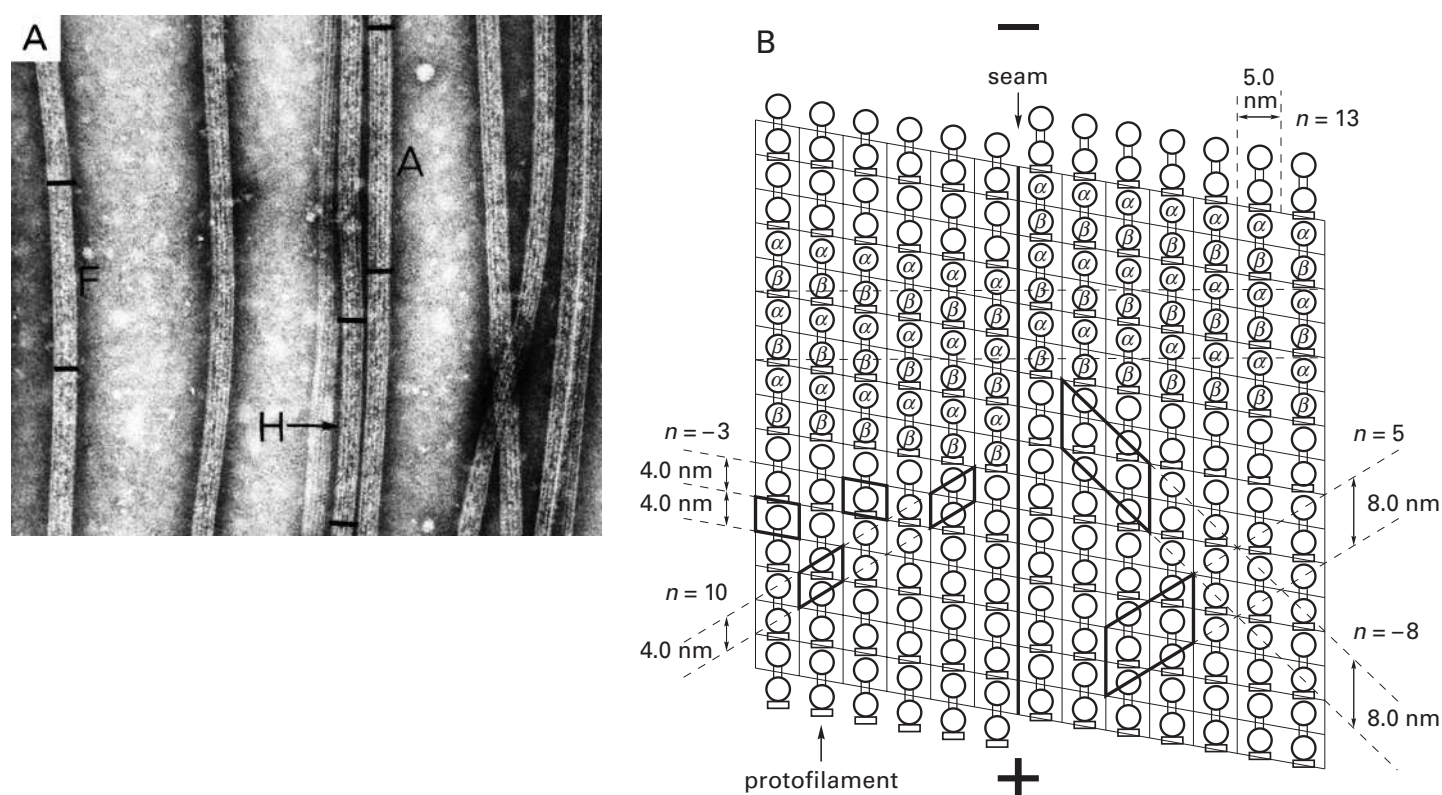


Figure 13–24: Helical surface lattice of $\alpha\beta$ heterodimers of tubulin producing a microtubule.⁴⁹¹ *Trichonympha agilis*, flagellated protists from the gut of the termite *Zootermopsis angusticollis*, were mixed with 1% phosphotungstate, pH 7, and applied to a film of amorphous carbon supported by a network of collodion. The excess negative stain was drained, the preparation was dried, and the negatively stained specimens were examined in the electron microscope. (A) Although the flagella of the *Trichonympha*, over most of their length, are composed of pairs of microtubules in which two microtubules are fused to each other, at the distal end of a flagellum these pairs dwindle to single microtubules. The electron micrograph is of a group of these individual, unpaired microtubules. Regions marked F, H, and A were chosen for image enhancement. The optical density of each of these images was digitized with an optical densitometer, and the Fourier transform of the digitized image was calculated. From examination of this calculated Fourier transform and from optical diffraction patterns of the images themselves, the reflections arising from the helical array of protomers were identified and indexed. These reflections defined the helical lattice in which the $\alpha\beta$ heterodimers of tubulin are arrayed in a microtubule. (B) A schematic diagram of that lattice is presented. The $\alpha\beta$ dimers of tubulin are aligned in 13 protofilaments parallel to the axis of the cylinder (see panel A). The cylinder on which the lattice is arrayed was cut along one of the lattice lines between two protofilaments on the side of the microtubule opposite to the seam and parallel to its axis, and the cylindrical surface was flattened onto the page. Individual subunits, if no distinction is made between α and β , lie on a triply threaded, left-handed screw ($n = -3$) and a decuply threaded, right-handed screw ($n = 10$). The two different unit cells arrayed along the resulting helices, respectively, are indicated by parallelograms. $\alpha\beta$ Heterodimers lie on a pentuply threaded right-handed screw ($n = 5$) and an octuply threaded left-handed screw ($n = -8$). Each unit cell along each of the resulting helices is identical and contains the equivalent of two $\alpha\beta$ heterodimers. The dimensions of each of these helices and of the parallel protofilaments are noted in nanometers. The horizontal dashed lines indicate the points of fusion for the flattened array when it is rolled into the cylinder. The heterodimers are in register in the adjacent protofilaments except at the seam. The locations of the binding sites for GTP on the β subunits are indicated by rectangles. The sites for the nonexchangeable GTP on the α subunits sit between the two subunits in the heterodimer. The end of the microtubule displaying only β subunits is the plus end; that displaying only α subunits is the minus end. Reprinted with permission from ref 491. Copyright 1974 Biochemical Society.

that of the Fourier transform of a triply threaded, left-handed ($n = -3$) screw the three helices of which are spaced at 4.0-nm intervals along the axis (Figure 13–24B). Because the array is built upon the 13 parallel protofilaments, the three left-handed helices ($n = -3$) of subunits with spacing of 4.0 nm create 10 right-handed helices ($n = 10$) with spacing of 4.0 nm, and together these sets of helices form a $(-3,10)$ lattice. These sets of three and ten parallel, contiguous helices are helices of unit cells each of which contains the equivalent of one individual subunit, where α subunits are indistinguishable from β subunits, but because α subunits and β subunits actually are different from each other, there are two different but indistinguishable unit cells arrayed translationally along each of these sets of helices. A set of $(8.0 \text{ nm})^{-1}$ reflections also appears in the Fourier transform of the images, and these reflections result from an octuply threaded, left-handed ($n = -8$) screw and the resulting pentuply threaded screw ($n = 5$) with spacing of 8.0 nm between the threads, a $(-8,5)$ lattice. In these sets of eight and five, the helices are helices of identical, translationally arrayed unit cells, each containing the equivalent of two $\alpha\beta$ heterodimers.

Because an α subunit is different from a β subunit and because they sit one on top of the other in the $\alpha\beta$ heterodimer, the two ends of a microtubule are different, and a microtubule is **polar**. This has been verified by the observation that growth at one end of a microtubule during assembly is slower than growth at the other end.⁵⁰¹ In the arrangement depicted in Figure 13–24B, one end would display only α subunits; and the other end, only β subunits.

The most peculiar structural feature of an extant microtubule, which was not a feature of the primordial microtubule, is its **seam** (Figure 13–24B). In all microtubules except those with 10 and 16 protofilaments, the protofilaments are in register, α subunit next to α subunit and β subunit next to β subunit, except at one junction, at which they are out of register. This junction is a seam of discontinuity that runs parallel or almost parallel to the axis of the microtubule along its surface.^{502–504} At the seam, there is a discontinuity in the $(-8,5)$ lattice of $\alpha\beta$ heterodimers but not in the $(-3,10)$ lattice of indistinguishable subunits.

When a solution of purified monomeric tubulin is brought to the proper conditions of temperature and pH and is mixed with the proper substrates, the tubulin spontaneously polymerizes to form microtubules. The process can be divided into two phases, nucleation and elongation.⁵⁰⁵ **Nucleation** is the sequence of events that leads to the formation of a large enough oligomer of tubulin to act as an origin from which a microtubule can then elongate by the consecutive, repetitive addition of more tubulin. **Elongation** is the addition of tubulin to one end or the other of a microtubule in such a way that each successive addition at that end is formally equivalent regardless of the length of the microtubule. Until

such a stage is reached, the steps in the assembly of a microtubule are steps in the process of nucleation.

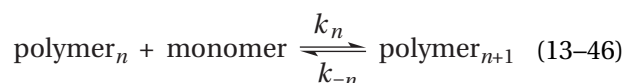
Under all experimental conditions, **spontaneous nucleation** of microtubules in an originally monodisperse solution of tubulin is a complicated process that involves a number of intermediates of peculiar structure.^{506,507} It also depends on the twelfth power of the concentration of tubulin.⁵⁰⁸ This is not surprising because the steps between monodisperse tubulin in free solution and an oligomer of tubulin large enough to offer the end of a cylinder of 10–16 rows equivalent to the end of the cylinder in an established microtubule are not easy to accomplish. Furthermore, even though it occurs with even the most highly purified preparations of tubulin,⁵⁰⁸ spontaneous nucleation seems to involve a minor component contaminating the preparation of purified tubulin.⁵⁰⁹ At concentrations of tubulin high enough to support spontaneous nucleation, it proceeds slowly and continuously, independent of any decreases in the concentration of free tubulin caused by its incorporation into elongating microtubules. All of these properties make it difficult to separate cleanly the kinetics of spontaneous nucleation from those of elongation.

These complexities of spontaneous nucleation in the laboratory, however, may be irrelevant to the polymerization of tubulin within a cell, the process for which the protein has evolved. In a living cell almost all of the microtubules originate in only one region of the cytoplasm.⁵¹⁰ In cells containing centrosomes, it is the pericentriolar material that serves as the origin. The pericentriolar material is a diffuse structure surrounding the centriole that lies in the center of the **centrosome**. In cells lacking centrosomes, the point of origin is associated with structures resembling centrosomes. Within the pericentriolar material, it is rings of a different isoform of tubulin, γ tubulin, that serve as the nuclei upon which $\alpha\beta$ tubulin polymerizes.⁵¹¹ The rings of γ tubulin embedded in purified centrosomes are able to initiate the formation of microtubules readily⁵¹² and do so at lower concentrations of free tubulin than are necessary for spontaneous nucleation.^{509,510} From these observations, it becomes clear that spontaneous nucleation in the absence of centrosomes is an adventitious process for which $\alpha\beta$ tubulin was probably not designed.

Elongation of microtubules in the absence of the complications of spontaneous nucleation can also be accomplished by adding **seeds**, which are short, uniform fragments of preformed microtubules, to solutions containing high enough concentrations of unpolymerized monomeric tubulin to support the elongation of those seeds.⁵⁰⁵ Seeds have already passed through the steps of spontaneous nucleation. The fragments of preformed microtubules used as seeds can be stabilized by cross-linking with ethylene glycol bis(succinimidylsuccinate)⁵¹³ or by preparing the seeds with guanylyl 5'-(β , γ -methylenediphosphonate). Solutions of tubulin of high purity are reasonably unsusceptible to sponta-

neous initiation⁵¹⁰ but readily support the elongation of seeds in a reaction that assumes its maximum rate immediately after the seeds are added to the solution.

Consider a polymerization in which monomers, such as monomeric tubulin, are successively adding to only one end of a polymer, such as a microtubule elongating at only one of its ends from a centrosome, by the reaction



where polymer_n is a polymer composed of n protomeric units and polymer_{n+1} is a polymer composed of $(n + 1)$ protomeric units. When the reaction has come to equilibrium, the dissociation constant, $K_{dn,\text{poly}}$, of a monomer from the end of a polymer $n + 1$ units in length is

$$K_{dn,\text{poly}} = \frac{[\text{polymer}_n]_{\text{eq}} [\text{monomer}]_{\text{eq}}}{[\text{polymer}_{n+1}]_{\text{eq}}} = \frac{k_{-n}}{k_n} \quad (13-47)$$

When all values of n are large enough so that only elongation is being considered, $K_{dn,\text{poly}}$, k_n , and k_{-n} , respectively, have the same mean values for all values of n . The ends of all the polymers are indistinguishable because each is elongating from the same one of its two ends. In the case of microtubules elongating from a centrosome, the other end is not elongating because it is anchored in the centrosome.

Two measures of the solution of a polymer can be separately defined. The **bulk concentration** of polymer, $[\text{polymer}]_b$, or bulk concentration of microtubules, $[\text{microtubule}]_b$, is equal to the molar concentration of protomers, or of protomeric tubulin, that are incorporated into polymers. The bulk concentration is directly proportional to the total length of polymer in the solution. In the case of microtubules the bulk concentration is usually determined by light scattering, which is linearly related to the total length of microtubule present. Normally, a significant fraction of the light is scattered by these solutions, and the absorbance, which is decreased by the amount of light scattered, is measured rather than the amplitude of the scattered light. The **number concentration** of polymer, $[\text{polymer}]_n$, or number concentration of microtubules, $[\text{microtubule}]_n$, is equal to the molar concentration of individual, intact molecules of polymer, or microtubules, regardless of their individual lengths. The number concentration of microtubules is measured by quantitative electron microscopy.⁵¹⁴ Individual microtubules in a field on an electron micrograph are counted, and their density is related to their density in the original solution. The number concentration of a polymer elongating at only one end is equal to the molar concentration of that end, $[\text{end}]$.

Assume still that elongation is occurring at only one end of a polymer and that

$$\frac{d[\text{polymer}]_b}{dt} = k_n [\text{end}] [\text{monomer}] - k_{-n} [\text{end}] \quad (13-48)$$

If the reaction has been initiated by adding origins of nucleation, such as centrosomes, to a solution of monomer, and if no spontaneous nucleation occurs, the molar concentration of ends at which elongation is proceeding must remain constant, and the initial rate for the formation of bulk polymer is defined by the relationship

$$\left(\frac{d[\text{polymer}]_b}{dt} \right)_0 = [\text{end}]_0 (k_n [\text{monomer}]_0 - k_{-n}) \quad (13-49)$$

where the subscripts indicate initial quantities. It has been shown, in agreement with this equation, that the initial rate at which the bulk concentration of microtubules increases is **directly proportional to the molar concentration of seeds** added to a series of reactions (Figure 13-25).⁵⁰⁵

The initial rate of formation of bulk polymer at only one end should also be directly proportional to the term $(k_n [\text{monomer}]_0 - k_{-n})$, and a plot of initial rate against initial concentration of monomer should be linear and pass through zero when $k_n [\text{monomer}]_0 = k_{-n}$. Immediately after addition of seeds, if the seeds were simply fragments of polymer equivalent to the polymer about to be formed by elongation, the bulk concentration of polymer $[\text{polymer}]_b$ should increase when $k_n [\text{monomer}]_0 > k_{-n}$ and decrease when $k_n [\text{monomer}]_0 < k_{-n}$. When the bulk concentration of polymer is decreasing rather than increasing, the individual molecules of the polymer forming the seeds would be depolymerizing by shedding monomer from their ends and decreasing in length. The **critical concentration** is the initial concentration of monomer at which neither net elongation nor net depolymerization occurs. If only one molecule of polymer were being observed, its initial rate of elongation (in monomers second⁻¹) at only one of its ends should be equal to $(k_n [\text{monomer}]_0 - k_{-n})$. When $k_n [\text{monomer}]_0 > k_{-n}$, that molecule of polymer should elongate; and when $k_n [\text{monomer}]_0 < k_{-n}$, it should depolymerize.

If monomer adds to a population of elongating polymers until the concentration of monomer free in solution and the concentration of monomer within polymers reach equilibrium with each other, the total bulk concentration of polymer will be linearly related to the total concentration of monomer, and the line relating these two variables will intersect the abscissa at the value of the critical concentration (Figure 13-26).⁵¹⁵ This behavior necessarily follows from the facts that the amount of monomer incorporated into polymer when equilibrium is reached is equal to $[\text{monomer}]_{\text{tot}} - [\text{monomer}]_{\text{eq}}$; that $[\text{monomer}]_{\text{eq}}$, the critical concentra-

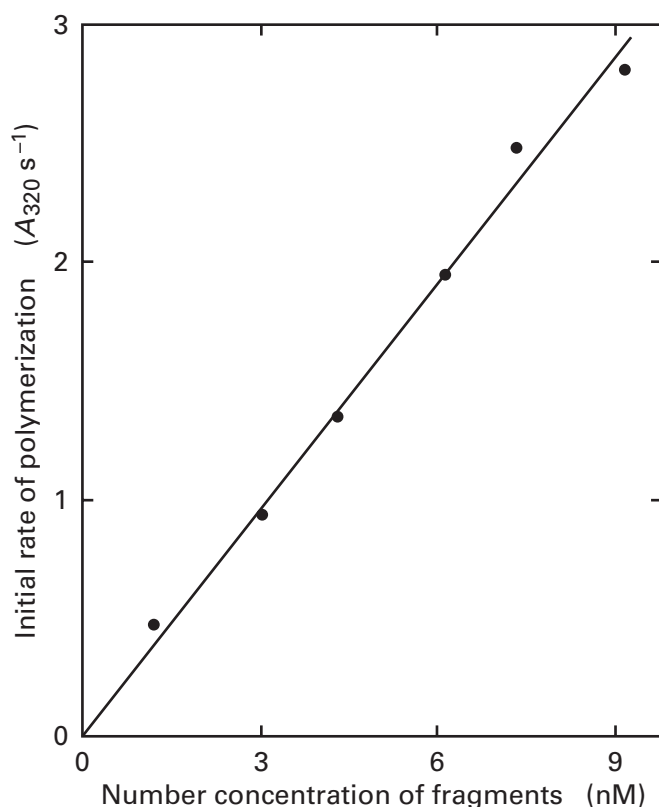


Figure 13-25: Initial rate of tubulin polymerization as a function of the initial concentration of fragmented whole microtubules.⁵⁰⁵ A preparation of microtubules ($70 \mu\text{M}$) that had been polymerized separately by spontaneous nucleation was sheared by passing it through a 22-gauge needle to produce fragments about $1 \mu\text{m}$ in length referred to as seeds. The seeds were immediately added to a solution of unpolymerized monomeric tubulin at a concentration high enough ($18 \mu\text{M}$) to elongate the seeds. The solution was 0.1 mM MgCl_2 and 0.5 mM GTP , pH 6.9. The elongation of the fragments was followed by an increase in absorbance at 320 nm due to light scattering. The initial rate of increase in absorbance ($A_{320} \text{ second}^{-1}$) is plotted against the initial number concentration (nanomolar) of sheared microtubules added to initiate elongation. Note that the initial molar concentration of monomeric tubulin is always more than 2000 times that of the seeds. Adapted with permission from ref 505. Copyright 1977 Academic Press.

tion, is the same for each point; and that when $[\text{monomer}]_{\text{eq}}$ equals the critical concentration, no increase in $[\text{polymer}]_{\text{b}}$ can occur.

Seeds were added to a series of solutions containing increasing concentrations of unpolymerized, monodisperse monomeric tubulin at concentrations below those at which significant spontaneous nucleation occurs.⁵⁰⁹ At various times after the addition, samples were taken from each mixture. They were quenched by quantitative cross-linking with glutaraldehyde, sedimented onto a specimen grid, and examined in the electron microscope.⁵¹⁴ The seeds that were used were fragments of **axonemes**, which are naturally occurring, rigid bundles of microtubules the polarity of which can be determined visually. The two ends of a fragment of an axoneme are arbitrarily termed the **plus end** and the **minus end**. The

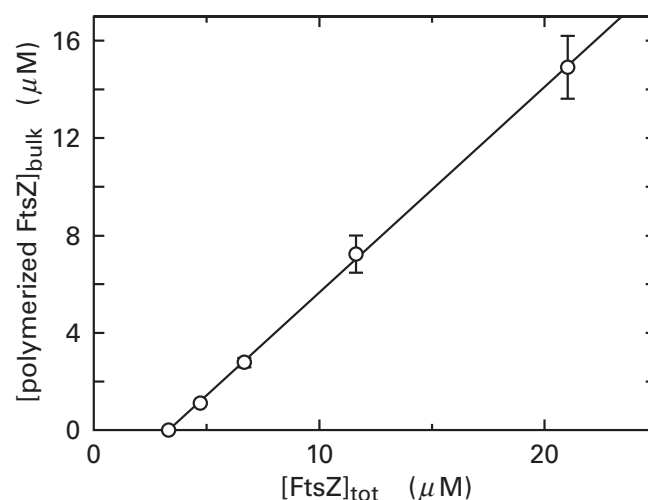


Figure 13-26: Demonstration of the critical concentration for the polymerization of the monomer of cell division protein FtsZ from *E. coli*,⁵¹⁵ a bacterial homologue of tubulin.^{589–591} Solutions containing the noted total concentrations of cell division protein FtsZ, which had been stripped of all nucleotides, were prepared in 50 mM KCl and 10 mM MgCl_2 , pH 6.5 at 55°C . After equilibrium was reached, polymer was collected by centrifugation, and its bulk concentration was determined by direct analysis of the amount of protein in the pellet. The bulk concentration (micromolar) of sedimentable polymer ($[\text{polymerized FtsZ}]_{\text{bulk}}$) is presented as a function of the total concentration (micromolar) of monomer in the solution ($[\text{FtsZ}]_{\text{tot}}$). The observed critical concentration is $3.4 \mu\text{M}$.

plus end and minus end were originally defined in terms of the appearance of an axoneme in an electron micrograph. It is now known that the end of a microtubule at the plus end of an axoneme displays only β subunits^{516,517} and the end at the minus end displays only α subunits (Figure 13-24B).⁵¹⁸ Elongation was found to proceed from both ends of the axonemes. The initial rate of increase in the length of the microtubules projecting from each end of the axoneme could be measured and plotted against the initial concentration of tubulin (Figure 13-27).⁵¹⁴ The initial rate of elongation is linearly related to the initial concentration of monomeric tubulin, $[\text{monomer}]_0$, as predicted by Equation 13-49, but k_{-n} is too small to be estimated accurately.*

When elongation occurs because seeds are added to a solution of monomer above the critical concentration, the free concentration of monomer should decrease as polymer is formed until equilibrium is reached and further elongation ceases. If elongation were occurring from only one end of the polymer, from Equation 13-48 it follows that

* At rates of elongation around $10 \mu\text{m min}^{-1}$, a rate beyond those observed in Figure 13-26, there is a change in the rate-limiting step of elongation, and the rate becomes independent of the concentration of monomeric tubulin,⁵⁰⁹ presumably because it is limited by the rate of some step that must occur between the addition of the last monomer of tubulin and the addition of the next monomer of tubulin to the growing end.

$$\frac{d[\text{polymer}]_b}{dt} = - \frac{d[\text{monomer}]}{dt} = [\text{end}] (k_n [\text{monomer}] - k_{-n}) \quad (13-50)$$

When this relationship is rearranged and integrated between $t = 0$ and t

$$\ln(k_n [\text{monomer}] - k_{-n}) = \ln(k_n [\text{monomer}]_0 - k_{-n}) - k_n [\text{end}] t \quad (13-51)$$

when $t = \infty$ and $[\text{monomer}] = [\text{monomer}]_{\text{eq}}$, $k_n [\text{monomer}]_{\text{eq}} = k_{-n}$. Therefore, the concentration of monomer at equilibrium should be equal to the critical concentration at which no elongation occurs when seeds are added to a solution of monomer.

But a microtubule has two ends, and both are elongating in the experiment. Consider any linear polymer such as a microtubule in which an arrangement of protomers, such as the 13 protomers of tubulin across the microtubule that are labeled α and β in Figure 13-24B, repeats precisely along the polymer to create its structure. Remove a complete set of these protomers, for example the 13 protomers of tubulin at the minus end, from one end of the polymer, and add them to the other end in such a way that the newly added protomers duplicate the arrangement that was at that end before. For example, add one of these monomers of tubulin to each of the 13 protofilaments at the plus end. The structure of the altered polymer is identical to that of the initial polymer because of the linear repeat with which it is created. Because free energy is a state function, the standard free energy change for this reaction must be zero. Because this is true for whatever structure is present at the two ends initially, the mean value for the dissociation constants, $K_{d,\text{tbn}}$, for tubulin at either end of a microtubule must be the same. Because $K_{d,\text{tbn}} = k_{-n}/k_n = [\text{tubulin}]_{\text{eq}}$ (Equation 13-47), the concentration of monomeric tubulin in equilibrium with either end, the critical concentration, is the same. If only mass action were governing the reaction, a microtubule on average could not elongate preferentially at one end while it is depolymerizing at the other. The two ends, however, can, and do, have different rate constants of elongation (Figure 13-27) and depolymerization.

One of the ingredients that is essential for the elongation of microtubules is **GTP**.⁵¹⁹ Either GTP or GDP is bound by monomeric tubulin at one site on each of the two homologous subunits in the $\alpha\beta$ heterodimer.^{498,520} In monomeric tubulin, the site for binding GTP on the surface of the α subunit is enclosed within the interface of the $\alpha\beta$ heterodimer. Consequently, because of the screw axis

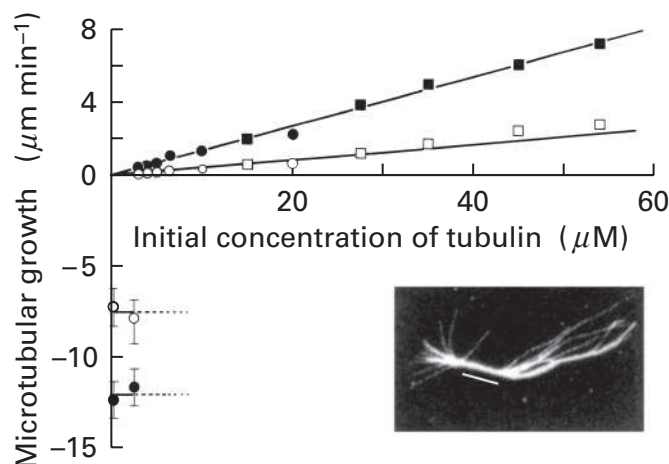


Figure 13-27: Rate of elongation (upper quadrant; \square , \blacksquare , \circ , \bullet ; micrometers minute⁻¹) and rate of depolymerization (lower quadrant; \bullet , \circ ; -micrometers minute⁻¹) of microtubules from the plus end (\bullet , \blacksquare) and the minus end (\circ , \square) of fragments of axonemes.⁵¹⁴ Fragments of axonemes from *Tetrahymena pyriformis* were added to final number concentrations of 10^7 mL^{-1} to solutions of monomeric tubulin at the noted concentrations (micromolar). Each of these solutions was 1 mM MgCl_2 , 1 mM EDTA, and 1 mM GTP, pH 6.8. At a series of time points, samples were withdrawn from these solutions and rapidly fixed with glutaraldehyde, and the products were sedimented onto electron microscopic grids or glass coverslips. The elongated axonemes were examined, following negative staining, in the electron microscope or, following staining with fluorescent anti-tubulin immunoglobulin G, in a fluorescence microscope (inset). The boundary between the original axoneme (white bar in inset) and the newly elongated microtubules was clearly defined because the newly elongated tubules splayed from the rigid cylinder of the bundle of microtubules making up the axoneme. Rates of elongation were calculated from direct measurements of the length (micrometers) of the newly elongated microtubules elongating from the ends of the axonemes as a function of time (minutes). The circles (\circ , \bullet) in the upper quadrant of the graph are measurements made by electron microscopy, and the squares (\square , \blacksquare) are measurements made by immunofluorescence. At concentrations of monomeric tubulin greater than $3 \mu\text{M}$, microtubules would elongate from axonemes, and the rate of elongation was linearly related to the concentration of initial unpolymerized monomeric tubulin. At initial concentrations of monomeric tubulin below $3 \mu\text{M}$, no elongation would occur. If microtubules, however, were grown on axonemes to $20 \mu\text{m}$ on the plus end and $7.5 \mu\text{m}$ on the minus end at a high monomeric tubulin concentration and the concentration of monomeric tubulin was then dropped to one of the noted concentrations less than $3 \mu\text{M}$, the microtubules would begin to depolymerize at the noted rates (circles with minus values of rates in the lower quadrant of the graph), which were independent of the concentration of unpolymerized tubulin. Reprinted with permission from *Nature*, ref 514. Copyright 1984 Macmillan Magazines Limited.

of pseudosymmetry, the site for binding GTP on the surface of the β subunit is on the opposite side of the β subunit from that interface, but at a position (indicated by the rectangles in Figure 13-24B) that becomes enclosed within the interface that is formed in a protofilament when a β subunit enters the elongating tubule at the minus end. At the positive end of a microtubule, each β subunit has a site for binding GTP that is displayed at its open end and

that is enclosed in the interface formed when monomeric tubulin associates with it during elongation at that end. A molecule of GTP bound to the α subunit of monomeric tubulin, which is within the interface that is stable and does not dissociate, remains unhydrolyzed, does not exchange with GTP in solution, and is an inert feature of the protein.^{520–522} It is the descendant of the GTP molecule that was at a dissociable interface in the primordial microtubule but now serves only a structural role uninvolved in the dynamics of elongation and depolymerization of an extant microtubule. Consequently, only the GTP or GDP and inorganic phosphate bound to the site on the β subunit is relevant to the roles of tubulin in the dynamic behavior of a microtubule.

At the open site on the β subunit in monomeric tubulin, the bound GTP exchanges readily with GTP or GDP in the solution, and molecules of bound GTP are **slowly hydrolyzed** to GDP and inorganic phosphate at that site but only after the incorporation of this site into a microtubule coincident with the addition of the monomeric GTP-tubulin to an elongating microtubule.⁵²² At 37 °C, the rate of this hydrolysis within a microtubule is 0.06 s^{-1} , and the rate for release of the inorganic phosphate into the solution is 0.02 s^{-1} .⁵²³ The molecules of GDP that are formed by this hydrolysis, however, remain trapped at the site while the protomeric GDP-tubulin is within the microtubule.^{522,524} This peculiar feature of elongation causes an elongating microtubule to have a region at its end in which all the tubulin has GTP bound to it because it has recently been added. Beyond this **cap of protomeric GTP-tubulin**, however, the frequency of protomeric GDP-tubulin increases until the main body of the microtubule is reached, which is entirely formed from protomeric GDP-tubulin. In the cytoplasm of a cell, in which the concentration of GTP is much greater than that of GDP, when a protomer of GDP-tubulin leaves a microtubule, the GDP rapidly dissociates from it and is replaced by GTP from the solution.

Although there is one measurement⁵²⁵ giving a critical concentration for GDP-tubulin of $3\text{ }\mu\text{M}$, which is contradicted by measurements from a similar preparation of tubulin⁵²⁶ giving a critical concentration of greater than $30\text{ }\mu\text{M}$, it is generally believed that “the critical concentration for assembly of Tu-GDP is apparently very high, effectively infinite for practical purposes.”⁵²⁷ The critical concentration for (GTP)-tubulin, however, is immeasurably small. This conclusion follows from the fact that the intercepts with the horizontal axis in Figure 13–27 are indistinguishable from zero and the fact that the measured critical concentration for tubulin to which guanylyl 5'-(β , γ -methylenediphosphonate) has been bound, which is an analogue of GTP that is very slowly hydrolyzed, is less than $0.2\text{ }\mu\text{M}$.⁵²⁸ Unfortunately, although the difference in critical concentrations between GDP-tubulin and GTP-tubulin is probably large, no direct measurement of that difference is available.

The **rate constants for elongation** of a microtubule

can be distinguished separately by the following convention. Those for elongation at the positive end can be identified with a plus sign (+); those for elongation at the negative end, with a minus sign (–); those for GTP-tubulin, with the letter T; and those for GDP-tubulin, with the letter D. From the results in Figure 13–27, $k_{n,T+} = 3.8 \times 10^6\text{ M}^{-1}\text{ s}^{-1}$ and $k_{n,T-} = 1.2 \times 10^6\text{ M}^{-1}\text{ s}^{-1}$ at 37 °C. Both $k_{n,T+}$ and $k_{n,T-}$ are too small ($\leq 1\text{ s}^{-1}$) to be measured accurately.⁵¹⁴ Under most circumstances in which elongation occurs in the presence of GTP ($[\text{monomer}] \geq 4\text{ }\mu\text{M}$), it is governed by $k_{n,T+}$ and $k_{n,T-}$. Even though the dissociation constants for GTP-tubulin to both ends of the microtubule must be the same, the rate constants for association and dissociation are not the same because the microtubule is a polar structure. It happens that a microtubule elongates about 3-fold more rapidly from its plus end than from its minus end.

When microtubules depolymerize, for example, upon dilution, the cap of protomeric GTP-tubulin near the end that was formed from recently added monomeric GTP-tubulin is rapidly lost, and depolymerization then proceeds by the dissociation of monomers of GDP-tubulin. If the dilution has been great enough, ends cannot be recapped and only $k_{n,D+}$ and $k_{n,D-}$ govern the rates of depolymerization. The values for these observed **rate constants for depolymerization** upon dilution (Figure 13–27),⁵¹⁴ under the same conditions in which $k_{n,T+}$ and $k_{n,T-}$ were measured, are $k_{n,D+} = 340\text{ s}^{-1}$ and $k_{n,D-} = 210\text{ s}^{-1}$ at 37 °C.

This arrangement of reactions—elongation by GTP-tubulin, hydrolysis of the GTP within the microtubule, and the consequent dissociation of mainly GDP-tubulin upon depolymerization—creates a peculiar **steady state** when the concentration of bulk polymer has reached its maximum level. At all times and at random, regions of protomeric GDP-tubulin are overtaking the now slowly elongating ends of individual microtubules and switching those microtubules from ones that are elongating to ones that are catastrophically depolymerizing (Figure 13–28).⁵¹⁴ The monomeric tubulin released during depolymerization picks up new molecules of GTP and reenters a microtubule that happens by chance still to be outdistancing its protomeric GDP-tubulin.

If the concentration of monomeric GTP-tubulin is less than a certain **apparent critical concentration**, the boundary of protomeric GDP-tubulin, whose rate of propagation is independent of the concentration of monomeric GTP-tubulin, will move along the microtubule faster than the rate of its elongation, which is dependent on the concentration of monomeric GTP-tubulin, and the microtubules will be switched (Figure 13–28) and begin to depolymerize with the rate constants $k_{n,D+}$ and $k_{n,D-}$. At concentrations of tubulin in excess of this apparent critical concentration, elongation will be faster than the rate at which the boundary of protomeric GDP-tubulin can move, and the microtubules will elongate with rate constants $k_{n,T+}$ and $k_{n,T-}$ as if they

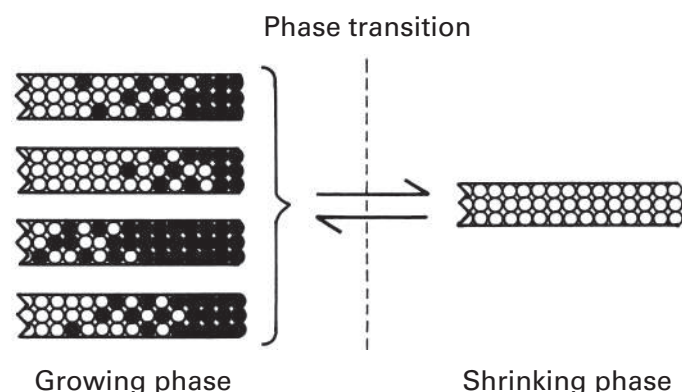


Figure 13-28: Schematic model describing the role of GTP in the elongation of microtubules.⁵¹⁴ The dark circles represent protomers of tubulin to which GTP is bound, and the open circles represent protomers of tubulin on which the GTP has hydrolyzed to GDP. As long as the tubule is elongating at a significant rate, the end of the microtubule is occupied by protomeric GTP-tubulin, and the end has a low critical concentration (Figure 13-27). In this state, it will not depolymerize catastrophically, and this is the growing phase. If elongation slows down because the concentration of monomeric GTP-tubulin decreases, at a certain point the spreading boundary of GTP hydrolysis will reach the end of the microtubule. The end of the microtubule will then be occupied mostly by protomeric GDP-tubulin and the end will then have a much higher critical concentration. It has passed through a phase transition into the shrinking phase and will rapidly depolymerize. Reprinted with permission from *Nature*, ref 514. Copyright 1984 Macmillan Magazines Limited.

contained only GTP-tubulin. At the apparent critical concentration, the rate of depolymerization from uncapped ends will be equal in magnitude to the rate of elongation from capped ends. This apparent critical concentration is not the reflection of an equilibrium process, as is the actual critical concentration (Figure 13-26), but the result of a complex combination of rate constants producing a steady state, and it should not be confused with a real critical concentration. The steady state is maintained by the continuous hydrolysis of GTP, and as long as GTP is available, equilibrium cannot be reached. When all the GTP has been hydrolyzed, all of the microtubules disappear because the critical concentration for GDP-tubulin is so large.

At concentrations of 10 μM free monomeric tubulin, which is the apparent critical concentration of GTP-tubulin at steady state in the presence of excess concentrations of GTP⁵²⁵ rather than at equilibrium,⁵¹⁰ the rates of depolymerization from uncapped ends containing protomeric GDP-tubulin are at least 10-fold greater than the rates of elongation from ends capped with protomeric GTP-tubulin. As a result, a microtubule that is depolymerizing from an uncapped end will rapidly and catastrophically disappear even if it is elongating at its other end. **Catastrophic depolymerization** of microtubules following dilution proceeds as a zero-order reaction (Figures 13-27 and 13-29)^{505,514} because the rate of depolymerization is defined by the equation

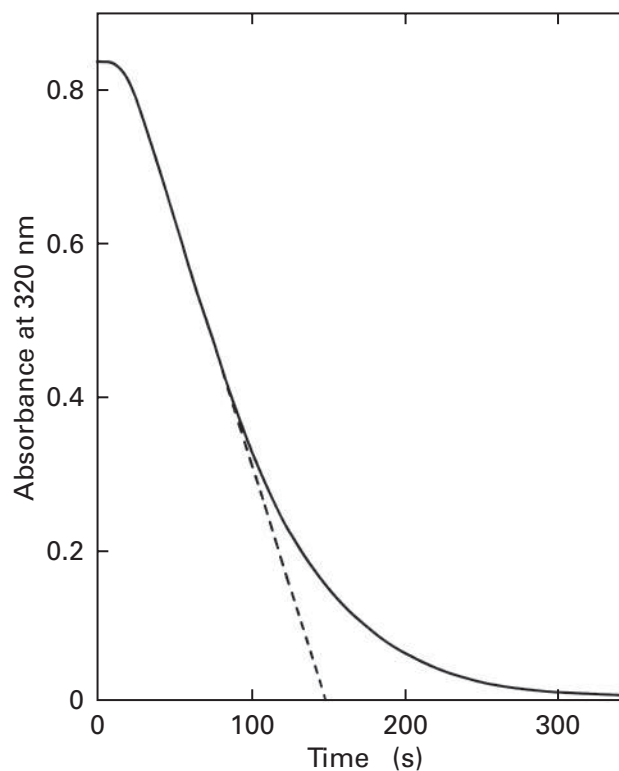


Figure 13-29: Kinetics of the depolymerization of microtubules.⁵⁰⁵ Tubulin (35 μM) was polymerized at 30 $^{\circ}\text{C}$ in 0.1 mM MgCl_2 and 0.5 mM GTP at pH 6.9 to steady state (30 min). When the temperature of a solution of microtubules is dropped, the microtubules depolymerize. When the temperature of this solution was brought to 5 $^{\circ}\text{C}$, the depolymerization could be followed by the decrease in the absorbance at 320 nm. After a lag coinciding with the time necessary to lower the temperature, the depolymerization followed zero-order kinetics (dashed line) until the number of microtubules in the solution began to decrease. Adapted with permission from ref 505. Copyright 1977 Academic Press.

$$\frac{d[\text{polymer}]_b}{dt} = k_{-n,D^+} [+ \text{end}] + k_{-n,D^-} [- \text{end}] \quad (13-52)$$

As long as [end] remains constant, the reaction remains zero-order. As [end] begins to decrease, when more and more microtubules cease to exist, the rate of depolymerization decreases.⁵⁰⁵ From electron micrographs of samples removed at various times from a population of depolymerizing microtubules, the decrease in the observed rate constant of depolymerization at the longer times could be quantitatively correlated to the decrease in the number concentration of microtubules and hence the molar concentration of ends.⁵²⁹

There are a number of observations that support this description of the polymerization of tubulin. Microtubules assembled from tubulin to which guanylyl 5'-(β , γ -methylenediphosphonate), a nonhydrolyzable analogue of GTP, has been bound, rather than GTP itself, exhibit only elongation and no rapid catastrophic depoly-

merization.⁵³⁰ If the concentration of GTP in the solution is suddenly dropped so that GTP dissociates from the protomeric tubulin at the end of a microtubule, the rate at which ends switch to catastrophic depolymerization increases 50-fold. When microtubules are elongating in a mixture of GTP-tubulin and GDP-tubulin, and the concentration of GDP in the solution is increased 100-fold, so that the GTP bound to protomers at the ends of elongating tubules exchanges for GDP, the rate at which ends switch to catastrophic depolymerization increases 10-fold.⁵³¹ When a uniform population of microtubules elongating rapidly at high concentrations of monomeric GTP-tubulin is diluted to a concentration slightly below the apparent critical concentration, the bulk concentration of microtubules, $[\text{microtubule}]_b$, immediately begins to decrease. This decrease, however, is entirely due to a decrease in the number concentration of microtubules rather than the mean length of the remaining microtubules. The microtubules that remain are still slowly elongating.⁵¹⁴ Those that have disappeared have lost the race with the advancing boundary of protomeric GDP-tubulin at one of their ends. When microtubules grown from seeds reach a steady-state bulk concentration, this concentration is maintained by a decrease in the number of microtubules and an elongation of those that remain.⁵¹⁴ Those that are still in the race are staying ahead at the expense of the losers.

Colchicine and podophyllotoxin are inhibitors of tubulin elongation. They bind to free $\alpha\beta$ heterodimers of tubulin, which then have a higher affinity for an elongating end of a microtubule. Once a few of the toxin-tubulin complexes have entered an elongating end, however, it is no longer able either to elongate further or to depolymerize catastrophically when the wave of GDP-tubulin reaches it,^{532,533} because it is capped by those complexes of tubulin and colchicine or podophyllotoxin. When podophyllotoxin is added to microtubules grown to steady state from seeds, the bulk concentration of microtubule, $[\text{microtubule}]_b$, decreases in two phases. One phase has the rate constant ($k_{-n,D+} + k_{-n,D-}$) of normal depolymerization, and the other phase is much slower. The rapid phase is the depolymerization from the ends that lose the race with the boundary of GDP-tubulin and begin to depolymerize before they can be capped by complexes of tubulin and podophyllotoxin. The slow phase is the slow depolymerization of microtubules that have become capped at both ends by podophyllotoxin before either end can begin to depolymerize. If fresh tubulin is added with the podophyllotoxin, the magnitude of the rapid phase is decreased as expected.

The elongation of a microtubule is a **race** between addition of monomeric GTP-tubulin at the end and the spread of the boundary of protomeric GDP-tubulin along the body of the microtubule (Figure 13–28). The race is lost when the GDP-tubulin reaches an end that cannot outdistance it, and the penalty for losing is catastrophic depolymerization. The observed rate constant for

hydrolysis of GTP within an elongating microtubule is slow (0.06 s^{-1}), so there is little chance that it will switch from elongation to catastrophic depolymerization while it is elongating rapidly; and when it has reached a goal, the structure with which it associates at that goal caps its end.

The purpose of this elaborate device seems to be the elimination of microtubules that have failed to find a goal and the maintenance of the origin of the network of microtubules in the cell.⁵¹⁴ In a microtubule that has not been able to find a **goal** and have its elongating end capped in recognition of its success, the boundary of GDP-tubulin will eventually catch up and the uncapped microtubule that has failed in its search will catastrophically depolymerize. Also, when a microtubule breaks into two pieces for any reason, the break will almost always occur in the region where protomeric GDP-tubulin is located. The broken tubule will catastrophically depolymerize from its broken ends, and the fragment attached to the centrosome and the other fragment will disappear. This prevents broken pieces from initiating microtubules unattached to centrosomes. The centrosome has a finite capacity to initiate microtubules, but as sites become empty after the catastrophic depolymerizations of the failures, new microtubules, the elongating ends of which are again outracing their destruction and which are in their turn searching for success, are initiated at those empty sites. As the centrosome can initiate microtubules at concentrations lower than those at which they initiate spontaneously, almost all microtubules end up originating at the centrosome.

The schematic drawings of Figure 13–24B and Figure 13–28 are somewhat misleading. The actual structure of the elongating end is a **flattened sheet of protofilaments** that has not yet been rolled up and joined at the seam of its microtubule. The rolling up of the sheet and zipping of the microtubule along the seam lags behind the elongation of the protofilaments forming the unrolled, flattened sheet at the very end.⁵³⁴

The assembly of **actin** into **thin filaments** (Figure 9–1B) is similar to that of tubulin into microtubules. Thin filaments of actin are polar. As with a microtubule, a thin filament elongates at both ends, but it elongates at one end about 7 times more rapidly than at the other.^{535,536} The end that elongates more rapidly is the plus end or the barbed end of the thin filament when it is decorated with fragments of myosin. This end is the anchored end of a thin filament within the cell, toward which a thick filament of myosin usually slides.

The elongation of actin has all of the characteristic features of the elongation of tubulin, albeit with distinct rate constants, but the nucleotide that controls the polymerization is **ATP** rather than GTP.⁵³⁷ The ATP binds to the actin, the ATP-actin complex is incorporated into the thin filament, and the ATP is then hydrolyzed to ADP slowly enough that the **hydrolysis of the ATP** significantly **lags behind the elongation** of the filament.^{538,539} The elongation of actin filaments displays a critical concentration

for monomeric actin.^{540–542} The ADP·actin complex is unable to elongate actin filaments at concentrations at which ATP·actin can,⁵³⁷ because the critical concentration of ADP·actin is much higher than that of ATP·actin.^{538,540,541} Actin has been crystallized with ATP bound and with ADP bound, and the two crystallographic molecular models have significantly different conformations, an observation that explains their different critical concentrations.⁵⁴³ The rate of depolymerization of filaments of actin with ADP·actin at their ends is about 10-fold greater than the rate of depolymerization of filaments of actin with ATP·actin at their ends,^{540,541,544} so catastrophic depolymerization occurs upon exposure of an end uncapped by ATP·actin.

The **nucleation of the polymerization of actin filaments** in the cell is accomplished by particular proteins or complexes of proteins. One complex of proteins responsible for nucleation of actin filaments is the Arp213 complex, which contains seven different subunits, each present in a single copy, two of which are homologues of actin and may serve as the actual sites of nucleation.⁵⁴⁵ There are also a number of individual, monomeric proteins that are able to initiate the polymerization of actin, such as villin,^{542,546–548} fragmin,⁵⁴⁹ gelsolin,⁵⁵⁰ and F-actin capping protein.⁵⁵¹ They do so by forming complexes with two or three, one, two, or two or three molecules of actin, respectively, and it is these complexes that nucleate polymerization. Consequently, the nuclei required to initiate polymerization of actin are much simpler than those required to initiate polymerization of tubulin. Purified preparations of actin, like purified preparations of tubulin, can spontaneously nucleate polymerization when nucleotide is added, but at least one of the contaminating proteins responsible for this self-nucleation is a covalent dimer of actin produced adventitiously during the purification.⁵⁵²

Reversibly assembled helical polymers such as microtubules and filaments of actin interact with an elaborate set of stabilizing and destabilizing proteins that control, sculpt, and employ the basic polymers according to the needs of the cell. In humans, there are more than 15 **microtubule-associated proteins** that have been identified by their ability to copolymerize with tubulin. The kinetochores of chromosomes bind to the elongating ends of microtubules in such a way that the ends can still elongate and then depolymerize, in the process pushing and then pulling the kinetochores and their attached chromosomes away from and then toward the centrosome.⁵¹³ In this process, it is the significant, favorable free energy of elongation resulting from the fact that the concentration of monomeric GTP·tubulin is well above its critical concentration that provides the free energy to drag the kinetochore along with the elongating end away from the centrosome, and it is the much more favorable free energy of depolymerization resulting from the fact that the concentration of monomeric GDP·tubulin is even more distant from its critical concentration that provides the free

energy to pull the kinetochore even more vigorously toward the centrosome. Dynein is a protein that slides along the outer surfaces of microtubules toward the centrosome while hydrolyzing MgATP, carrying structures to which it in turn is attached in that direction.^{553–555} Kinesins slide along microtubules also while hydrolyzing MgATP.⁵⁵⁶

Thin filaments of actin are often sculpted to precisely regulated lengths and shapes for certain purposes. There are a number of proteins responsible for this function.^{557–561} For example, **F-actin capping protein** is a widely distributed $\alpha\beta$ heterodimer that binds to elongating barbed ends of actin filaments and stably caps them by preventing them from either elongating further or catastrophically depolymerizing.^{551,562–566} Its wide distribution suggests that it is the major capping protein in animal cells. F-Actin capping protein is located in the Z line of skeletal muscle,⁵⁶⁷ a structure in which are embedded the barbed plus ends of the actin filaments that are organized in regular arrays of precise length found in this tissue. The length of these actin filaments in skeletal muscle is defined by the length of a single molecule of the long fibrous protein nebulin,⁵⁶⁸ which acts as a **molecular ruler**^{569,570} that binds tightly along the length of the actin filaments.⁵⁷¹ It is the Z line upon which the actin filament pulls as the thick filament of myosin slides along it in that direction. Capped plus ends of actin filaments are also embedded in structures at the cellular membrane organized around the protein vinculin.^{572,573} The minus ends of the actin filaments in the regular arrays in cardiac muscle are capped by tropomodulin.⁵⁷⁴

Thick filaments of myosin are the oligomeric proteins that slide along thin filaments of actin and pull upon them while hydrolyzing MgATP. Thick filaments are helical polymeric proteins noncovalently assembled from a monomer known as myosin. Myosin is composed from two identical α polypeptides ($n_{aa} = 1940$) and several shorter polypeptides ($n_{aa} = 150–200$). The carboxy-terminal 1100 aa of the two α polypeptides are entwined around each other to form a two-stranded, **α -helical coiled coil** (Figure 6–29) 150 nm in length⁵⁷⁵ that has two globular, detachable domains known as heads, each of which is formed from the amino-terminal 800 aa of one of the α polypeptides, at one of its ends (Figure 13–30A).^{575–578} The shorter polypeptides are incorporated into these **globular heads**.

The individual coiled coils of the myosin molecules are segments of rope that are assembled into a helical cable from which the myosin heads protrude (Figure 13–30B).⁵⁷⁷ The segments of rope add to the elongating cable at each of its two ends with opposite orientation. In each direction along the cable, the molecules of myosin add so that the empty carboxy-terminal ends of their segments of rope point toward the middle of the cable and the amino-terminal ends of the segments of rope to which the myosin heads are attached point away from the middle (Figure 13–30C).⁵⁷⁹ The absence of myosin

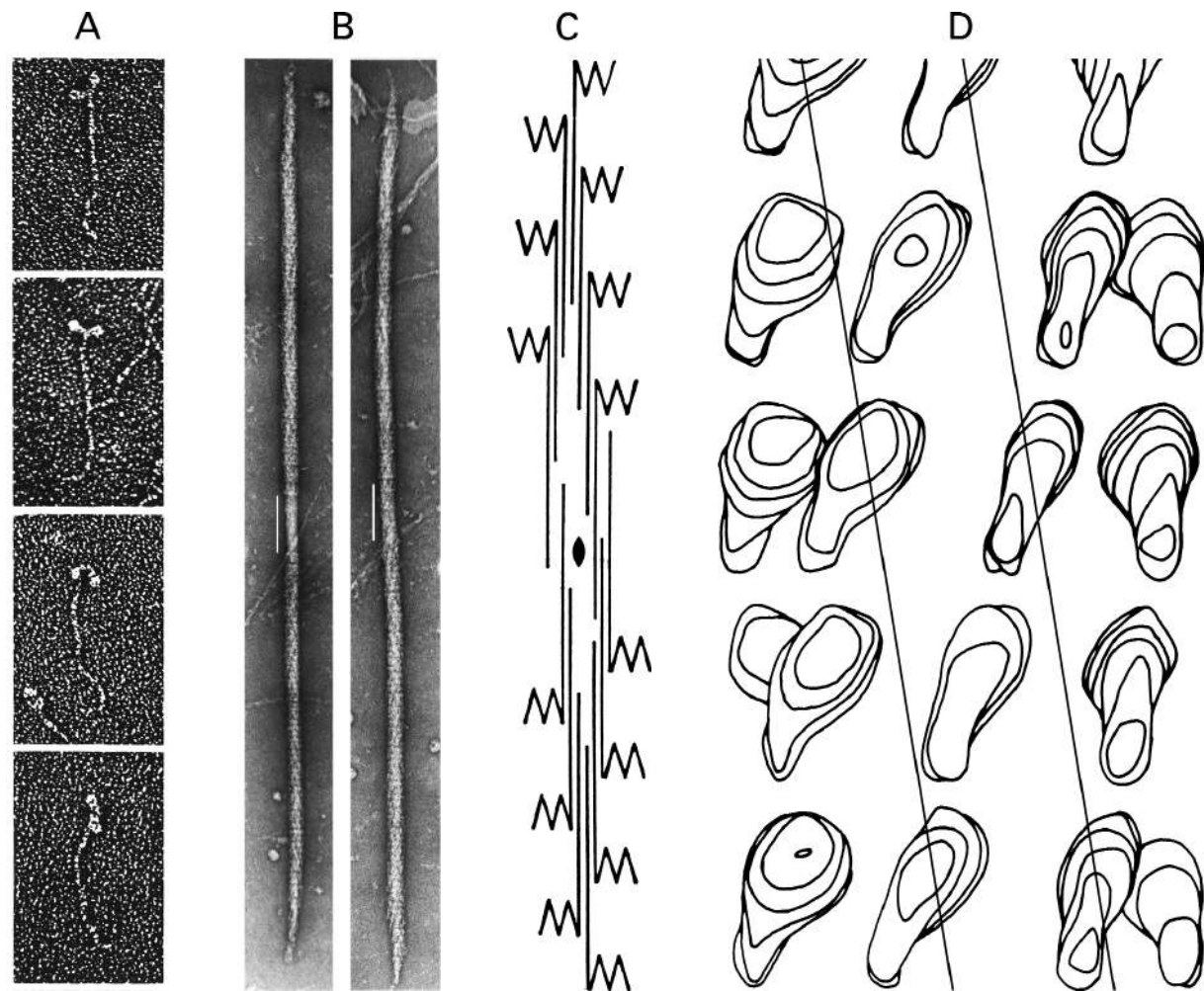


Figure 13-30: Structures of myosin and thick filaments. (A) Gallery of electron micrographs of myosin molecules.⁵⁷⁵ Myosin from thick filaments in rabbit skeletal muscle, which had been disassembled at high ionic strength (0.5 M KCl), was purified by precipitation at low ionic strength, ammonium sulfate precipitation, and anion-exchange chromatography. A solution of purified myosin at $50 \mu\text{g mL}^{-1}$ in 0.6 M ammonium formate was sprayed onto the surface of freshly cleaved mica, and the water and the volatile salt of ammonium formate were evaporated from the surface. The adsorbed molecules of myosin were coated with platinum as the mica was rotated in a beam of platinum vapor. The film of platinum was removed from the mica, transferred to a grid, and viewed in an electron microscope. Magnification 120000 \times . Reprinted with permission from ref 575. Copyright 1978 Academic Press. (B) Thick filaments from muscle of *Placopectin magellanicus*.⁵⁷⁷ Strips of muscle were chopped finely and homogenized in a solution containing MgATP to dissociate thick and thin filaments. A drop of this homogenate was placed on a carbon film and negatively stained with 3% uranyl acetate. Thick filaments were located in the specimen in the electron microscope and photographs were taken. The white bars indicate the bare zones on the thick filaments. Magnification 60000 \times . Adapted with permission from ref 577. Copyright 1983 Academic Press. (C) Diagrammatic representation of the way in which molecules of myosin are assembled to form a thick filament.⁵⁷⁸ Each continuous set of line segments is an individual molecule of myosin; the two globular heads (panel A) are represented by a W and the tail by a line. All of the tails point toward the center of the filament, and at the center the orientation reverses. Because the pairs of heads are directed distally, the bare zone (the smooth central portion of each of the thick filaments indicated by white bars in panel B) has no heads protruding from it. The structure assembled in this way has a 2-fold rotational axis of pseudosymmetry at its center. Reprinted with permission from ref 578. Copyright 1969 American Association for the Advancement of Science. (D) Helical surface lattice of globular myosin heads upon a thick filament.⁵⁷⁷ The optical density of electron micrographs, such as those in panel B, was digitized, and the Fourier transform of the digitized optical density was calculated. Discrete reflections arising from longitudinal spacings of 14.5 and 29.0 nm along the thick filament and helical spacings of 48.0 nm (panel B) were observed in the Fourier transform. Reflections on the reciprocal helical lattice in the Fourier transform were selected and used to calculate a three-dimensional distribution of electron scattering density. The image presented is of the front four helical strands of the seven-start right-handed helical lattice on one thick filament. This helical pattern produces the strong reflections of a 48.0 nm helical repeat. The globular heads are arranged in circular disks 14.5 nm in height and seven heads in circumference that are stacked helically to produce the lattice. The lines drawn in the figure indicate the probable orientation of two thin filaments of actin relative to the surface lattice of the thick filament myosin. Reprinted with permission from ref 577. Copyright 1983 Academic Press.

heads where the segments of rope pointing in opposite directions overlap in the middle of the cable creates a **bare zone** 150 nm in length.⁵⁷⁷ Because of this pattern of assembly, thick filaments (Figure 13–30B),⁵⁷⁷ unlike thin filaments, which have two distinct ends, have two ends that are identical but of opposite orientation. This necessarily produces a 2-fold rotational axis of pseudosymmetry normal to the axis of the thick filament in the center of the bare zone (Figure 13–30C).

Upon the surface of the thick filament distal to the bare zone, the myosin heads are arranged in a **helical surface lattice**, reflecting the underlying helical symmetry of the cable (Figure 13–30D). This helical surface lattice is right-handed and septuply threaded, and myosin heads protrude from each of the seven constituent helices at intervals that are vertically in register⁵⁷⁷ to create horizontal rings, or crowns, each spaced at 14.4-nm intervals.^{577,580} The seven globular protrusions⁵⁸¹ around each crown are each single myosin heads because each crown accounts for the total molecular mass of about 3.5 molecules of myosin.^{580,582} Because seven is an odd number, no two adjacent globular heads in the same crown can be from the same molecule of myosin. The pair of heads from the same molecule of myosin must be consecutive to each other within the same helix. Therefore, along each of the seven threads of the helical lattice, the heads must alternate in the pattern lower head of myosin i , upper head of myosin i , lower head of myosin $i+1$, upper head of myosin $i+1$, and so forth. If upper heads are in register across the seven helices and lower heads are in register, crowns of lower heads and crowns of upper heads would alternate along the thick filament. Such an alternating pattern has been observed.⁵⁷⁷

Thick filaments of myosin assemble spontaneously from monomers of myosin⁵⁷⁹ but do not become helical polymeric proteins of indefinite length such as fibrin, tubulin, or actin. Their **final length** is between 1000 and 3000 nm when they are polymerized under experimental situations⁵⁸³ or between 1600 and 2000 nm when they are polymerized within a contractile tissue such as skeletal muscle (Figure 13–30B).^{577,584}

Myosin from *Acanthamoeba castellanii* spontaneously assembles into minifilaments that are bipolar thick filaments composed of only eight myosin monomers. During the assembly of these minifilaments, monomers first form antiparallel dimers, antiparallel dimers form antiparallel tetramers, and tetramers then form octamers.⁵⁸⁵ The formation of such an antiparallel dimer may also be the initial step in the formation of the larger types of thick filaments.^{586,587}

When monomeric actin is induced to polymerize in the presence of thick filaments of myosin, thin filaments of actin form around each thick filament of myosin.⁵⁸³ The thin filaments are positioned around the thick filaments at seven evenly spaced intervals. Presumably, the assembly of this 7-fold array is dictated by the underlying seven helices of the myosin heads. The pitch of the seven

thin filaments of this actin, however, is much steeper than that of the seven primary helices on the surface of the thick filament, the assembled thin filaments of actin are almost parallel to the axis of the thick filament, and the pitch of the septuply threaded helical array of actin filaments is left-handed instead of right-handed. This alignment could be explained if the thin filaments of actin were in contact only with every other crown or every fourth crown along the thick filament and stepped up one helix for each contact (Figure 13–30D).

The measured rise for each subunit in a thin filament of actin is 2.8 nm and the measured rotation for each subunit is 166°. For a thin filament to span two crowns and step up one helix (Figure 13–30D) would require 29.7 nm, the distance covered by 11 subunits of actin if the rise for each subunit in a thin filament were actually 2.7 nm rather than 2.8 nm. The eleventh protomer further along a thin filament would be pointed in exactly the same direction toward the thick filament as a subunit of actin already attached to a myosin head in that thick filament if the rotation for each subunit of actin in a thin filament were 164° instead of 166°. A similarly successful but not so remarkable fit of the dimensions can be made if the thin filament of actin makes contact only with every fourth crown and steps up one helix. The **coincidences between the dimensions of the thin filament and the thick filament** are reminders that the more primordial of the two served as the template for the evolution of the other.

With the assembly of helical polymeric proteins such as microtubules, thick filaments of myosin, and sculpted thin filaments of actin and the assembly of oligomeric proteins such as ribosomes, protein chemistry enters the microscopic realm and becomes cell biology. Another set of striking microscopic cellular features that are of importance in cell biology are membranes.

Suggested Reading

Mitchison, T., & Kirschner, M. (1984) Microtubule assembly nucleated by isolated centrosomes and dynamic instability of microtubule growth, *Nature* 312, 232–242.

Problem 13–8: Make a xerographic copy of Figure 13–24B. Cut out the surface lattice and roll the paper into a cylinder. Follow the various helices over the cylinder and identify how many individual strands each of them has.

Problem 13–9: The observed rate constants for microtubule assembly listed in this table were obtained from the data in Figure 13–27.

rate constant	value	rate constant	value
$k_{n,T+}$	$3.8 \times 10^6 \text{ M}^{-1} \text{ s}^{-1}$	$k_{-n,T+}$	0.4 s^{-1}
$k_{n,T-}$	$1.2 \times 10^6 \text{ M}^{-1} \text{ s}^{-1}$	$k_{-n,T-}$	1.1 s^{-1}
$k_{-n,D+}$	340 s^{-1}	$k_{-n,D-}$	210 s^{-1}

- (A) Describe which rate constant was derived from which aspect of this figure. Which rate constants are in error and why?
- (B) If the concentration of monomeric tubulin were 10 μM , the steady-state apparent critical concentration, at what rate (second^{-1}) would a tubule capped with GTP-tubulin be elongating at its minus end?
- (C) If the plus end of this tubule were capped with GDP-tubulin, at what rate (second^{-1}) would it be depolymerizing?

References

1. Dill, K.A., & Shortle, D. (1991) *Annu. Rev. Biochem.* 60, 795–825.
2. Weinreb, P.H., Zhen, W., Poon, A.W., Conway, K.A., & Lansbury, P.T., Jr. (1996) *Biochemistry* 35, 13709–13715.
3. Uversky, V.N., Gillespie, J.R., & Fink, A.L. (2000) *Proteins: Struct., Funct., Genet.* 41, 415–427.
4. Rariy, R.V., & Klibanov, A.M. (1997) *Proc. Natl. Acad. Sci. U.S.A.* 94, 13520–13523.
5. Tanford, C. (1968) Protein denaturation, *Adv. Protein Chem.* 23, 121–282.
6. Edelhoch, H. (1967) *Biochemistry* 6, 1948–1954.
7. Nozaki, Y., & Tanford, C. (1963) *J. Biol. Chem.* 238, 4074–4081.
8. Nozaki, Y., & Tanford, C. (1970) *J. Biol. Chem.* 245, 1648–1652.
9. Wetlaufer, D.B., Malik, S., Stoller, L., & Coffin, R.L. (1964) *J. Am. Chem. Soc.* 86, 508–514.
10. Nandi, P.K., & Robinson, D.R. (1984) *Biochemistry* 23, 6661–6668.
11. Creighton, T.E. (1979) *J. Mol. Biol.* 129, 235–264.
12. Parker, M.J., Spencer, J., & Clarke, A.R. (1995) *J. Mol. Biol.* 253, 771–786.
13. Makhatadze, G.I., & Privalov, P.L. (1992) *J. Mol. Biol.* 226, 491–505.
14. Lee, J.C., & Timasheff, S.N. (1974) *Biochemistry* 13, 257–265.
15. Courtenay, E.S., Capp, M.W., & Record, M.T., Jr. (2001) *Protein Sci.* 10, 2485–2497.
16. Timasheff, S.N., & Xie, G. (2003) *Biophys. Chem.* 105, 421–448.
17. Breslow, R., & Guo, T. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87, 167–169.
18. Roseman, M., & Jencks, W.P. (1975) *J. Am. Chem. Soc.* 97, 631–640.
19. Herskovits, T.T., Jaillet, H., & Gadegbeku, B. (1970) *J. Biol. Chem.* 245, 4544–4550.
20. Pace, C.N., & Marshall, H.F., Jr. (1980) *Arch. Biochem. Biophys.* 199, 270–276.
21. Steiner, D.F., & Clark, J.L. (1968) *Proc. Natl. Acad. Sci. U.S.A.* 60, 622–629.
22. Roxby, R., & Tanford, C. (1971) *Biochemistry* 10, 3348–3352.
23. Yao, M., & Bolen, D.W. (1995) *Biochemistry* 34, 3771–3781.
24. Bolen, D.W., & Santoro, M.M. (1988) *Biochemistry* 27, 8069–8074.
25. Pace, C.N., Laurents, D.V., & Thomson, J.A. (1990) *Biochemistry* 29, 2564–2572.
26. Fitch, C.A., Karp, D.A., Lee, K.K., Stites, W.E., Lattman, E.E., & Garcia-Moreno, E.B. (2002) *Biophys. J.* 82, 3289–3304.
27. Langsetmo, K., Fuchs, J.A., & Woodward, C. (1991) *Biochemistry* 30, 7603–7609.
28. Inoue, M., Yamada, H., Hashimoto, Y., Yasukochi, T., Hamaguchi, K., Miki, T., Horiuchi, T., & Imoto, T. (1992) *Biochemistry* 31, 8816–8821.
29. Anderson, D.E., Becktel, W.J., & Dahlquist, F.W. (1990) *Biochemistry* 29, 2403–2408.
30. Giletto, A., & Pace, C.N. (1999) *Biochemistry* 38, 13379–13384.
31. Oliveberg, M., Arcus, V.L., & Fersht, A.R. (1995) *Biochemistry* 34, 9424–9433.
32. Stites, W.E., Gittis, A.G., Lattman, E.E., & Shortle, D. (1991) *J. Mol. Biol.* 221, 7–14.
33. Hermans, J., Jr., & Acampora, G. (1967) *J. Am. Chem. Soc.* 89, 1547–1552.
34. Robertson, A.D., & Baldwin, R.L. (1991) *Biochemistry* 30, 9907–9914.
35. Arcus, V.L., Vuilleumier, S., Freund, S.M., Bycroft, M., & Fersht, A.R. (1995) *J. Mol. Biol.* 254, 305–321.
36. Liu, Z.P., Rizo, J., & Gierasch, L.M. (1994) *Biochemistry* 33, 134–142.
37. Burton, S.J., Quirk, A.V., & Wood, P.C. (1989) *Eur. J. Biochem.* 179, 379–387.
38. Mainfroid, V., Terpstra, P., Beauregard, M., Frere, J.M., Mande, S.C., Hol, W.G., Martial, J.A., & Goraj, K. (1996) *J. Mol. Biol.* 257, 441–456.
39. Carra, J.H., & Privalov, P.L. (1997) *Biochemistry* 36, 526–535.
40. Edge, V., Allewell, N.M., & Sturtevant, J.M. (1985) *Biochemistry* 24, 5899–5906.
41. Manly, S.P., Matthews, K.S., & Sturtevant, J.M. (1985) *Biochemistry* 24, 3842–3846.
42. Brandts, J.F., Hu, C.Q., Lin, L.N., & Mos, M.T. (1989) *Biochemistry* 28, 8588–8596.
43. Alber, T., Sun, D.P., Wilson, K., Wozniak, J.A., Cook, S.P., & Matthews, B.W. (1987) *Nature* 330, 41–46.
44. Novokhatny, V.V., Kudinov, S.A., & Privalov, P.L. (1984) *J. Mol. Biol.* 179, 215–232.
45. Flanagan, M.T., & Hesketh, T.R. (1974) *Eur. J. Biochem.* 44, 251–259.
46. Yang, M., Liu, D., & Bolen, D.W. (1999) *Biochemistry* 38, 11216–11222.
47. Baskakov, I.V., & Bolen, D.W. (1998) *Biochemistry* 37, 18010–18017.
48. Benz, F.W., & Roberts, G.C. (1975) *J. Mol. Biol.* 91, 367–387.
49. Hoeltzli, S.D., & Frieden, C. (1994) *Biochemistry* 33, 5502–5509.
50. Perl, D., Welker, C., Schindler, T., Schroder, K., Marahiel, M.A., Jaenicke, R., & Schmid, F.X. (1998) *Nat. Struct. Biol.* 5, 229–235.
51. Steif, C., Weber, P., Hinz, H.J., Flossdorf, J., Cesareni, G., & Kokkinidis, M. (1993) *Biochemistry* 32, 3867–3876.
52. Lumry, R., Biltonen, R., & Brandts, J.F. (1966) *Biopolymers* 4, 917.

53. Brandts, J.F., & Hunt, L. (1967) *J. Am. Chem. Soc.* 89, 4826–4838.
54. Aune, K.C., & Tanford, C. (1969) *Biochemistry* 8, 4579–4585.
55. Plaza del Pino, I.M., Pace, C.N., & Freire, E. (1992) *Biochemistry* 31, 11196–11202.
56. Yu, Y., Makhatadze, G.I., Pace, C.N., & Privalov, P.L. (1994) *Biochemistry* 33, 3312–3319.
57. Shen, L.L., & Hermans, J., Jr. (1972) *Biochemistry* 11, 1836–1841.
58. Ginsburg, A., & Carroll, W.R. (1965) *Biochemistry* 4, 2159–2174.
59. Jasanoff, A., Davis, B., & Fersht, A.R. (1994) *Biochemistry* 33, 6350–6355.
60. Rudolph, R., Siebendritt, R., Nessler, G., Sharma, A.K., & Jaenicke, R. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87, 4625–4629.
61. Frech, C., Wunderlich, M., Glockshuber, R., & Schmid, F.X. (1996) *Biochemistry* 35, 11386–11395.
62. Gualfetti, P.J., Bilsel, O., & Matthews, C.R. (1999) *Protein Sci.* 8, 1623–1635.
63. Herold, M., & Kirschner, K. (1990) *Biochemistry* 29, 1907–1913.
64. Brazhnikov, E.V., Chirgadze, Y., Dolgikh, D.A., & Ptitsyn, O.B. (1985) *Biopolymers* 24, 1899–1907.
65. Hughson, F.M., Wright, P.E., & Baldwin, R.L. (1990) *Science* 249, 1544–1548.
66. Kuwajima, K., Nitta, K., Yoneyama, M., & Sugai, S. (1976) *J. Mol. Biol.* 106, 359–373.
67. Wong, K.P., & Tanford, C. (1973) *J. Biol. Chem.* 248, 8518–8523.
68. Chaudhuri, T.K., Arai, M., Terada, T.P., Ikura, T., & Kuwajima, K. (2000) *Biochemistry* 39, 15643–15651.
69. Sasahara, K., Demura, M., & Nitta, K. (2000) *Biochemistry* 39, 6475–6482.
70. Timm, D.E., de Haseth, P.L., & Neet, K.E. (1994) *Biochemistry* 33, 4667–4676.
71. Apiyo, D., Jones, K., Guidry, J., & Wittung-Stafshede, P. (2001) *Biochemistry* 40, 4940–4948.
72. Zhuang, P., Eisenstein, E., & Howell, E.E. (1994) *Biochemistry* 33, 4237–4244.
73. Grimsley, J.K., Scholtz, J.M., Pace, C.N., & Wild, J.R. (1997) *Biochemistry* 36, 14366–14374.
74. Predki, P.F., & Regan, L. (1995) *Biochemistry* 34, 9834–9839.
75. Liang, H., Sandberg, W.S., & Terwilliger, T.C. (1993) *Proc. Natl. Acad. Sci. U.S.A.* 90, 7010–7014.
76. Silinski, P., Allingham, M.J., & Fitzgerald, M.C. (2001) *Biochemistry* 40, 4493–4502.
77. Xie, D., Gulnik, S., & Erickson, J.W. (2000) *J. Am. Chem. Soc.* 122, 11533–11534.
78. Pace, N.C., & Tanford, C. (1968) *Biochemistry* 7, 198–208.
79. Johnson, C.M., Oliveberg, M., Clarke, J., & Fersht, A.R. (1997) *J. Mol. Biol.* 268, 198–208.
80. Milne, J.S., Xu, Y., Mayne, L.C., & Englander, S.W. (1999) *J. Mol. Biol.* 290, 811–822.
81. Privalov, P.L., Tiktopulo, E.I., Venyaminov, S., Griko Yu, V., Makhatadze, G.I., & Khechinashvili, N.N. (1989) *J. Mol. Biol.* 205, 737–750.
82. Pace, C.N., & Laurents, D.V. (1989) *Biochemistry* 28, 2520–2525.
83. Sturtevant, J.M. (1977) *Proc. Natl. Acad. Sci. U.S.A.* 74, 2236–2240.
84. Doig, A.J., & Williams, D.H. (1991) *J. Mol. Biol.* 217, 389–398.
85. Privalov, P.L., & Makhatadze, G.I. (1990) *J. Mol. Biol.* 213, 385–391.
86. Johnson, C.M., & Fersht, A.R. (1995) *Biochemistry* 34, 6795–6804.
87. Myers, J.K., Pace, C.N., & Scholtz, J.M. (1995) *Protein Sci.* 4, 2138–2148.
88. Baldwin, R.L. (1986) *Proc. Natl. Acad. Sci. U.S.A.* 83, 8069–8072.
89. Privalov, P.L., & Makhatadze, G.I. (1992) *J. Mol. Biol.* 224, 715–723.
90. Hackel, M., Hinz, H.J., & Hedwig, G.R. (1999) *J. Mol. Biol.* 291, 197–213.
91. Makhatadze, G.I., & Privalov, P.L. (1990) *J. Mol. Biol.* 213, 375–384.
92. Clark, N.S., Dodd, I., Mossakowska, D.E., Smith, R.A., & Gore, M.G. (1996) *Protein Eng.* 9, 877–884.
93. Brandts, J.F. (1964) *J. Am. Chem. Soc.* 86, 4302–4314.
94. Chen, B.L., & Schellman, J.A. (1989) *Biochemistry* 28, 685–691.
95. Wong, K.B., Freund, S.M., & Fersht, A.R. (1996) *J. Mol. Biol.* 259, 805–818.
96. Zhang, J., Peng, X., Jonas, A., & Jonas, J. (1995) *Biochemistry* 34, 8631–8641.
97. Brandts, J.F., Oliveira, R.J., & Westort, C. (1970) *Biochemistry* 9, 1038–1047.
98. Hawley, S.A. (1971) *Biochemistry* 10, 2436–2442.
99. Zipp, A., & Kauzmann, W. (1973) *Biochemistry* 12, 4217–4228.
100. Gavish, B., Gratton, E., & Hardy, C.J. (1983) *Proc. Natl. Acad. Sci. U.S.A.* 80, 750–754.
101. Paladini, A.A., Jr., & Weber, G. (1981) *Biochemistry* 20, 2587–2593.
102. Kellis, J.T., Jr., Nyberg, K., & Fersht, A.R. (1989) *Biochemistry* 28, 4914–4922.
103. Pace, C.N. (1975) *CRC Crit. Rev. Biochem.* 3, 1–43.
104. Tanford, C. (1970) *Adv. Protein Chem.* 24, 1–95.
105. Santoro, M.M., & Bolen, D.W. (1988) *Biochemistry* 27, 8063–8068.
106. Puett, D. (1973) *J. Biol. Chem.* 248, 4623–4634.
107. Greene, R.F., Jr., & Pace, C.N. (1974) *J. Biol. Chem.* 249, 5388–5393.
108. Staniforth, R.A., Burston, S.G., Smith, C.J., Jackson, G.S., Badcoe, I.G., Atkinson, T., Holbrook, J.J., & Clarke, A.R. (1993) *Biochemistry* 32, 3842–3851.
109. Schindler, T., Herrler, M., Marahiel, M.A., & Schmid, F.X. (1995) *Nat. Struct. Biol.* 2, 663–673.
110. Myers, J.K., & Oas, T.G. (2001) *Nat. Struct. Biol.* 8, 552–558.
111. Santoro, M.M., & Bolen, D.W. (1992) *Biochemistry* 31, 4901–4907.
112. Jamin, M., & Baldwin, R.L. (1996) *Nat. Struct. Biol.* 3, 613–618.
113. Horng, J.C., Cho, J.H., & Raleigh, D.P. (2005) *J. Mol. Biol.* 345, 163–173.
114. McNutt, M., Mullins, L.S., Raushel, F.M., & Pace, C.N. (1990) *Biochemistry* 29, 7572–7576.
115. Goto, Y., & Hamaguchi, K. (1982) *J. Mol. Biol.* 156, 891–910.

116. Bai, Y., Sosnick, T.R., Mayne, L., & Englander, S.W. (1995) *Science* 269, 192–197.
117. Chamberlain, A.K., Handel, T.M., & Marqusee, S. (1996) *Nat. Struct. Biol.* 3, 782–787.
118. Huyghues-Despointes, B.M., Scholtz, J.M., & Pace, C.N. (1999) *Nat. Struct. Biol.* 6, 910–912.
119. Scholtz, J.M., Barrick, D., York, E.J., Stewart, J.M., & Baldwin, R.L. (1995) *Proc. Natl. Acad. Sci. U.S.A.* 92, 185–189.
120. Ibarra-Molero, B., & Sanchez-Ruiz, J.M. (1996) *Biochemistry* 35, 14689–14702.
121. Baldwin, E., Xu, J., Hajiseyedi, O., Baase, W.A., & Matthews, B.W. (1996) *J. Mol. Biol.* 259, 542–559.
122. Mendel, D., Ellman, J.A., Chang, Z., Veenstra, D.L., Kollman, P.A., & Schultz, P.G. (1992) *Science* 256, 1798–1802.
123. Wynn, R., & Richards, F.M. (1993) *Protein Sci.* 2, 395–403.
124. Salahuddin, A., & Tanford, C. (1970) *Biochemistry* 9, 1342–1347.
125. Aune, K.C., & Tanford, C. (1969) *Biochemistry* 8, 4586–4590.
126. Pace, C.N., & Vanderburg, K.E. (1979) *Biochemistry* 18, 288–292.
127. Knapp, J.A., & Pace, C.N. (1974) *Biochemistry* 13, 1289–1294.
128. Neira, J.L., Itzhaki, L.S., Otzen, D.E., Davis, B., & Fersht, A.R. (1997) *J. Mol. Biol.* 270, 99–110.
129. Akasaka, A., Haruki, M., Oobatake, M., & Kanaya, S. (1995) *Biochemistry* 34, 8115–8122.
130. Van den Burg, B., Vriend, G., Veltman, O.R., Venema, G., & Eijssink, V.G. (1998) *Proc. Natl. Acad. Sci. U.S.A.* 95, 2056–2060.
131. Rehage, A., & Schmid, F.X. (1982) *Biochemistry* 21, 1499–1505.
132. Llinas, M., Gillespie, B., Dahlquist, F.W., & Marqusee, S. (1999) *Nat. Struct. Biol.* 6, 1072–1078.
133. Linse, S., Teleman, O., & Drakenberg, T. (1990) *Biochemistry* 29, 5925–5934.
134. Wagner, G. (1983) *Q. Rev. Biophys.* 16, 1–57.
135. Wang, Q.W., Kline, A.D., & Wuthrich, K. (1987) *Biochemistry* 26, 6488–6493.
136. Wand, A.J., Roder, H., & Englander, S.W. (1986) *Biochemistry* 25, 1107–1114.
137. Goedken, E.R., & Marqusee, S. (2001) *J. Mol. Biol.* 314, 863–871.
138. Chamberlain, A.K., & Marqusee, S. (1998) *Biochemistry* 37, 1736–1742.
139. Laurents, D.V., Scholtz, J.M., Rico, M., Pace, C.N., & Bruix, M. (2005) *Biochemistry* 44, 7644–7655.
140. Heinz, D.W., Baase, W.A., & Matthews, B.W. (1992) *Proc. Natl. Acad. Sci. U.S.A.* 89, 3751–3755.
141. Lin, M.C. (1970) *J. Biol. Chem.* 245, 6726–6731.
142. Taniuchi, H., & Anfinsen, C.B. (1969) *J. Biol. Chem.* 244, 3864–3875.
143. Sachs, D.H., Schechter, A.N., Eastlake, A., & Anfinsen, C.B. (1974) *Nature* 251, 242–244.
144. Alexandrescu, A.T., Abeygunawardana, C., & Shortle, D. (1994) *Biochemistry* 33, 1063–1072.
145. Wang, Y., & Shortle, D. (1995) *Biochemistry* 34, 15895–15905.
146. Peters, R.J., Shiau, A.K., Sohl, J.L., Anderson, D.E., Tang, G., Silen, J.L., & Agard, D.A. (1998) *Biochemistry* 37, 12058–12067.
147. Ikemura, H., & Inouye, M. (1988) *J. Biol. Chem.* 263, 12959–12963.
148. Eder, J., Rheinhecker, M., & Fersht, A.R. (1993) *Biochemistry* 32, 18–26.
149. Zhu, X.L., Ohta, Y., Jordan, F., & Inouye, M. (1989) *Nature* 339, 483–484.
150. Winther, J.R., & Sorensen, P. (1991) *Proc. Natl. Acad. Sci. U.S.A.* 88, 9330–9334.
151. Klee, W.A. (1968) *Biochemistry* 7, 2731–2736.
152. Wyckoff, H.W., Hardman, K.D., Allewell, N.M., Inagami, T., Johnson, L.N., & Richards, F.M. (1967) *J. Biol. Chem.* 242, 3984–3988.
153. Taniuchi, H., & Anfinsen, C.B. (1971) *J. Biol. Chem.* 246, 2291–2301.
154. Sancho, J., & Fersht, A.R. (1992) *J. Mol. Biol.* 224, 741–747.
155. Eder, J., & Kirschner, K. (1992) *Biochemistry* 31, 3617–3625.
156. Lindsay, C.D., & Pain, R.H. (1991) *Biochemistry* 30, 9034–9040.
157. Rochet, J.C., Oikawa, K., Hicks, L.D., Kay, C.M., Bridger, W.A., & Wolodko, W.T. (1997) *Biochemistry* 36, 8807–8820.
158. Goldenberg, D.P., & Creighton, T.E. (1983) *J. Mol. Biol.* 165, 407–413.
159. Luger, K., Hommel, U., Herold, M., Hofsteenge, J., & Kirschner, K. (1989) *Science* 243, 206–210.
160. Yang, Y.R., & Schachman, H.K. (1993) *Proc. Natl. Acad. Sci. U.S.A.* 90, 11980–11984.
161. Buchwalder, A., Szadkowski, H., & Kirschner, K. (1992) *Biochemistry* 31, 1621–1630.
162. Kreitman, R.J., Puri, R.K., & Pastan, I. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 6889–6893.
163. Mullins, L.S., Wesseling, K., Kuo, J.M., Garrett, J.B., & Raushel, F.M. (1994) *J. Am. Chem. Soc.* 116, 5529–5533.
164. Graf, R., & Schachman, H.K. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 11591–11596.
165. Hennecke, J., Sebbel, P., & Glockshuber, R. (1999) *J. Mol. Biol.* 286, 1197–1215.
166. Iwakura, M., Nakamura, T., Yamane, C., & Maki, K. (2000) *Nat. Struct. Biol.* 7, 580–585.
167. Haber, E. (1964) *Proc. Natl. Acad. Sci. U.S.A.* 52, 1099–1106.
168. Whitney, P.L., & Tanford, C. (1965) *Proc. Natl. Acad. Sci. U.S.A.* 53, 524–532.
169. Painter, R.G., Sage, H.J., & Tanford, C. (1972) *Biochemistry* 11, 1338–1345.
170. Kauzman, W. (1959) *Adv. Protein Chem.* 14, 1–63.
171. Chothia, C. (1976) *J. Mol. Biol.* 105, 1–12.
172. Dill, K.A. (1990) *Biochemistry* 29, 7133–7155.
173. Takano, K., Scholtz, J.M., Sacchettini, J.C., & Pace, C.N. (2003) *J. Biol. Chem.* 278, 31790–31795.
174. Privalov, P.L., & Makhatadze, G.I. (1993) *J. Mol. Biol.* 232, 660–679.
175. Flory, P.J. (1949) *J. Chem. Phys.* 17, 303–310.
176. Dill, K.A. (1985) *Biochemistry* 24, 1501–1509.
177. Villafranca, J.E., Howell, E.E., Oatley, S.J., Xuong, N.H., & Kraut, J. (1987) *Biochemistry* 26, 2182–2189.
178. Pjura, P.E., Matsumura, M., Wozniak, J.A., & Matthews, B.W. (1990) *Biochemistry* 29, 2592–2598.

179. Siedler, F., Rudolph-Bohner, S., Doi, M., Musiol, H.J., & Moroder, L. (1993) *Biochemistry* 32, 7488–7495.
180. Johnson, R.E., Adams, P., & Rupley, J.A. (1978) *Biochemistry* 17, 1479–1484.
181. Imoto, T., & Rupley, J.A. (1973) *J. Mol. Biol.* 80, 657–667.
182. Mitra, S., & Lawton, R.G. (1979) *J. Am. Chem. Soc.* 101, 3097–3110.
183. Lin, S.H., Konishi, Y., Denton, M.E., & Scheraga, H.A. (1984) *Biochemistry* 23, 5504–5512.
184. Matsumura, M., Becktel, W.J., Levitt, M., & Matthews, B.W. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 6562–6566.
185. Zhang, T., Bertelsen, E., & Alber, T. (1994) *Nat. Struct. Biol.* 1, 434–438.
186. Chan, H.S., & Dill, K.A. (1989) *J. Chem. Phys.* 90, 492–509.
187. Ikeguchi, M., Sugai, S., Fujino, M., Sugawara, T., & Kuwajima, K. (1992) *Biochemistry* 31, 12695–12700.
188. Eder, J., & Wilmanns, M. (1992) *Biochemistry* 31, 4437–4444.
189. Robinson, C.R., & Sauer, R.T. (2000) *Biochemistry* 39, 12494–12502.
190. Goto, Y., & Hamaguchi, K. (1982) *J. Mol. Biol.* 156, 911–926.
191. Lau, K.F., & Dill, K.A. (1989) *Macromolecules* 22, 3986–3997.
192. Dill, K.A., Alonso, D.O., & Hutchinson, K. (1989) *Biochemistry* 28, 5439–5449.
193. Flory, P.J. (1953) *Principles of Polymer Chemistry*, Cornell University Press, Ithaca, NY.
194. Flory, P.J., & Fisk, S. (1966) *J. Chem. Phys.* 44, 2243–2248.
195. Sanchez, I.C. (1979) *Macromolecules* 12, 980–988.
196. Shaw, G.S., Hodges, R.S., & Sykes, B.D. (1990) *Science (Washington, D.C.)* 249, 280–283.
197. Chen, L.H., Kenyon, G.L., Curtin, F., Harayama, S., Bembenek, M.E., Hajipour, G., & Whitman, C.P. (1992) *J. Biol. Chem.* 267, 17716–17721.
198. Achari, A., Hale, S.P., Howard, A.J., Clore, G.M., Gronenborn, A.M., Hardman, K.D., & Whitlow, M. (1992) *Biochemistry* 31, 10449–10457.
199. Macias, M.J., Gervais, V., Civera, C., & Oschkinat, H. (2000) *Nat. Struct. Biol.* 7, 375–379.
200. Spector, S., Kuhlman, B., Fairman, R., Wong, E., Boice, J.A., & Raleigh, D.P. (1998) *J. Mol. Biol.* 276, 479–489.
201. Chan, H.S., & Dill, K.A. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87, 6388–6392.
202. Nishii, I., Kataoka, M., & Goto, Y. (1995) *J. Mol. Biol.* 250, 223–238.
203. Buchner, J., Renner, M., Lilie, H., Hinz, H.J., Jaenicke, R., Kiefhabel, T., & Rudolph, R. (1991) *Biochemistry* 30, 6922–6929.
204. Kuwajima, K. (1977) *J. Mol. Biol.* 114, 241–258.
205. Xie, D., Bhakuni, V., & Freire, E. (1991) *Biochemistry* 30, 10673–10678.
206. Stellwagen, E., & Babul, J. (1975) *Biochemistry* 14, 5135–5140.
207. Davis-Searles, P.R., Morar, A.S., Saunders, A.J., Erie, D.A., & Pielak, G.J. (1998) *Biochemistry* 37, 17048–17053.
208. Matthews, J.M., Norton, R.S., Hammacher, A., & Simpson, R.J. (2000) *Biochemistry* 39, 1942–1950.
209. Dolgikh, D.A., Gilmanshin, R.I., Brazhnikov, E.V., Bychkova, V.E., Semisotnov, G.V., Venyaminov, S., & Ptitsyn, O.B. (1981) *FEBS Lett.* 136, 311–315.
210. Ohgushi, M., & Wada, A. (1983) *FEBS Lett.* 164, 21–24.
211. Baum, J., Dobson, C.M., Evans, P.A., & Hanley, C. (1989) *Biochemistry* 28, 7–13.
212. Bu, Z., Neumann, D.A., Lee, S.H., Brown, C.M., Engelman, D.M., & Han, C.C. (2000) *J. Mol. Biol.* 301, 525–536.
213. Nolting, B., Jiang, M., & Sligar, S.G. (1993) *J. Am. Chem. Soc.* 115, 9879–9882.
214. Jeng, M.F., Englander, S.W., Elove, G.A., Wand, A.J., & Roder, H. (1990) *Biochemistry* 29, 10433–10437.
215. Schulman, B.A., Redfield, C., Peng, Z.Y., Dobson, C.M., & Kim, P.S. (1995) *J. Mol. Biol.* 253, 651–657.
216. Eliezer, D., & Wright, P.E. (1996) *J. Mol. Biol.* 263, 531–538.
217. Hughson, F.M., Barrick, D., & Baldwin, R.L. (1991) *Biochemistry* 30, 4113–4118.
218. Chakraborty, S., Ittah, V., Bai, P., Luo, L., Haas, E., & Peng, Z. (2001) *Biochemistry* 40, 7228–7238.
219. Oas, T.G., & Kim, P.S. (1988) *Nature* 336, 42–48.
220. Baldwin, R.L. (1989) *Trends Biochem. Sci.* 14, 291–294.
221. Jeng, M.F., & Englander, S.W. (1991) *J. Mol. Biol.* 221, 1045–1061.
222. Gualfetti, P.J., Iwakura, M., Lee, J.C., Kihara, H., Bilsel, O., Zitzewitz, J.A., & Matthews, C.R. (1999) *Biochemistry* 38, 13367–13378.
223. Peng, Z.Y., & Kim, P.S. (1994) *Biochemistry* 33, 2136–2141.
224. Otzen, D.E., & Oliveberg, M. (2001) *J. Mol. Biol.* 313, 479–483.
225. Yamasaki, K., Ogasahara, K., Yutani, K., Oobatake, M., & Kanaya, S. (1995) *Biochemistry* 34, 16552–16562.
226. Raschke, T.M., & Marqusee, S. (1997) *Nat. Struct. Biol.* 4, 298–304.
227. Jennings, P.A., & Wright, P.E. (1993) *Science* 262, 892–896.
228. Nishimura, C., Riley, R., Eastman, P., & Fink, A.L. (2000) *J. Mol. Biol.* 299, 1133–1146.
229. Su, Z.D., Arooz, M.T., Chen, H.M., Gross, C.J., & Tsong, T.Y. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 2539–2544.
230. Elove, G.A., Chaffotte, A.F., Roder, H., & Goldberg, M.E. (1992) *Biochemistry* 31, 6876–6883.
231. Kuwajima, K., Garvey, E.P., Finn, B.E., Matthews, C.R., & Sugai, S. (1991) *Biochemistry* 30, 7693–7703.
232. Fujiwara, K., Arai, M., Shimizu, A., Ikeguchi, M., Kuwajima, K., & Sugai, S. (1999) *Biochemistry* 38, 4455–4463.
233. Morozova-Roche, L.A., Jones, J.A., Noppe, W., & Dobson, C.M. (1999) *J. Mol. Biol.* 289, 1055–1073.
234. Morgan, C.J., Miranker, A., & Dobson, C.M. (1998) *Biochemistry* 37, 8473–8480.
235. Qi, P.X., Sosnick, T.R., & Englander, S.W. (1998) *Nat. Struct. Biol.* 5, 882–884.
236. Parker, M.J., & Marqusee, S. (1999) *J. Mol. Biol.* 293, 1195–1210.
237. Rocek, J., Westheimer, F.H., Eschenmoser, A., Moldovanyi, L., & Schreiber, J. (1962) *Helv. Chim. Acta* 45, 2554–2567.
238. Roder, H., & Colon, W. (1997) *Curr. Opin. Struct. Biol.* 7, 15–28.

239. Parker, M.J., Dempsey, C.E., Lorch, M., & Clarke, A.R. (1997) *Biochemistry* 36, 13396–13405.
240. Schreiber, G., & Fersht, A.R. (1993) *Biochemistry* 32, 11195–11203.
241. Sauder, J.M., MacKenzie, N.E., & Roder, H. (1996) *Biochemistry* 35, 16852–16862.
242. Herning, T., Yutani, K., Taniyama, Y., & Kikuchi, M. (1991) *Biochemistry* 30, 9882–9891.
243. Chen, B.L., Baase, W.A., Nicholson, H., & Schellman, J.A. (1992) *Biochemistry* 31, 1464–1476.
244. Kyte, J. (1995) *Mechanism in Protein Chemistry*, pp 461–473, Garland, New York.
245. Eliezer, D., Jennings, P.A., Wright, P.E., Doniach, S., Hodgson, K.O., & Tsuruta, H. (1995) *Science* 270, 487–488.
246. Arai, M., Ikura, T., Semisotnov, G.V., Kihara, H., Amemiya, Y., & Kuwajima, K. (1998) *J. Mol. Biol.* 275, 149–162.
247. Jones, B.E., Beechem, J.M., & Matthews, C.R. (1995) *Biochemistry* 34, 1867–1877.
248. Kuwajima, K., Yamaya, H., Miwa, S., Sugai, S., & Nagamura, T. (1987) *FEBS Lett.* 221, 115–118.
249. Hooke, S.D., Radford, S.E., & Dobson, C.M. (1994) *Biochemistry* 33, 5867–5876.
250. Bycroft, M., Matouschek, A., Kellis, J.T., Jr., Serrano, L., & Fersht, A.R. (1990) *Nature* 346, 488–490.
251. Roder, H., Elove, G.A., & Englander, S.W. (1988) *Nature* 335, 700–704.
252. Udgaonkar, J.B., & Baldwin, R.L. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87, 8197–8201.
253. Parker, M.J., & Marqusee, S. (2001) *J. Mol. Biol.* 305, 593–602.
254. O'Neill, J.C., Jr., & Robert Matthews, C. (2000) *J. Mol. Biol.* 295, 737–744.
255. Dabora, J.M., Pelton, J.G., & Marqusee, S. (1996) *Biochemistry* 35, 11951–11958.
256. Raschke, T.M., Kho, J., & Marqusee, S. (1999) *Nat. Struct. Biol.* 6, 825–831.
257. Segel, D.J., Bachmann, A., Hofrichter, J., Hodgson, K.O., Doniach, S., & Kiefhaber, T. (1999) *J. Mol. Biol.* 288, 489–499.
258. Heidary, D.K., O'Neill, J.C., Jr., Roy, M., & Jennings, P.A. (2000) *Proc. Natl. Acad. Sci. U.S.A.* 97, 5866–5870.
259. Akiyama, S., Takahashi, S., Ishimori, K., & Morishima, I. (2000) *Nat. Struct. Biol.* 7, 514–520.
260. Kuwata, K., Shastry, R., Cheng, H., Hoshino, M., Batt, C.A., Goto, Y., & Roder, H. (2001) *Nat. Struct. Biol.* 8, 151–155.
261. Park, S.H., Shastry, M.C., & Roder, H. (1999) *Nat. Struct. Biol.* 6, 943–947.
262. Yeh, S.R., Ropson, I.J., & Rousseau, D.L. (2001) *Biochemistry* 40, 4205–4210.
263. Capaldi, A.P., Shastry, M.C., Kleanthous, C., Roder, H., & Radford, S.E. (2001) *Nat. Struct. Biol.* 8, 68–72.
264. Hagen, S.J., & Eaton, W.A. (2000) *J. Mol. Biol.* 301, 1019–1027.
265. Nolting, B., Golbik, R., Neira, J.L., Soler-Gonzalez, A.S., Schreiber, G., & Fersht, A.R. (1997) *Proc. Natl. Acad. Sci. U.S.A.* 94, 826–830.
266. Ballew, R.M., Sabelko, J., & Gruebele, M. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 5759–5764.
267. Phillips, C.M., Mizutani, Y., & Hochstrasser, R.M. (1995) *Proc. Natl. Acad. Sci. U.S.A.* 92, 7292–7296.
268. Sosnick, T.R., Shtilerman, M.D., Mayne, L., & Englander, S.W. (1997) *Proc. Natl. Acad. Sci. U.S.A.* 94, 8545–8550.
269. Chan, C.K., Hu, Y., Takahashi, S., Rousseau, D.L., Eaton, W.A., & Hofrichter, J. (1997) *Proc. Natl. Acad. Sci. U.S.A.* 94, 1779–1784.
270. Shastry, M.C., & Roder, H. (1998) *Nat. Struct. Biol.* 5, 385–392.
271. Plaxco, K.W., Millett, I.S., Segel, D.J., Doniach, S., & Baker, D. (1999) *Nat. Struct. Biol.* 6, 554–556.
272. Bieri, O., Wirz, J., Hellrung, B., Schutkowski, M., Drewello, M., & Kiefhaber, T. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 9597–9601.
273. Jacob, M., Geeves, M., Holtermann, G., & Schmid, F.X. (1999) *Nat. Struct. Biol.* 6, 923–926.
274. Teilum, K., Kragelund, B.B., Knudsen, J., & Poulsen, F.M. (2000) *J. Mol. Biol.* 301, 1307–1314.
275. Ladurner, A.G., & Fersht, A.R. (1999) *Nat. Struct. Biol.* 6, 28–31.
276. Matouschek, A., Kellis, J.T., Jr., Serrano, L., Bycroft, M., & Fersht, A.R. (1990) *Nature* 346, 440–445.
277. Gassner, N.C., Baase, W.A., Lindstrom, J.D., Lu, J., Dahlquist, F.W., & Matthews, B.W. (1999) *Biochemistry* 38, 14451–14460.
278. Jacobs, M.D., & Fox, R.O. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 449–453.
279. Radford, S.E., Dobson, C.M., & Evans, P.A. (1992) *Nature* 358, 302–307.
280. Andersson, D., Hammarstrom, P., & Carlsson, U. (2001) *Biochemistry* 40, 2653–2661.
281. Xu, Y., Mayne, L., & Englander, S.W. (1998) *Nat. Struct. Biol.* 5, 774–778.
282. Stewart, D.E., Sarkar, A., & Wampler, J.E. (1990) *J. Mol. Biol.* 214, 253–260.
283. MacArthur, M.W., & Thornton, J.M. (1991) *J. Mol. Biol.* 218, 397–412.
284. Garel, J.R., & Baldwin, R.L. (1973) *Proc. Natl. Acad. Sci. U.S.A.* 70, 3347–3351.
285. Garel, J.R., Nall, B.T., & Baldwin, R.L. (1976) *Proc. Natl. Acad. Sci. U.S.A.* 73, 1853–1857.
286. Brandts, J.F., Halvorson, H.R., & Brennan, M. (1975) *Biochemistry* 14, 4953–4963.
287. Lin, L.N., & Brandts, J.F. (1983) *Biochemistry* 22, 559–563.
288. Schmid, F.X., & Baldwin, R.L. (1978) *Proc. Natl. Acad. Sci. U.S.A.* 75, 4764–4768.
289. Schmid, F.X., Grafl, R., Wrba, A., & Beintema, J.J. (1986) *Proc. Natl. Acad. Sci. U.S.A.* 83, 872–876.
290. Hensens, R.W., Gerber, A.D., Cooper, M.R., & Herzog, W.R., Jr. (1980) *J. Biol. Chem.* 255, 7075–7078.
291. Schmid, F.X., & Baldwin, R.L. (1979) *J. Mol. Biol.* 133, 285–287.
292. Schultz, D.A., & Baldwin, R.L. (1992) *Protein Sci.* 1, 910–916.
293. Kim, P.S., & Baldwin, R.L. (1980) *Biochemistry* 19, 6124–6129.
294. Schmid, F.X., & Blaschek, H. (1981) *Eur. J. Biochem.* 114, 111–117.
295. Schmid, F., & Blaschek, H. (1984) *Biochemistry* 23, 2128–2133.

296. Martinez-Oyanedel, J., Choe, H.W., Heinemann, U., & Saenger, W. (1991) *J. Mol. Biol.* 222, 335–352.
297. Mayr, L.M., Odefey, C., Schutkowski, M., & Schmid, F.X. (1996) *Biochemistry* 35, 5550–5561.
298. Mullins, L.S., Pace, C.N., & Raushel, F.M. (1993) *Biochemistry* 32, 6152–6156.
299. Kiefhaber, T., Quaas, R., Hahn, U., & Schmid, F.X. (1990) *Biochemistry* 29, 3061–3070.
300. Kiefhaber, T., Grunert, H.P., Hahn, U., & Schmid, F.X. (1992) *Proteins: Struct., Funct., Genet.* 12, 171–179.
301. Kelley, R.F., & Richards, F.M. (1987) *Biochemistry* 26, 6765–6774.
302. Walkenhorst, W.F., Green, S.M., & Roder, H. (1997) *Biochemistry* 36, 5795–5805.
303. Maki, K., Ikura, T., Hayano, T., Takahashi, N., & Kuwajima, K. (1999) *Biochemistry* 38, 2213–2223.
304. Bilsel, O., Zitzewitz, J.A., Bowers, K.E., & Matthews, C.R. (1999) *Biochemistry* 38, 1018–1029.
305. Jackson, S.E., & Fersht, A.R. (1991) *Biochemistry* 30, 10436–10443.
306. van Nuland, N.A., Chiti, F., Taddei, N., Raugei, G., Ramponi, G., & Dobson, C.M. (1998) *J. Mol. Biol.* 283, 883–891.
307. Mayr, L.M., Landt, O., Hahn, U., & Schmid, F.X. (1993) *J. Mol. Biol.* 231, 897–912.
308. Pappenberger, G., Aygun, H., Engels, J.W., Reimer, U., Fischer, G., & Kiefhaber, T. (2001) *Nat. Struct. Biol.* 8, 452–458.
309. Schiene-Fischer, C., & Fischer, G. (2001) *J. Am. Chem. Soc.* 123, 6227–6231.
310. Krebs, H., Schmid, F.X., & Jaenicke, R. (1983) *J. Mol. Biol.* 169, 619–635.
311. Chazin, W.J., Kordel, J., Drakenberg, T., Thulin, E., Brodin, P., Grundstrom, T., & Forsen, S. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 2195–2198.
312. Fischer, G., Wittmann-Liebold, B., Lang, K., Kiefhaber, T., & Schmid, F.X. (1989) *Nature* 337, 476–478.
313. Takahashi, N., Hayano, T., & Suzuki, M. (1989) *Nature* 337, 473–475.
314. Fischer, G., Bang, H., & Mech, C. (1984) *Biomed. Biochim. Acta* 43, 1101–1111.
315. Fischer, G., & Bang, H. (1985) *Biochim. Biophys. Acta* 828, 39–42.
316. Lang, K., Schmid, F.X., & Fischer, G. (1987) *Nature* 329, 268–270.
317. Lang, K., & Schmid, F.X. (1988) *Nature* 331, 453–455.
318. Beachinger, H.P. (1987) *J. Biol. Chem.* 262, 17144–17148.
319. Siekierka, J.J., Hung, S.H., Poe, M., Lin, C.S., & Sigal, N.H. (1989) *Nature* 341, 755–757.
320. Harding, M.W., Galat, A., Uehling, D.E., & Schreiber, S.L. (1989) *Nature* 341, 758–760.
321. Schonbrunner, E.R., Mayer, S., Tropschug, M., Fischer, G., Takahashi, N., & Schmid, F.X. (1991) *J. Biol. Chem.* 266, 3630–3635.
322. Matouschek, A., Rospert, S., Schmid, K., Glick, B.S., & Schatz, G. (1995) *Proc. Natl. Acad. Sci. U.S.A.* 92, 6319–6323.
323. Stoller, G., Rucknagel, K.P., Nierhaus, K.H., Schmid, F.X., Fischer, G., & Rahfeld, J.U. (1995) *EMBO J.* 14, 4939–4948.
324. Tropschug, M., Nicholson, D.W., Hartl, F.U., Kohler, H., Pfanner, N., Wachter, E., & Neupert, W. (1988) *J. Biol. Chem.* 263, 14433–14440.
325. Gasser, C.S., Gunning, D.A., Budelier, K.A., & Brown, S.M. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87, 9519–9523.
326. Scholz, C., Mucke, M., Rape, M., Pecht, A., Pahl, A., Bang, H., & Schmid, F.X. (1998) *J. Mol. Biol.* 277, 723–732.
327. Mucke, M., & Schmid, F.X. (1992) *Biochemistry* 31, 7848–7854.
328. Veeraraghavan, S., & Nall, B.T. (1994) *Biochemistry* 33, 687–692.
329. Schindler, T., & Schmid, F.X. (1996) *Biochemistry* 35, 16833–16842.
330. Chiti, F., Taddei, N., van Nuland, N.A., Magherini, F., Stefani, M., Ramponi, G., & Dobson, C.M. (1998) *J. Mol. Biol.* 283, 893–903.
331. Otzen, D.E., Kristensen, O., Proctor, M., & Oliveberg, M. (1999) *Biochemistry* 38, 6499–6511.
332. Mayor, U., Johnson, C.M., Daggett, V., & Fersht, A.R. (2000) *Proc. Natl. Acad. Sci. U.S.A.* 97, 13518–13522.
333. Krantz, B.A., & Sosnick, T.R. (2000) *Biochemistry* 39, 11696–11701.
334. Burton, R.E., Huang, G.S., Daugherty, M.A., Fullbright, P.W., & Oas, T.G. (1996) *J. Mol. Biol.* 263, 311–322.
335. Huang, G.S., & Oas, T.G. (1995) *Proc. Natl. Acad. Sci. U.S.A.* 92, 6878–6882.
336. Spector, S., & Raleigh, D.P. (1999) *J. Mol. Biol.* 293, 763–768.
337. Vidugiris, G.J., Markley, J.L., & Royer, C.A. (1995) *Biochemistry* 34, 4909–4912.
338. Jacob, M., Holtermann, G., Perl, D., Reinstein, J., Schindler, T., Gieves, M.A., & Schmid, F.X. (1999) *Biochemistry* 38, 2882–2891.
339. Tan, Y.J., Oliveberg, M., & Fersht, A.R. (1996) *J. Mol. Biol.* 264, 377–389.
340. Taddei, N., Chiti, F., Fiaschi, T., Bucciantini, M., Capanni, C., Stefani, M., Serrano, L., Dobson, C.M., & Ramponi, G. (2000) *J. Mol. Biol.* 300, 633–647.
341. Srivastava, A.K., & Sauer, R.T. (2000) *Biochemistry* 39, 8308–8314.
342. Burton, R.E., Huang, G.S., Daugherty, M.A., Calderone, T.L., & Oas, T.G. (1997) *Nat. Struct. Biol.* 4, 305–310.
343. Otzen, D.E., Itzhaki, L.S., elMasry, N.F., Jackson, S.E., & Fersht, A.R. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 10422–10425.
344. Jager, M., Nguyen, H., Crane, J.C., Kelly, J.W., & Gruebele, M. (2001) *J. Mol. Biol.* 311, 373–393.
345. Lee, J.C., Gray, H.B., & Winkler, J.R. (2001) *Proc. Natl. Acad. Sci. U.S.A.* 98, 7760–7764.
346. Parker, M.J., & Marqusee, S. (2000) *J. Mol. Biol.* 300, 1361–1375.
347. Kiefhaber, T. (1995) *Proc. Natl. Acad. Sci. U.S.A.* 92, 9029–9033.
348. Wildegger, G., & Kiefhaber, T. (1997) *J. Mol. Biol.* 270, 294–304.
349. Matagne, A., Radford, S.E., & Dobson, C.M. (1997) *J. Mol. Biol.* 267, 1068–1074.
350. Jennings, P.A., Finn, B.E., Jones, B.E., & Matthews, C.R. (1993) *Biochemistry* 32, 3783–3789.
351. Iwakura, M., Jones, B.E., Falzone, C.J., & Matthews, C.R. (1993) *Biochemistry* 32, 13566–13574.
352. Cayley, P.J., Dunn, S.M., & King, R.W. (1981) *Biochemistry* 20, 874–879.

353. Ionescu, R.M., Smith, V.F., O'Neill, J.C., Jr., & Matthews, C.R. (2000) *Biochemistry* 39, 9540–9550.
354. Takahashi, S., Yeh, S.R., Das, T.K., Chan, C.K., Gottfried, D.S., & Rousseau, D.L. (1997) *Nat. Struct. Biol.* 4, 44–50.
355. Guidry, J., & Wittung-Stafshede, P. (2000) *J. Mol. Biol.* 301, 769–773.
356. Silow, M., Tan, Y.J., Fersht, A.R., & Oliveberg, M. (1999) *Biochemistry* 38, 13006–13012.
357. Nawrocki, J.P., Chu, R.A., Pannell, L.K., & Bai, Y. (1999) *J. Mol. Biol.* 293, 991–995.
358. Goldberg, M.E., Rudolph, R., & Jaenicke, R. (1991) *Biochemistry* 30, 2790–2797.
359. Vaucheret, H., Signon, L., Le Bras, G., & Garel, J.R. (1987) *Biochemistry* 26, 2785–2790.
360. Pelham, H.R. (1986) *Cell* 46, 959–961.
361. Pelham, H. (1988) *Nature* 332, 776–777.
362. Reading, D.S., Hallberg, R.L., & Myers, A.M. (1989) *Nature* 337, 655–659.
363. Hemmingsen, S.M., Woolford, C., van der Vies, S.M., Tilly, K., Dennis, D.T., Georgopoulos, C.P., Hendrix, R.W., & Ellis, R.J. (1988) *Nature* 333, 330–334.
364. Ostermann, J., Horwich, A.L., Neupert, W., & Hartl, F.U. (1989) *Nature* 341, 125–130.
365. Kubota, H., Hynes, G., & Willison, K. (1995) *Eur. J. Biochem.* 230, 3–16.
366. Braig, K., Otwinowski, Z., Hegde, R., Boisvert, D.C., Joachimiak, A., Horwich, A.L., & Sigler, P.B. (1994) *Nature* 371, 578–586.
367. Ditzel, L., Lowe, J., Stock, D., Stetter, K.O., Huber, H., Huber, R., & Steinbacher, S. (1998) *Cell* 93, 125–138.
368. Llorca, O., Smyth, M.G., Carrascosa, J.L., Willison, K.R., Radermacher, M., Steinbacher, S., & Valpuesta, J.M. (1999) *Nat. Struct. Biol.* 6, 639–642.
369. Wang, J., & Boisvert, D.C. (2003) *J. Mol. Biol.* 327, 843–855.
370. Chen, S., Roseman, A.M., Hunter, A.S., Wood, S.P., Burston, S.G., Ranson, N.A., Clarke, A.R., & Saibil, H.R. (1994) *Nature* 371, 261–264.
371. Chaudhry, C., Farr, G.W., Todd, M.J., Rye, H.S., Brunger, A.T., Adams, P.D., Horwich, A.L., & Sigler, P.B. (2003) *EMBO J.* 22, 4877–4887.
372. Xu, Z., Horwich, A.L., & Sigler, P.B. (1997) *Nature* 388, 741–750.
373. Goloubinoff, P., Christeller, J.T., Gatenby, A.A., & Lorimer, G.H. (1989) *Nature* 342, 884–889.
374. Buchner, J., Schmidt, M., Fuchs, M., Jaenicke, R., Rudolph, R., Schmid, F.X., & Kiefhaber, T. (1991) *Biochemistry* 30, 1586–1591.
375. van der Vies, S.M., Viitanen, P.V., Gatenby, A.A., Lorimer, G.H., & Jaenicke, R. (1992) *Biochemistry* 31, 3635–3644.
376. Fisher, M.T. (1992) *Biochemistry* 31, 3955–3963.
377. Fisher, M.T. (1993) *J. Biol. Chem.* 268, 13777–13779.
378. Itzhaki, L.S., Otzen, D.E., & Fersht, A.R. (1995) *Biochemistry* 34, 14581–14587.
379. Bhutani, N., & Udgaonkar, J.B. (2001) *J. Mol. Biol.* 314, 1167–1179.
380. Goldberg, M.S., Zhang, J., Sondek, S., Matthews, C.R., Fox, R.O., & Horwich, A.L. (1997) *Proc. Natl. Acad. Sci. U.S.A.* 94, 1080–1085.
381. Coyle, J.E., Texter, F.L., Ashcroft, A.E., Masselos, D., Robinson, C.V., & Radford, S.E. (1999) *Nat. Struct. Biol.* 6, 683–690.
382. Zahn, R., Perrett, S., & Fersht, A.R. (1996) *J. Mol. Biol.* 261, 43–61.
383. Zahn, R., Spitzfaden, C., Ottiger, M., Wuthrich, K., & Pluckthun, A. (1994) *Nature* 368, 261–265.
384. Walter, S., Lorimer, G.H., & Schmid, F.X. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 9425–9430.
385. Robinson, C.V., Gross, M., Eyles, S.J., Ewbank, J.J., Mayhew, M., Hartl, F.U., Dobson, C.M., & Radford, S.E. (1994) *Nature* 372, 646–651.
386. Nieba-Axmann, S.E., Ottiger, M., Wuthrich, K., & Pluckthun, A. (1997) *J. Mol. Biol.* 271, 803–818.
387. Chen, J., Walter, S., Horwich, A.L., & Smith, D.L. (2001) *Nat. Struct. Biol.* 8, 721–728.
388. Zahn, R., Perrett, S., Stenberg, G., & Fersht, A.R. (1996) *Science* 271, 642–645.
389. Gervasoni, P., Staudenmann, W., James, P., Gehrig, P., & Pluckthun, A. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 12189–12194.
390. Landry, S.J., & Gierasch, L.M. (1991) *Biochemistry* 30, 7359–7362.
391. Falke, S., Fisher, M.T., & Gogol, E.P. (2001) *J. Mol. Biol.* 308, 569–577.
392. Braig, K., Simon, M., Furuya, F., Hainfeld, J.F., & Horwich, A.L. (1993) *Proc. Natl. Acad. Sci. U.S.A.* 90, 3978–3982.
393. Langer, T., Pfeifer, G., Martin, J., Baumeister, W., & Hartl, F.U. (1992) *EMBO J.* 11, 4757–4765.
394. Chaudhuri, T.K., Farr, G.W., Fenton, W.A., Rospert, S., & Horwich, A.L. (2001) *Cell* 107, 235–246.
395. Sakikawa, C., Taguchi, H., Makino, Y., & Yoshida, M. (1999) *J. Biol. Chem.* 274, 21251–21256.
396. Wang, J.D., Michelitsch, M.D., & Weissman, J.S. (1998) *Proc. Natl. Acad. Sci. U.S.A.* 95, 12163–12168.
397. Zahn, R., Buckle, A.M., Perrett, S., Johnson, C.M., Corrales, F.J., Golbik, R., & Fersht, A.R. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 15024–15029.
398. Buckle, A.M., Zahn, R., & Fersht, A.R. (1997) *Proc. Natl. Acad. Sci. U.S.A.* 94, 3571–3575.
399. Mendoza, J.A., Rogers, E., Lorimer, G.H., & Horowitz, P.M. (1991) *J. Biol. Chem.* 266, 13044–13049.
400. Makio, T., Takasu-Ishikawa, E., & Kuwajima, K. (2001) *J. Mol. Biol.* 312, 555–567.
401. Kubo, T., Mizobata, T., & Kawata, Y. (1993) *J. Biol. Chem.* 268, 19346–19351.
402. Lin, Z., & Eisenstein, E. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 1977–1981.
403. Badcoe, I.G., Smith, C.J., Wood, S., Halsall, D.J., Holbrook, J.J., Lund, P., & Clarke, A.R. (1991) *Biochemistry* 30, 9195–9200.
404. Jackson, G.S., Staniforth, R.A., Halsall, D.J., Atkinson, T., Holbrook, J.J., Clarke, A.R., & Burston, S.G. (1993) *Biochemistry* 32, 2554–2563.
405. Wynn, R.M., Davie, J.R., Zhi, W., Cox, R.P., & Chuang, D.T. (1994) *Biochemistry* 33, 8962–8968.
406. Viitanen, P.V., Donaldson, G.K., Lorimer, G.H., Lubben, T.H., & Gatenby, A.A. (1991) *Biochemistry* 30, 9716–9723.
407. Boisvert, D.C., Wang, J., Otwinowski, Z., Horwich, A.L., & Sigler, P.B. (1996) *Nat. Struct. Biol.* 3, 170–177.

408. Beissinger, M., Rutkat, K., & Buchner, J. (1999) *J. Mol. Biol.* 289, 1075–1092.
409. Todd, M.J., Viitanen, P.V., & Lorimer, G.H. (1994) *Science* 265, 659–666.
410. Ranson, N.A., Dunster, N.J., Burston, S.G., & Clarke, A.R. (1995) *J. Mol. Biol.* 250, 581–586.
411. Hunt, J.F., Weaver, A.J., Landry, S.J., Gierasch, L., & Deisenhofer, J. (1996) *Nature* 379, 37–45.
412. Weissman, J.S., Hohl, C.M., Kovalenko, O., Kashi, Y., Chen, S., Braig, K., Saibil, H.R., Fenton, W.A., & Horwich, A.L. (1995) *Cell* 83, 577–587.
413. Burston, S.G., Ranson, N.A., & Clarke, A.R. (1995) *J. Mol. Biol.* 249, 138–152.
414. Rye, H.S., Burston, S.G., Fenton, W.A., Beechem, J.M., Xu, Z., Sigler, P.B., & Horwich, A.L. (1997) *Nature* 388, 792–798.
415. Shtilerman, M., Lorimer, G.H., & Englander, S.W. (1999) *Science* 284, 822–825.
416. Flynn, G.C., Chappell, T.G., & Rothman, J.E. (1989) *Science* 245, 385–390.
417. Gisler, S.M., Pierpaoli, E.V., & Christen, P. (1998) *J. Mol. Biol.* 279, 833–840.
418. Mayer, M.P., Schroder, H., Rudiger, S., Paal, K., Laufen, T., & Bukau, B. (2000) *Nat. Struct. Biol.* 7, 586–593.
419. McCarty, J.S., Buchberger, A., Reinstein, J., & Bukau, B. (1995) *J. Mol. Biol.* 249, 126–137.
420. Zhu, X., Zhao, X., Burkholder, W.F., Gragerov, A., Ogata, C.M., Gottesman, M.E., & Hendrickson, W.A. (1996) *Science* 272, 1606–1614.
421. Newton, G.L., Arnold, K., Price, M.S., Sherrill, C., Delcardayre, S.B., Aharonowitz, Y., Cohen, G., Davies, J., Fahey, R.C., & Davis, C. (1996) *J. Bacteriol.* 178, 1990–1995.
422. Hwang, C., Sinskey, A.J., & Lodish, H.F. (1992) *Science* 257, 1496–1502.
423. Frech, C., & Schmid, F.X. (1995) *J. Mol. Biol.* 251, 135–149.
424. Creighton, T.E. (1983) in *Functions of Glutathione: Biochemical, Physiological, Toxicological, and Clinical Aspects* (Larsson, A., Ed.) pp 205–222, Raven Press, New York.
425. Walker, K.W., Lyles, M.M., & Gilbert, H.F. (1996) *Biochemistry* 35, 1972–1980.
426. De Lorenzo, F., Goldberger, R.F., Steers, E., Jr., Givol, D., & Anfinsen, B. (1966) *J. Biol. Chem.* 241, 1562–1567.
427. Lyles, M.M., & Gilbert, H.F. (1991) *Biochemistry* 30, 613–619.
428. Narhi, L.O., Hua, Q.X., Arakawa, T., Fox, G.M., Tsai, L., Rosenfeld, R., Holst, P., Miller, J.A., & Weiss, M.A. (1993) *Biochemistry* 32, 5214–5221.
429. Edman, J.C., Ellis, L., Blacher, R.W., Roth, R.A., & Rutter, W.J. (1985) *Nature* 317, 267–270.
430. Bardwell, J.C., McGovern, K., & Beckwith, J. (1991) *Cell* 67, 581–589.
431. Zapun, A., Bardwell, J.C., & Creighton, T.E. (1993) *Biochemistry* 32, 5083–5092.
432. Maskos, K., Huber-Wunderlich, M., & Glockshuber, R. (2003) *J. Mol. Biol.* 325, 495–513.
433. Nakamoto, H., & Bardwell, J.C. (2004) *Biochim. Biophys. Acta* 1694, 111–119.
434. Darby, N.J., Freedman, R.B., & Creighton, T.E. (1994) *Biochemistry* 33, 7937–7947.
435. Ruoppolo, M., & Freedman, R.B. (1995) *Biochemistry* 34, 9380–9388.
436. Gilbert, H.F. (1997) *J. Biol. Chem.* 272, 29399–29402.
437. Quan, H., Fan, G., & Wang, C.C. (1995) *J. Biol. Chem.* 270, 17078–17080.
438. Schonbrunner, E.R., & Schmid, F.X. (1992) *Proc. Natl. Acad. Sci. U.S.A.* 89, 4510–4513.
439. Lundstrom, J., & Holmgren, A. (1990) *J. Biol. Chem.* 265, 9114–9120.
440. Krause, G., Lundstrom, J., Barea, J.L., Pueyo de la Cuesta, C., & Holmgren, A. (1991) *J. Biol. Chem.* 266, 9494–9500.
441. Miranker, A., Radford, S.E., Karplus, M., & Dobson, C.M. (1991) *Nature* 349, 633–636.
442. Teschner, W., Rudolph, R., & Garel, J.R. (1987) *Biochemistry* 26, 2791–2796.
443. Lillo, M.P., Szpikowska, B.K., Mas, M.T., Sutin, J.D., & Beechem, J.M. (1997) *Biochemistry* 36, 11273–11281.
444. Wong, S.C., Burton, P.M., & Josse, J. (1970) *J. Biol. Chem.* 245, 4353–4357.
445. Hermann, R., Rudolph, R., Jaenicke, R., Price, N.C., & Scobbie, A. (1983) *J. Biol. Chem.* 258, 11014–11019.
446. Hermann, R., Jaenicke, R., & Price, N.C. (1985) *Biochemistry* 24, 1817–1821.
447. Mateu, M.G., Sanchez Del Pino, M.M., & Fersht, A.R. (1999) *Nat. Struct. Biol.* 6, 191–198.
448. Jaenicke, R., Rudolph, R., & Heider, I. (1979) *Biochemistry* 18, 1217–1223.
449. Gleason, W.B., Fu, Z., Birktoft, J., & Banaszak, L. (1994) *Biochemistry* 33, 2078–2088.
450. Leistler, B., Herold, M., & Kirschner, K. (1992) *Eur. J. Biochem.* 205, 603–611.
451. Kim, D.H., Jang, D.S., Nam, G.H., Yun, S., Cho, J.H., Choi, G., Lee, H.C., & Choi, K.Y. (2000) *Biochemistry* 39, 13084–13092.
452. Milla, M.E., & Sauer, R.T. (1994) *Biochemistry* 33, 1125–1133.
453. Waldburger, C.D., Jonsson, T., & Sauer, R.T. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 2629–2634.
454. Burns, D.L., & Schachman, H.K. (1982) *J. Biol. Chem.* 257, 8648–8654.
455. Burns, D.L., & Schachman, H.K. (1982) *J. Biol. Chem.* 257, 8638–8647.
456. Yamato, S., & Murachi, T. (1979) *Eur. J. Biochem.* 93, 189–195.
457. Kervinen, J., Dunbrack, R.L., Jr., Litwin, S., Martins, J., Scarrow, R.C., Volin, M., Yeung, A.T., Yoon, E., & Jaffe, E.K. (2000) *Biochemistry* 39, 9018–9029.
458. Zhang, Z.Y., Poorman, R.A., Maggiora, L.L., Heinrikson, R.L., & Kezdy, F.J. (1991) *J. Biol. Chem.* 266, 15591–15594.
459. Vimard, C., Orsini, G., & Goldberg, M.E. (1975) *Eur. J. Biochem.* 51, 521–527.
460. Flynn, G.C., Beckers, C.J., Baase, W.A., & Dahlquist, F.W. (1993) *Proc. Natl. Acad. Sci. U.S.A.* 90, 10826–10830.
461. Lane, A.N., Paul, C.H., & Kirschner, K. (1984) *EMBO J.* 3, 279–287.
462. Hyde, C.C., Ahmed, S.A., Padlan, E.A., Miles, E.W., & Davies, D.R. (1988) *J. Biol. Chem.* 263, 17857–17871.
463. Bothwell, M.A., & Schachman, H.K. (1980) *J. Biol. Chem.* 255, 1962–1970.

464. Yang, Y.R., Syvanen, J.M., Nagel, G.M., & Schachman, H.K. (1974) *Proc. Natl. Acad. Sci. U.S.A.* 71, 918–923.
465. Jacobson, G.R., & Stark, G.R. (1973) *J. Biol. Chem.* 248, 8003–8014.
466. Bothwell, M.A., & Schachman, H.K. (1980) *J. Biol. Chem.* 255, 1971–1977.
467. Bates, D.L., Danson, M.J., Hale, G., Hooper, E.A., & Perham, R.N. (1977) *Nature* 268, 313–316.
468. Wagenknecht, T., Francis, N., & DeRosier, D.J. (1983) *J. Mol. Biol.* 165, 523–539.
469. Reed, L.J., Pettit, F.H., Eley, M.H., Hamilton, L., Collins, J.H., & Oliver, R.M. (1975) *Proc. Natl. Acad. Sci. U.S.A.* 72, 3068–3072.
470. DeRosier, D.J., & Oliver, R.M. (1971) *Cold Spring Harbor Symp. Quant. Biol.* 36, 199–203.
471. Brosius, J., Palmer, M.L., Kennedy, P.J., & Noller, H.F. (1978) *Proc. Natl. Acad. Sci. U.S.A.* 75, 4801–4805.
472. Held, W.A., Mizushima, S., & Nomura, M. (1973) *J. Biol. Chem.* 248, 5720–5730.
473. Held, W.A., Ballou, B., Mizushima, S., & Nomura, M. (1974) *J. Biol. Chem.* 249, 3103–3111.
474. Herold, M., & Nierhaus, K.H. (1987) *J. Biol. Chem.* 262, 8826–8833.
475. Wimberly, B.T., Brodersen, D.E., Clemons, W.M., Jr., Morgan-Warren, R.J., Carter, A.P., Vornrhein, C., Hartsch, T., & Ramakrishnan, V. (2000) *Nature* 407, 327–339.
476. Morrison, C.A., Garrett, R.A., & Bradbury, E.M. (1977) *Eur. J. Biochem.* 78, 153–159.
477. Rohde, M.F., O'Brien, S., Cooper, S., & Aune, K.C. (1975) *Biochemistry* 14, 1079–1087.
478. Franz, A., Georgalis, Y., & Giri, L. (1979) *Biochim. Biophys. Acta* 578, 365–371.
479. Doolittle, R.F. (1984) *Annu. Rev. Biochem.* 53, 195–229.
480. Williams, R.C. (1981) *J. Mol. Biol.* 150, 399–408.
481. Yang, Z., Kollman, J.M., Pandi, L., & Doolittle, R.F. (2001) *Biochemistry* 40, 12515–12523.
482. Brown, J.H., Volkmann, N., Jun, G., Henschen-Edman, A.H., & Cohen, C. (2000) *Proc. Natl. Acad. Sci. U.S.A.* 97, 85–90.
483. Laudano, A.P., & Doolittle, R.F. (1980) *Biochemistry* 19, 1013–1019.
484. Laudano, A.P., & Doolittle, R.F. (1978) *Proc. Natl. Acad. Sci. U.S.A.* 75, 3085–3089.
485. Spraggon, G., Everse, S.J., & Doolittle, R.F. (1997) *Nature* 389, 455–462.
486. Everse, S.J., Spraggon, G., Veerapandian, L., Riley, M., & Doolittle, R.F. (1998) *Biochemistry* 37, 8637–8642.
487. Madrazo, J., Brown, J.H., Litvinovich, S., Dominguez, R., Yakovlev, S., Medved, L., & Cohen, C. (2001) *Proc. Natl. Acad. Sci. U.S.A.* 98, 11967–11972.
488. Hantgan, R.R., & Hermans, J. (1979) *J. Biol. Chem.* 254, 11272–11281.
489. Chen, R., & Doolittle, R.F. (1971) *Biochemistry* 10, 4487–4491.
490. Doolittle, R.F., Cassman, K.G., Cottrell, B.A., & Friezner, S.J. (1977) *Biochemistry* 16, 1715–1719.
491. Amos, L., & Klug, A. (1974) *J. Cell Sci.* 14, 523–549.
492. Wade, R.H., Chretien, D., & Job, D. (1990) *J. Mol. Biol.* 212, 775–786.
493. Kirschner, M.W., Williams, R.C., Weingarten, M., & Gerhart, J.C. (1974) *Proc. Natl. Acad. Sci. U.S.A.* 71, 1159–1163.
494. Feit, H., Slusarek, L., & Shelanski, M.L. (1971) *Proc. Natl. Acad. Sci. U.S.A.* 68, 2028–2031.
495. Ludueana, R.F., Shooter, E.M., & Wilson, L. (1977) *J. Biol. Chem.* 252, 7006–7014.
496. Ponstingl, H., Krauhs, E., Little, M., & Kempf, T. (1981) *Proc. Natl. Acad. Sci. U.S.A.* 78, 2757–2761.
497. Krauhs, E., Little, M., Kempf, T., Hofer-Warbinek, R., Ade, W., & Ponstingl, H. (1981) *Proc. Natl. Acad. Sci. U.S.A.* 78, 4156–4160.
498. Lowe, J., Li, H., Downing, K.H., & Nogales, E. (2001) *J. Mol. Biol.* 313, 1045–1057.
499. Nogales, E., Wolf, S.G., & Downing, K.H. (1998) *Nature* 391, 199–203.
500. Nettles, J.H., Li, H., Cornett, B., Krahn, J.M., Snyder, J.P., & Downing, K.H. (2004) *Science* 305, 866–869.
501. Bergen, L.G., & Borisy, G.G. (1980) *J. Cell Biol.* 84, 141–150.
502. Mandelkow, E.M., Schultheiss, R., Rapp, R., Muller, M., & Mandelkow, E. (1986) *J. Cell Biol.* 102, 1067–1073.
503. Song, Y.H., & Mandelkow, E. (1993) *Proc. Natl. Acad. Sci. U.S.A.* 90, 1671–1675.
504. Kikkawa, M., Ishikawa, T., Nakata, T., Wakabayashi, T., & Hirokawa, N. (1994) *J. Cell Biol.* 127, 1965–1971.
505. Johnson, K.A., & Borisy, G.G. (1977) *J. Mol. Biol.* 117, 1–31.
506. Kirschner, M.W., Honig, L.S., & Williams, R.C. (1975) *J. Mol. Biol.* 99, 263–276.
507. Scheele, R.B., & Borisy, G.G. (1978) *J. Biol. Chem.* 253, 2846–2851.
508. Fygenson, D.K., Braun, E., & Libchaber, A. (1994) *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* 50, 1579–1588.
509. Caudron, N., Valiron, O., Usson, Y., Valiron, P., & Job, D. (2000) *J. Mol. Biol.* 297, 211–220.
510. Mitchison, T., & Kirschner, M. (1984) *Nature* 312, 232–237.
511. Moritz, M., Braunfeld, M.B., Sedat, J.W., Alberts, B., & Agard, D.A. (1995) *Nature* 378, 638–640.
512. Bergen, L.G., Kuriyama, R., & Borisy, G.G. (1980) *J. Cell Biol.* 84, 151–159.
513. Koshland, D.E., Mitchison, T.J., & Kirschner, M.W. (1988) *Nature* 331, 499–504.
514. Mitchison, T., & Kirschner, M. (1984) *Nature* 312, 237–242.
515. Huecas, S., & Andreu, J.M. (2004) *FEBS Lett.* 569, 43–48.
516. Mitchison, T.J. (1993) *Science* 261, 1044–1047.
517. Hirose, K., Fan, J., & Amos, L.A. (1995) *J. Mol. Biol.* 251, 329–333.
518. Fan, J., Griffiths, A.D., Lockhart, A., Cross, R.A., & Amos, L.A. (1996) *J. Mol. Biol.* 259, 325–330.
519. Weisenberg, R.C. (1972) *Science* 177, 1104–1105.
520. Weisenberg, R.C., Borisy, G.G., & Taylor, E.W. (1968) *Biochemistry* 7, 4466–4479.
521. Desai, A., & Mitchison, T.J. (1998) *Bioessays* 20, 523–527.
522. Weisenberg, R.C., Deery, W.J., & Dickinson, P.J. (1976) *Biochemistry* 15, 4248–4254.
523. Melki, R., Carlier, M.F., & Pantaloni, D. (1990) *Biochemistry* 29, 8921–8932.
524. Margolis, R.L. (1981) *Proc. Natl. Acad. Sci. U.S.A.* 78, 1586–1590.

525. Karr, T.L., Podrasky, A.E., & Purich, D.L. (1979) *Proc. Natl. Acad. Sci. U.S.A.* 76, 5475–5479.
526. Jameson, L., & Caplow, M. (1980) *J. Biol. Chem.* 255, 2284–2292.
527. Vandecandelaere, A., Martin, S.R., & Bayley, P.M. (1995) *Biochemistry* 34, 1332–1343.
528. Hyman, A.A., Salser, S., Drechsel, D.N., Unwin, N., & Mitchison, T.J. (1992) *Mol. Biol. Cell* 3, 1155–1167.
529. Karr, T.L., Kristofferson, D., & Purich, D.L. (1980) *J. Biol. Chem.* 255, 8560–8566.
530. Dye, R.B., & Williams, R.C., Jr. (1996) *Biochemistry* 35, 14331–14339.
531. Caplow, M., & Shanks, J. (1995) *Biochemistry* 34, 15732–15741.
532. Margolis, R.L., & Wilson, L. (1977) *Proc. Natl. Acad. Sci. U.S.A.* 74, 3466–3470.
533. Panda, D., Daijo, J.E., Jordan, M.A., & Wilson, L. (1995) *Biochemistry* 34, 9921–9929.
534. Chretien, D., Fuller, S.D., & Karsenti, E. (1995) *J. Cell Biol.* 129, 1311–1328.
535. Woodrum, D.T., Rich, S.A., & Pollard, T.D. (1975) *J. Cell Biol.* 67, 231–237.
536. Pollard, T.D., & Mooseker, M.S. (1981) *J. Cell Biol.* 88, 654–659.
537. Straub, F.B., & Feuer, G. (1950) *Biochim. Biophys. Acta* 4, 455–470.
538. Pieper, U., & Wegner, A. (1996) *Biochemistry* 35, 4396–4402.
539. Carlier, M.F., Pantaloni, D., & Korn, E.D. (1984) *J. Biol. Chem.* 259, 9983–9986.
540. Lal, A.A., Brenner, S.L., & Korn, E.D. (1984) *J. Biol. Chem.* 259, 13061–13065.
541. Pollard, T.D. (1984) *J. Cell Biol.* 99, 769–777.
542. Weber, A., Northrop, J., Bishop, M.F., Ferrone, F.A., & Mooseker, M.S. (1987) *Biochemistry* 26, 2537–2544.
543. Otterbein, L.R., Graceffa, P., & Dominguez, R. (2001) *Science* 293, 708–711.
544. Walsh, T.P., Weber, A., Higgins, J., Bonder, E.M., & Mooseker, M.S. (1984) *Biochemistry* 23, 2613–2621.
545. Robinson, R.C., Turbedsky, K., Kaiser, D.A., Marchand, J.B., Higgs, H.N., Choe, S., & Pollard, T.D. (2001) *Science* 294, 1679–1684.
546. Mooseker, M.S., Graves, T.A., Wharton, K.A., Falco, N., & Howe, C.L. (1980) *J. Cell Biol.* 87, 809–822.
547. Glenney, J.R., Jr., Kaulfus, P., & Weber, K. (1981) *Cell* 24, 471–480.
548. Weber, A., Northrop, J., Bishop, M.F., Ferrone, F.A., & Mooseker, M.S. (1987) *Biochemistry* 26, 2528–2536.
549. Hasegawa, T., Takahashi, S., Hayashi, H., & Hatano, S. (1980) *Biochemistry* 19, 2677–2683.
550. Bryan, J., & Coluccio, L.M. (1985) *J. Cell Biol.* 101, 1236–1244.
551. Caldwell, J.E., Heiss, S.G., Mermall, V., & Cooper, J.A. (1989) *Biochemistry* 28, 8506–8514.
552. Selden, L.A., Kinosian, H.J., Estes, J.E., & Gershman, L.C. (2000) *Biochemistry* 39, 64–74.
553. Gibbons, I.R., & Fronk, E. (1979) *J. Biol. Chem.* 254, 187–196.
554. Paschal, B.M., Shpetner, H.S., & Vallee, R.B. (1987) *J. Cell Biol.* 105, 1273–1282.
555. Paschal, B.M., & Vallee, R.B. (1987) *Nature* 330, 181–183.
556. Vale, R.D., Reese, T.S., & Sheetz, M.P. (1985) *Cell* 42, 39–50.
557. Yin, H.L., Hartwig, J.H., Maruyama, K., & Stossel, T.P. (1981) *J. Biol. Chem.* 256, 9693–9697.
558. Glenney, J.R., Jr., Kaulfus, P., Matsudaira, P., & Weber, K. (1981) *J. Biol. Chem.* 256, 9283–9288.
559. Bretscher, A., & Weber, K. (1979) *Proc. Natl. Acad. Sci. U.S.A.* 76, 2321–2325.
560. Yonezawa, N., Nishida, E., Iida, K., Yahara, I., & Sakai, H. (1990) *J. Biol. Chem.* 265, 8382–8386.
561. Safer, D., Elzinga, M., & Nachmias, V.T. (1991) *J. Biol. Chem.* 266, 4029–4032.
562. Isenberg, G., Aeby, U., & Pollard, T.D. (1980) *Nature* 288, 455–459.
563. Kilimann, M.W., & Isenberg, G. (1982) *EMBO J.* 1, 889–894.
564. Casella, J.F., Maack, D.J., & Lin, S. (1986) *J. Biol. Chem.* 261, 10915–10921.
565. Kuhlman, P.A., & Fowler, V.M. (1997) *Biochemistry* 36, 13461–13472.
566. Maun, N.A., Speicher, D.W., DiNubile, M.J., & Southwick, F.S. (1996) *Biochemistry* 35, 3518–3524.
567. Casella, J.F., Craig, S.W., Maack, D.J., & Brown, A.E. (1987) *J. Cell Biol.* 105, 371–379.
568. Wang, K., & Wright, J. (1988) *J. Cell Biol.* 107, 2199–2212.
569. Kruger, M., Wright, J., & Wang, K. (1991) *J. Cell Biol.* 115, 97–107.
570. Labeit, S., & Kolmerer, B. (1995) *J. Mol. Biol.* 248, 308–315.
571. Jin, J.P., & Wang, K. (1991) *J. Biol. Chem.* 266, 21215–21223.
572. Geiger, B., Tokuyasu, K.T., Dutton, A.H., & Singer, S.J. (1980) *Proc. Natl. Acad. Sci. U.S.A.* 77, 4127–4131.
573. Small, J.V. (1985) *EMBO J.* 4, 45–49.
574. Gregorio, C.C., Weber, A., Bondad, M., Pennise, C.R., & Fowler, V.M. (1995) *Nature* 377, 83–86.
575. Elliott, A., & Offer, G. (1978) *J. Mol. Biol.* 123, 505–519.
576. Slayter, H.S., & Lowey, S. (1967) *Proc. Natl. Acad. Sci. U.S.A.* 58, 1611–1618.
577. Vibert, P., & Craig, R. (1983) *J. Mol. Biol.* 165, 303–320.
578. Huxley, H.E. (1969) *Science* 164, 1356–1365.
579. Huxley, H.E. (1963) *J. Mol. Biol.* 7, 281–308.
580. Knight, P.J., Erickson, M.A., Rodgers, M.E., Beer, M., & Wiggins, J.W. (1986) *J. Mol. Biol.* 189, 167–177.
581. Craig, R., Padron, R., & Alamo, L. (1991) *J. Mol. Biol.* 220, 125–132.
582. Morimoto, K., & Harrington, W.F. (1974) *J. Mol. Biol.* 83, 83–97.
583. Hayashi, T., Silver, R.B., Ip, W., Cayer, M.L., & Smith, D.S. (1977) *J. Mol. Biol.* 111, 159–171.
584. Morimoto, K., & Harrington, W.F. (1973) *J. Mol. Biol.* 77, 165–175.
585. Sinard, J.H., & Pollard, T.D. (1990) *J. Biol. Chem.* 265, 3654–3660.
586. Katsura, I., & Noda, H. (1971) *J. Biochem. (Tokyo)* 69, 219–229.
587. Davis, J.S. (1985) *Biochemistry* 24, 5263–5269.
588. Osapay, K., Theriault, Y., Wright, P.E., & Case, D.A. (1994) *J. Mol. Biol.* 244, 183–197.
589. Lowe, J., & Amos, L.A. (1998) *Nature* 391, 203–206.
590. Erickson, H.P. (1995) *Cell* 80, 367–370.
591. Mukherjee, A., & Lutkenhaus, J. (1994) *J. Bacteriol.* 176, 2754–2758.

Chapter 14

Membranes

Implicit in the cellular theory is the existence of a boundary between the cytoplasm of a cell and its surroundings, be they seawater or the extracellular fluid in a highly organized tissue. The boundary is a defined physical structure known as the **plasma membrane**, and it is a thin, continuous, closed bag marking the boundary of the cell. In electron micrographs of thin sections through a cell, the plasma membrane appears as a continuous closed curve that designates the perimeter of the cytoplasm.

In many microorganisms, such as algae, fungi, and bacteria, the plasma membrane is surrounded on its outer surface by an **outer membrane** or a **cell wall**. The cells of higher plants are also surrounded by a thick cell wall. Usually, the cells of animals, when they are located in organized tissues, are surrounded by networks of collagen and mucopolysaccharide. All of these integuments encasing these various cells are tough polymeric materials that provide support and security for the plasma membrane, which is the formal boundary between the cytoplasm and the environment, between the living and the inert.

When a thin section of a eukaryotic cell is examined in the electron microscope, the most striking feature of the image is the collection of closed curves that represent systems of **intracellular membranes** cut in cross section. These intracellular membranes are the endoplasmic reticulum, the Golgi membranes, and the membranes of the mitochondria, the nucleus, the lysosomes, the peroxisomes, the chloroplasts, the endosomes, and the vacuoles of the cell. Each of these structures at any instant is a closed, continuous, often highly irregular bag enclosing its respective volume of fluid, which is isolated by its membrane from the cytoplasm of the cell. The aqueous solution of proteins found within any one of these bags is unique from that in the cytoplasm surrounding it and is characteristic of the particular organelle. The membranes creating each of these organelles have the same structure as the plasma membrane, although each is distinct in its chemical composition. Because almost every membrane in the cell separates the cytoplasm from another space, the two sides of a membrane are defined relative to the cytoplasm, and they are referred to as **cytoplasmic** and **extracytoplasmic**, respectively. Among the more interesting ambiguities in this situation are the inner membranes of the mitochondria and the thylakoids of chloroplasts, which are the descendants of smaller cells that were incorporated into larger cells.

Consequently, their extracytoplasmic volumes were at one time cytoplasm, of which vestiges remain.

Each of the membranes composing a cell can be purified from the homogenate of a eukaryotic tissue by **cell fractionation**.¹ Originally these purifications were performed in water,¹ but it was subsequently noted that the organelles retained their appearance more successfully in concentrated solutions of sucrose.² Tissues are usually homogenized in a solution of 0.25 M sucrose, and the membranes are isolated by centrifugation on gradients composed of solutions of increasing concentrations of sucrose, which is a solute that stabilizes proteins by salting in, or of other solutes that change the density or osmolarity of the solution. During homogenization, the mitochondria, lysosomes, peroxisomes, chloroplasts, nuclei, and Golgi membranes remain intact and can be identified by their characteristic morphology.^{3,4} Plasma membranes and endoplasmic reticulum are disintegrated by the homogenization and become rounded fragments, often goblet-shaped or vesicular in morphology, and these fragments are known as **microsomes**.⁵ Microsomes of rough endoplasmic reticulum are readily identified by their adherent ribosomes.³

The various membranes and intact organelles suspended in the homogenate of a eukaryotic cell differ in their size and shape, their ratio of protein, lipid, and carbohydrate, and their composition of fixed acid-bases. Therefore, they can be separated from each other by differences in their **sedimentation coefficients** and their **buoyant densities** (Figure 14-1)⁶ or their net charges at a particular pH. Homogenates are often submitted to sequential centrifugations at different centrifugal forces for different durations. Such **differential centrifugations**¹ separate only crudely on the basis of sedimentation coefficient because large differences in sedimentation coefficient between any two organelles are necessary if one of the organelles is to form a pellet exclusively at one centrifugal force and duration while the other forms a pellet exclusively at a higher force or longer duration. **Rate sedimentation**⁷ is a technique in which a narrow band of sample is layered onto a gradient that changes only gradually in sucrose concentration, and the components are separated owing to their differences in sedimentation coefficient as they move through the gradient under the influence of a centrifugal field. Rate sedimentation provides much higher resolution than differential centrifugation.

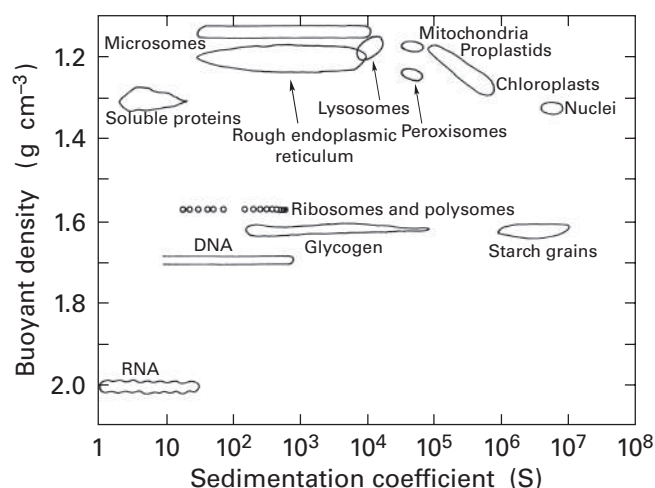


Figure 14-1: Buoyant density and sedimentation coefficient of the organelles and fragments of membrane found in a homogenate of a eukaryotic cell.⁶ Each of the boundaries surrounds the distribution of buoyant densities (grams centimeter⁻³) and sedimentation coefficients of the particular organelle or fragment of membrane. Uniform organelles such as mitochondria, lysosomes, nuclei, and peroxisomes have relatively tight distributions of buoyant density and sedimentation coefficient. Smooth endoplasmic reticulum and plasma membrane become microsomes upon homogenization. Microsomes and rough endoplasmic reticulum, because they are heterogeneous fragments of membrane broken from much larger continuous structures, have fairly uniform buoyant densities but a wide range of sedimentation coefficients. The diagram illustrates that each of these classes of particles occupies a unique region in the two-dimensional space, and this allows each class of membranes to be isolated by a method exploiting differences in sedimentation coefficient in combination with a method exploiting differences in buoyant density. Usually, the boundaries for microsomes of plasma membrane and smooth endoplasmic reticulum coincide, so these two types of membrane fragments are difficult to separate. Adapted with permission from ref 6. Copyright 1974 Academic Press.

Unfortunately, the population of a given organelle in a tissue usually has a significant variation (Figure 14-1)⁶ in its sedimentation coefficient, and additional steps that separate membranes by other independent properties are often required for complete purification. During **isopycnic centrifugation**⁴ the sample is layered onto a much steeper gradient of density formed usually either by varying sucrose concentration or by varying Ficoll concentration.⁸ Ficoll is a polysaccharide that does not have the high osmolarity of sucrose. The gradients are submitted to centrifugation until all of the components have traveled to their respective buoyant densities, at which point they cease to move. The various membranes suspended in a homogenate can also be separated on the basis of their charge by **free-flow electrophoresis**.⁹ In certain instances, where the fixed anionic functional groups on a particular type of membrane are properly oriented, these membranes can be precipitated exclusively with divalent cations such as magnesium or calcium.¹⁰ Such a **precipitation** has been used to separate microsomes derived from endoplasmic reticulum from microsomes derived from plasma membrane.¹¹

During purification, the various classes of membranes can be followed by assaying for particular **marker enzymes**. Each type of organelle has at least one enzymatic activity that is almost exclusively confined to it and can be used as a measure of its concentration.^{3,4,12,13} The ability of certain membranes to bind very specific ligands has also been used to follow their purification. For example, the plasma membranes of animal cells bind the protein wheat germ agglutinin, the peptide hormone insulin, and the toxin from *Vibrio cholerae* with high specificity; the binding of any one of these three ligands can be used to identify plasma membranes.¹⁴ The final identification of the purified suspension of membranes as the organelle of interest, however, must always be made by examining thin sections of pellets of the purified material by **electron microscopy**.

All of these procedures have been used to develop methods for the purification of **mitochondria**,⁴ **peroxisomes**,⁴ **lysosomes**,⁴ **Golgi membranes**,¹³ **rough endoplasmic reticulum**,¹⁵ and **chloroplasts**.¹⁶ Ironically, it is the plasma membrane of most cells that is the most difficult membrane to purify because upon homogenization it fragments and becomes very similar to the much more abundant fragments of smooth endoplasmic reticulum. For this reason, the plasma membrane of the erythrocyte, a cell lacking any other membranes, has often been used as a model for an animal plasma membrane. Plasma membranes, however, have been purified from a number of animal tissues, including liver,¹⁷ kidney,⁹ adipose tissue,⁸ and brain,¹⁸ and from cells grown in tissue culture, such as murine fibroblasts (L-cells),¹⁹ A431 cells,¹¹ and HeLa cells.²⁰ Plasma membranes from fungi²¹ or bacteria, such as *Escherichia coli*,²² are prepared from spheroplasts, which are individual cells that have been enzymatically stripped of their outer membranes or cell walls. The spheroplasts are ruptured and the smooth plasma membranes are isolated from the homogenate.

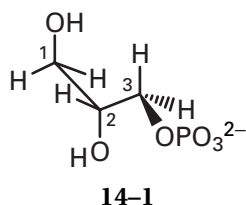
Any of the membranes comprising an organelle or derived from a larger cellular structure can be freed from the soluble proteins it encloses by lysis and sedimentation. Such a membrane is constituted from lipids, carbohydrate, and proteins. All of the carbohydrate is covalently attached to protein in the form of glycoprotein or to lipid in the form of glycolipid. The component lipids, glycolipids, proteins, and glycoproteins are both heterogeneous mixtures, and the fraction of the mass that is protein can vary up to 75%. The basic structure upon which biological membranes are based is a bilayer of amphipathic lipids in which some neutral lipid is dissolved.

Suggested Reading

Price, C.A. (1974) Plant cell fractionation, *Methods Enzymol.* 31, 501-519.

The Bilayer

The basic structural element of a biological membrane is a bilayer of **phospholipids** (Figure 14-2). Bilayers can be formed from a wide variety of lipids, but the plasma membrane of the typical eukaryotic cell is formed from phospholipids and cholesterol. The most prevalent phospholipids are **phosphatidylcholine** (Figure 14-2A) and **phosphatidylethanolamine** (Figure 14-2B). Phosphatidylcholine and phosphatidylethanolamine are **glycerophospholipids** constructed from a molecule of *sn*-glycerol 3-phosphate



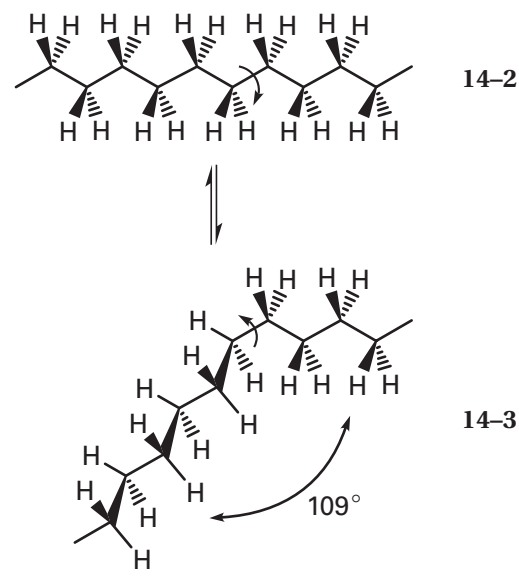
in which carbon 2 has the *R* configuration. Either trimethylethanolammonium (choline; Figure 14-2A) or ethanolamine (Figure 14-2B), respectively, is esterified to the phosphate, forming a phosphate diester. The two remaining hydroxyls of the glycerol 3-phosphate are acylated with a pair of fatty acids.

Phosphatidylcholine and phosphatidylethanolamine are amphipathic lipids. An **amphipathic lipid** is an elongated molecule, usually of biological origin, that is composed of a substantial portion of unadulterated hydrocarbon at one of its ends and a portion of hydrophilic functional groups at its other end. One end of each molecule of phosphatidylcholine or phosphatidylethanolamine, the end containing the phosphate diester and the choline or ethanolamine, is hydrophilic. The other end, the end containing the linear hydrocarbon of the fatty acids, is hydrophobic. Linear hydrocarbons are the most hydrophobic functional groups among biological molecules (Table 5-9).²³

Most glycerophospholipids have two fatty acids attached to the glycerol in **ester linkages**. The fatty acids found in naturally occurring phospholipids seem to be chosen almost at random from the mixture of fatty *S*-acylcoenzymes A produced by the particular organism in which the membrane is located. Saturated fatty acids are esterified mainly to carbon 1 of the *sn*-glycerol 3-phosphate, and unsaturated fatty acids are esterified mainly to carbon 2 (Figure 14-2), so that the naturally occurring phospholipids end up with a roughly equal ratio of saturated and unsaturated hydrocarbons, inescapably intermixed.

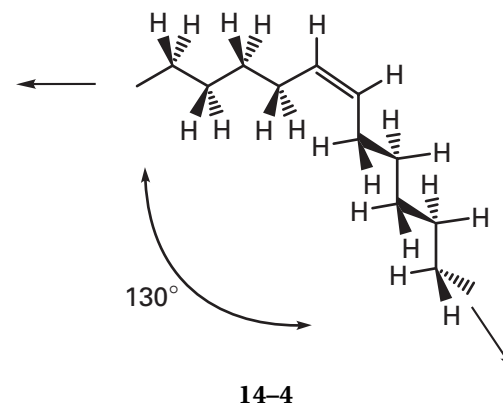
The **saturated fatty acids** are mainly linear carboxylic acids that vary in length from 12 to 24 carbons. The most frequently encountered saturated fatty acids in biological membranes are palmitic acid (*n*-hexadecanoic acid; Figure 14-2C,F) and stearic acid (*n*-octadecanoic acid; Figure 14-2B). The most stable conformation of a

linear hydrocarbon is all-*trans* (14-2), but the introduction of a *gauche* (14-3) conformation at one of the carbon-carbon bonds requires only about 5 kJ mol⁻¹ if the hydrocarbon is unhindered. Therefore, at 25 °C about 10% of the unhindered carbon-carbon single bonds in a saturated fatty acid should be *gauche*. The *gauche* conformation transiently introduces an elbow with an angle of about 109° into the chain:



A second *gauche* conformation can reorient the chain to its original direction.

Unsaturated fatty acids contain one or more carbon-carbon double bonds. A *trans* double bond would not put an elbow in the hydrocarbon, but almost every carbon-carbon double bond in naturally occurring fatty acids is *cis*. One *cis* double bond introduces a permanent elbow with an angle of about 120-130° into an otherwise unsaturated linear hydrocarbon:



In the most stable conformation, one of the hydrogens on one of the two methylenes adjacent to the double bond fits between the two hydrogens of the other. The most common unsaturated fatty acids with only one carbon-carbon double bond are palmitoleic acid (*cis*-hexadec-9-enoic acid; Figure 14-2F) and oleic acid (*cis*-

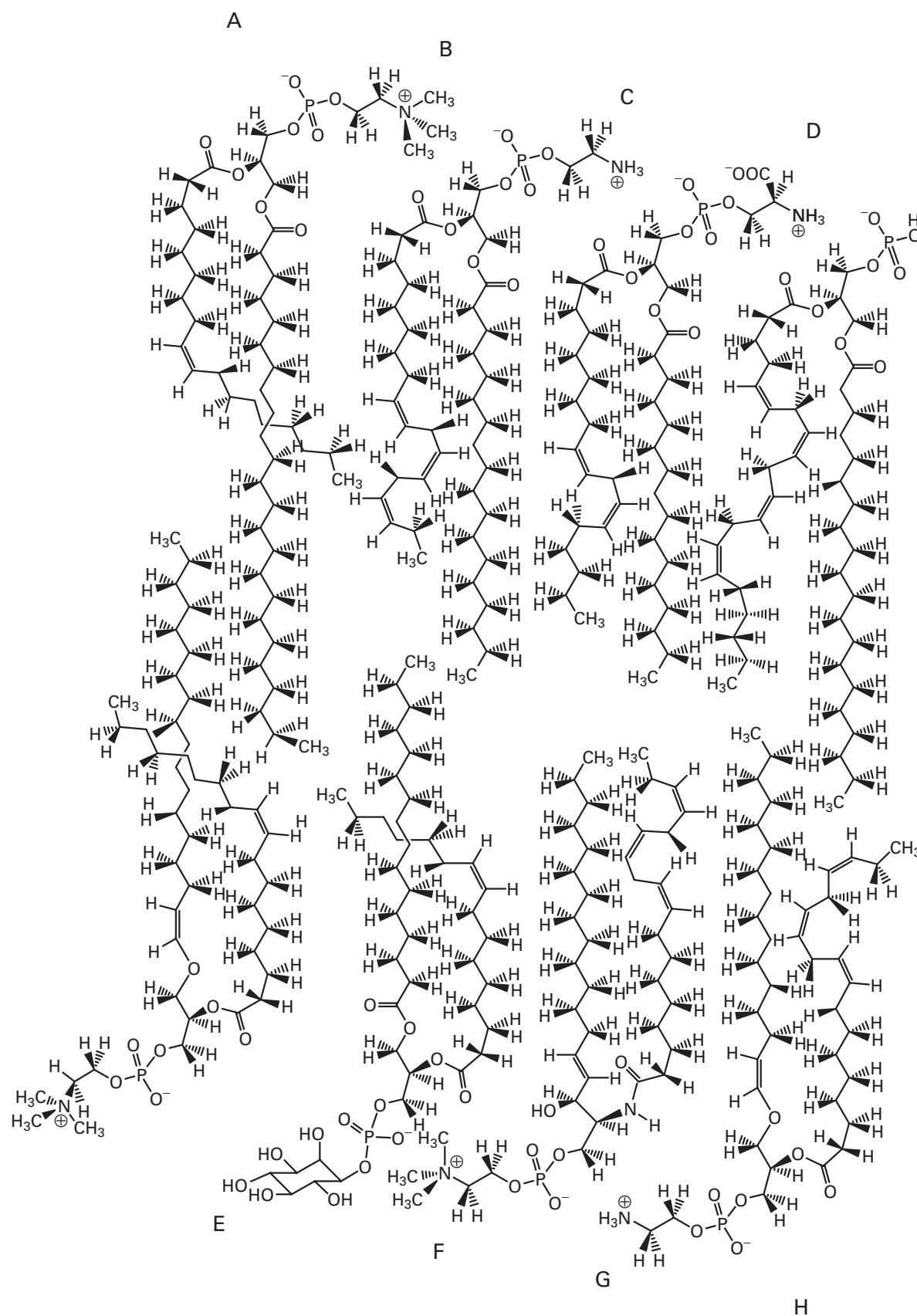
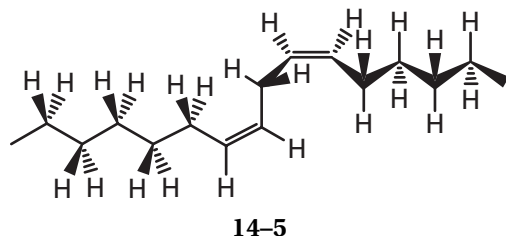


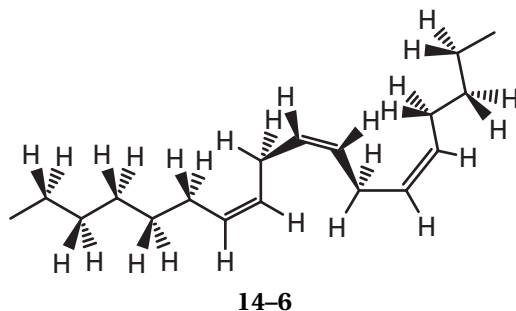
Figure 14-2: Representatives of the types of amphipathic lipids found in the bilayers of biological membranes, drawn as if they were in a bilayer. (A) 1-Lignoceroyl-2-oleoylphosphatidylcholine. (B) 1-Stearoyl-2- α -linolenoylphosphatidylethanolamine. (C) 1-Palmitoyl-2-linoleoylphosphatidylserine. (D) 1-Arachidoyl-2-arachidonoylphosphatidic acid. (E) 1-Octadec-1'-enyl-2-oleoylglycero-3-phosphocholine (a plasmalogen). (F) 1-Palmitoyl-2-palmitoleoylphosphatidylinositol. (G) *N*- α -Linolenoylsphingosine-1-phosphocholine (a sphingomyelin). (H) 1-Hexadec-1'-enyl-2- α -linolenoyl-3-phosphoethanolamine (a plasmalogen).

octadec-9-enoic acid; Figure 14-2A,E), and the latter predominates.

When an unsaturated fatty acid contains two or three double bonds, they are spaced three carbons apart with a saturated carbon between them. This prevents the double bonds from conjugating with each other and becoming a rigid planar structure. Two *cis* double bonds spaced in this way produce a sinuous curve in the alkyl chain but do not change its ultimate direction:

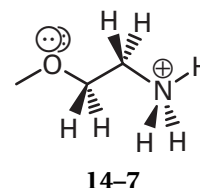


They do, however, shorten its ultimate length by the equivalent of two carbon atoms, and the volume lost at the end is expressed as a bulge at the location of the unsaturation. The most common unsaturated fatty acid with two double bonds is linoleic acid (*cis,cis*-octadeca-9,12-dienoic acid; Figure 14-2C). Three *cis* double bonds produce an even longer sinuous curve that does place a permanent elbow in the alkyl chain:



The most common **polyunsaturated fatty acid** with three carbon-carbon double bonds is linolenic acid in its two geometric isomers, α -linolenic acid (*cis,cis,cis*-octadeca-9,12,15-trienoic acid; Figure 14-2B,G,H) and γ -linolenic acid (*cis,cis,cis*-octadeca-6,9,12-trienoic acid). Arachidonic acid (*cis,cis,cis,cis*-icosa-5,8,11,14-tetraenoic acid; Figure 14-2D) is a less common polyunsaturated fatty acid that serves as the precursor to prostaglandins, and its frequency is regulated for that purpose. Almost all of the unsaturation commences at carbon 9 in the usual naturally occurring mixture of the various fatty acids, so the portion of the hydrocarbon closest to the glyceryl group in a phospholipid is fully saturated and that farthest away contains unsaturated positions and is more geometrically disordered (Figure 14-2).

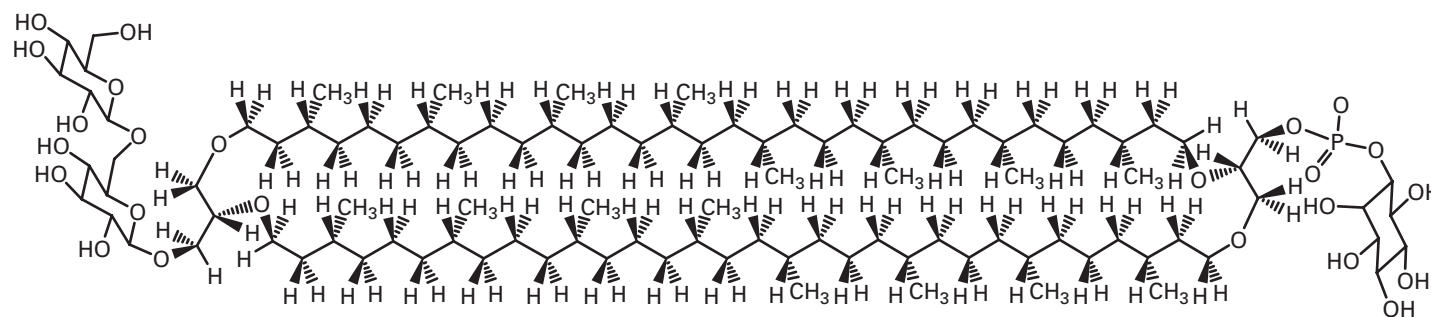
The **head groups** are the polar alcohols esterified to the phosphoric acid in naturally occurring phospholipids. The majority of the head groups are based on ethanolamine:



In phosphatidylethanolamine (Figure 14-2B), the ethanolamine is unaltered. In phosphatidylcholine (Figure 14-2A), the ethanolamine is triply methylated on nitrogen. In **phosphatidylserine** (Figure 14-2C), the ethanolamine is carboxylated. Two glycerophospholipids not based on ethanolamine are **phosphatidic acid** (Figure 14-2D), which lacks a second ester on the phosphoric acid, and **phosphatidylinositol**, in which the alcohol is *myo*-inositol (Figure 14-2F). Phosphatidylinositol is a minor phospholipid present at less than 5% in membranes. It provides the covalently attached anchor for phosphatidylinositol-linked proteins (Figure 3-17). Because it is also an intermediate in the biosynthesis of the second messenger *myo*-inositol 1,4,5-trisphosphate, its levels are independently regulated.

There are several other phospholipids that are variations on the theme developed by the glycerophospholipids. The **plasmalogens** (Figure 14-2E,H) have an enol ether at carbon 1 of the *sn*-glycerol 3-phosphate rather than an acylated oxygen. Upon treatment with BF_3 in methanol, the enol ether is released as the dimethyl acetal of a fatty aldehyde.²⁴ The enol ethers are usually derived from *n*-hexadecanal (Figure 14-2H) or *n*-octadecanal (Figure 14-2E), but minor amounts of the derivatives of unsaturated fatty aldehydes are also present.¹⁹ Either ethanolamine (Figure 14-2H) or choline (Figure 14-2E) is esterified, respectively, to the *sn*-glycerol 3-phosphate of plasmalogens.

There are a number of phospholipids and glycolipids in the membranes of archaebacteria that have tetraisoprenyl alcohols, such as 3,7,11,15-tetramethylhexadecanol, in **ether linkage** to two of the oxygens of glycerol.²⁵ The head group attached to the third oxygen can be a monosaccharide, such as glucose, in direct acetal linkage; a disaccharide, such as glucosyl (β 1-6) glucose, in acetal linkage; or a 1-phospho-*myo*-inositol, a phosphoserine, or a phosphoethanolamine in a phosphodiester linkage.²⁶⁻²⁸ One of these glycolipids and one of these phosphoinositides can also be fused at both of the terminal methyls of their tetraisoprenyl alkanes to produce a two-headed phosphoglycolipid.²⁷



14-8

At its two ends **14-8** is a paradigm for the types of head groups in archaeobacterial isopranylether lipids. In the membranes of the protist *Leishmania donovani* there is a different type of glycoposphoetherlipid in which either tetracosanol or hexacosanol forms an alkylether at carbon 1 of glycerol, the 2-hydroxy group of the glycerol is unsubstituted, and the head group on the 3-hydroxy group is a 1-phospho-*myo*-inositol to which a branched heptasaccharide is attached, and a (-6Gal β 1-4Man- α 1-phosphate)₁₆ is attached to the heptasaccharide.^{29,30}

A **sphingomyelin** has a primary alkene of 15 carbons replacing one of the hydrogens on carbon 1 of *sn*-glycerol 3-phosphate, the oxygen on carbon 1 is not acylated, and a nitrogen replaces the oxygen on carbon 2 and forms an amide rather than an ester with a fatty acid (Figure 14-2G). Upon saponification, **sphingosine** [(2*S*,3*R*)2-amino-3-hydroxyoctadec-4-en-1-ol], the fundamental skeleton on which sphingomyelin is constructed, is released. Sphingomyelins have choline for their head group.

Glycosphingolipids are also derived from sphingosine. As in sphingomyelin, the amino group on carbon 2 of the sphingosine in a glycosphingolipid is acylated. An acylated, unglycosylated sphingosine is a **ceramide**. A **cerebroside** is a glycosphingolipid in which either a glucose or a galactose is attached to the 1-hydroxyl of the ceramide in acetal linkage. Glycosphingolipids, however, can also have oligosaccharides attached to the 1-hydroxyl of the ceramide. Depending on the sequence of the immediately attached core oligosaccharide, the resulting glycosphingolipid is a ganglioside, a globoside, a lactoside, or of some other name (Table 14-1). As in the oligosaccharides on glycoproteins, the core of the oligosaccharide defining each type of glycosphingolipid can be incompletely finished, but the core oligosaccharides are usually further elaborated by adding *N*-acetylgalactosamine, galactoses, and as many as four or five sialic acids, and these modifications produce a dizzying array of microheterogeneous oligosaccharides more complex than that of the oligosaccharides on glycoproteins.

The composition (Table 14-2) of the lipids in plasma membranes from human erythrocytes³² or from murine fibroblasts (L-cells) grown in tissue culture¹⁹ are typical of **plasma membranes from animal cells**. The distribution of fatty acids among the various phospholipids from

Table 14-1: Oligosaccharides Forming the Core of the Various Glycosphingolipids³¹

type	sequence ^a
galaside	Gal(α 1,4)Gal-ceramide
schistoside	GalNAc(β 1,4)Glc-ceramide
molluside	Man(α 1,3)Man(β 1,4)Glc-ceramide
arthroside	GlcNAc(β 1,3)Man(β 1,4)Glc-ceramide
mucoside	Gal(β 1,3)Gal(β 1,4)Gal(β 1,4)Glc-ceramide
ganglioside	Gal(β 1,3)GalNAc(β 1,4)Gal(β 1,4)Glc-ceramide
globoside	GalNAc(β 1,3)Gal(α 1,4)Gal(β 1,4)Glc-ceramide
isogloboside	GalNAc(β 1,3)Gal(α 1,3)Gal(β 1,4)Glc-ceramide
lactoside	Gal(β 1,3)GlcNAc(β 1,3)Gal(β 1,4)Glc-ceramide
neolactoside	Gal(β 1,4)GlcNAc(β 1,3)Gal(β 1,4)Glc-ceramide

^aThese sequences form the core of the oligosaccharide that is attached to the ceramide. Other monosaccharides, in particular sialic acids, are attached to these cores.

Table 14-2: Composition of Amphipathic Lipids in Plasma Membranes from Human Erythrocytes³² and Murine Fibroblasts (L-Cells)¹⁹

	percentage of total lipid	
	erythrocytes	L-cells
amphipathic lipid		
phosphatidylcholine	16	23
sphingomyelin	16	14
phosphatidylethanolamine	16	9
phosphatidic acid	2	9
phosphatidylserine	8	3
phosphatidylinositol	2	3
choline plasmalogen	2	3
ethanolamine plasmalogen	NR ^a	2
ganglioside	4	NR
neutral lipid		
cholesterol	27	20
triglyceride	0	13

^aNot reported.

the plasma membranes of murine fibroblasts (Table 14-3) illustrates the heterogeneity of the collection. Membranes from fungi, such as the yeast *Saccharomyces cerevisiae*, have a similar ratio of phosphatidylcholine to phosphatidylethanolamine but greater amounts of both

Table 14–3: Fatty Acid Composition of Major Phospholipids and Neutral Lipid of Plasma Membranes of L-Cells¹⁹

fatty acids ^b	composition of fatty acids in each type of lipid ^a (%)						
	total neutral lipid	phosphatidylserine	sphingomyelin	phosphatidylcholine	phosphatidylethanolamine	phosphatidylinositol	phosphatidic acid
14:0	4	0.5	trace	1	1	0.6	
16:0	13	8	3	31	29	24	41
16:1	1	1		2	8	9	5
18:0	46	8	3	20	10	13	15
18:1	23	1	2	5	14	16	2
18:2	1	2	0.3		2	3	1
18:3	4	49	60	27	25	20	27
22:0		2	1		0.3	1	1
20:4	4		0.4	0.2		trace	trace
24:0	3	28	30	14	11	13	8
unsaturated fatty acids	33	53	63	34	48	48	35
long-chain fatty acids ^c	7	30	32	14	11	15	9
polyunsaturated fatty acids	9	51	61	27	27	22	28

^aData represent the composition of fatty acids (percent) present in total neutral lipid and the several types of phospholipid. Each of the compositions is the average of those from two membrane preparations. Neutral lipid and phospholipid from the plasma membranes were separated by silicic acid chromatography. Phospholipids were further separated by two-dimensional thin-layer chromatography, and the lipids were visualized by spraying with bromthymol blue. Iodine was not used in order to avoid possible losses of unsaturated fatty acids. Fatty acid methyl esters were prepared from the fatty acids in each preparation and analyzed by gas-liquid chromatography. ^bFatty acids are designated by the number of carbons and the number of double bonds. ^cLong-chain fatty acids are defined as fatty acids containing 20 or more carbon atoms.

phosphatidylinositol and phosphatidylserine. Fungi lack plasmalogens and sphingolipids.³³

The composition of the lipids varies among the various types of membranes in an animal cell: plasma membrane, endoplasmic reticulum, and mitochondria.³⁴ Mitochondria have less neutral lipid, sphingomyelin, and phosphatidylserine and more phosphatidylethanolamine than plasma membranes but about the same amount of phosphatidylcholine and plasmalogen.

The **plasma membranes of eubacteria**, such as the bacterium *E. coli*,^{35,36} are composed mainly of phosphatidylethanolamine (70%) but also contain small amounts of phosphatidic acid and phosphatidylserine as well as two unusual phospholipids, phosphatidylglycerol, where a glycerol is esterified to the phosphate of phosphatidic acid, and diphosphatidylglycerol, where a single molecule of glycerol is esterified at its two ends with two respective phosphatidic acids. These latter two phospholipids are found exclusively in prokaryotes and mitochondria, which are the direct descendants of prokaryotes. The lipids of the plasma membranes from the bacterium *Mycoplasma laidlawii*, however, have high percentages (45%) of the glycolipids 3-[*O*- α -D-glucopyranosyl]-1,2-diacyl-*sn*-glycerol and 3-[*O*- α -D-glucopyranosyl-(1,2)-*O*- α -D-glucopyranosyl]-1,2-diacyl-*sn*-glycerol.³⁷ This bacterium also contains phosphatidylglucose, a homologue of phosphatidylinositol.

In Gram-negative bacteria,* the plasma membrane

is surrounded by an **outer membrane**. The outer surface of this outer membrane, the surface exposed to the hostility of the environment, is formed mostly of a glycolipid called **lipopolysaccharide**. Instead of glycerol, the central element of lipopolysaccharide is a repeating polymer, (-phosphate-4-glucosamine(β 1,6)-glucosamine-1-)_n. 3-Keto fatty acids and 3-hydroxy fatty acids are linked to the 3-hydroxy and the 2-amino positions of each glucosamine.³⁸ To the glucosamines are also attached long oligosaccharides composed of mannose, glucose, galactose, *N*-acetylglucosamine, abequose, L-rhamnose, L-glycero-D-manno-heptose, 3-deoxy-D-manno octulosonic acid, and ethanolamine.³⁹ The inner surface of this outer membrane, however, is formed from normal glycerophospholipids.

Phosphatidylcholine, purified from a natural source such as eggs of *Gallus gallus*, spontaneously forms bilayers when it is suspended in water.⁴⁰ Initially, these bilayers are gathered in sets of nested spherical shells known as **multibilayer vesicles** (Figure 14–3).⁴⁰ Each of the shells in a multibilayer vesicle is a thin, closed, continuous bag. If such a suspension is submitted to sonication, the multibilayers eventually fragment and become small, unilamellar spherical vesicles.⁴¹ These vesicles are so small that they are not representative of the bilayer in biological membranes because of their excessive curvature. Larger, sealed, spherical unilamellar vesicles with uniform diameters of around 100 nm⁴² or with heterogeneous diameters as large as 50 μ m⁴³ can also be prepared from glycerophospholipids suspended in aqueous solution. A flat planar membrane, which is a single bilayer of phospholipid, can be formed

* Gram-negative bacteria are bacteria that do not stain with a particular dye because the dye cannot penetrate their outer membranes.

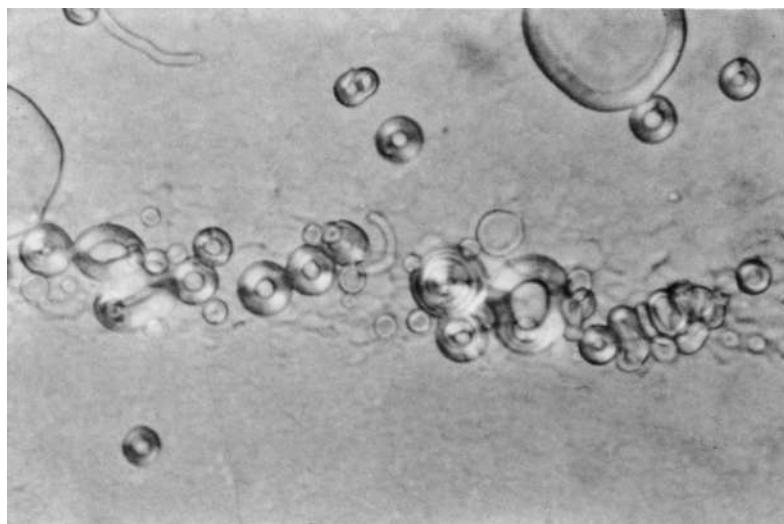


Figure 14-3: Spherical multibilayers of phosphatidylcholine.⁴⁰ Phosphatidylcholine was purified from the yolks of eggs from *G. gallus* by chromatography on alumina and then silicic acid with solvents of chloroform in methanol. The pure solid phosphatidylcholine was suspended in 0.15 M sodium chloride, and the resulting suspensions were examined in a polarizing microscope. The structures observed are small hollow spheres of phospholipid. The pattern of alternating dark and light sectors around the wall of the sphere within the plane of the page results from the fact that each sphere is formed from many concentric spherical shells, nested each within the other. Each of these spherical shells is a single bilayer of phosphatidylcholine. Reprinted with permission from ref 40. Copyright 1965 Academic Press.

across a small circular hole separating two aqueous compartments.^{44,45} Oriented bilayers of phosphatidylcholine can also be produced by evaporating a solution of the phospholipid in chloroform-methanol onto a flat surface such as mica and hydrating it with moist helium.⁴⁶ In each of these forms it is believed that the basic structural element, the bilayer of phospholipids, is the same and that the bilayer is a thin (4–5 nm) fluid film of phospholipid that can assume all of these different forms.

When bilayers of phosphatidylcholine are stacked upon mica as flat, planar, parallel sheets and this stack is placed in a beam of X-radiation, it produces a **diffraction pattern** (Figure 14-4A)⁴⁶ that is characterized by a set of sharp meridional arcs and two broad symmetrical equatorial reflections.⁴⁶ The equatorial reflections arise from the diffraction of the array of the linear hydrocarbons of the phospholipids oriented normal to the plane of the specimen. The meridional arcs arise from diffraction by the planes stacked one upon the other parallel to the orienting surface. The diffraction pattern of the meridional arcs can be transformed into a distribution of electron density along an axis normal to the orienting surface of the mica (Figure 14-4B). Because the stack of flat bilayers is a regularly repeating structure, the recurring variation of **electron density** in this dimension produces the diffraction pattern.

The repeating pattern in the properly phased Fourier transform of the meridional diffraction pattern consists of two regions of high electron density symmetrically sandwiching a region of low electron density (Figure 14-4B). This sandwich is the bilayer of phosphatidylcholine. The two regions with the highest electron density on either surface of the bilayer have been assigned to the glycerol, the phosphate, and the choline of the phosphatidylcholine (Figure 14-2). These functional groups, because they contain oxygen, nitrogen, and, in particular, phosphorus, have high electron density. The central region of the bilayer has been assigned to the hydrocarbon of the fatty acids.

A more recent calculation of the distribution of electron density in oriented bilayers of synthetic 1,2-dioleoylphosphatidylcholine, based on more extensive phasing of the meridional reflections, gave the same profile as that in Figure 14-4B.⁴⁷ The analysis of the profile of electron density, however, could be extended significantly in these later studies, because profiles of scattering density for both X-radiation and neutrons were calculated from the diffraction of X-radiation and the diffraction of neutrons, respectively, by the same sample of 1,2-dioleoylphosphatidylcholine. The scattering lengths for hydrogen, carbon, nitrogen, oxygen, and phosphorus differ dramatically relative to each other when these atoms are scattering X-radiation as opposed to when they are scattering neutrons. In particular, there are large differences in the relative scattering lengths for hydrogen. Because the different functional groups within a molecule of the phospholipid have different atomic compositions, the differences in scattering length and the significant differences between the profiles for the scattering of X-radiation and for the scattering of neutrons could be used to dissect the profile of electron density into the components that produce it. This dissection defines the **mean locations of choline, phosphate, glycerol, ester, carbon-carbon double bond, and hydrocarbon**.⁴⁷ These functional groups were calculated to be situated symmetrically at 2.2, 2.0, 1.9, 1.6, 0.8, and 0–1.6 nm, respectively, from the center of the bilayer. The width of this particular synthetic bilayer from choline to choline is 4.4 nm.

As the widths of the layers of water between each bilayer in such a stack is decreased by decreasing the vapor pressure of the water in the chamber (Figure 14-4B), greater and greater decreases in vapor pressure are required to elicit the same change in width once the separation between the maxima of electron density goes below about 0.5 nm.⁴⁸ In the dissection of the map of electron density into its components, it was estimated that the choline head groups extend 0.2 nm beyond

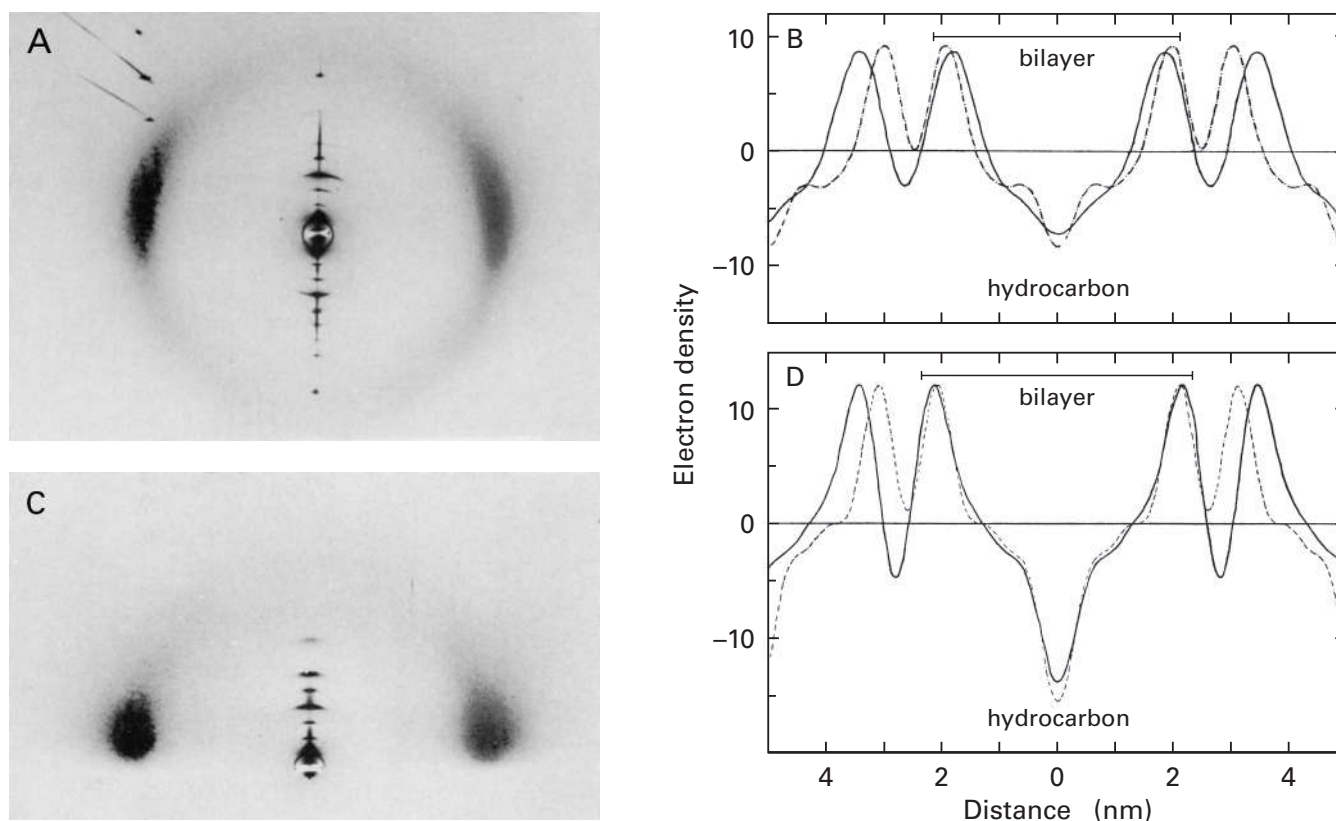


Figure 14-4: Diffraction patterns (A and C) and their respective computed Fourier transforms (B and D) for multilayers of pure phosphatidylcholine (A and B) from yolks of eggs from *G. gallus* or an equimolar mixture of the same phosphatidylcholine and cholesterol (C and D).⁴⁶ Lipids dissolved in chloroform in methanol were smeared on mica sheets, and the solvent was evaporated under a stream of moist helium. The multilayers that resulted were submitted to diffraction in a beam of X-radiation (A and C). The two symmetrically displayed, diffuse but intense diffractions on the equators (0.46 nm) in panels A and C are from the spacings of the linear hydrocarbons of the phospholipids oriented normal to the plane of the specimen. The sharp reflections on the meridian (the vertical axis of the pattern) are the reflections arising from the set of planes produced by the stacking of the bilayers. With the appropriate choice of phase, the amplitudes of the meridional reflections can be submitted to Fourier transform to obtain profiles of electron density (B and D) along an axis normal to the plane of the specimen. Presumably this axis is normal to the flat sheets producing the multilayer. The electron density in arbitrary units is presented as a function of the distance (nanometers) from the center of the bilayer. In panel B, the profile of electron density is given for lipids equilibrated with moist helium of 57% relative humidity (dashed lines; 14% water), or 100% relative humidity (solid line; 21% water), and in panel D, the profile of electron density is given for lipids equilibrated with moist helium of 57% relative humidity (dashed lines; 13% water) or 100% humidity (solid line; 22% water). It was the expectation of the investigators that the width and structure of the bilayer (within the double arrow) would remain constant while the distance between bilayers would increase as the water content increased. This expectation was used to assign the phases so its fulfillment is inconsequential. Adapted with permission from *Nature*, ref 46. Copyright 1971 Macmillan Magazines Limited.

maxima of electron density on the two sides of the aqueous space.⁴⁷ It has been proposed that the cause of the resistance to decreasing the separation between the bilayers is **steric repulsion** between these cholines on the apposed surfaces.⁴⁸ The observed distance at which repulsion sets in is consistent with the calculated location of the cholines. That the repulsion at these short distances arises from the collision of the head groups is supported by the fact that incorporation of cholesterol into the bilayers, which spreads apart the head groups and permits them to interdigitate significantly, decreases the repulsion.⁴⁹

When the amount of water in a sample of hydrated natural phosphatidylcholine from eggs is systematically increased from 10% to 45%, changes in the structure of the bilayers occur.^{50,51} The width of the bilayers decrease from 4.5 to 3.8 nm,⁵⁰ and the cross-sectional area for

each phospholipid in one of the two monolayers increases^{50,51} from 0.55 to 0.68 nm². It is thought that these changes are responses to the steric effects coincident to the **hydration** of the hydrophilic functional groups on the external surfaces. As hydration increases, it pushes apart the adjacent molecules of phosphatidylcholine in each monolayer of the bilayer and produces the observed changes. Above a certain level of hydration (>40% water), when all hydrophilic groups are fully hydrated, the thickness of the bilayer no longer decreases.

Because it is heterogeneous, naturally occurring phosphatidylcholine will not crystallize. Synthetically prepared dimyristoylphosphatidylcholine, however, has been crystallized, and a **crystallographic molecular model** has been constructed (Figure 14-5).⁵² The crystalline material consists of stacks of bilayers whose dis-

tribution of electron density along an axis normal to the planes of the bilayers (vertical axis in Figure 14-5) would resemble the distribution seen in Figure 14-4B. The molecules of phospholipid are distributed symmetrically about the center of the bilayer just as they are in a bilayer of fluid natural phospholipids. In fact, in the crystal, there are crystallographic screw axes of symmetry between the methyl groups at the ends of the hydrocarbons. Each plane of adjacent phospholipid molecules oriented in the same direction forms one of the two **monolayers** of one of the bilayers. As in the fluid, uncrys-

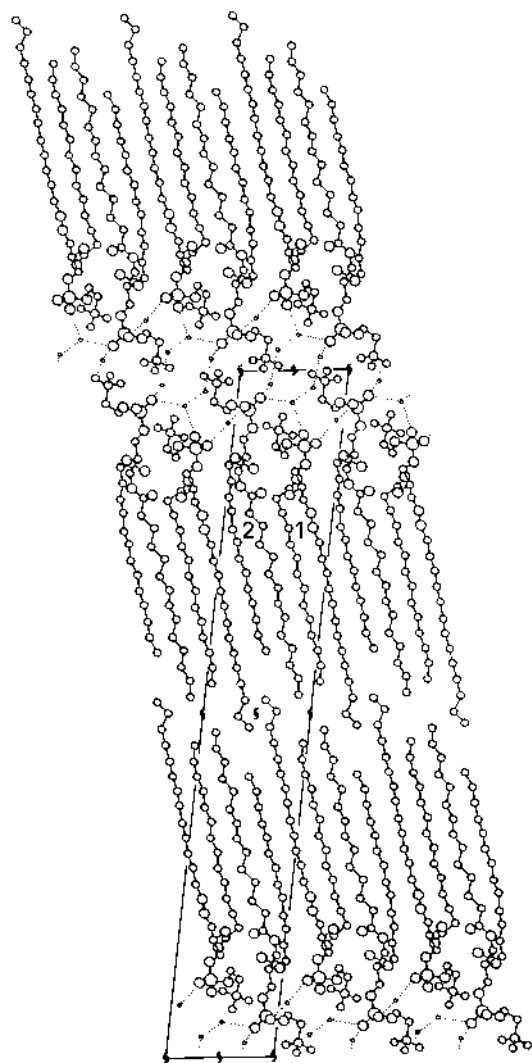


Figure 14-5: Arrangement of the molecules of phosphatidylcholine in the crystallographic molecular model of dimyristoylphosphatidylcholine dihydrate.⁵² The asymmetric unit in the unit cell of the $P2_1$ space group is formed from two molecules of dimyristoylphosphatidylcholine (1 and 2). The unit cell is outlined, and the 2-fold screw axes of symmetry normal to the plane of the page are designated. Because the crystal is the dihydrate, grown from a mixture of ether, ethanol, and water, each asymmetric unit has four water molecules (●) associated with it in the hydrophilic region between the bilayers. This corresponds to a water content of 5%. Reprinted with permission from *Nature*, ref 52. Copyright 1979 Macmillan Magazines Limited.

talline bilayer, the functional groups of the phospholipid are encountered in the order choline, phosphate, glycerol, ester, hydrocarbon from the outermost surface to the interior. Because the fatty acids are homogenous, their hydrocarbon has solidified into a crystalline array that is hexagonally packed. Dilauroylphosphatidylethanolamine crystallizes from acetic acid in bilayers, and the vertical hexagonal packing of the fatty acids in its crystallographic molecular model is readily discerned from a view normal to the plane of one of the bilayers (Figure 14-6).⁵³ Such crystallographic models provide a starting point for a discussion of the structure of a bilayer of amphipathic lipids. It must be remembered, however, that they represent homogeneous lipids in which the alkane is fully saturated and solid.

Four types of **conformation at the glyceryl backbone** of phospholipids have been observed in crystallographic molecular models (Figure 14-7), and there is evidence from nuclear magnetic resonance spectra that, in the liquid state, the phospholipids fluctuate among these conformations.⁵⁴ One interesting aspect of these structures is that in the conformations represented by dimyristoylphosphatidylcholine (DMPC) and dilauroyl-*N,N*-dimethylphosphatidylethanolamine (DLPEM₂), the acyl carbon of the fatty acid on carbon 3 of the glyceryl group is buried more deeply in the bilayer than the acyl carbon of the fatty acid on carbon 2, while in the two conformations represented by dilauroylphosphatidic acid (DLPA) and dimyristoylphosphatidylglycerol (DMPG), it is the acyl carbon of the fatty acid on carbon 2 that is buried more deeply. This means that, in the rapid transitions among these conformations that occur in bilayers of liquid phospholipid, the linear hydrocarbons slide back and forth past each other in a direction normal to the plane of the bilayer. As these sliding movements occur, each of the acyl oxygens on the two fatty acids comes in turn to the surface of the bilayer (Figure 14-7). Because of the sliding movements, the equilibrium among these conformations can be shifted by changing the structure of the surroundings in which the far ends of the fatty acids are located.⁵⁵

In these **sliding fluctuations**, the positions of the charged phosphate and the charged nitrogen on phosphatidylcholine must average to about the same mean location relative to the surface of the bilayer. This follows from the fact that vesicles of phosphatidylcholine have zero electrophoretic mobility, which demonstrates that on the average the negative charges on their phosphates must reside in the same plane parallel to the surface of the bilayer as the positive charges on their cholines.⁵⁶ The dielectric properties of bilayers of phosphatidylcholine are also consistent with this disposition. In the crystallographic molecular model of dilauroylphosphatidylethanolamine (Figure 14-6), the ammoniums of the ethanolamines form hydrogen bonds to the oxygens of the phosphates, bringing the two opposite charges into the same plane parallel to the surface of the bilayer.

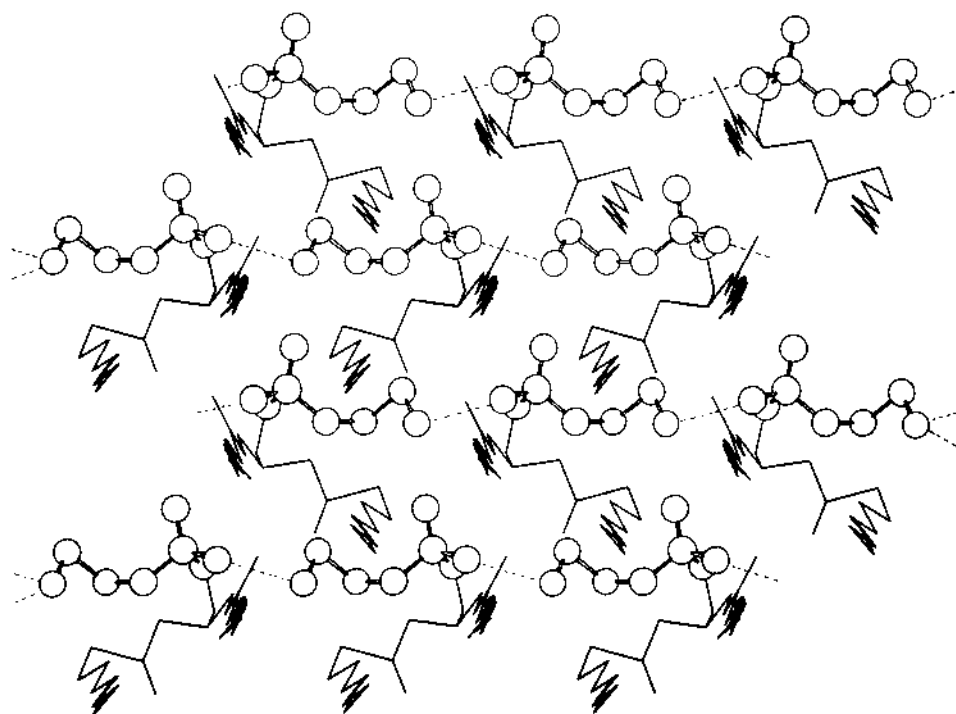


Figure 14-6: View of the crystallographic molecular model of dilauroylphosphatidylethanolamine looking down on the hydrophilic surface of the bilayer.⁵³ Only one monolayer of the bilayer of phospholipid is shown in the drawing. One of the oxygens of each phosphate forms a hydrogen bond with the ammonium group of an adjacent ethanolamine. The dilauroylphosphatidylethanolamine was crystallized from glacial acetic acid, and in the crystals there was one molecule of acetic acid (not shown) for each molecule of phospholipid. The fact that the linear hydrocarbons are all normal to the surface of the monolayer is apparent, and the drawing gives a representation of the view from above the surface of a bilayer of phospholipid. Reprinted with permission from ref 53. Copyright 1974 National Academy of Sciences.

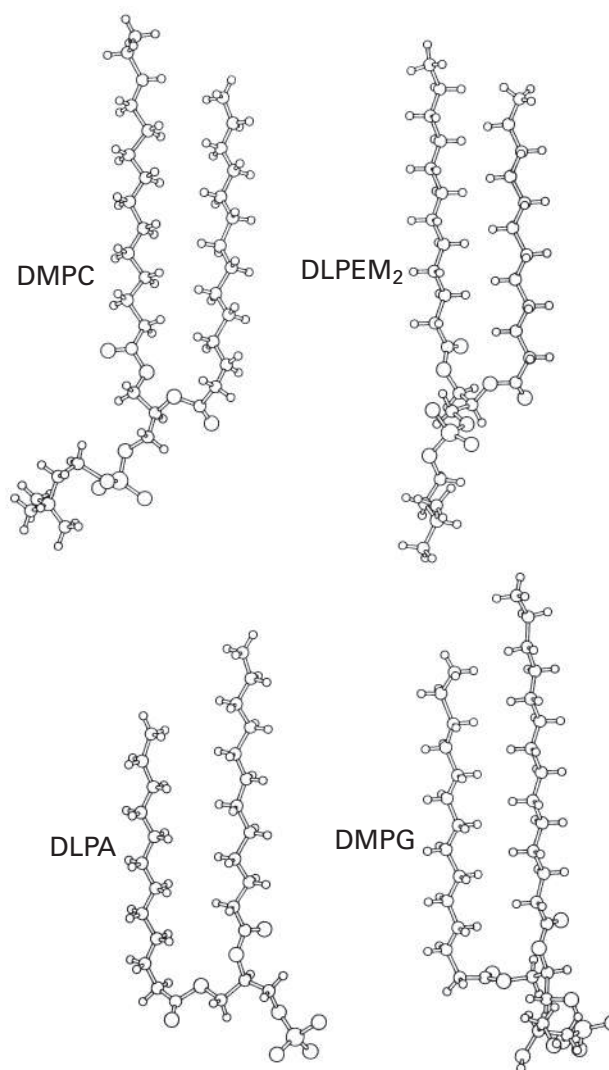


Figure 14-7: Four conformations available to a phospholipid in a bilayer.⁵⁴ These four drawings are taken from the crystallographic molecular models of crystalline dimyristoylphosphatidylcholine (DMPC), dilauroyl-*N,N*-dimethylphosphatidylethanolamine (DLPEM₂), dilauroylphosphatidic acid (DLPA), and dimyristoylphosphatidylglycerol (DMPG). Within each structure the two fatty acids are the same length, so the relative positions of the two fatty acids are most readily ascertained by looking at their ends. Within each structure the fatty acid to the right is the one on carbon 2 of the glycerol, and the fatty acid to the left is the one on carbon 3. In the upper two conformations, the fatty acid on carbon 3 of the glycerol is deeper in the bilayer of phospholipid than that on carbon 2, and in the lower two conformations the fatty acid on carbon 2 is deeper in the bilayer of phospholipid than that on carbon 3. In the upper two structures the acyl oxygen of the fatty acid on carbon 2 of the glycerol is at the surface of the bilayer; in the lower two structures the acyl oxygen of the fatty acid on carbon 3 is at the surface of the bilayer. Reprinted with permission from ref 54. Copyright 1988 American Chemical Society.

Phosphatidylcholine and phosphatidylethanolamine are zwitterionic and neutral, but phosphatidylserine, phosphatidylglycerol, phosphatidylinositol, and half a molecule of diphosphatidylglycerol each have a net charge number of -1 . Consequently, natural bilayers have a net **negative surface potential**. In a bilayer of pure phosphatidylserine, the potential at the surface is about -80 mV and falls off as a function of the distance from the surface as predicted by the Gouy-Chapman equation for an ionic double layer.^{57,58} At a distance of 1 nm from the surface in a solution of ionic strength of 0.1 M, the potential has dropped to -30 mV. The magnitude of the surface potential varies with the mole fraction of negatively charged lipid in a bilayer and the ionic strength of the solution⁵⁹ and affects the adsorption of small charged molecules⁵⁹⁻⁶¹ and proteins⁶² to the bilayer in a predictable manner.

The **electrostatic repulsion** of the negatively charged phospholipids decreases the stability of a bilayer. In fact, at low ionic strength, bilayers of pure dimyristoylphosphatidylglycerol are unstable enough that the monomer has a measurable solubility in aqueous solution (10^{-15} – 10^{-16} mole fraction),⁶³ which is almost unheard of. The **solubilities** of neutral phospholipids or monoanionic phospholipids with longer fatty acyl groups are so small that they cannot be measured. Bilayers of entirely monoanionic phospholipid are unstable enough that they do not occur naturally. There is, however, a mutant of *E. coli* in which the synthesis of phosphatidylethanolamine has been deleted; and this curiosity, the phospholipids of which are almost entirely phosphatidylglycerol and diphosphatidylglycerol, will grow as long as the concentration of Mg^{2+} in the medium is high enough to decrease significantly the surface potential of its plasma membrane.⁶⁴

The bilayers observed in crystallographic molecular models are solids in which the hydrocarbon is frozen; the bilayers of the mixture of phospholipids purified from a natural source or the bilayer present in a biological membrane is liquid. The **transition between solid and liquid** resembles the melting of paraffin, and it can be observed by cooling a bilayer to solidify it and then raising the temperature gradually to melt it. As phospholipids from most natural sources remain fluid even at low temperatures, the transition is usually followed either in homogeneous, synthetic phospholipids or in biological membranes the composition of which is highly enriched in one particular fatty acid. For example, when the bacterium *M. laidlawii* is grown on medium supplemented with a chosen fatty acid, up to 70% of the fatty acids in its membranes are the supplemented fatty acid.⁶⁵

The transition between solid and liquid can be observed by diffraction of X-radiation. In a solid bilayer the hydrocarbon of the phospholipids is in a **hexagonal array** (Figure 14-6), and the spacing between the linear, all-*trans* alkanes produces a strong sharp equatorial

reflection (see Figure 14-4A) at $(0.415 \text{ nm})^{-1}$ characteristic of crystalline paraffins.⁶⁶ The cross-sectional area of such a hexagonal array of solid paraffin hydrocarbons is 0.40 nm^2 for every two alkyl chains,⁶⁶ and this agrees closely with the cross-sectional areas of 0.39 – 0.41 nm^2 for one complete molecule of phospholipid in each monolayer of the crystallographic molecular models.^{52,53,67} The width of a solid bilayer of phospholipid with only saturated fatty acids is consistent⁶⁶ with the width of two layers of slightly tilted ($\leq 20^\circ$) alkanes of the appropriate length in the all-*trans* configuration (Figure 14-5).

When a solid bilayer is melted to a liquid bilayer by raising the temperature, its width decreases by 0.5 – 1.0 nm .^{66,68-70} If at the same time the hydrocarbon is expanding to the extent that paraffins expand as they become liquid, the **cross-sectional area for each phospholipid** in one of the monolayers of the bilayer must increase^{66,68} to 0.55 – 0.70 nm^2 . This expansion of the cross-sectional area of the bilayer, among other factors, reflects the establishment of the normal disorder of the liquid state of a paraffin. In this state, *gauche* conformations, which necessarily shorten the distance that can be covered by a hydrocarbon, become common features that lead to the narrowing of the bilayer.

In such molten bilayers of fully saturated phospholipid, the hydrocarbons remain oriented preferentially with their long axis aligned with an axis normal to the plane of the bilayer. This conformation has been demonstrated by neutron diffraction of oriented bilayers of dipalmitoylphosphatidylcholine in which different carbons along the palmitates have been labeled with deuterium, an atom that scatters neutrons strongly. In the solid phase at low hydration, where all of the hydrocarbons are in hexagonal array normal to the plane of the membrane, the location of the deuteriums in the distribution of scattering density is easily distinguished (Figure 14-8).⁶⁹ In such solid bilayers, the deuteriums appear at the expected distances from the center (Table 14-4). When bilayers of the various dipalmitoylphosphatidylcholines are melted, the deuteriums in the fluid

Table 14-4: Distance of Various Carbons from the Center of a Bilayer⁶⁹

carbon deuterated	distance from center ^a (nm)	
	solid bilayer	fluid bilayer
C15	0.20 ± 0.1	0.19 ± 0.1
C14	0.36 ± 0.1	0.36 ± 0.1
C9	0.94 ± 0.1	0.81 ± 0.1
C5	1.21 ± 0.2	1.05 ± 0.1
C4	1.53 ± 0.2	1.22 ± 0.1

^aDetermined by neutron diffraction of bilayers of dipalmitoylphosphatidylcholine selectively deuterated at the noted positions.

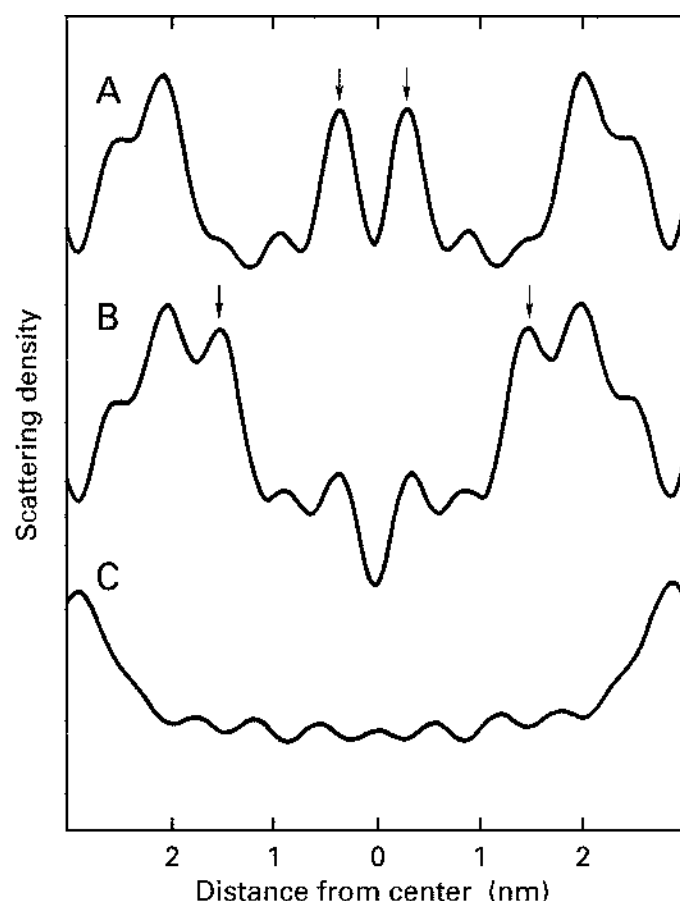


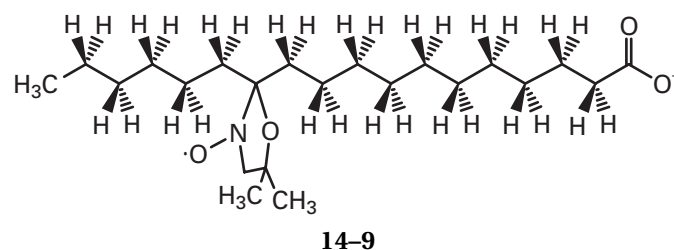
Figure 14-8: Distribution of neutron scattering density across bilayers of phospholipid in which hydrogen has been replaced by deuterium at specific locations along the linear alkyl groups of the phospholipid.⁶⁹ Stacked, parallel bilayers of the di-(15,15-dideuteriopalmityl)phosphatidylcholine (A), di-(5,5-dideuteriopalmityl)phosphatidylcholine (B), or dipalmitoylphosphatidylcholine hydrated with $^2\text{H}_2\text{O}$ (C) were prepared on quartz slides. Each of these multilayers was brought to 20 °C, which is below the melting point of dipalmitoylphosphatidylcholine under these circumstances, and allowed to diffract neutrons. From the meridional reflections and appropriate phases, a distribution of neutron scattering density (relative units) normal to the plane of the membranes as a function of distance from the center (nanometers) could be calculated by Fourier transformation. The positions of the deuterated carbons in the first two samples are clearly observed (arrows). The position of the $^2\text{H}_2\text{O}$ in the third sample was defined by a difference map of neutron scattering density (C) between a specimen hydrated with H_2O and one hydrated with $^2\text{H}_2\text{O}$. Reprinted with permission from *Nature*, ref 69. Copyright 1978 Macmillan Magazines Limited.

hydrocarbon remain similarly distributed (Table 14-4), but each moves closer to the center. This rearrangement is consistent with the narrowing of the bilayer that occurs upon melting. In fluid bilayers of phospholipids in which all of the fatty acids are the same length, the methyl groups at the ends of the fatty acids from the two monolayers end up adjacent to each other (Figure 14-5), but in bilayers in which the two fatty acids on the phospholipids are of significantly different length, the longer

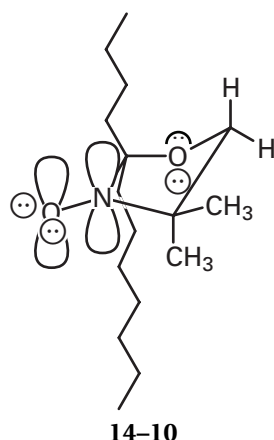
fatty acids interdigitate so that the methyl groups of the longer fatty acids on one monolayer end up adjacent to the methyl groups of the shorter fatty acids on the other.⁷⁰

The transition between solid and liquid in a bilayer composed of mixtures of various homogeneous, synthetic phospholipids has also been studied. When a suspension of bilayers composed of only one phospholipid such as dipalmitoylphosphatidylcholine is melted, a single sharp transition that occurs completely over 2–3 °C is observed. It can be monitored in a **calorimeter** as the absorption of heat resulting from the **heat of fusion**.⁷¹ When two phospholipids are mixed, the transition occurs over a broader range of temperatures somewhere between the temperatures of the transitions of the separated components.^{72–74} At temperatures in the range over which the transition is occurring, regions of fluid phase are in equilibrium with regions of solid phase laterally separated from each other in the plane of the bilayer.⁷² In many instances, the two component phospholipids are not miscible with each other as solids and a separate solid phase of one or the other remains laterally isolated in the bilayer.⁷² For example, mixtures of up to 40% dipalmitoylphosphatidylethanolamine in dimyristoylphosphatidylcholine contain significant regions of unmixed dimyristoylphosphatidylcholine in the solid phase.⁷² These results suggest that **regions of immiscible, unmelted phospholipid** might form in natural bilayers under certain circumstances. The heterogeneous mixtures of phospholipids normally found in natural circumstances, however, appear to form bilayers that are fully liquid and fully miscible at all physiological temperatures.

The fluid bilayers in normal biological membranes and in bilayers formed spontaneously from the amphipathic lipids extracted from normal biological membranes have been studied by following the **electron spin resonance** of probes incorporated into them. **Nitroxyl fatty acid 14-9**⁷⁵ is an example of such a probe:



The nitroxyl radical is in a five-membered ring similar to that of the 1-oxyl-2,2,5,5-tetramethylpyrrolin-3-yl radical (**12-11**). The unpaired electron in the nitroxyl radical is located in a π molecular orbital the principal axis of which is aligned parallel to the axis of the all-*trans* linear hydrocarbon:⁷⁶



The spectrum of 3-carbamoyl-1-oxyl-2,2,5,5-tetramethylpyrroline freely tumbling in aqueous solution displays three sharp peaks of equal intensity (Figure 12-35),^{77,78} reflecting the rapid isotropic motion of the molecule and the full coupling of the unpaired electron and the nitrogen nucleus. When nitroxyl fatty acid **14-9**, however, is incorporated into an oriented multibilayer of phosphatidylcholine from eggs of *G. gallus*, the absorbances of the unpaired electron in the spin-labeled fatty acid within the multibilayer are much broader because the motions reorienting its nitroxyl radical are much less rapid (Figure 14-9).^{77,78} The decrease in the intensity of the symmetrically displayed hyperfine peaks resulting from the change in bonding of the nitroxyl radical (**12-12**) that occurs within the nonpolar environment of the interior of the bilayer is also apparent.

The spectra also have become anisotropic because the motion of the nitroxyl radical has become **anisotropic**. The easiest way to demonstrate this is to take two spectra from the specimen oriented so that the magnetic field is either normal to the planes of the multibilayers (Figure 14-9B) or parallel to the planes of the multibilayers (Figure 14-9C). In the one case the splitting of the peaks is larger than that of the isotropic spectrum, and in the other it is smaller. From theoretical simulations of these spectra, it could be concluded that, in the bilayers of phosphatidylcholine, nitroxyl fatty acid **14-9** is oriented with the axis of its hydrocarbon perpendicular to the planes of the multibilayers and rotates exclusively or almost exclusively about this axis. These conclusions are consistent with expectations based on the structure of a bilayer of phospholipid (Figure 14-2) and the amphipathic structure of nitroxyl fatty acid **14-9**. When such nitroxyl radicals are incorporated into multilamellar vesicles of phosphatidylcholine suspended in water (Figure 14-3), the spectrum that results is a composite of the perpendicular and parallel spectra seen in Figure 14-9, panels B and C, respectively.⁷⁹

That **molecular motions of the linear alkane** in a fluid bilayer of phospholipid increase in proceeding from the acyl carbon to the center can be demonstrated with spin-labeled phospholipids.⁷⁹ A series of phosphatidyl-

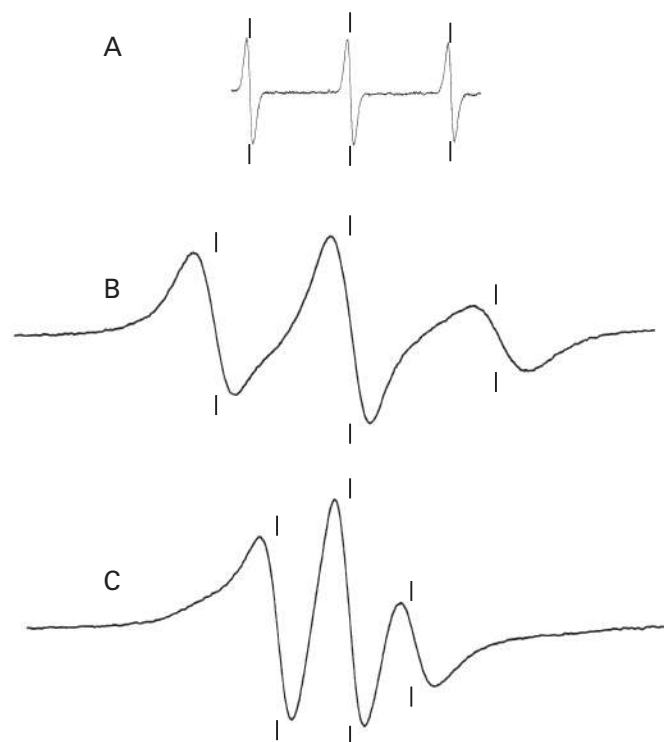


Figure 14-9: Electron spin resonance spectra of azacyclic *N*-oxides. (A) 2,2,5,5-Tetramethyl-3-(aminocarbonyl)azacyclopentane *N*-oxide (see **12-11**) dissolved in water.⁷⁷ (B, C) 2-(10-Carboxydecyl)-2-hexyl-5,5-dimethyl-3-azetidine *N*-oxide (**14-9**) incorporated into multibilayers of phosphatidylcholine from eggs of *G. gallus*. The magnetic field of the spectrometer was oriented perpendicular (B) or parallel (C) to the plane of the cover slip.⁷⁸ Nitroxyl fatty acid **14-9** and phosphatidylcholine were dissolved in chloroform and the mixture was evaporated to dryness. Water was added and a portion of the opalescent suspension that resulted was spread on a cover slip. The water was evaporated at 39 °C to produce the multibilayers oriented by the plane of the cover slip. In all of the spectra, the derivative of the absorbance is presented on the vertical axis. The microwave frequency was held at a constant value while the strength of the magnetic field was varied continuously. The magnetic flux density (tesla) is the variable on the horizontal axis. The distance between the two absorbances at the ends of the spectrum in panel A is 0.0028 T. The vertical lines mark the positions of maximum absorption (zero slope) of microwave energy. Reprinted with permission from refs 77 and 78. Copyright 1965 and 1969 National Academy of Sciences.

cholines were synthesized in which the acyl groups on carbon 1 of the *sn*-glycerol 3-phosphate were derived from either palmitic acid or stearic acid, and the acyl groups on carbon 2 were derivatives of palmitic acid or stearic acid, respectively, on which the cyclic dimethyl nitroxyl radical was positioned at the 5th, 8th, 12th, and 16th carbon.⁷⁹ An **order parameter**, *S*, can be defined, which is a number that quantifies the confinement of the rotational motion of these cyclic nitroxyl radicals to one particular axis. When *S* = 1, the ring rotates about a fixed axis in space; and when *S* = 0, its rotational motion is isotropic (Figure 14-9A).

When the various labeled phospholipids were

incorporated into multibilayers of natural phosphatidylcholine, the order parameter S was observed to decrease as the cyclic nitroxyl radical was situated farther from the acyl carbon. For cyclic nitroxyl radicals at the 5th, 8th, 12th, and 16th carbon of the labeled fatty acid, the order parameters S were 0.68, 0.50, 0.33, and 0.16, respectively.⁸⁰ In experiments of this type, caution must be taken that the probe does not affect the behavior of the alkane to which it is attached. For example, when a similar experiment was performed with the fluorescent probe 7-nitro-2,1,3-benzoxadiazol-4-yl covalently attached to the phospholipid, the hydrophilicity of the probe drew the end of the fatty acyl group to which it was attached to the surface of the membrane, pulling it away from the center of the core.⁸¹ The order parameters observed with the much less hydrophilic nitroxyl group, however, have been confirmed by using deuterium as a probe, which produces the least possible perturbation of the fatty acyl group.

Measurements of the order parameter as a function of the position along the linear alkane of the fatty acyl groups in a phospholipid have been made by **deuterium nuclear magnetic resonance**.⁸² A series of synthetic phospholipids were prepared that contained either palmitic acid at both carbon 1 and carbon 2 or palmitic acid at carbon 1 and oleic acid at carbon 2 of the *sn*-glycerol 3-phosphate. In each member of the series, deuterium atoms were placed synthetically on a specific carbon in the palmitic acids. Because a deuteron, like a nucleus of ¹⁴nitrogen, is quadripolar, a deuterium nuclear magnetic resonance spectrum can also be used to estimate an order parameter S for the degree of anisotropy experienced by the carbon to which it is attached.^{67,83} Order parameters S for dipalmitoylphosphatidylcholine, 1-palmitoyl-2-oleoylphosphatidylcholine, and dipalmitoylphosphatidylserine, gathered from bilayers of these phospholipids held at temperatures an equivalent distance above each of their melting points, have been presented as a function of the carbon on which the deuteriums were located (Figure 14-10).⁸⁴ The confinement experienced by a carbon in the liquid hydrocarbon of the bilayer decreases as the distance from the acyl carbon increases. Carbons at the very core of the bilayer are able to assume almost every orientation, while carbons near the periphery are confined in their orientations. These observations relate to an apparent **stereochemical paradox** in the structure of a bilayer of phospholipids from natural sources.

Beyond the eighth carbons of the fatty acids attached to carbon 2 of either the *sn*-glycerol 3-phosphates or the sphingosines in a natural bilayer, within the **core of the hydrocarbon**, the carbon-carbon double bonds begin (Figure 14-2). This fact has two consequences. First, permanent elbows incompatible with straight linear alkanes necessarily disrupt the alignments of the hydrocarbons of the fatty acids. Second, the average length for each carbon is decreased by the multiple

double bonds. Both the disorder introduced by single and triple *cis* double bonds and the shortening of the chains caused by all of the double bonds necessarily increase the mean cross-sectional area parallel to the plane of the bilayer for each chain of hydrocarbon in this region distal to the glyceryl groups located on the two sides of the bilayer. Added to this effect is the disorder that naturally occurs farther away from the surfaces of the bilayer even in saturated phospholipids (Figure 14-10).

Before the ninth carbons of the acyl groups on carbon 2 of the *sn*-glycerol 3-phosphates and sphingosines, however, essentially all of the hydrocarbon is

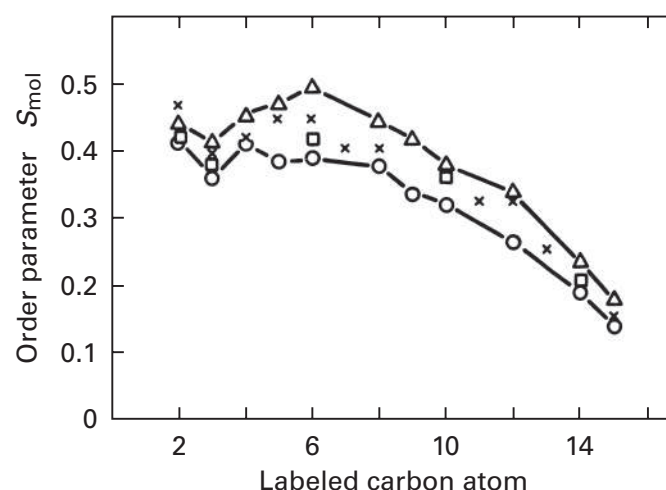


Figure 14-10: Variation of the molecular order parameter, S_{mol} , estimated from deuterium nuclear magnetic resonance spectra as a function of the position of the deuterium along the hydrocarbon of 1,2-di(dideuteriopalmityl)phosphatidylcholine (○), 1,2-di(dideuteriopalmityl)phosphatidylserine (□), or 1-(dideuteriopalmityl)-2-oleoylphosphatidylcholine (△).⁸⁴ A series of selectively deuterated palmitic acids were synthesized, each with two deuteriums at a different carbon along the chain. From these dideuteriopalmic acids a series of dipalmitoylphosphatidylcholines was synthesized, each with two palmitic acids in which deuterium occupied the same position. These lipids were separately suspended in water at a temperature 19 °C above their melting points (41 °C) and deuterium nuclear magnetic resonance spectra were recorded, from which order parameters, S_{mol} , were calculated (○). A similar series of dipalmitoylphosphatidylserines was also prepared and a similar analysis performed at 51 °C (□). Samples of each of the dipalmitoylphosphatidylcholines were digested with phospholipase A, and the resulting 2-lysophosphatidylcholines were esterified with oleic acid to produce a series of 1-(dideuteriopalmityl)-2-oleoylphosphatidylcholines in each of which two deuteriums occupied a different position, respectively, in the palmitic acid. Deuterium nuclear magnetic resonance spectra were taken of suspensions of these phospholipids at a temperature 16 °C above their melting points (−5 °C). From these spectra, order parameters, S_{mol} , were calculated (△). A series of palmitic acids selectively deuterated at specific positions were separately fed to *M. laidlawii* bacteria to enrich (70%) the membranes of these cells in the added fatty acid, and the order parameters for samples of each of these selectively deuterated membranes were also determined (×). Order parameters are plotted as a function of the position of the labelled carbon in the respective fatty acids. Reprinted with permission from ref 84. Copyright 1978 Elsevier Science Publishers.

linear, saturated alkane, and more ordered (Figure 14–10). In a bilayer composed only of phospholipids and sphingomyelins, these regions of **alkane proximal to the glyceryl groups** are nevertheless necessarily required to have the same mean cross-sectional area for each phospholipid as exists in the distal regions at the core. All of these considerations require that the solution to this paradox incorporate a large cross-sectional area, low density, and high disorder in the regions distal to the glyceryl groups at the core of the hydrocarbon and a large cross-sectional area, high density, and low disorder in the two symmetrical regions of alkane proximal to the glyceryl groups in a bilayer.

One solution to this paradox would be that the linear alkanes in the two proximal regions are tilted.⁸⁰ Tilting the alkane increases its cross-sectional area parallel to the plane of the bilayer while retaining the density of the condensed, hexagonal array. This possibility has been examined with a series of synthetic phosphatidylcholines that each had a dimethyl cyclic nitroxyl radical attached at a different carbon in the saturated fatty acyl group on carbon 2 of their *sn*-glycerol 3-phosphates. These labeled phosphatidylcholines were incorporated into bilayers of phosphatidylcholine from eggs of *G. gallus*. When the dimethyl cyclic nitroxyl radical was on the 5th or the 8th carbon, its principal axis was tilted 30° relative to the plane of the bilayer, but when it was on the 12th or the 16th carbon, its principal axis was oriented on average normal to the plane of the bilayer.⁸⁰ Tilted hydrocarbon chains have been directly observed in crystallographic molecular models of bilayers of phospholipids.⁶⁷ A tilting of the alkyl chains in the two regions on the two sides of the bilayer proximal to the glyceryl groups would also explain the increase in cross-sectional area and decrease in width that occurs upon the melting of bilayers of homogeneous phospholipids.⁶⁶ The implausible feature of this explanation is that all of the alkyl chains in a fluid bilayer would have to tilt in the same direction within one of these regions for it to be correct.

Such a **coordinated tilting of the hydrocarbon** over large areas of a solid bilayer has been observed by diffraction of X-radiation. In bilayers of dipalmitoylphosphatidylcholine and distearoylphosphatidylcholine below the temperature at which they melt, the equatorial reflection in the X-ray diffraction pattern that arises from the aligned chains of alkane displays the fine structure of a sharp reflection superimposed upon a broader reflection.⁸⁵ This distribution of reflected intensity has been shown to arise from hydrocarbons tilted relative to the axis normal to the bilayer, and the degree of tilt can be calculated from the diffraction pattern. As the degree of hydration in these solids was increased from 6% to 30%, the tilt of the hydrocarbons increased from 17° to 40°. Presumably the **steric effects of the hydration** force the hydrophilic functional groups of the phospholipids to take up a greater surface area for each molecule of phos-

pholipid, and the linear hydrocarbons adjust to the required increase in their cross-sectional area by tilting. It has also been observed that a dimethyl cyclic nitroxyl radical, attached at the fifth carbon of a fatty acid on carbon 2 of the *sn*-glycerol 3-phosphate of dipalmitoylphosphatidylcholine, appeared to be in a more polar environment in bilayers of natural phospholipids than nitroxyl radicals attached farther down the fatty acid.⁸⁶ If the hydrocarbons were tilted in this region, their surfaces should be more exposed to the aqueous phase.

These observations, however, illustrate a difficulty in interpreting many of the physical studies on bilayers. The bilayers used in these experiments were vesicles prepared by **sonication**. It has been shown by nuclear magnetic resonance spectroscopy that vesicles of small diameter (30–90 nm) prepared by sonication display anomalous physical properties because of their high curvature.⁸⁷ It has been pointed out that these anomalies arise from the fact that the high curvature unavoidably forces a portion of the hydrocarbon adjacent to the hydrophilic functional groups of the phospholipids to occupy locations on the outer surface of the vesicle in contact with the water,⁸⁸ and this would explain why the environment of the nitroxyl radical in this situation appears to be so polar. Naturally occurring bilayers, however, rarely have such high curvature. The tension within such small vesicles produced by sonication seems to be significant. When the kinetic barrier is overcome by adding appropriate catalysts, small vesicles of phospholipid spontaneously fuse among themselves to produce much larger single-walled structures.⁸⁹

The **paradox of the cross-sectional areas** has been phrased in terms of the structure of bilayers of phospholipids from natural sources because this is the most critical situation. Natural phospholipids have unsaturated fatty acids that necessarily disrupt the hydrocarbon in the core of their bilayers (Figure 14–2). Consistent with the stereochemical consequences of the *cis* double bonds, the most obvious discontinuity in the plot of the order parameter *S* against position (Figure 14–10) occurs after the sixth carbon of the palmitate on 1-palmitoyl-2-oleoylphosphatidylcholine (see Figure 14–2A). A similar, although a much less abrupt, discontinuity, however, also seems to be present in the plot for bilayers formed from fully saturated phospholipids such as dipalmitoylphosphatidylcholine. Even in this case, the disorder increases most precipitously beyond the eighth carbon.

It has been argued that there is no stereochemical paradox associated with the first seven saturated carbons of the fatty acyl groups in a bilayer of phospholipid in which there is the normal complement of unsaturated fatty acyl groups if the carbon–carbon bonds connecting these carbons can support enough *gauche* configurations to make this region fluid enough to fill the volume allotted to it.⁹⁰ At first glance, this would seem difficult to accomplish. The difference in standard free energy between a *trans* and a *gauche* configuration in a linear

At the maximum, the width of the bilayer has increased by 19%.⁹⁸ At the same time, however, the **cross-sectional area** for each molecule of phosphatidylcholine in a monolayer of the bilayer decreases from 0.62 to 0.48 nm², if it is assumed that each molecule of cholesterol contributes 0.37 nm² to the surface area.⁹⁸ These are the changes expected if cholesterol straightens the posture of the alkane in the regions of the bilayer proximal to the glyceryl groups. The minimum value of 0.48 nm² for the cross-sectional area is not far from the value of 0.40 nm² for the cross-sectional area of a pair of hexagonally arrayed all-*trans* linear alkanes in a solid paraffin. Because the molecules of cholesterol are spacing the molecules of phosphatidylcholine, the phosphocholine head groups are more widely separated, and the steric effects of their hydration are no longer significant. If the distance between the esters in a bilayer of dioleoylphosphatidylcholine is 3.2 nm, the width of the hydrocarbon should be 3.0 nm.⁴⁷ If this bilayer is representative of one composed of only phospholipid and if the addition of the normal amount of cholesterol increases the width of the hydrocarbon in such a membrane of phospholipids by 20%, the width of the hydrocarbon in a membrane of phospholipids and cholesterol in a cell should be about 3.6 nm.

There is evidence from calorimetric studies,⁷¹ X-ray diffraction,⁹⁹ deuterium nuclear magnetic resonance,¹⁰⁰ and pressure-area functions of monolayers^{101,102} that complementary interactions occur between cholesterol and phospholipids, causing them to segregate into **distinct phases**. For example, between mole fractions of 0.08 and 0.28 mole % cholesterol at 30 °C, the lipids in a bilayer of dimyristoylphosphatidylcholine and cholesterol separate into a phase enriched in cholesterol and a phase depleted in cholesterol. Above 0.28 mole %, the cholesterol and the dimyristoylphosphatidylcholine are miscible and form a single phase.¹⁰³ Below the melting point of pure dimyristoylphosphatidylcholine, the phase enriched in cholesterol remains fluid while the phase depleted in cholesterol with which it coexists solidifies. The phases enriched in cholesterol that separate from mixtures of cholesterol and phospholipid have volumes that are smaller than the sum of the volumes of the separate components¹⁰¹ and have much broader phase transitions.⁷¹ In these distinct phases, the alkane of the phospholipid is more ordered than it is in the absence of cholesterol but reorients more rapidly,¹⁰⁰ results consistent with a decrease in the frequency of *gauche* conformations and the elimination of any tilting of that alkane.

These separate phases enriched in cholesterol usually have distinct molar ratios between the two components. The **stoichiometry between cholesterol and phospholipid** in these phases varies depending on the type of phospholipid with which the cholesterol is mixed.¹⁰² In mixtures between synthetic phospholipids and cholesterol the mole fraction of cholesterol in one of these separate phases is usually between 0.25 and

0.4,^{99,100} but with mixtures of natural phospholipids it may be higher, judging from the normal ranges in the cholesterol composition in biological membranes.

One possibility is that the stoichiometry between cholesterol and phospholipid established in one of these phases for a particular phospholipid or mixture of phospholipids is determined by the adjustment of volumes within the bilayer that is accomplished upon the dissolution of the cholesterol.¹⁰¹ The difference in cross-sectional area between the fused rings and the isoprenoid of the cholesterol cancels the imbalance between the cross-sectional area for the regions of alkane proximal to the glyceryl groups and the cross-sectional area for the hydrocarbon in the distal region beyond the eighth carbons of the fatty acyl groups. To effect this cancellation completely, there should be an optimal molar ratio between cholesterol and phospholipid. The fact that, within these separate phases formed between phospholipid and cholesterol, the molecules of cholesterol are evenly spaced, with each molecule of cholesterol surrounded by about four molecules of phospholipid,¹⁰⁴ is consistent with the adjustment of the volumes being the force establishing the stoichiometry and the very existence of these phases.

The distribution of electron density across a bilayer of amphipathic lipids in a membrane from a eukaryotic cell, represented by an equimolar mixture of cholesterol and phosphatidylcholine (Figure 14-4D), displays three regions.

First, the two symmetrical boundaries of high electron density, formed by the hydrophilic head groups, the glyceryl group, and the esters, sandwich the hydrocarbon and provide interfaces compatible with the water on either side. The distance between the two maxima of electron density that designate these two interfaces in a bilayer of phospholipid and cholesterol is 5.0–5.4 nm.^{47,98} In a natural membrane, these surfaces are irregular (Figure 14-5) and are formed by the phosphocholines, phosphoethanolamines, phosphoserines, phosphoinositols, and oligosaccharides. These surfaces are constantly changing in appearance owing to the fluid state of the bilayer.

Second, the two symmetric regions of hydrocarbon proximal to the glyceryl groups, formed by the first seven saturated carbons of the fatty acyl chains (Figure 14-2) and the fused rings of the cholesterol (**14-11**), have a lower electron density (Figure 14-4D) only because they are hydrocarbon. They are densely packed, with the alkane of the fatty acyl groups predominantly in its fully extended, all-*trans* configuration aligned normal to the plane of the bilayer, supported by the fused rings of the cholesterol and in turn spacing the molecules of cholesterol in a fairly uniform distribution. In one of the monolayers of a bilayer formed from an equimolar mixture of cholesterol and natural phosphatidylcholine, the cross-sectional area for each phospholipid is 0.48 nm², if the cross-sectional area for each cholesterol is 0.37 nm². The

width of each of the two symmetrically displayed proximal regions of hydrocarbon is about 1 nm. They commence at the level of the two acyl carbons attached to the glycerol and extend into the bilayer to the level at which the unsaturation of the fatty acids and the isoprenoid tail of the cholesterol commence.

Third, within these two symmetric boundaries, the central core of the bilayer contains the unsaturated hydrocarbon of the fatty acids and the disordered alkane of the fatty acids and the isoprenoid tail of the cholesterol. It has the lowest electron density (Figure 14-4D) because the disorder of the hydrocarbon in this region increases the frequency at which vacant space is encountered. It is believed that the hydrocarbon in the central core of the bilayer has most of the properties of liquid paraffin. The width of the central core, about 1.5 nm, brings the width of the entire sheet of hydrocarbon to about 3.6 nm.

When an amphipathic lipid, such as natural phosphatidylcholine from eggs of *G. gallus*, is spread at an **interface between air and water**, it forms a **monolayer** with its hydrophilic functional groups directed toward the water and its hydrophobic hydrocarbon directed toward the air. The area for each phospholipid in this monolayer is a function of the **surface pressure**. This pressure is exerted mechanically by changing systematically the area of the surface, and it is measured with a torsion balance. Above a certain pressure, or in other words below a certain area, the monolayer becomes so compressed that phospholipid molecules leave it and form small patches of bilayer adhering to the monolayer.

Before this breakdown, however, the area for each molecule of phospholipid at the interface is a monotonic inverse function of the surface pressure (Figure 14-11).^{105,106} The explanation for this behavior is that, at low pressure, the tendency of the hydrocarbon to be maximally disordered causes the monolayer to have a large surface area, which is about 3-fold greater than that of hexagonally packed linear alkane oriented normal to the interface. The molecules of phospholipid, at zero pressure, do not lie flat upon the surface, presumably because this would bring all of their hydrophobic hydrocarbon into contact with water. The observed **surface area at zero pressure** is a balance between the entropy that would spread the lipids and the hydrophobic effect that would contract them. As the surface is compressed and its free energy is thereby increased, the hydrocarbons become more and more aligned, and of lower and lower entropy. The surface area of a molecule of phosphatidylcholine in a normal fluid bilayer of natural phosphatidylcholine is about 0.7 nm^2 , which corresponds to a surface pressure of about 37 dyn cm^{-1} in a monolayer at an air-water interface.

The same measurements can be made on a monolayer of natural phosphatidylcholine at an interface between an alkane, such as *n*-hexadecane, and water (Figure 14-11).¹⁰⁶ At each surface pressure, the mono-

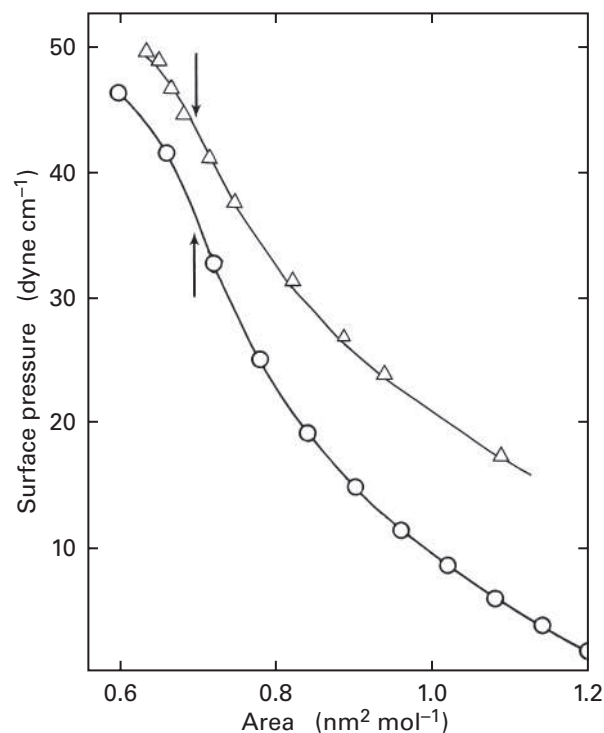


Figure 14-11: Relationship of surface pressure (dynes centimeter⁻¹) and area [nanometers² (mole of phosphatidylcholine)⁻¹] for monolayers of phosphatidylcholine purified from eggs of *G. gallus* at an interface between air and water (○)¹⁰⁵ or between *n*-hexadecane and water (△).¹⁰⁶ Phospholipid was spread at the interfaces from a solution in *n*-hexane; and, in the case of the interface with air, the hexane evaporated immediately, leaving behind the monolayer of lipid. In each case, a monolayer of phosphatidylcholine was produced. The area of the interface could be varied by movable boundaries, and the interfacial pressure could be measured directly by a torsion balance. Areas for a molecule of phospholipid were calculated on the assumption that all of the phospholipid added to the system had been incorporated into the monolayer. Arrows mark areas of $0.7 \text{ nm}^2 \text{ mol}^{-1}$. Adapted with permission from refs 105 and 106. Copyright 1960 Biochemical Society and 1971 Springer-Verlag.

layer in this situation has a greater surface area than when it is backed by air. The reason for this is that, because of van der Waals forces, liquid alkane is more compatible with the hydrocarbon side of the monolayer than is air; and when the monolayer is backed by alkane, there is not an interfacial free energy causing it to contract spontaneously and minimize its surface area at the interface between hydrocarbon and air as well as at the interface between phospholipid and water. Consequently, the pressure needed to compress this monolayer to 0.7 nm^2 (molecule of phosphatidylcholine)⁻¹ is 44 dyn cm^{-1} .

A droplet of alkane in water has a surface tension of about 50 dyn cm^{-1} , which reflects the free energy of the hydrophobic effect. As amphipathic lipid is added to a droplet of alkane, its surface tension rapidly decreases and reaches zero before the mole fraction of the amphipathic lipid reaches 1.¹⁰⁷ When its surface tension reaches zero, the surface area of the droplet will begin to

expand indefinitely, and at a mole fraction of amphipathic lipid equal to 1, it will be a bilayer.¹⁰⁸ That the **surface tension of a bilayer of amphipathic lipid** is zero has been verified experimentally.¹⁰⁶ The initial surface tension of the droplet of alkane is a direct measurement of the cohesive force of the hydrophobic effect. In the bilayer, this cohesive force is still in operation, trying to minimize its surface area, but it is counterbalanced by the hydration of the head groups.

From these various considerations, it follows that a bilayer of amphipathic lipids immersed in an aqueous solution represents a **compromise among a number of forces**. The hydrocarbon of the fatty acyl groups is hydrophobic and is withdrawn as successfully as possible from contact with the water by the cohesive force of the **hydrophobic effect**. The most successful stereochemical solution to this problem would be all-*trans* alkane in the regions proximal to the water oriented normal to the surface in hexagonal array. The presence of *cis* double bonds, the **steric effects of hydration**, and the **entropy of the liquid state** defeat this solution (Figures 14-10 and 14-11) and cause the cross-sectional area for each amphipathic lipid to be greater than the value of 0.40 nm² for hexagonal packing. This stereochemically enforced spreading of the bilayer must expose some of the hydrocarbon in the regions proximal to the water.⁸⁸ The hydrophilic head groups of the amphipathic lipids are facing the aqueous phase. If they were buried or sterically excluded from contact with water, their **free energies of hydration** would be lost, which would be unfavorable (Figures 5-8 and 5-18). In the compromise among the various forces, a bilayer of unadulterated natural phosphatidylcholine ends up with about half the surface area (0.7 nm²) for each molecule in one of its two monolayers as that for a molecule in a monolayer of phosphatidylcholine at an air-water interface (Figure 14-11) at zero surface pressure. Presumably, this difference between a bilayer and a monolayer arises from the greater cohesion, due to van der Waals forces, that can be established within an interior of liquid hydrocarbon as opposed to a thin layer of hydrocarbon at an interface with air and from the fact that there are two interfaces with the water, one on each side of the bilayer.

A bilayer represents one example from a spectrum of different **structures that can be formed by amphipathic compounds** such as amphipathic lipids, soaps, and detergents. An amphipathic compound usually contains one or more hydrocarbons that are each covalently attached at one of their ends to the others and to one or more hydrophilic functional groups. When an amphipathic compound is added to an aqueous solution, it forms noncovalent, multimolecular complexes referred to as either micelles or bilayers. In these complexes, all of the hydrophilic functional groups of the constituent molecules reside on the surface at the interface or interfaces with the aqueous phase so that they can be hydrated by the water. The hydrocarbon occupies the interior of the

complex sequestered from the aqueous phase by the hydrophobic effect. The **molar volume of the hydrocarbon** in the complex is determined simply by the partial molar volume of the hydrocarbon from which it is composed.

The final **molar surface area** at an interface of the complex with the aqueous phase, however, is determined by the balance between two opposing forces.^{23,109} The hydrophilic functional groups have an inescapable atomic cross-sectional area for their covalent structure that forces them to be spaced at least a minimum distance apart on the surface of the complex. This spacing is increased by the layers of hydration that are noncovalently associated with each hydrophilic functional group and any mutual electrostatic repulsion driving them apart. The farther apart the hydrophilic functional groups are spaced to relieve these repulsive forces, the more of the hydrocarbon to which they are covalently attached is drawn out to the surface to come in contact with water. This exposure of the hydrocarbon to water is resisted by the hydrophobic effect. The balance between the intermolecular repulsion among the hydrophilic functional groups and the hydrophobic effect determines the ultimate molar surface area of the complex.

There is a third geometric constraint on the complex. Because every hydrocarbon is covalently attached to one or more hydrophilic functional groups, and every hydrophilic functional group must remain in contact with the aqueous phase, no carbon in the interior of the complex can be located more than a **maximum distance from the aqueous phase**. If the hydrocarbon were fully extended linear alkane with all of its carbon-carbon bonds in the all-*trans* conformation, that maximum distance would be the maximum length of the amphipathic molecule. A certain amount of the hydrocarbon, however, having been dragged out of the interior by the repulsive forces among the hydrophilic functional groups, is required to occupy the surface of the complex, and the hydrocarbon in the interior is rarely fully extended because it is fluid and because it must mix to fill the interior. Therefore, the maximum distance any carbon can be from the aqueous phase is significantly less than the maximum length of the fully extended hydrocarbon found in the amphipathic molecule. One dimension of the complex formed from molecules of the amphipathic compound must always be less than or equal to twice this maximum distance. If it were not, the complex would contain a region farther from the aqueous phase than any matter can be located.

The dimensions of the complex that an amphipathic compound can form are dictated by this maximum dimension. The **shapes** available are a sphere the radius of which is less than or equal to the maximum dimension; an ellipsoid of revolution, prolate or oblate, the minor axis of which is less than or equal to the maximum dimension; a cylindrical rod of indefinite length the diameter of which is less than or equal to the maxi-

mum dimension; a bilayer of indefinite area the width of which is less than or equal to the maximum dimension; or cylindrical rods, ellipsoids of revolution, or spheres of water embedded uniformly in a volume otherwise filled with the amphipathic compound and spaced such that no distance between two adjacent surfaces of these aqueous inclusions is greater than the maximum dimension.

The choice among these different geometric alternatives in a given situation is determined by the ratio between the molar surface area, which is determined independently by the repulsion among the hydrophilic functional groups and the hydrophobic effect, and the molar volume, which is determined by the molecular structure of the particular amphipathic compound. The hydrocarbon of the compound must also be able sterically to fill the volume allotted to it by a particular shape;⁸⁸ certain volumes are too anisotropic to be filled by real hydrocarbon, which is made from atoms of carbon and hydrogen joined by covalent bonds of precise bond angle and bond length and around which only particular rotations are permitted, even though these same volumes can be filled with imaginary hydrocarbon, which is a uniform continuum that can fill any shape drawn on a sheet of paper.

Several examples will illustrate the outcome of the competition among the various free energies. When one of the fatty acyl chains is removed from phosphatidylcholine to form **lysophosphatidylcholine**, the product forms **ellipsoidal micelles**²³ rather than bilayers because the internal volume of a bilayer about half the width of that formed by phosphatidylcholine that would be dictated by the appropriate molar surface area and molar volume cannot be filled uniformly by the hydrocarbon available, but an ellipsoidal micelle, with its smaller and more isotropic volume for each unit of surface area, can be filled readily. When suspended in 6% ethanol in water at 40 °C, distearoylphosphatidylcholine forms a phase in which the alkyl chains of its fatty acids interdigitate, rather than butting up against each other, to produce a narrower bilayer.^{110,111} The ethanol promotes the increased exposure of the hydrocarbon to the water required by the increased cross-sectional area for each phospholipid. Dodecyl sulfate in 0.3 M lithium chloride forms spherical micelles because the electrostatic repulsion among the sulfates is sufficient to produce the largest possible ratio of molar surface area to molar volume and the linear, fully saturated hydrocarbon is flexible enough to fill the appropriate spherical volume uniformly.¹¹²

A heterogeneous mixture of phospholipids extracted from mammalian brain, when hydrated at 37 °C to a low content of water ($\leq 20\%$), forms a **reversed hexagonal phase**¹¹³ in which parallel cylinders of water spaced 4.5 nm apart are embedded in a volume otherwise filled with phospholipid.^{114,115} In this case, the poorly hydrated hydrophilic functional groups of the

phospholipids produce such a small molar surface area that when combined with the unavoidable molar volume of the acyl groups, it would produce a bilayer wider than the maximum dimension permitted these phospholipids. Pure phosphatidylethanolamine, presumably because its head group is more compact than that of phosphatidylcholine, forms bilayers under certain circumstances and reversed hexagonal phases under other circumstances. The reverse hexagonal phase of phosphatidylethanolamine becomes more stable relative to the bilayer as the temperature is raised, the length of the fatty acyl groups is increased, the unsaturation of the fatty acyl groups is increased, or when the fatty acyl groups are branched.¹¹⁶ All of these alterations increase the cross-sectional area of the hydrocarbon and favor the reversed hexagonal phase with its lower ratio of surface area in contact with water to mean cross-sectional area for hydrocarbon.

The fact that the particular phospholipids synthesized by living organisms form bilayers spontaneously rather than one of these other structures is as much a result of evolution by natural selection as the fact that the polypeptides synthesized by living organisms happen to fold.

Suggested Reading

- Wiener, M.C., & White, S.H. (1992) Structure of a fluid dioleoylphosphatidylcholine bilayer determined by joint refinement of X-ray and neutron diffraction data. III. Complete structure, *Biophys. J.* 61, 434–447.
- Hubbell, W.L., & McConnell, H.M. (1971) Molecular motion in spin-labeled phospholipids and membranes, *J. Am. Chem. Soc.* 93, 314–326.

The Proteins

Even a homogeneous suspension of biological membranes, highly purified by a series of centrifugations so that only identical membranes from the same source within the cells are present, contains a diverse collection of proteins. These proteins fall into several categories. All membranes when they are present in the cell are closed continuous sacs, containing solutions of soluble proteins. Even if the final suspension of purified membranes has been submitted to lysis and centrifugation, **entrapped soluble proteins** may still be enclosed in small vesicles of membrane and contaminate the preparation. For example, it is difficult to obtain from erythrocytes a suspension of plasma membranes completely devoid of hemoglobin.

Peripheral membrane-bound proteins¹¹⁷ are proteins that are not physically embedded in the bilayer of phospholipid but are associated with the membrane either through interfaces with proteins that are embedded in the bilayer or through superficial interactions with the bilayer of phospholipid of the membrane. Peripheral

membrane-bound proteins can be dissociated by treatments that do not dissolve the bilayer of the membrane. If they are associated with more firmly attached proteins, they can often be removed from the membrane by mild treatments that are normally used to dissociate the subunits of multimeric proteins. Examples of such treatments are increasing or decreasing either the pH or the ionic strength, removing divalent cations, using mild denaturants, or some combination of these treatments.^{118–120} For example, the cytoskeletal proteins spectrin and actin can be released from the plasma membranes of erythrocytes by chelating divalent cations.¹²¹

Some peripheral membrane-bound proteins **associate directly with the head groups** of the phospholipids in a bilayer. For example, in the presence of Ca^{2+} , isoform V of annexin associates at diffusion-controlled rates ($1 \times 10^{10} \text{ M}^{-1} \text{ s}^{-1}$) with large unilamellar vesicles of phospholipid¹²² and releases from these vesicles rapidly upon chelation of the Ca^{2+} . In contrast to annexin, which displays no preference for the head group of the phospholipid,¹²² the peripheral association of protein kinase C with vesicles of phospholipid requires that they contain significant concentrations of phosphatidylserine,¹²³ and the dissociation constant between the enzyme and the phospholipid bilayer decreases as the concentration of 1,2-diacylglycerol in the membranes is increased.^{124,125} That the interaction of protein kinase C is mainly with the hydrophilic surface of the bilayer is suggested by the fact that the interaction requires the naturally occurring enantiomers of both the phosphatidylserine and the diacylglycerol.¹²⁶ That no association occurs with bilayers formed from the unnatural enantiomers demonstrates that it is not the surface charge of the bilayer that is being recognized by protein kinase C but the head groups themselves. The pleckstrin domain of phospholipase C binds specifically to one molecule of phosphatidylinositol 4,5-bisphosphate within a bilayer of phospholipid,¹²⁷ and prothrombin and protein Z both also seem to bind to the head group of only one of the phospholipids in a bilayer.¹²⁸

In contrast to these proteins that recognize the head groups of the phospholipids specifically, the peripheral association of choline-phosphate cytidylyltransferase with a bilayer of phospholipid requires only that its **surface charge** be negative because any negative phospholipid promotes its binding.¹²⁹

An **anchored membrane-bound protein** is a protein a portion of whose primary covalent structure, uninvolved in its function, is immersed within the hydrocarbon of the bilayer of phospholipid and serves only to anchor the protein to the membrane. The portion of the protein embedded in the bilayer is not engaged in the native structure of the globular domain or globular domains to which it is covalently joined and can often be removed endopeptidolytically or by genetic manipulation to produce a protein that is soluble and that still displays all of the functions of the membrane-bound form.

Often the embedded portion in an anchored membrane-bound protein is a short segment of polypeptide at its **amino terminus or carboxy terminus** that appears to have been tacked on to an otherwise soluble protein to confine it to the surface of the membrane. Carboxypeptidase E is attached to the membranes of secretory granules from the adrenal medulla through an **amphipathic α helix** (see Figure 6–8) about 20 aa in length at its carboxy terminus.¹³⁰ The hydrophobic surface of this α helix is submerged in the hydrocarbon of the bilayer of phospholipid. A more common type of anchor, however, is a segment of polypeptide at one of the termini that spans the bilayer of the membrane. For example, bovine polypeptide *N*-acetylgalactosaminyltransferase has the amino-terminal sequence MRKFAYCKVVLATSLIWVLLDMFLLLYFSECNKCDEKKER–.¹³¹ The segment from Valine 9 to Phenylalanine 28 spans the bilayer of phospholipid that forms one of the Golgi membranes to which the protein is anchored, and the rest of the protein is a normal water-soluble, globular structure.¹³² This enzyme is a member of a large group of glycosyltransferases, each of which is anchored to the Golgi membranes.¹³³ Cytochrome c_1 is a cytochrome that is anchored in the mitochondrial membrane by a **membrane-spanning segment** at its carboxy terminus. When this segment is removed, a completely soluble form of the cytochrome is produced that is functionally intact.¹³⁴ Cytochrome c_1 , however, in addition to possessing the carboxy-terminal anchor, is also, under normal circumstances, a subunit of ubiquinol-cytochrome-*c* reductase, a large, heterooligomeric, membrane-spanning complex.

Many proteins are directed to certain organelles in the cell by amino-terminal **signal sequences**. The signal sequence directing bovine dopamine- β -monooxygenase to secretory granules of the adrenal medulla is a hydrophobic segment 20 aa in length, and if the signal sequence is not removed as it normally is, it becomes a membrane-spanning segment anchoring this protein in the bilayer of phospholipid.¹³⁵

Sometimes a **separate domain** at one of the termini of a protein is responsible for anchoring it within the bilayer of phospholipid. For example, the domain of 60 aa at the amino terminus of the receptor Tom 20 is an anchor embedded in the outer mitochondrial membrane,^{136,137} and the domain of 103 aa at the carboxy terminus of 3-hydroxybutyrate dehydrogenase¹³⁸ is an anchor embedded in the inner mitochondrial membrane; each of these domains anchors the otherwise soluble protein into its respective membrane. Hydroxymethylglutaryl-CoA reductase, which is an anchored membrane-bound protein by virtue of the fact that a detachable domain with full catalytic activity can be removed from the membrane by endopeptidolytic cleavage, has an embedded anchor almost 400 amino acids in length containing seven hydrophobic segments of greater than 20 aa each.¹³⁹

The segment of the polypeptide anchoring a protein in a bilayer of phospholipid can also be in its interior.¹⁴⁰ The amino acid sequence between Methionine 177 and Alanine 229 in (S)-mandelate dehydrogenase from *Pseudomonas putida*, a protein 393 aa in length, anchors it in the plasma membrane of the bacterium. When this segment is replaced by sequence 20 aa in length that occupies the homologous location in (S)-2-hydroxy-acid oxidase, a closely related but completely soluble protein, the resulting chimera no longer associates with the membrane, is fully active, and can be readily crystallized.^{141,142} The three respective **interior segments** of coagulation factor VIII¹⁴³ and coagulation factor V¹⁴⁴ responsible for anchoring each of them in a bilayer of phospholipid form loops directing hydrophobic side chains into the hydrocarbon of the membrane.

Proteins that have been posttranslationally modified with glycosylphosphatidylinositol (Figure 3-17) are anchored in their respective membranes by the covalently attached phosphatidylinositol, which spontaneously takes its place within the bilayer of phospholipid.^{145,146} The set of **glycosylphosphatidylinositol-linked (GPI-linked) proteins** is a heterogeneous collection. Their respective functions are unrelated to each other, so their common mode of attachment is probably fortuitous. They are, however, all inserted into the extracytoplasmic surfaces of the plasma membranes of the respective cells in which they are found. Immediately after they have been synthesized, these proteins are anchored temporarily in the membrane by a carboxy-terminal segment of their polypeptide that is rich in hydrophobic amino acids. The ultimate carboxy terminus of the posttranslationally modified protein is a glycine, alanine, cysteine, serine, or asparagine 15-30 amino acids in from the initial carboxy terminus. Through an enzymatically catalyzed transamidation, the carboxy terminus of this amino acid is transferred from the carboxy-terminal segment of 15-30 amino acids to which it was originally attached to the amine of the ethanolamine phosphate connected through the oligosaccharide to the phosphatidylinositol (Figure 3-17).¹⁴⁷

A glycosylphosphatidylinositol-linked protein can be identified by its ability to be released from the surface of the cell by glycosylphosphatidylinositol diacylglycerol-lyase¹⁴⁸ or by site-directed mutation.¹⁴⁹ Once released, the resulting globular, soluble protein can be purified and crystallized.^{149,150} In fact, glycosylphosphatidylinositol-linked proteins are often isoforms from species of proteins the other members of which are water-soluble and make no contact with membranes. Examples of such glycosylphosphatidylinositol-linked proteins are the variant surface glycoprotein on the exterior of cells of *Trypanosoma brucei*,¹⁵¹ the receptor for the F_c domain of immunoglobulin G on the exterior surface of neutrophilic lymphocytes,^{152,153} one isoform of acetylcholinesterase on the exterior surfaces of several types of animal cells,¹⁵⁴⁻¹⁵⁶ isoform IV of carbonate dehydratase on the exterior surfaces of cells from lung and

kidney,^{157,158} and the cell surface glycoprotein a-2 on the exterior surfaces of hematopoietic cells.^{159,160} Sometimes two isoforms of one of these proteins, the glycosylphosphatidylinositol-linked isoform and an isoform anchored to the membrane by a carboxy-terminal membrane-spanning segment of polypeptide, are produced in the same tissue.^{160,161}

Each of the various types of anchor is **more or less firmly embedded** in the bilayer of phospholipid. The segments of polypeptide at the amino or carboxy termini that span the bilayer, either once or several times, are permanently affixed to it. Amphipathic α helices, either at one of the termini¹³⁰ or in the middle of the protein,¹⁴² are much less firmly embedded and can be dissociated under appropriate circumstances.¹³⁰ Proteins that dip loops of their polypeptide in the bilayer of phospholipid are also less firmly embedded, display requirements such as negative surface charge for competent association, and have measurable dissociation constants.¹⁶² A protein with an anchor of glycosylphosphatidylinositol is permanently embedded in the membrane because its covalently attached phosphatidylinositol, with two fatty acids of the normal length, is so hydrophobic that its dissociation constant from the bilayer of phospholipid is immeasurably small.⁶³ Proteins posttranslationally modified by isoprenylation, however, because they have only one covalently attached hydrocarbon (Figure 3-16), are less firmly anchored in the membrane. Those with a geranylgeranyl modification ($C_{20}H_{33}$) have dissociation constants^{163,164} for a bilayer of phospholipid of 0.1-40 μ M; and those with a farnesyl modification ($C_{15}H_{25}$), 1-150 μ M. The particular value of the dissociation constant depends on the number of basic amino acids in the carboxy-terminal sequence of the protein and the surface potential of the bilayer of phospholipid.

Proteins that have been posttranslationally modified by **acylation with fatty acids** such as palmitic acid, oleic acid, and stearic acid (Table 3-1)¹⁶⁵ are usually bound to the cytoplasmic surface of the plasma membrane.¹⁶⁶ They are bound tightly because these proteins usually have several sites of fatty acylation. In some cases, however, it is unclear whether such fatty acylation is directed almost exclusively to an otherwise membrane-bound protein after it has associated irreversibly with the membrane or the fatty acylation itself promotes the association of the protein with the membrane. For example, proteolipid protein from myelin is a membrane-bound protein with six cysteines in its amino acid sequence that are fatty acylated, but it also has several membrane-spanning segments.¹⁶⁷ The cytochrome subunit of the photosynthetic reaction center from *Rhodospseudomonas viridis*, however, has a diglyceride in ether linkage with its amino-terminal cysteine that does seem to be the only portion of the protein anchoring it in the membrane.¹⁶⁸

Proteins that are posttranslationally modified by **myristoylation** of their amino termini (tetradecanoyl in

Figure 3–16) may or may not associate with membranes. Whether or not they do seems to depend on their amino-terminal sequence and the accessibility of the myristate. The amino-terminal segments of several normally *N*-myristoylated proteins associate with bilayers of phospholipids even when they are not myristoylated^{169,170} either by forming an amphipathic α helix or by accumulating at membranes with negative surface potential.¹⁷¹ Other *N*-myristoylated proteins, however, lose their ability to associate with the membrane when they are not myristoylated.¹⁷² Yet other proteins either associate with membranes or fail to associate depending on whether or not their myristoyl groups are fully exposed on their surface or buried in their interior.¹⁷³ The variability in behavior is probably the result of the fact that the dissociation constant for just a myristoylated amino terminus from a bilayer of phospholipid is only about 0.1 μ M.¹⁷⁴

Many **enzymes** catalyze reactions in which phospholipids, steroids, other membrane-bound proteins, or other molecules embedded in a membrane are substrates. To find their substrates, these enzymes must associate with the membrane in which those substrates are located. Such enzymes often bind tightly to those membranes and process their substrates by **scooting** across their surfaces.^{175–177} They can be firmly anchored in the membranes with which they associate by embedding one or more segments of their polypeptides in the bilayer of phospholipid.¹⁷⁸ For example, the polypeptide between Isoleucine 50 and Asparagine 98 of prostaglandin-endoperoxide synthase forms four short α helices that are embedded in one of the monolayers of the bilayer of amphipathic lipids¹⁴⁰ forming the membrane. The membrane to which it is anchored contains the phospholipids to which are acylated the arachidonoyl groups that are the substrates for this enzyme. Enzymes that must move from one membrane to another to perform their function, however, are only loosely associated with the respective bilayers of phospholipids. For example, isoform 2 of sterol carrier protein associates desultorily with membranes by an amino-terminal segment forming two amphipathic α helices.¹⁷⁹ Enzymes that catalyze reactions with membrane-bound substrates are often anchored in the membrane only by a portion of their structure uninvolved in the actual catalysis. Unlike anchored membrane-bound proteins, however, they would be unable to perform their function were the anchor to be removed and they were no longer able to associate with a membrane, because they would be unable to find their substrates.

An **integral membrane-bound protein** is a protein a portion of the polypeptide of which is permanently embedded in the bilayer of phospholipid constituting its membrane, and that portion of its polypeptide within that membrane is essential for its function. There is a family of proteins, exemplified by the receptor for epidermal growth factor,^{180,181} the members of which contain only one short hydrophobic segment of their

polypeptide embedded within membrane. This segment is in the middle of the amino acid sequence^{182–184} and spans the bilayer of phospholipid once. On the two ends of this membrane-spanning segment, there are globular domains on the cytoplasmic and extracytoplasmic sides of the membrane. The role of these proteins is to transmit across the membrane to their cytoplasmic domains the information that a circulating hormone is bound to their extracytoplasmic domains. Upon receipt of the information, a protein tyrosine kinase, catalyzed by the cytoplasmic domain, is activated. Neither domain by itself is capable of displaying hormone-dependent protein tyrosine kinase activity,¹⁸³ so in this case, the single membrane-spanning segment performs a much greater role than merely anchoring the protein in the membrane.

Usually, however, a significant portion of the native structure of the polypeptide or polypeptides of an integral membrane-bound protein¹¹⁷ is **within the hydrocarbon of the bilayer** of amphipathic lipids forming the membrane. Integral membrane-bound proteins can never be detached in a functional form from the bilayer by endopeptidolytic cleavage or site-directed mutation and often lose their native structure or precipitate from solution or both when the bilayer is completely dissolved by detergents. As with any protein, the native structure of the portion of an integral membrane-bound protein that is within the membrane is determined by the solvent in which it is dissolved. This solvent is a sheet of liquid paraffin about 3.6 nm wide possessing covalently attached oligosaccharides, anions, and zwitterions at both of its surfaces and in contact on each of its surfaces with an aqueous solution.

Examples of integral membrane-bound proteins that are required by their function to have significant portions of their mass within the hydrocarbon itself are proteins that form **channels** through the membrane for the transport of polar, water-soluble metabolites across the membrane, proteins that catalyze the **active transport** of metallic cations and metabolites against their gradients of concentration across the membrane, large complexes of subunits that catalyze **electron transport** and the concomitant active transport of protons across the membrane, and proteins that change their conformation or oligomeric state upon the **reception of information** to transfer that information across the membrane. Integral membrane-bound proteins can also be enzymes the substrates for which are dissolved in and confined to the membrane.¹⁸⁵

Integral membrane-bound proteins vary in **size** from diacylglycerol kinase of *E. coli*, with a single folded polypeptide 121 aa in length,¹⁸⁵ to the ryanodine receptor of animal sarcoplasmic reticulum, with a single folded polypeptide about 5000 aa in length,¹⁸⁶ or NADH dehydrogenase (ubiquinone), a complex of 45 different subunits for a total of about 9000 aa.¹⁸⁷

The distinction between an anchored membrane-bound protein and an integral membrane-bound protein

is not a clean one. For example, a membrane-bound protein known as glycophorin is found in the plasma membrane of erythrocytes. The protein is 131 amino acids long, and its embedded anchor¹⁸⁸ is located to the carboxy-terminal side of Glutamate 72. This embedded anchor spans the plasma membrane.¹⁸⁹ The 35 carboxy-terminal amino acids on the cytoplasmic side of the membrane are rich in proline and seem to be structureless and functionless, simply acting as a barb that cannot be pulled across the membrane. The protein is about 60% carbohydrate by weight,¹⁸⁸ and this carbohydrate is entirely linked as oligosaccharides¹⁹⁰ to the extracytoplasmic amino-terminal portion of the protein through 15 *O*-glycosidic linkages to threonines and serines and one *N*-glycosidic linkage.¹⁸⁸ The function of glycophorin is to serve as the source of most of the oligosaccharide on the extracytoplasmic surface of the erythrocyte, so its function would be lost if its anchor were removed.

Membrane-bound proteins are inserted into natural membranes such that every copy of the same protein is oriented in the same direction relative to the cytoplasm. The earliest observations addressing this point explicitly confirmed this assumption of **vectorial insertion**.^{117,191,192} For example, nucleophilic amino acids in 10 of the thermolytic peptides on the peptide map of a digest of band 3 anion transport protein from human erythrocytes could not be modified with *N*-formyl-[³⁵S]sulfinylmethionyl methylphosphate, a polar reagent that cannot pass through an intact membrane, when the native protein was in sealed, intact erythrocytes, even though they could be readily modified when the erythrocytes were broken open.¹⁹¹ The explanation of this observation is that every copy of anion carrier is oriented the same way in the membrane, each presenting the same unique surface to the cytoplasmic space of the cell as well as a different, also unique face to the extracytoplasmic space, and that the cytoplasmic surface is inaccessible to the impermeant reagent in an intact cell. Since these early studies, many examples of vectorial insertion have been verified, and no example of a membrane-bound protein the copies of which are oriented at random in a natural membrane has been verified.

A property related to the vectorial insertion of every protein in a biological membrane is the asymmetric distribution of the oligosaccharides on the **glycoproteins** and glycolipids embedded in the plasma membranes of cells. Almost all¹⁹³ of the oligosaccharide bound to the plasma membrane of an animal cell is located upon its extracytoplasmic surface.^{194,195} This feature is a corollary of the fact that almost no¹⁹⁶ glycoproteins are found in the cytoplasm, only in extracytoplasmic spaces, and a direct result of the fact that the glycosyltransferases that synthesize the oligosaccharides on these glycoproteins are in the extracytoplasmic lumens of the Golgi membranes. Membrane-bound proteins are synthesized on ribosomes bound to the endoplasmic reticulum and incorporated in their proper orientation by the machin-

ery responsible for their insertion into the membranes of the endoplasmic reticulum. The membrane in which they have been incorporated is then sent to the Golgi membranes where the oligosaccharides are added. Then vesicles, which bud off the Golgi membranes and which maintain the vectorial orientation of the membrane-bound proteins and their attached oligosaccharide, transport them to the plasma membrane. These vesicles then fuse with the plasma membrane so that their extracytoplasmic surfaces, on which the membrane-bound oligosaccharides reside, and their extracytoplasmic lumens, in which the soluble glycoproteins are located, remain extracytoplasmic.

In addition to glycosylation, membrane-bound proteins are posttranslationally modified as often as are water-soluble proteins. For example, they can be modified to contain covalently attached coenzymes or they can be phosphorylated.

When a suspension of purified membranes is examined by electrophoresis in solutions of dodecyl sulfate, a large collection of different polypeptides, each present at its own characteristic concentration, is observed (Figure 14-12).¹⁹⁷ Each of these polypeptides is a component of one of the many native proteins bound to the membranes, and the protein it constitutes is responsible for a particular biochemical function. Therefore, a biological membrane, although often much less complex, resembles cytoplasm in being a **heterogeneous solution of a large number of different proteins**, each present at a different concentration and each with a specific function. Many of the functions performed by membrane-bound proteins have been identified, and biochemical assays have been developed for determining their presence and their concentration. In many instances, the protein responsible for one of these functions has been identified and purified, and its cDNA has been cloned and sequenced.

Because membrane-bound proteins are often more difficult to purify than soluble proteins, indirect procedures are often used to **identify the gene** encoding them to assist in their purification. Classical genetics can be used with bacteria and fungi to identify the gene encoding a membrane-bound protein responsible for a particular function.¹⁹⁸⁻²⁰¹ It is also possible to select a cDNA encoding a membrane-bound protein that is responsible for a particular function, such as the transport of a particular metabolite²⁰² or a change in the conductance of the membrane in response to a neurotransmitter,²⁰³ by screening the expression of a library of cDNA in oocytes of *Xenopus laevis*. If a cDNA has been identified before a membrane-bound protein has been purified, that cDNA can be modified to assist in its purification. For example, the protein can be expressed with a sequence of histidines at its carboxy terminus to permit its purification by affinity adsorption.²⁰⁴ If the protein is an anchored membrane-bound protein, a site for endopeptidolytic cleavage can be inserted to ensure that the protein can be

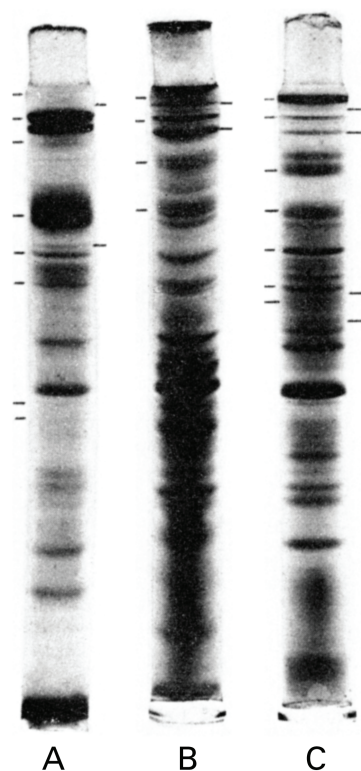


Figure 14-12: Polyacrylamide gels displaying the collection of polypeptides found in the plasma membranes of erythrocytes from *R. norvegicus* (A), the plasma membranes of cells from liver of *R. norvegicus* (B), and the portion of the plasma membrane of kidney cells from *R. norvegicus* referred to as the brush border (C).¹⁹⁷ Purified membranes from these various tissues were dissolved in a solution of sodium dodecyl sulfate, which unfolded and coated each polypeptide with a layer of dodecyl sulfate. The polypeptides were then separated by electrophoresis on gels of polyacrylamide cast in a solution of sodium dodecyl sulfate. Each band represents a different polypeptide. The dashes indicate polypeptides that are glycosylated. Reprinted with permission from ref 197. Copyright 1971 *Journal of Biological Chemistry*.

released from the membrane by digestion.²⁰⁵ In most cases, however, a membrane-bound protein of interest is purified before the cDNA for it is available.

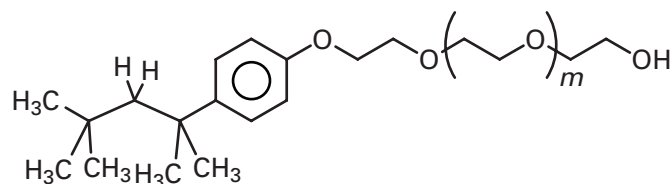
The purification of the particular membrane-bound protein identified by a biochemical assay proceeds in two stages. In the first stage, a biological source is chosen that contains the highest possible concentration of the protein. This involves assaying the biochemical activity in different tissues from different species or trying to increase the concentration of the protein in its active form by genetic manipulation of a microorganism or cultured eukaryotic cells. Membranes that contain the protein in high concentration are then separated by the procedures of **cell fractionation** from other membranes in the biological source that contain little or none of the protein. These purified membranes are lysed to release the entrapped soluble proteins and submitted to treatments that release peripheral membrane-bound proteins without inactivating or releasing the protein of

interest. The product of these manipulations is a suspension of membranes in which are embedded the protein of interest in the highest possible concentration. At the completion of this first stage, the membrane-bound protein being purified may be essentially homogeneous. For example, fragments of membrane the only protein of which (90%)²⁰⁶ is Na^+/K^+ -exchanging ATPase can be purified²⁰⁷ from a region of the mammalian kidney, the only function of which is to transport sodium and potassium. In the kidney these membranes are paved with this protein. Many of the membrane-bound proteins that have been purified to homogeneity are those that are already present at high density in such suspensions of appropriately purified and extracted membranes.

Once **membranes enriched in a particular protein** have been obtained, the next step in its purification is to release that protein from them. The few proteins that are linked to the membrane by glycosylphosphatidylinositol can be released by glycosylphosphatidylinositol diacylglycerol-lyase and then purified as water-soluble proteins. Anchored membrane-bound proteins can often be released from a membrane by mild endopeptidolytic digestion while retaining their full biological activity. The majority of membrane-bound proteins, however, cannot be released from the membrane so easily, and the second stage in their purification is usually to dissolve the membranes without unfolding the protein and then purify the dissolved protein as if it were a soluble protein. A non-ionic or zwitterionic detergent is used to **dissolve the membranes**, because nonionic or zwitterionic detergents bind only to the hydrophobic membrane-spanning portions of integral membrane-bound proteins and are unable to bind tightly all along a polypeptide and unfold it as does dodecyl sulfate.^{208,209} Ionic detergents, such as those used to wash laundry, are much harsher than non-ionic detergents, which are used for washing dishes by hand or for shampoos.

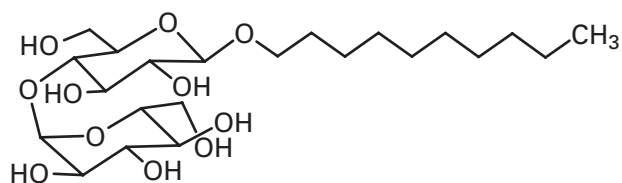
Nonionic or zwitterionic detergents are amphipathic compounds that have neutral or zwitterionic hydrophilic functional groups attached to one end of their hydrocarbon. A common class of **nonionic detergents**, the **alkyl oligo(ethylene oxide) ethers** (Brij series), is synthesized from linear, saturated primary alcohols produced commercially by the reduction of linear fatty acids 12–20 carbons in length. Because these fatty acids are usually from biological sources, the alcohols usually have an even number of carbons. Ethylene oxide is polymerized at random to the hydroxyl of one of these alcohols to form the detergent $\text{CH}_3(\text{CH}_2)_m\text{CH}_2\text{O}(\text{CH}_2\text{CH}_2\text{O})_n\text{H}$. The length of the alcohol ($m + 2$) is defined by the synthesis, but when the detergent is prepared for commercial use, the ethylene oxide is simply polymerized at random to produce a random mixture of hydrophilic extensions of different lengths within the same batch of synthetic detergent. As demand grew for structurally homogeneous detergents, the mixtures of these random polymers were separated chromatograph-

ically into their pure components to produce detergents such as $n\text{-C}_{12}\text{H}_{25}\text{O}(\text{CH}_2\text{CH}_2\text{O})_8\text{H}$ (abbreviated C_{12}E_8) and $n\text{-C}_8\text{H}_{17}\text{O}(\text{CH}_2\text{CH}_2\text{O})_4\text{H}$ (abbreviated C_8E_4). Another series of structurally heterogeneous oligo(ethyleneoxide) detergents are the **Tritons**:



14-12

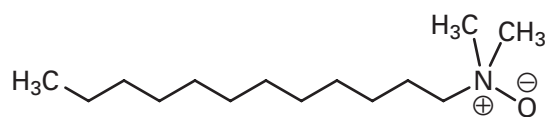
A broad class of nonionic detergents each member of which can be synthesized directly in pure form are the **alkyl glycosides**. A pure saccharide such as glucose or maltose is coupled synthetically to a pure long-chain alcohol such as octanol, decanol, or dodecanol in a glycosidic linkage at the carbonyl carbon of the saccharide. An example of such a structurally homogeneous detergent would be decyl $\beta\text{-D}$ -maltoside:



14-13

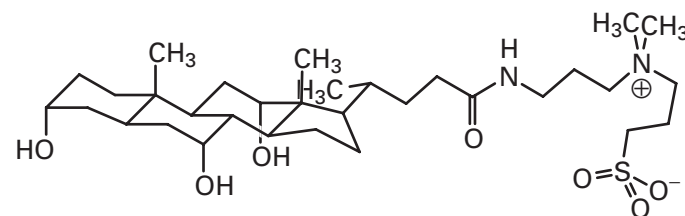
The exclusive coupling of the alcohol to the only carbonyl carbon of the saccharide in acetal linkage permits the direct synthesis of a detergent that is structurally homogeneous except at the anomeric carbon, and the two anomers can then be separated chromatographically. Either glucose or maltose can be chosen as the hydrophilic group and an alcohol of length between 8 and 14 carbons can be chosen as the hydrophobic group to generate a wide selection of different detergents. A set of related, naturally occurring detergents are the **saponins**, which are glycosides of triterpenes such as oleanolic acid. In the saponins, monosaccharides, disaccharides, and trisaccharides are coupled in acetal linkage to hydroxyls and carboxylates on the triterpene to produce biosynthetically a dramatically heterogeneous mixture.²¹⁰

A structurally homogeneous class of **zwitterionic detergents** that can be synthesized directly are the N -oxides of linear dimethylalkyl amines such as N,N -dimethyldodecylamine N -oxide:



14-14

3-[(3-Cholamidopropyl)dimethylammonio]-1-propane-sulfonate (CHAPS)



14-15

is a zwitterionic detergent that is synthesized from cholic acid, which itself is a mild ionic detergent, as is 7-deoxycholic acid.

Each of these detergents, the pure and the impure, because it contains only a single amphipathic compound or is a mixture of several related amphipathic compounds, forms elliptical micelles when it is dissolved in water. If the detergent is pure, however, the micelles that it forms in aqueous solution are of a uniform size (Table 14-5). Each detergent has a **critical micelle concentration** (Table 14-5). When its concentration is below the critical micelle concentration, the detergent is present in solution as free, independent molecules; and when its concentration is above the critical micelle concentration, there is a mixture of free molecules of detergent at the critical micelle concentration and micelles of detergent accounting for the excess over the critical micelle concentration. As the concentration of detergent is increased above the critical micelle concentration, the concentration of free detergent remains constant while the concentration of micelles increases.

At high enough concentrations, a nonionic detergent in aqueous solution is able to form **mixed micelles with phospholipids and cholesterol** and thereby dissolve membranes.²¹²⁻²¹⁴ There are several stages in this dissolution of a bilayer of phospholipid by a nonionic detergent.²¹⁵⁻²¹⁷ At low concentrations of free detergent, below its critical micelle concentration, there is a partition coefficient governing the distribution of detergent between the bilayer of phospholipid and the water, the membranes remain intact, and the detergent incorporates into the bilayer just as if it were an amphipathic lipid.²¹⁸ As the concentration of detergent is increased, its incorporation into the intact membrane begins to saturate, the apparent partition coefficient begins to drop, and the permeability of the membranes abruptly increases, as if fissures or pores were opening, but the bilayer of phospholipid still remains intact.^{217,218} As the concentration of the detergent is increased even further, the amount of bound detergent suddenly increases dramatically and the membranes dissolve and are replaced by mixed micelles of detergent and phospholipid.

These mixed micelles are completely formed before the free concentration of detergent reaches its critical micelle concentration in the absence of phospholipid

Table 14-5: Micelles of Nonionic and Zwitterionic Detergents²¹¹

detergent	mean aggregation number ^a	molar mass of micelle (g ⁻¹ mol ⁻¹)	critical micelle concentration (mM)
<i>n</i> -C ₈ H ₁₇ O(CH ₂ CH ₂ O) ₅ H	32	11,000	6.0
<i>n</i> -C ₁₀ H ₂₁ O(CH ₂ CH ₂ O) ₆ H	76	32,000	0.46
<i>n</i> -C ₁₀ H ₂₁ O(CH ₂ CH ₂ O) ₈ H			0.28
<i>n</i> -C ₁₂ H ₂₅ O(CH ₂ CH ₂ O) ₆ H	105	50,000	0.065
<i>n</i> -C ₁₂ H ₂₅ O(CH ₂ CH ₂ O) ₈ H	120	65,000	0.056
<i>n</i> -C ₁₄ H ₂₉ O(CH ₂ CH ₂ O) ₈ H			0.0052
<i>n</i> -C ₁₆ H ₃₃ O(CH ₂ CH ₂ O) ₈ H			0.00047
Triton X-100	140	90,000	0.21
<i>n</i> -C ₈ H ₁₇ N(CH ₃) ₂ O			200
<i>n</i> -C ₁₀ H ₂₁ N(CH ₃) ₂ O			7.5
<i>n</i> -C ₁₂ H ₂₅ N(CH ₃) ₂ O	76	17,300	0.4
<i>n</i> -decyl β-D-maltoside			1.4
<i>n</i> -dodecyl β-D-maltoside	98	50,000	0.14
<i>n</i> -octyl β-D-glucoside	84	25,000	25
<i>n</i> -decyl β-D-glucoside			4.2
<i>n</i> -dodecyl β-D-glucoside			0.14
CHAPS (14-15)	4-14	6000	6.2

^aMean aggregation number is the average number of molecules of detergent in a micelle.

(Table 14-5), an observation suggesting that the critical micelle concentration of the mixed micelles is lower than that of pure micelles of detergent. The **abrupt dissolution of the bilayer** of phospholipid to form these mixed micelles occurs at a fixed ratio of detergent to lipid rather than at a particular concentration of detergent,²¹⁶ an observation suggesting that there is an optimal ratio of detergent to phospholipid for the formation of these mixed micelles. An intermediate stage in some instances between a suspension of bilayers and a solution of elliptical mixed micelles seems to be the formation of long tubular micelles.²¹⁹

When natural membranes containing membrane-bound proteins are dissolved with a nonionic detergent, the same stages are passed through, and ideally, if the removal of the bilayer of phospholipid by this process does not unfold the protein being purified, the nonionic detergent forms a **toroidal micelle** surrounding the segments of the protein formerly embedded in the membrane and replacing the hydrocarbon of the bilayer of phospholipid with the hydrocarbon of the detergent. This toroidal micelle presents the hydrophilic functional groups of the detergent to the aqueous phase while its hydrocarbon surrounds and supports the hydrophobic portions of the protein formerly embedded in the membrane. The previously membrane-spanning segment or segments of polypeptide end up in the center of the toroid; the inner surface of the toroid is formed from the hydrocarbon of the detergent flush against the membrane-spanning segments of the protein, and the outer surface of the toroid is the hydrophilic portion of the detergent directed outward into the water (Figure 14-13). Such toroidal micelles have been crystallographically

observed surrounding molecules of integral membrane-bound proteins crystallized from solutions of the protein produced with detergent.²²⁰

As with phospholipids, there seems to be a required **ratio between the detergent and the protein and lipid** present in the original membrane for the protein to be dissolved completely.²²¹ This minimum ratio of concentrations is that required for there to be at least one

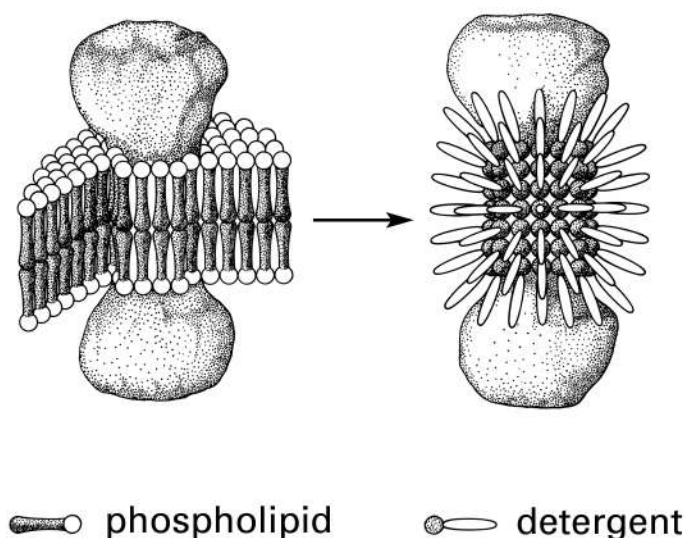


Figure 14-13: Diagrammatic representations of the mechanism by which a nonionic detergent dissolves an integral membrane-bound protein. A toroidal micelle is formed within which the bilayer of phospholipids is replaced by the hydrocarbon of the detergent. Figure courtesy of Steven Clarke, Department of Chemistry and Biochemistry, University of California at Los Angeles.

micelle of detergent, as measured in the absence of the membranes (Table 14-5), for each molecule of protein.²¹¹ Furthermore, there must be sufficient detergent in the solution to maintain its free concentration at a level equal to its critical micelle concentration in the absence of membranes (Table 14-5).

One peculiarity of these mixed micelles of protein and detergent is that at ratios of detergent where the protein is not completely dissolved and the solution contains micelles with one molecule of protein, micelles with two molecules of protein, micelles with three molecules of protein and so forth, these micelles neither fuse with each other nor dissociate into smaller aggregates.²²²⁻²²⁴ Once the ratio of these states of aggregation is established in the initial rapid dissolution, it remains fixed as long as the total concentration of detergent remains fixed.

There are several drawbacks to the necessity of dissolving the membranes in detergent to dissolve the proteins within them. The actual dissolution of the membrane occurs over a narrow range of free concentration of detergent,²¹⁷ so it is not possible to control the process very successfully. For reasons that are unknown, when ratios of detergent to protein increase beyond those needed to dissolve the membranes, the biological function of the protein is often impaired or disappears entirely.²²¹ Consequently, changes in the concentration of detergent or the ratio of detergent to protein that automatically occur during chromatography often lead to loss of activity. For this reason, it is often necessary to **screen a large number of different detergents** to discover by trial and error the one that preserves the biological activity of the protein over the broadest possible range of concentrations of both protein and detergent. Fortunately, there are a large set of homogeneous non-ionic detergents to explore with a wide range of critical micelle concentrations (Table 14-5).

Often it is difficult to find a detergent and the proper conditions to produce a monodisperse solution of the membrane-bound protein of interest without producing its inactivation,²²⁵ yet a **monodisperse solution of the protein** is essential if chromatography by molecular exclusion is to be used to purify it.²²¹ The large sizes of the micelles of the detergents limit the range of molecular sizes that can be separated by chromatography by molecular exclusion because no complex between a micelle and a molecule of protein can be smaller than the micelle itself. Many membrane-bound proteins are sialoglycoproteins so their charge is microheterogeneous, a situation causing difficulties for chromatography by ion exchange. For all of these reasons, the purification of a membrane-bound protein is more difficult than that of a soluble protein.

As with a soluble protein, a **biochemical assay** is used to follow the purification of a membrane-bound protein. If the protein is an enzyme, its enzymatic activity can be measured in a standard assay.²²⁶⁻²²⁸ If one of the sub-

strates for the enzyme is a phospholipid or another lipid incorporated in a membrane, the enzyme is reassociated with vesicles of phospholipid containing that substrate before assay.²²⁹ If the membrane-bound protein is responsible for transmitting across the membrane the information that a hormone is bound at its extracytoplasmic surface, the binding of that hormone or of a synthetic agonist or antagonist that binds tightly to the site at which the hormone binds can be used as an assay for that protein.²³⁰ If the membrane-bound protein is responsible for transporting a metabolite across the membrane, the binding of an inhibitor of that transport with high affinity for the protein can be used as an assay.²³¹

If a membrane-bound protein catalyzes the transport of a particular metabolite across its native membrane, it is also possible to reconstitute that protein into sealed vesicles of phospholipid and then assay the accumulation of that metabolite into the vesicles.²³²⁻²³⁴ The membrane-bound protein, detergent, and phospholipid are mixed together at concentrations necessary to produce a disperse solution. The detergent is then removed from that solution; and if the **reconstitution** has been successfully performed, sealed vesicles of phospholipid will form in the bilayers of which the protein of interest is inserted in a functional state. The ability of that protein to transport its specific substrate across the membrane into the vesicles is then assayed. Rapid methods for performing the separate reconstitutions of multiple samples can be used to assay the fractions from chromatographic separations for the protein responsible for the particular transport of interest. For example, proteins responsible for the transport of citrate,²³⁵ oxalate,²³⁶ and ornithine,²³⁷ respectively, could be purified to homogeneity by assaying them with reconstitution. There is even one instance in which an integral membrane-bound protein can be renatured and then reconstituted from a completely denatured state.^{238,239}

When the membranes have been dissolved in a solution of detergent in such a way that the activity of the membrane-bound protein of interest is preserved and an assay for that protein has been developed, the protein can often be purified chromatographically. For example, (*R*)-pantolactone dehydrogenase (flavin) was purified 80-fold from membranes of *Nocardia asteroides* by ammonium sulfate precipitation and six **chromatographic steps** after it was dissolved in a solution of 0.5% Brij 35.²⁴⁰ Both chromatography by molecular exclusion and chromatography by ion-exchange run in solutions of nonionic detergent, are used to purify membrane-bound proteins dissolved with detergent,^{241,242} as well as chromatography on hydroxyapatite^{243,244} and chromatography by adsorption on solid phases modified with specific functional groups (Table 1-2) such as phenylboronate²⁴⁵ or particular dyes.^{246,247} If one has been able to produce immunoglobulins specific for a particular membrane-bound protein before that protein has been purified, those immunoglobulins can be used to purify a mem-

Table 14–6: Secondary Structure Spanning the Bilayer in Crystallographic Molecular Models of Integral Membrane-Bound Proteins

protein	detergent or phase used during crystallization	n_{aa}^a	subunits ^b	secondary structure spanning membrane	number of membrane-spanning segments ^a
bacterial photosynthetic reaction center ^{321–323}	<i>N,N</i> -dimethyldodecylamine <i>N</i> -oxide ³²⁴	1200	$\alpha\beta\gamma\delta$	α helices	11
rhodopsins					
bacteriorhodopsin ^{325–328}	bicontinuous cubic phase of 1-oleoyl- <i>rac</i> -glycerol and water ³²⁹	250	α_3	α helices	7
halorhodopsin ³³⁰	bicontinuous cubic phase of 1-oleoyl- <i>rac</i> -glycerol and water	250	α_3	α helices	7
mammalian rhodopsin ³³¹	nonyl β -D-glucoside and 1,2,3-heptanetriol ³³²	350	α	α helices	7
complexes for electron transport					
mammalian cytochrome- <i>c</i> oxidase ³³³	<i>n</i> -C ₁₂ H ₂₅ O(CH ₂ CH ₂ O) _{<i>n</i>} H (Brij 35) ³³⁴ or decyl β -D-maltoside	1800	$(\alpha\beta\gamma\delta\epsilon\zeta\eta\theta\iota\kappa\lambda\mu\nu)_2$	α helices	28
cytochrome- <i>c</i> oxidase from <i>Paracoccus denitrificans</i> ³³⁵	undecyl β -D-maltoside	811	$\alpha\beta$	α helices	14
mammalian ubiquinol-cytochrome- <i>c</i> reductase ^{336–339}	dodecyl β -D-maltoside, decanoyl- <i>N</i> -methylglucamide, diheptanoylphosphatidylcholine, or octyl β -D-glucoside	2200	$(\alpha\beta\gamma\delta\epsilon\zeta\eta\theta\iota\kappa\lambda)_2$	α helices	13
bacterial cytochrome <i>o</i> ubiquinol oxidase ³⁴⁰	octyl β -D-glucoside	1290	$\alpha\beta\gamma\delta$	α helices	25
bacterial succinate dehydrogenase ³⁴¹	<i>n</i> -C ₁₂ H ₂₅ O(CH ₂ CH ₂ O) ₉ H (C ₁₂ E ₉)	1100	$(\alpha\beta\gamma\delta)_2$	α helices	6
bacterial lipid A export ATP-binding/permease protein MsbA ³⁴²	dodecyl α -D-maltoside	580	α_2	α helices	6
ion channels					
bacterial potassium channel KcsA from <i>Streptomyces lividans</i> ^{343,344}	decyl β -D-maltoside	160	α_4	α helices	8 ^d
acetylcholine receptor from <i>Torpedo marmorata</i> ^{345,346}	image reconstruction from tubular helical surface lattice	2333	$\alpha_2\beta\gamma\delta$	α helices	20
bacterial large-conductance mechanosensitive channel ³⁴⁷	dodecyl β -D-maltoside	150	α_5	α helices	10 ^d
aquaporin ^{348,349}	octyl β -D-glucoside or nonyl β -D-glucoside	270	α_4	α helices	7 ^e
mammalian endoplasmic reticulum Ca ²⁺ -transporting ATPase ^{350,351}	<i>n</i> -C ₁₂ H ₂₅ O(CH ₂ CH ₂ O) ₈ H (C ₁₂ E ₈)	1000	α	α helices	10
bacterial outer membrane porins					
porin ³⁵²	<i>n</i> -C ₈ H ₁₇ O(CH ₂ CH ₂ O) ₄ H (C ₈ E ₄) and <i>N,N</i> -dimethyldodecylamine <i>N</i> -oxide	300	α_3	β barrel ^c	16 ^f
malto porin ³⁵³	decyl β -D-maltoside ³⁵⁴	420	α_3	β barrel ^c	18 ^f
outer membrane protein F ³⁵⁵	<i>n</i> -C ₈ H ₁₇ O(CH ₂ CH ₂ O) _{<i>n</i>} H	340	α_3	β barrel ^c	16 ^f
sucrose porin ³⁵⁶	octyl β -D-glucoside	480	α_3	β barrel ^c	18 ^f
bacterial ferrichrome-iron receptor ^{357,358}	<i>N,N</i> -dimethyldodecylamine <i>N</i> -oxide	720	α	β barrel ^c	22 ^f
bacterial outer membrane protein A ³⁵⁹	<i>n</i> -C ₈ H ₁₇ O(CH ₂ CH ₂ O) ₄ H	325	α	β barrel	8 ^f
bacterial outer membrane protein TolC ³⁶⁰	mixture of hexyl, heptyl, octyl, and dodecyl β -D-glucosides	470	α_3	β barrel	12 ^g
bacterial α -hemolysin ³⁶¹	octyl β -D-glucoside	290	α_7	β barrel	14 ^g

^aTotal number of amino acids and total number of membrane-spanning segments of the noted secondary structure in each protomer of the protein unless otherwise noted. ^bComposition of subunits in complete oligomer. ^cAll β barrels are antiparallel. ^dNumber of α helices in the continuous cylinder of α helices formed by the complete oligomer. ^eOne of the membrane-spanning α helices is formed from two smaller α helices that butt against each other, each of which passes only halfway through the membrane. ^fNumber of strands in the continuous β barrel formed by each subunit. ^gNumber of strands in the continuous β barrel formed by the complete oligomer.

brane-bound protein by **immunoadsorption**²⁴⁸⁻²⁵⁰ because nonionic detergents usually cannot denature immunoglobulins.

After a membrane-bound protein has been purified to homogeneity so that only one protein remains in the sample, reconstitution is often used to prove that the purified protein is responsible for the biological activity of interest. Examples of some of the purified proteins that have been shown by reconstitution to be responsible for a specific function are ones that catalyze respectively the passive transport of water,²⁵¹⁻²⁵³ the passive transport of glucose,²⁴² the passive transport of melibiose,²⁵⁴ the passive transport of halide ions,²⁵⁵ the voltage-activated passive transport of sodium ions,²⁵⁶ the calcium-activated passive transport of potassium ions,²⁵⁷ the inositol 1,4,5-triphosphate-activated passive transport of calcium ions,²⁵⁸ the ATP-driven active transport of sodium and potassium ions,²³³ and the ATP-driven active transport of covalent conjugates between glutathione and other molecules²⁵⁹ across the membrane, as well as ones that form large, nonspecific pores.²⁶⁰ It is also possible to demonstrate that a particular protein is responsible for a particular type of transport by expressing mRNA encoding that protein in oocytes of *X. laevis* and then demonstrating that the oocytes have become able to display the particular type of transport.^{261,262}

In addition to reconstitution into sealed vesicles of phospholipid, purified membrane-bound proteins can be transferred into **bicelles**.²⁶³ Bicelles are small flat disks, each a bilayer of dimyristoylphosphatidylcholine. The rim of each disk is a continuous ring of detergent, either 3-[(3-cholamidopropyl)dimethylammonio]-2-hydroxy-1-propanesulfonate or dihexanoylphosphatidylcholine.^{264,265} The diameter of these circular disks can be varied by changing the ratio of detergent to phospholipid.²⁶⁵ It is also possible to replace the micelle of detergent surrounding the membrane-spanning portion of a purified membrane-bound protein with a copolymer of acrylate, *N*-octylacrylamide, and *N*-isopropylacrylamide, which is an amphipathic, polymeric detergent.²⁶⁶

Once an integral membrane-bound protein has been purified, amino acid sequences from peptides can be used to design probes for screening libraries of **cDNA**. Once their cDNAs are available, integral membrane-bound proteins are often expressed from their cDNA so that, even though they are produced at low levels, they can be modified by **site-directed mutation** and other genetic manipulations to identify amino acids critical for one of their functions²⁶⁷⁻²⁷⁰ or for other studies.²⁷¹ The cDNA for a membrane-bound glycoprotein from an animal must be expressed in animal cells to produce the properly glycosylated protein.²⁰⁵ The expression of a particular cDNA that has been identified indirectly with a particular function is often used to prove that the protein for which it encodes actually is responsible for that function.^{272,273}

Many anchored membrane-bound proteins have been released from the membrane endopeptidolytically or have been dissolved with nonionic detergents as biologically active proteins and purified to homogeneity by the normal methods of chromatography or affinity adsorption. A few examples of such purified anchored membrane-bound proteins are cytochrome *b*₅,^{274,275} HLA histocompatibility antigens,²⁷⁶ HLA-linked B-cell antigen,²⁷⁷ dipeptidyl-peptidase IV,^{278,279} membrane alanine aminopeptidase,²⁸⁰ sucrose α -glucosidase/oligo-1,6-glucosidase,²⁸¹ dolichyl-phosphate β -D-mannosyltransferase,²⁸² unspecific monooxygenase,²⁸³⁻²⁸⁵ ATP diphosphatase,²⁸⁶ and the hemagglutinin of influenza virus.²⁸⁷

When anchored membrane-bound proteins are purified intact in the presence of nonionic detergent, they will recombine with bilayers of phospholipid when the detergent is removed^{275,281,288} and become anchored again, but when they are removed from the membrane by endopeptidolytic cleavage, the detached biochemically active, globular domains have no affinity for bilayers of phospholipids. Many of the **endopeptidolytically released detachable domains** have been crystallized, and crystallographic molecular models have been constructed from the maps of electron density.²⁸⁹⁻²⁹¹ The portions of anchored membrane-bound proteins that reside outside the membrane have also been expressed by themselves, after being genetically detached from their anchors, and crystallized, and crystallographic molecular models have been constructed for these **genetically released detachable domains**.^{292,293}

The **crystallographic molecular models of the detached domains** of anchored membrane-bound proteins are indistinguishable from those of normal water-soluble proteins. The terminal region of the polypeptide at which the cleavage releasing the detachable domain from the membrane occurred is usually disordered and featureless in the map of electron density. From all of these observations, it can be concluded that such an anchored membrane-bound protein is simply a water-soluble protein that is leashed to the bilayer of phospholipid by a flexible segment of its polypeptide attached in turn to the embedded anchor. An anchored membrane-bound protein may be attached to the membrane not only by a transmembrane anchor at one of its termini but also by adsorption to the bilayer through additional interactions on its surface. In such an instance the anchor must be removed and site-directed mutations within this region on its surface must be performed to produce a fully water-soluble protein capable of being crystallized.²⁹⁴

The **embedded anchor** left behind in the bilayer of phospholipid when a membrane-bound protein anchored at one of its termini is released by endopeptidolytic digestion almost always has at least one segment of sequence composed almost exclusively of the most hydrophobic amino acids. The entire stretch of polypep-

tide left behind in the membrane may be quite long, but the length of each of the hydrophobic segments is usually only about 20–25 aa long. The amino acid sequence of the hydrophobic segment from equine cytochrome *b*₅ is –WWTNWVIPAISAVVVALMY–,²⁹⁵ that from one of the human HLA histocompatibility antigens is –VPIVGI VAGLVLLVAVVTGAVVAVMW–,²⁹⁶ that from sucrose α -glucosidase/oligo-1,6-glucosidase from *O. cuniculus* is –LIVLFVIVFIIAIALI AVL A–,²⁹⁷ that from the hemagglutinin of an influenza virus is –WILWISFAISCFLLCVVL GFIMWAS–,²⁹⁸ and that from human glycophorin A is –ITLIIFGVMAGVIGTILLISYGI–.²⁹⁹ Such hydrophobic segments from anchored membrane-bound proteins are the most hydrophobic sequences of their length found in any protein.³⁰⁰ These sequences are usually flanked at both ends by regions containing normal or above-normal frequencies of polar and charged hydrophilic amino acids.

It is these hydrophobic segments of anchored membrane-bound proteins that span the hydrocarbon of the bilayer of phospholipid. That each of these **single, isolated membrane-spanning segments** of amino acid sequence is completely surrounded by liquid hydrocarbon explains their extreme hydrophobicity. The short hydrophilic segments that end up in most of these proteins on the opposite side of the membrane from the large globular, detachable domains probably act simply as barbs that cannot be pulled through the bilayer of phospholipid, but other roles, such as the relay of information across the membrane, have been proposed for some of them.

The hydrophobic segment of polypeptide that spans the membrane and serves to attach an anchored membrane-bound protein to the bilayer of phospholipid will spontaneously assume an α helix over most or all of its length, as judged by circular dichroic spectra, when it is incorporated into micelles of detergent^{301,302} or bilayers of phospholipid.³⁰³ It is believed that these hydrophobic anchors when they are attached to the native protein are also uninterrupted α helices spanning the hydrocarbon of the biological membranes in which they are normally found. An α helix is the logical structure for a segment of polypeptide to assume when it is immersed in liquid hydrocarbon in the total absence of water or any other donor or acceptor of hydrogen bonds. Within itself, an α helix satisfies all of the hydrogen-bond donors on the polypeptide (Figure 4–16A). The width of the hydrocarbon in a bilayer of naturally occurring amphipathic lipids and cholesterol is about 3.6 nm (Figure 14–4D).^{47,98} As the rise for each amino acid in an α helix is 0.15 nm, it should require about 24 aa to span the hydrocarbon. The fact that the lengths of the hydrophobic segments of anchored membrane-bound proteins are usually greater than 20 aa is further support for the proposal that these hydrophobic segments are α -helical in their normal situation.

The **distribution of amino acids** in the hydropho-

bic segments of the naturally occurring anchors that span the membrane in one isolated α helix elucidate the hydrophobic imperatives of a bilayer of phospholipid.³⁰⁴

As one might expect, isoleucines, leucines, valines, alanines, and phenylalanines represent 74% of the amino acids in these segments, a percentage 2.5 times greater than their percentage in water-soluble proteins. Cysteine, glycine, and methionine have frequencies equal to those for these amino acids in water-soluble proteins. Serines and threonines, although they occur half as frequently as they do in water-soluble proteins, are present within these fully engulfed membrane-spanning segments, but the donors on their hydroxyls can be automatically satisfied by the intramolecular hydrogen bonds in which these two amino acids often participate with the empty lone pairs on the acyl oxygens of the amino acids three or four positions ahead of them in an α helix (Figure 6–7).

Tryptophans and tyrosines are present in these membrane-spanning segments at about two-thirds the frequency with which they are present in water-soluble proteins, but they occur exclusively at the **ends of the segments** where their lone hydrogen-bond donors can remain in contact with water as they do almost always in a crystallographic molecular model of a soluble protein. If, however, a membrane-spanning segment is hydrophobic enough, it can drag a tryptophan into the middle of a bilayer of phospholipid.^{305,306} Prolines are also confined to the ends of such membrane-spanning segments, usually at the amino-terminal end, so that the α helix can cross the bilayer unbroken. The remaining polar amino acids, histidine, lysine, glutamine, aspartate, asparagine, glutamine, and arginine, constitute less than 1% of the amino acids in these segments and are always found at the ends, so that the hydrophilic portions of their side chains can remain in contact with the water or the polar headgroups of the phospholipids.

A number of peptides incorporating such hydrophobic sequences have been synthesized. For example, the peptide acetyl-KK(LA)₁₂KK- α -amide forms a stable α helix that spans the bilayers of phospholipid in vesicles of dipalmitoylphosphatidylcholine.³⁰⁷ The central portion of such a **synthetic hydrophobic peptide** is a rigid α helix from which the α -amido protons cannot exchange with protons in the water on the two sides of the bilayer, but the α -amido protons in the portions of the peptide exposed on the two sides of the bilayer exchange rapidly.³⁰⁸ There is a symbiotic relationship between the length of the hydrophobic sequence and the width of the bilayer. The peptide forms a fully miscible solution with the bilayer only when the length of the hydrophobic α helix matches the width of the bilayer,^{306,307} and the hydrophobic α helix can, to a certain extent, adjust the width of the bilayer to match its length.³⁰⁹ Cholesterol exerts an influence on this symbiosis to the extent that it increases the width of the bilayer.³⁰⁶

If a single lysine, aspartate, asparagine, glutamate, glutamine, or histidine is positioned during the synthesis in the center of an α -helical, polyleucyl, membrane-spanning peptide, the leucines will drag that side chain into the center of the bilayer^{310,311} because there is more than enough standard free energy in the hydrophobic effect to do so. The side chain of the lone lysine, the lone histidine, the lone aspartate, or the lone glutamate, however, enters the hydrocarbon as the neutral unprotonated or protonated form, respectively, so that the debit of standard free energy is only for its neutralization (Equation 5-66). That each enters as the neutral form, which is the only form of its acid-base that contains both a donor and an acceptor for hydrogen bonding, is supported by the fact that the peptides containing a single histidine, aspartic acid, or glutamic acid, as well as those containing a single asparagine or glutamine, readily form dimers and higher oligomers when they are incorporated into micelles or bilayers.³¹¹⁻³¹⁴ These oligomers result from hydrogen bonding within the hydrocarbon between the polar side chains.

These results clearly demonstrate that **hydrogen bonds**, which will not form in water because of competition from donors and acceptors on the molecules of water, are as stable within a phase of hydrocarbon, removed from contact with water, as they are in organic solvents (Table 5-3). When a hydrogen-bond donor and acceptor enter a hydrogen bond during the folding of a polypeptide in aqueous solution, the standard enthalpy change for the reaction is zero because the reaction proceeds with no net change in the number of hydrogen bonds (Equation 5-40) and little change in their net intrinsic stability (Equation 5-48). A hydrogen-bond donor or acceptor in the middle of an otherwise hydrophobic segment spanning a membrane, however, is held within the hydrocarbon by the hydrophobic α helix that was formed by the segment when it entered the membrane. The price of withdrawing the hydrogen-bond donors and acceptors from the water and stripping them of their hydration has already been paid by the hydrophobic effect that immersed the membrane-spanning segment in the first place. When a hydrogen-bond donor and acceptor form a hydrogen bond between these α helices within the hydrocarbon, the standard enthalpy change for the reaction is -12 to -20 kJ mol⁻¹ (Table 5-2).

When a single tryptophan is positioned in a sequence of leucines longer than is necessary to span a membrane, that tryptophan shifts the membrane-spanning polyleucyl α helix across the membrane until that tryptophan ends up close enough to the surface of the bilayer of phospholipid to thrust the hydrogen-bond donor in its side chain out of the hydrocarbon,³¹⁵ but a tyrosine does not display the same compulsion.

Many integral membrane-bound proteins have also been purified,^{221,242,316-320} some of these have been crystallized, and those crystals have provided crystallographic molecular models (Table 14-6).

It is generally assumed that, as for its purification, the most effective strategy for **crystallizing an integral membrane-bound protein** is to begin with the protein dissolved in a solution of a nonionic or zwitterionic detergent that is structurally homogeneous, such as one of the alkyl glycosides, one of the alkyl oligo(ethylene oxide) ethers, or *N,N*-dimethyldodecylamine *N*-oxide (Table 14-6). There are examples, however, of integral membrane-bound proteins crystallizing from solutions of structurally heterogeneous mixtures of detergents, such as the crystallization of mammalian cytochrome-*c* oxidase from a random mixture of dodecyl poly(ethylene oxide) ethers (Brij 35; average number of ethylene oxides equals 23)³³⁴ or the outer membrane protein F from *E. coli* from a random mixture of octyl poly(ethylene oxide) ethers.³⁵⁵ Sometimes a mixture of several structurally homogeneous nonionic detergents is intentionally prepared, as for the crystallization of the bacterial outer membrane protein TolC.³⁶⁰ Mammalian ubiquinol-cytochrome-*c* reductase will crystallize from a solution of methyl 6-*O*-(*N*-heptylcarbamoyl)- α -D-glucopyranoside, octyl β -D-glycoside, octanoyl-*N*-methylglucamide, octanoylsucrose, or octyl β -D-maltoside³³⁶ but also from a solution of dodecyl β -D-maltoside, decanoyl-*N*-methylglucamide, or diheptanoylphosphatidylcholine (Table 14-6). All of these observations seem to suggest that any one of several detergents could be used to obtain readily crystals of a particular integral membrane protein, but the efforts of many investigators over many years that have been expended to produce crystals of only a few integral membrane-bound proteins belie this suggestion.

Molecules of a particular protein must be dissolved in a **continuous isotropic phase** so that they can diffuse over the full extent of that phase to associate with each other and form a macroscopic crystal. Molecules of an integral membrane-bound protein dissolved in a solution of nonionic detergent are in an isotropic solution and each can encounter all of the others. A **bicontinuous cubic phase** of lipid and water is an isotropic phase that forms from particular aqueous suspensions of lipid.³⁶² In such a phase, a three-dimensional network formed from a single, continuous bilayer of amphipathic lipid encloses a single continuous three-dimensional network of aqueous channels. Such a bicontinuous cubic phase forms spontaneously under the proper conditions from a mixture of water and 1-oleoylglycerol.³²⁹ If bacteriorhodopsin from *Halobacterium salinarium* is incorporated into the lipid of such a bicontinuous cubic phase, the molecules of protein can diffuse over the full extent of the lipid phase to find each other and crystallize.³²⁹

When integral membrane-bound proteins are in vesicles of phospholipid that are suspended in an aqueous solution, each molecule of protein can associate productively only with other molecules of protein within that vesicle. Under certain circumstances, however, it is possible to produce stacked planar bilayers by fusion of

vesicles containing a particular integral membrane-bound protein, and the molecules of protein may be able to crystallize within each bilayer, and these two-dimensional arrays may then be able to stack regularly upon each other to produce a macroscopic crystal. Macroscopic crystals of bacteriorhodopsin suitable for crystallography have also been prepared in this way.^{363,364}

The membrane-bound proteins from bacteria that have been crystallized successfully are often purified from bacteria that have been genetically altered to overexpress that one particular protein.^{340,342,343,353,357,359,365} For example, succinate dehydrogenase from *E. coli* was produced by the **overexpression** of a gene that had been introduced on a plasmid; and in the resulting bacteria, 50% of the protein in the plasma membranes was succinate dehydrogenase.³⁶⁶ The protein responsible for glycerol transport in *E. coli* was overexpressed with a sequence of six consecutive histidines on its carboxy terminus so that it could be purified by affinity adsorption.³⁴⁹ In order to discover a large-conductance mechanosensitive channel that would crystallize, the proteins from nine different prokaryotes were each overexpressed and purified.³⁴⁷ To obtain a bacterial outer membrane containing only one porin, the gene for outer membrane protein F was expressed from a plasmid in a strain of *E. coli* lacking all of its porins.³⁵⁵

So far, however, integral membrane-bound proteins from **plants and animals** can seldom²⁰⁴ be expressed in a functional form at high levels. For example, when isoform 4A4 of unspecific monooxygenase from lung of *Oryctolagus cuniculus* was expressed in *E. coli*, less than 0.1 mg of the monooxygenase that had been incorporated into the bacterial plasma membrane could be purified from each liter of culture.²⁸³ When human glucose transporter was expressed in *E. coli*, the most sensitive methods of immunoblotting were used to detect its presence in membranes isolated from the cells.³⁶⁷ When bovine opsin was expressed in the mammalian cell line HEK293, only about 2 mg of the protein that had been incorporated into the plasma membrane of the cells could be isolated from each liter of culture.³⁶⁸ And when porcine Na⁺/K⁺-exchanging ATPase was expressed in the yeast *Pichia pastoris*, only about 1 mg of the protein was present in the unfractionated plasma membranes obtained from each liter of culture.³⁶⁹ It seems that bacteria are unable to insert animal proteins into their plasma membranes efficiently and that eukaryotic expression systems such as yeast or animal cells have trouble inserting extra protein into their membranes, perhaps because they are already crowded or perhaps because the systems for inserting them have only limited capacity. Consequently, most of the integral membrane-bound proteins from animals that have been crystallized have been purified from naturally occurring membranes that normally contain high concentrations of those proteins and can be obtained in large quantities from whole tissues.

From an examination of the crystallographic molecular models of integral membrane-bound proteins, there seem to be only two successful strategies that have been discovered by evolution through natural selection for immersing a protein in a membrane so completely that a major portion of its structure spans the bilayer of phospholipid, exposing respective portions of its surface to each side. Either the portion of the protein within the hydrocarbon of the bilayer is a **bundle of α helices** or it is a **β barrel**. The reason that these two arrangements are exclusive is that they seem to be the only two ways to provide acceptors for most if not all of the amido nitrogen-hydrogens in the backbone of those segments of polypeptide spanning the hydrocarbon of the bilayer. This conclusion is reinforced by the fact that bacterial outer membrane protein A spans the membrane with an eight-stranded β barrel that is almost a perfect cylinder so that every amido nitrogen-hydrogen participates in a hydrogen bond. Eight-stranded β barrels in soluble proteins are usually flattened so that the hydrogen bonds of the backbone in the two flattened regions can be straighter, but the α -amido nitrogen-hydrogens in the creases at the two edges of the flattened cylinder are able to form hydrogen bonds with water. Such an arrangement would be impossible within a bilayer of phospholipid, so the cylinder cannot be flattened.

At the moment, it appears that the integral membrane-bound proteins spanning the membrane with bundles of α helices are confined exclusively to the plasma membranes and intracellular membranes of cells, be they eukaryotic or bacterial, and the integral membrane-bound proteins spanning the membrane with β barrels are confined almost exclusively to the **bacterial outer membrane**, which is a bilayer formed from an outer monolayer of lipopolysaccharide and an inner monolayer of phospholipid. One of the few exceptions to this rule is bacterial α -hemolysin (Table 14-6), which is a toxin excreted from *Staphylococcus aureus* that forms a β barrel of 14 strands, two from each of its seven subunits, within the plasma membrane of a foreign cell to punch a hole in it and kill that cell.

Bacteriorhodopsin from *H. salinarium* (Figure 14-14),³⁷⁰ Ca²⁺-transporting ATPase from endoplasmic reticulum of *O. cuniculus* (Figure 14-15),³⁵¹ the membrane-spanning domain of potassium channel KcsA from *Streptomyces lividans* (Figure 14-16),^{343,344} photosynthetic reaction center from *R. viridis* (Figure 14-17),³²¹ and ubiquinol-cytochrome-*c* reductase from mitochondria of *S. cerevisiae* (Figure 14-18)^{371,372} are paradigms of the α -helical class of integral membrane-bound proteins; and porin OmpF from *E. coli* (Figure 14-19)³⁵⁵ and ferrichrome-iron receptor from *E. coli* (Figure 14-20)³⁵⁷ are paradigms of the β -barrel class of bacterial integral membrane-bound proteins. These crystallographic molecular models illustrate the variety of the structures assumed by integral membrane-bound proteins.

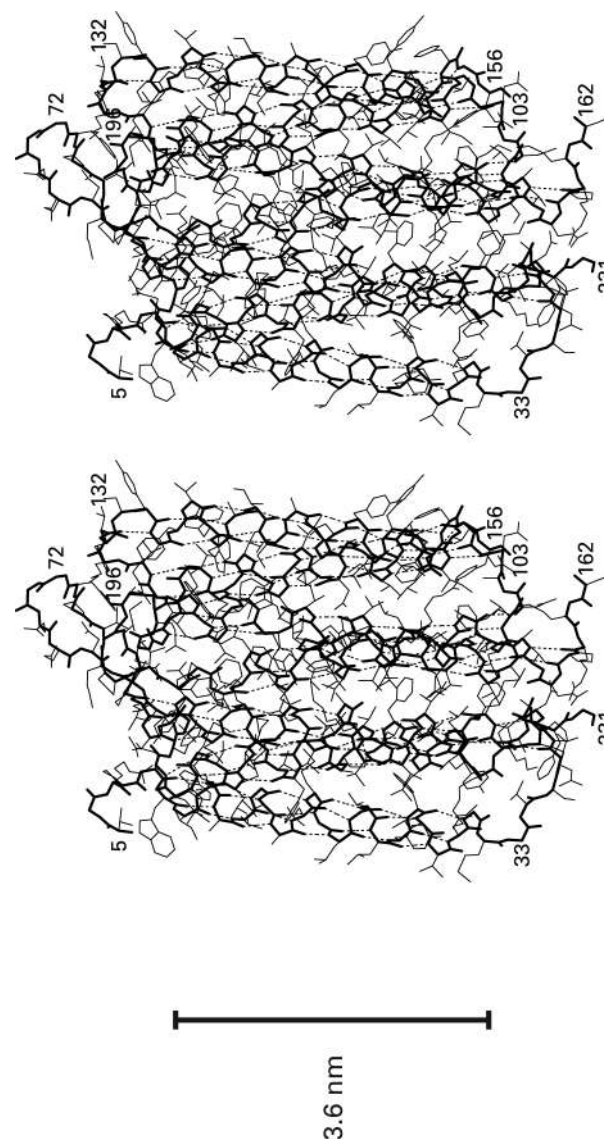
The extant **crystallographic molecular models of**

integral membrane-bound proteins demonstrate nicely the range of structures that are possible (Table 14-6). They vary in size from the monomer of mammalian rhodopsin (350 aa) to the dimer of heteroundecamers of mammalian ubiquinol-cytochrome-*c* reductase (4400 aa). In some of these proteins, such as the bacteriorhodopsin (Figure 14-14), the bacterial porins (Figure 14-19), or the aquaporins, most of the protein is within the membrane; in others, such as bacterial succinate dehydrogenase, mitochondrial ubiquinol-cytochrome-*c* reductase (Figure 14-18), or bacterial α -hemolysin, less than 15% of the protein is within the membrane. Some of the proteins are monomers (Figures 14-15 and 14-20), some are rotationally symmetric homooligomers (Figures 14-16 and 14-19), some are heterooligomers (Figures 14-17 and 14-18), and some are homooligomers of heterooligomers. The heterooligomeric protomer of mammalian cytochrome-*c* oxidase, with 13 different unrelated subunits with lengths ranging from 45 to 510 aa, is one of the most complex heterooligomeric proteins in existence, if matrices of polymeric proteins are not counted.

In some of the heterooligomers, such as, ironically, cytochrome-*c* oxidase, every one of the subunits has at least one membrane-spanning segment; in others, such as ubiquinol-cytochrome-*c* reductase (Figure 14-18) and bacterial succinate dehydrogenase, only about half the subunits have **membrane-spanning segments**. In these latter proteins, the subunits without membrane-spanning segments are globular structures associated by typical heterologous protein-protein interfaces with the subunits that do have them. When dissociated from such a complex, these subunits with no contact to the bilayer can sometimes be crystallized as water-soluble proteins.³⁷³ Two of the four nonidentical subunits of heterooligomeric photosynthetic reaction center from *R. viridis* (Figure 14-17) are homologous in sequence and have superposable arrangements of their five membrane-spanning α helices arranged around a 2-fold rotational axis of pseudosymmetry; the third subunit has a single membrane-spanning anchor; and the fourth subunit has no membrane-associated segments of polypeptide but does have a 1,2-diacyl-3-deoxyglyceryl group that is attached by a thioether linkage to its amino-terminal cysteine¹⁶⁸ and that is embedded in the membrane.

The eukaryotic complexes for electron transport often contain one or more short subunits that each contain one membrane-spanning segment and that appear to perform only a structural role in the complex. For example, ubiquinol-cytochrome-*c* reductase from *S. cerevisiae* (Figure 14-18) has two such membrane-spanning subunits of 94 and 65 aa, while bovine cytochrome-*c* oxidase (Table 14-6) has six of 84, 73, 56, 56, 47, and 46 aa, respectively. Beyond the 30 aa needed to span the membrane, the smaller of these subunits have little else left.

Figure 14-14: Skeletal drawing of the crystallographic molecular model of an individual subunit of the α_3 homotrimer of bacteriorhodopsin from *H. salinarum*.³⁷⁰ The protein used for the crystallography was overexpressed in *H. salinarum*. Membranes were dissolved in octyl β -glucoside, and the protein was purified chromatographically. The purified protein in a solution of octyl β -glucoside was mixed with 1-monooleoyl-*rac*-glycerol (monoolein) and a concentrated aqueous solution of sodium phosphate at pH 5.6 to incorporate the protein into a bicontinuous cubic phase of monoolein and water in which it crystallized.³²⁹ The complete crystallographic molecular model is drawn with the polypeptide backbone in thick line segments and side chains in thin line segments. The numbering is that of the mature posttranslationally modified protein. The crystallographic molecular model is oriented so that the plane of the membrane is horizontal and the cytoplasmic surface of the protein is at the bottom. A bar 3.6 nm high is placed next to the protein to represent the hydrocarbon of the bilayer. The exact location of the membrane is unknown because the protein was crystallized from monoolein.⁷⁷⁸ This drawing was produced with MolScript.⁷⁷⁸



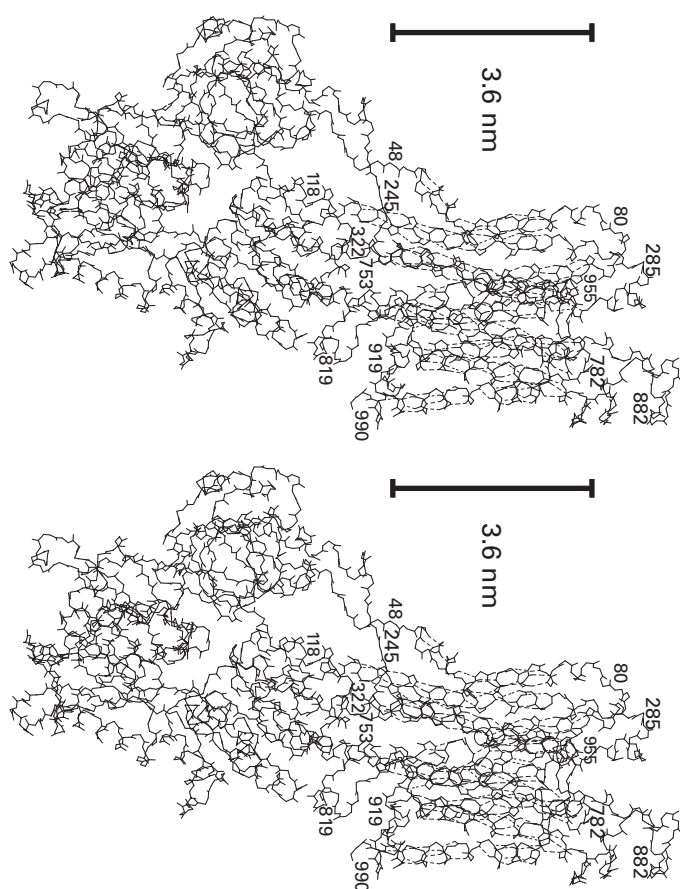


Figure 14-15: Skeletal drawing of the polypeptide backbone of the crystallographic molecular model of Ca^{2+} -transporting ATPase from *O. cuniculus*.³⁵¹ Purified endoplasmic reticulum from skeletal muscle was dissolved in a solution of $n\text{-C}_{12}\text{H}_{25}\text{O}(\text{CH}_2\text{CH}_2\text{O})_8\text{H}$, and the Ca^{2+} -transporting ATPase in this solution was purified by affinity adsorption. Crystals were grown by dialyzing the final purified protein dissolved in detergent against 0.8 M sodium butyrate, 2.75 M glycerol, 10 mM CaCl_2 , 3 mM MgCl_2 , 2.5 mM sodium azide, 0.2 mM dithiothreitol, and 20 mM *N*-(2-sulfoethyl)morpholine, pH 6.1. There are 10 membrane-spanning α -helices in the molecular model, and at one or the other end of each, an amino acid is numbered with its position in the amino acid sequence of the protein. The plane of the membrane is horizontal, and the cytoplasmic surface of the protein is at the bottom. The protein has a large globular cytoplasmic domain formed almost entirely by two long segments of polypeptide located between the second and the third membrane-spanning α -helices (from Alanine 118 to Aspartate 245) and between the fourth and the fifth membrane-spanning α -helices (from Glycine 322 to Isoleucine 753). Only the hydrogen bonds in the membrane-spanning α -helices are drawn. The exact location of the membrane is unknown because the protein was crystallized from a solution of detergent. The two cations of Ca^{2+} that are transported across the membrane by the protein were trapped in the crystals on their way across and were observed in the map of electron density. They are the open circles in the drawing, each drawn with a radius equal to the ionic radius of a Ca^{2+} cation. Many of the membrane-spanning α -helices in the structure are broken. This drawing was produced with MolScript.⁷⁷⁸

The membrane-spanning portion of the integral membrane-bound proteins represented by the crystallographic molecular models can be formed entirely by segments of secondary structure from one polypeptide, such as those in Ca^{2+} -transporting ATPase (Figure 14-15) mammalian rhodopsin, or bacterial outer membrane protein A; from segments of secondary structure from several different subunits, both heterologous and homologous, as in photosynthetic reaction center (Figure 14-17); from segments of secondary structure from many different heterologous subunits, as in mammalian cytochrome-*c* oxidase; or from identical segments of secondary structure from the identical subunits in a homooligomer, as in the membrane-spanning domain of potassium channel KcsA (Figure 14-16). In the homooligomers responsible for transport of specific molecules or inorganic ions across the membrane, each subunit can form within the membrane its own channel for the substrate, as in the bacterial porins and the aquaporins; or each subunit in the oligomer can contribute an identical set of secondary structures that are arrayed around an *n*-fold rotational axis of symmetry normal to the membrane to form together the sole channel in the complete oligomer, as in the membrane-spanning domain of potassium channel KcsA (Figure 14-16), the bacterial outer membrane protein TolC, and bacterial α -hemolysin.

Many of the integral membrane-bound proteins are responsible for transporting metabolites or inorganic ions, either actively or passively, or for transporting information across the membrane. Those responsible for passively transporting metabolites nonspecifically, such as the porin OmpF (Figure 14-19),³⁵⁵ have a fairly wide **hydrophilic water-filled channel** passing through their center; in the porin OmpF the channel at its narrowest is only 0.7 nm \times 1.1 nm. Those responsible for passively transporting particular molecules or ions usually have an obvious channel passing through them, within which there is a region in which the selection for those molecules is performed. For example, potassium channel KcsA (Figure 14-16) has a water-filled, cylindrically symmetric channel passing through it at one end of which is a constriction in which there is a symmetrically displayed set of acyl oxygens from the polypeptide backbone that select the potassium ions.^{343,344} In proteins that actively transport cations, such as Ca^{2+} -transporting ATPase (Figure 14-15),³⁵⁰ cytochrome-*c* oxidase,³³⁵ and cytochrome *o* ubiquinol oxidase,³⁴⁰ the passageway through which those cations pass is much narrower, less obvious, and convoluted so that the entrance into the center of the channel and the exit from the center of the channel by those ions can be controlled and coupled to conformational transitions in the protein. Those proteins responsible only for the passage of information such as mammalian rhodopsin³³¹ present no passageway for any metabolites, water, or cations by filling the membrane with solid protein; otherwise they would produce leaks.

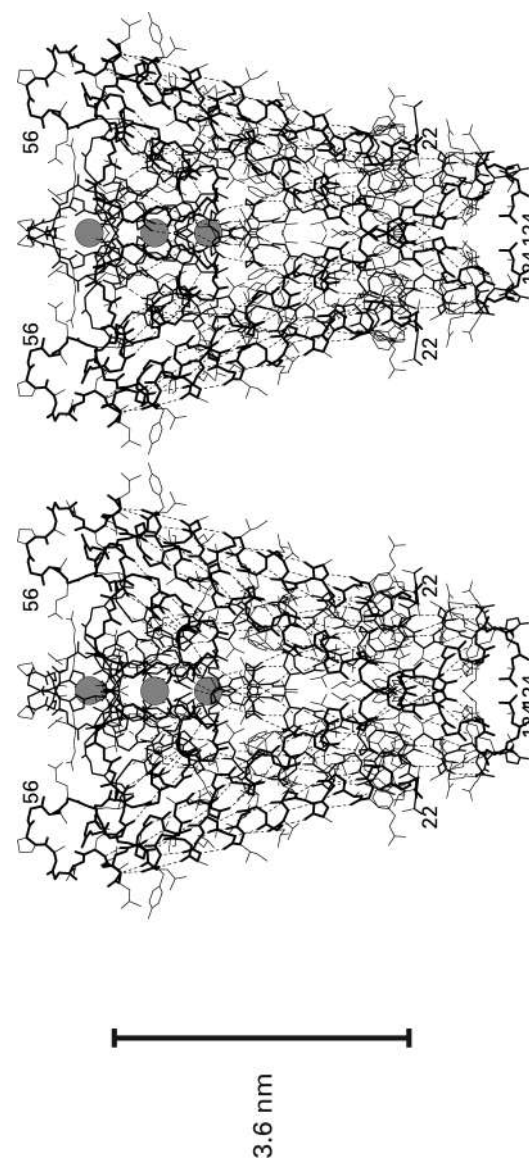
Many integral membrane-bound proteins have large **globular domains** on one side of the membrane or the other. For example, proteins that act as receptors that sense the presence of molecules outside the cell, such as acetylcholine receptor,³⁴⁵ have globular domains on the extracytoplasmic surface of the membrane that are responsible for this function. Proteins that catalyze the active transport of cations, such as mammalian endoplasmic reticulum Ca^{2+} -transporting ATPase (Figure 14–15),^{350,351} have globular domains on the cytoplasmic surface of the membrane that convert the binding and hydrolysis of cytoplasmic MgATP into the movement of those cations against their gradients of concentration. Most of these globular domains resemble globular soluble proteins in the details of their structure, so it is only the structural details of the portions of these molecular models that are immersed within the bilayer of phospholipid and located in its immediate vicinity that are peculiar to these proteins.

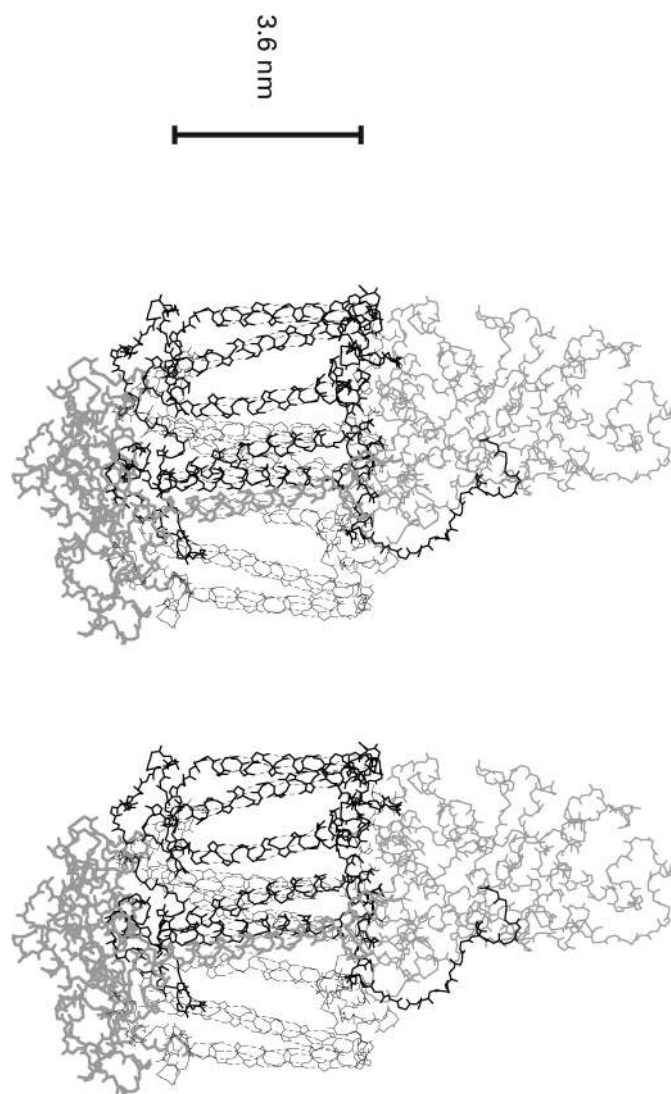
The side chains of the amino acids forming the continuous surface of the protein in direct contact with the bilayer of phospholipid³⁷⁴ can be considered as a **sheath**

that encloses the protein and forms the interface between it and the membrane. The most dramatic illustration of one of these sheaths is the layer of side chains protruding from the continuous β barrel of 22 strands completely enclosing the interior of ferrichrome–iron receptor from *E. coli* (Figure 14–20). Within this particular sheath, there is a typical globular structure of β sheets and α helices,³⁵⁷ insulated from the hydrocarbon of the bilayer of the outer membrane by the sheath. The side chains of leucine (23), valine (14), tryptophan (13), phenylalanine (11), glycine (9), alanine (9), isoleucine (8), tyrosine (7), methionine (4), and proline (2) constitute 74% of this sheath. Polar side chains of glutamine (4), lysine (2), arginine (2), aspartate (2), glutamate (1), asparagine (1), serine (1), and histidine (1) that are located at the ends of the sheath in contact with the head groups of the phospholipid make up only 10%.

Although the sheath surrounding the protein is not

Figure 14–16: Skeletal drawing of the crystallographic molecular model of the membrane-spanning domain of potassium channel KcsA from *Streptomyces lividans*.³⁴⁴ Potassium channel KcsA is an α_4 homotetramer with cyclic symmetry of point group $4(C_4)$. Its four subunits are arrayed around a 4-fold rotational axis of symmetry normal to the plane of the membrane. Plasma membranes from *S. lividans* were dissolved in *n*-decyl β -D-maltoside, and potassium channel KcsA was purified from this solution.³⁴³ The carboxy-terminal 35 amino acids following Phenylalanine 125 were removed from each subunit of the purified tetramer by digesting the protein with chymotrypsin. Each of the amino-terminal domains of the subunits in the α_4 tetramer that resulted from the digestion was bound to the Fab fragment of a monoclonal immunoglobulin directed against this domain, and the complex between the Fab fragments and the tetramer of the four domains was purified and crystallized. Only the crystallographic molecular model of the membrane-spanning tetramer of the four truncated subunits is drawn. The side chains of all of the truncated subunits are drawn with thin line segments, but two different thicknesses of line segments were used to draw the polypeptide backbones of the subunits so that they could be distinguished. Only two of the subunits are numbered. There are two membrane-spanning α helices in each truncated subunit. In between these membrane-spanning α helices in each subunit is a loop of polypeptide that joins its three other quadrants in the center of the tetramer around the 4-fold rotational axis of symmetry to form a channel that selects only potassium ions for transport across the membrane. The gray circles represent three potassium ions that were found in this channel on their way across the protein in the map of electron density. The channel that selects the potassium ions is lined by the acyl oxygens of the four segments of polypeptide from the Threonines 75 to the Aspartates 80. The cations approach the channel by passing through a larger, water-filled vestibule connected to the cytoplasm of the cell and formed by the eight membrane-spanning α helices. This vestibule is in the center of the lower half of the model. The cytoplasmic end of the protein is at the bottom of the drawing. This drawing was produced with MolScript.⁷⁷⁸





so clearly visible in the regions of an integral membrane-bound protein that spans a membrane with a bundle of α helices, it can be delineated by scoring the exposure of the side chains in these α helices to the bilayer of phospholipid. For example, the sheath surrounding the membrane-spanning α helices in the photosynthetic reaction center (Figure 14-17) from *Rhodobacter sphaeroides* is formed from side chains that have greater than 20% of their respective surface areas exposed to the bilayer of phospholipid.³⁷⁴ Of those side chains that have greater than 50% of their respective surface areas exposed to the bilayer, 91% are leucines (15), isoleucines (12), phenylalanines (12), valines (7), alanines (6), glycines (5), tyrosines (5), tryptophans (4), and methionines (3); of those with 20–50% exposed, 71% are from this same set of side chains. In keeping with the α -helical secondary structure, only one proline, that at the end of one of the α helices, is exposed to the bilayer. The two glutamines and the one lysine that have more than 50% of their surface area exposed are at the edges of the membrane, as are the three glutamates and the single asparagine, glu-

Figure 14-17: Skeletal drawing of the polypeptide backbone of the crystallographic molecular model of photosynthetic reaction center from *Rhodospseudomonas viridis*.³²¹ A purified preparation of photosynthetic membranes was isolated from broken cells by differential centrifugation followed by isopycnic centrifugation (see Figure 14-1). The membranes were dissolved in *N*-dodecyl-*N*,*N*-dimethylamine *N*-oxide (14-14) and submitted to chromatography by molecular exclusion in a solution of the same detergent. Fractions containing reaction center were identified by its characteristic absorbance at 830 nm. These fractions were pooled, and crystals of reaction center were produced in concentrated solutions of ammonium sulfate. The protein is composed of four different subunits. The C subunit (drawn with thin gray line segments) is a cytochrome with four hemes that sits on the extracytoplasmic side (top) of the protein. The L subunit (thick black line segments) and the M subunits (thin black line segments) are homologous, share a common ancestor, have superposable structures, and each contains five membrane-spanning α helices arrayed about a 2-fold rotational axis of pseudosymmetry normal to the plane of the membrane. These membrane-spanning α helices form two cages that enclose the bacteriochlorophylls and bacteriopheophytins contained within the protein. The H subunit (thick gray line segments) is an anchored membrane-bound protein with a single membrane-spanning α helix. The globular domain of the H subunit grips the cytoplasmic surface (bottom) of the heterodimer of the L subunit and the M subunit. Only the hydrogen bonds in the membrane-spanning α helices are drawn. This drawing was produced with MolScript.⁷⁷⁸

tamine, and arginine that have 20–50% of their respective surface areas exposed. The side chains of these amino acids are in contact with the polar head groups of the phospholipids.

These observations illustrate the fact that the portion of each of the integral membrane-bound proteins that is immersed in the bilayer is surrounded by a sheath that is significantly enriched in hydrophobic amino acids. This sheath forms a **boundary between protein and lipid** that is compatible with the hydrocarbon in the middle of the membrane and the head groups at the two surfaces and dissolves the protein in the bilayer of phospholipid and cholesterol just as the polar surfaces of globular, water-soluble proteins dissolve them in water. The distribution of hydropathy over the surface of this sheath determines the **depth** at which the protein floats within the bilayer of phospholipid and its orientation relative to the plane of the membrane.³⁷⁴

In integral membrane-bound proteins that are situated in plasma membranes and intracellular membranes and that consequently span these membranes with a bundle of α helices, the few lysines, arginines, aspartates, asparagines, glutamates, glutamines, and tyrosines that they contain are usually located at the ends of these membrane-spanning α helices, and the **polar or charged nitrogens and oxygens** in their side chains reach out of the hydrocarbon of the bilayer of phospholipid into the polar interfaces on each side, as if their side chains were **snorkels**.³⁷⁵⁻³⁷⁷ Lysines, arginines, aspartates, and glutamines are located almost twice as often at the amino-terminal end as at the carboxy-terminal end of a

membrane-spanning α helix. An amino-terminal location assists in the emergence of their side chains from the hydrocarbon because the two rotamers ($\chi_1 = +66^\circ$ and $\chi_1 = -65^\circ$) at the β carbon that point their side chains in the amino-terminal direction are about twice as populated as the one ($\chi_1 = -177^\circ$) that points them towards the carboxy terminus.^{378–381} Lysines and arginines both in single, membrane-spanning α -helical anchors and membrane-spanning bundles of α helices in integral membrane-bound proteins³⁸² are about twice as likely to be located at the cytoplasmic ends of the α helices.³⁸³ This tendency is due in part to the negative surface charge on the cytoplasmic side of naturally occurring membranes.³⁸⁴

Within the sheath of hydrophobic side chains, the structure of an integral membrane-bound protein resembles the interior of a globular, water-soluble protein except for the fact that all membrane-spanning segments of integral membrane-bound proteins from plasma membranes and intracellular membranes, even

those well within the sheath, are α helices. Proteins that form channels for metabolites and inorganic ions often have significant aqueous pores passing most of the way through them (Figure 14–16), but proteins that do not have such pores still have locations occupied by molecules of water within the β barrel³⁵⁹ or within the bundle of α helices³²⁶ that spans the membrane. As in water-soluble proteins, these **locations occupied by molecules of water** can be clustered, or they can be entirely surrounded by donors and acceptors from the protein. There are also sites occupied by inorganic cations³³⁵ or anions³³⁰ in the membrane-spanning regions.

In those integral membrane-bound proteins that span the membrane with a bundle of α helices that contain no aqueous channels, the density with which their atoms are packed is about 5% greater than that of a water-soluble protein formed from a bundle of α helices,³⁸⁵ but the **packing of the α helices** is similar. Although the α helices of bacteriorhodopsin (Figure

Figure 14–18: Skeletal drawing of the polypeptide backbone of an individual protomer of the $(\alpha\beta\gamma\delta\epsilon\zeta\eta\theta\kappa)_2$ homodimer in the crystallographic molecular model of ubiquinol-cytochrome-*c* reductase from mitochondria of *Saccharomyces cerevisiae*.³⁷¹ Mitochondrial membranes from *S. cerevisiae* were dissolved in a solution of *n*-dodecyl β -D-maltoside, and ubiquinol-cytochrome-*c* reductase was purified chromatographically from this solution.³⁷² The protein was crystallized in a complex with a fragment of a monoclonal immunoglobulin containing only the V_H and V_L domains (Figure 11–1) in a solution of *n*-undecyl β -D-maltoside by use of poly(ethylene glycol) as a precipitant. The cytoplasmic, extramitochondrial surface of the protein is at the bottom of the drawing; the extracytoplasmic, intramitochondrial surface is at the top. The protein contains 10 subunits, only nine of which appeared in the map of electron density. The two intramitochondrial core proteins (431 and 352 aa; thin gray and thin black line segments, respectively) form a large, globular structure devoid of coenzymes and peripheral to the membrane (upper right corner of the drawing). Cytochrome *b* (385 aa; thick gray line segments) contains two hemes and is the major membrane-spanning subunit of the protein. It is almost entirely immersed in the bilayer. Cytochrome c_1 (248 aa; intermediate black line segments) contains one heme and is an anchored membrane-bound protein with one membrane-spanning segment (center of the drawing) and an extramitochondrial globular domain (bottom of the drawing). The Rieske iron-sulfur subunit (185 aa; thin black line segments) contains one iron-sulfur cluster and is another anchored membrane-bound protein, but, unlike cytochrome c_1 , its globular, extramitochondrial domain (lower right corner of the drawing) is unattached to the rest of the structure. Subunit VI (147 aa; thin black line segments) is a peripheral membrane-bound protein that is bound to the globular, extramitochondrial domain of cytochrome c_1 (lower left corner of the drawing) and that has no contact with the bilayer. Subunit VII (127 aa; thin black line segments) is another peripheral membrane-bound protein that is bound to the intramitochondrial surface of cytochrome *b* (upper left corner of the drawing). Subunits VIII and IX (94 and 65 aa, respectively; thick black line segments) each span the membrane with a single α helix (to the left and to the right in the drawing). Again, only the hydrogen bonds in the α helices that span the membrane are drawn. This drawing was produced with MolScript.⁷⁷⁸

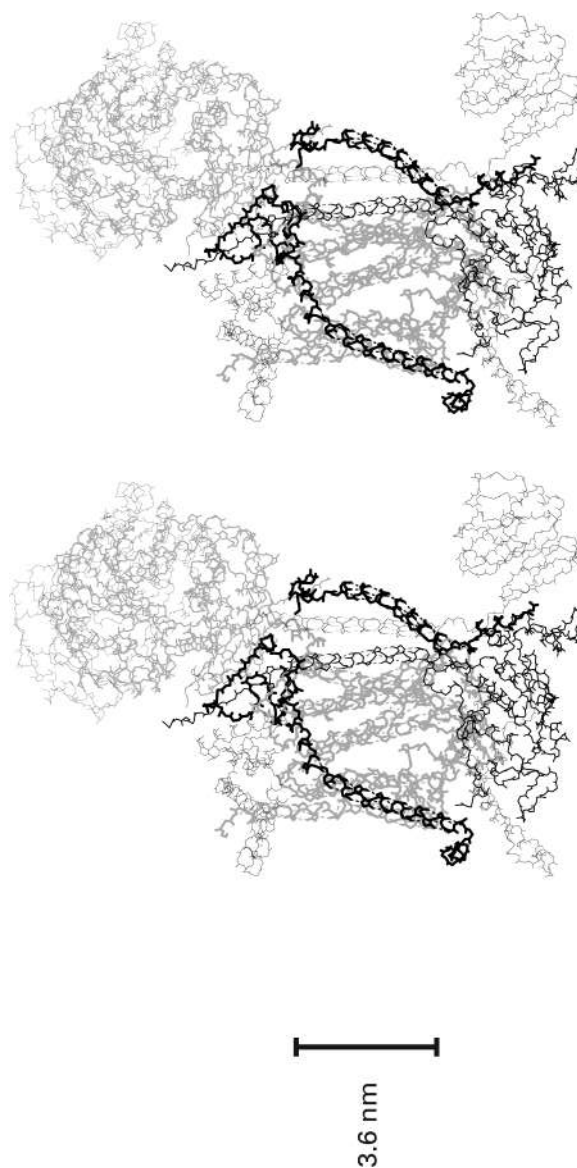
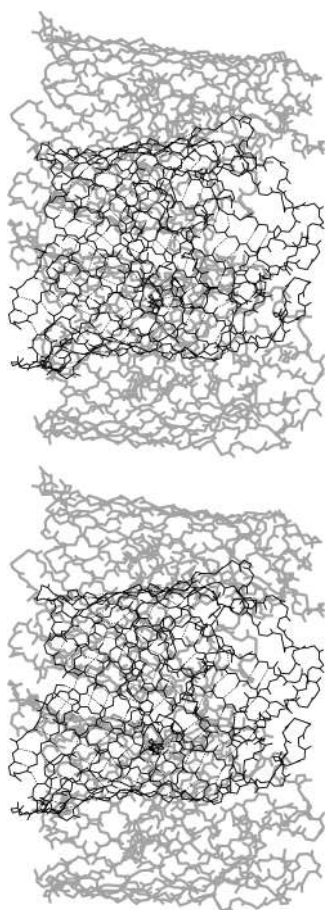


Figure 14-19: Skeletal drawing of the polypeptide backbone of the crystallographic molecular model of porin Ompf from *E. coli*.³⁵⁵ A mixture of fragments of all of the membranes from a homogenate of *E. coli* were first extracted with 0.5% octyl poly(ethylene oxide), and the membranes that remained were then dissolved in 3.0% octyl poly(ethylene oxide). The porin Ompf was purified chromatographically from this latter solution and crystallized by use of poly(ethylene glycol) as a precipitant. The α_3 trimer of the porin is drawn. The subunit to the front is drawn with black line segments, and the hydrogen bonds within the membrane are included only in this subunit. The two symmetrically related subunits in the back are drawn with gray line segments. The 3-fold rotational axis of symmetry is vertical and normal to the plane of the membrane, which is horizontal. Notice that the β sheet of each subunit enclosed within the interfaces that form the trimer is not high enough to span the membrane as do the β sheets on the outside of the trimeric protein, so the protein makes sense only as a trimer. The extracellular surface of the protein is at the top of the drawing. Because the protein spans the outer membrane of the bacterium, the surface to the bottom of the drawing faces the periplasm, which is the space between the outer membrane and the plasma membrane of the bacterium. This drawing was produced with MolScript.⁷⁷⁸



14-14) are all almost parallel to each and normal to the plane of the membrane,³²⁵ in most of the proteins of this class, the α helices are tilted with respect to each other (Figures 14-15, 14-16, 14-17, and 14-18). For example, in the photosynthetic reaction center from

R. sphaeroides,³⁷⁴ the average angle of tilt is $+22^\circ$. About 50% of the angles Ω between two adjacent membrane-spanning α helices (Figure 6-23) in crystallographic molecular models of integral membrane-bound proteins have values between $+10^\circ$ and $+30^\circ$,^{386,387} but these angles can be either positive or negative, even within the same protein.^{331,337}

In a β barrel, the β strands are automatically held in rigid orientation by the regular array of interstrand hydrogen bonds, but when a bundle of α helices assembles in the membrane they are held together by an irregular array of adventitious interhelical hydrogen bonds. For example, in bacteriorhodopsin, which has a bundle of seven α helices spanning the membrane (Figure 14-14), there are 31 hydrogen bonds that interconnect 12 pairs of these α helices³²⁶ and that are in part responsible for their relative orientations and the tertiary structure they assume. Most of these **hydrogen bonds** are between donors and acceptors found at the ends of the α helices, locations in which polar and charged side chains are frequently found, but 10 of them are in the middle of the membrane within the regions of the hydrocarbon of the bilayer of phospholipid (Table 14-7). All of the ionizable side chains should be in the neutral form of their acid-base (Lysine 216 is the imine of a retinol) so that the carboxy groups of the aspartic and glutamic acids have both a donor and three acceptors. Many of these interhelical hydrogen bonds are from a donor or acceptor on the protein to a molecule of water and from there to a donor or acceptor on the protein. These bridging molecules of water are found at locations within the bilayer.

As has been noted previously, each of these hydrogen bonds has a significantly negative enthalpy of formation that is released during the assembly of these

Table 14-7: Hydrogen Bonds in the Interior of the Membrane between Donors and Acceptors on Different Membrane-Spanning α Helices in Bacteriorhodopsin from *H. salinarium*³²⁶

α helices joined ^a	acceptor ^b	donor
C-D	L87O	D115OD1
C-D	D115OD1	T90OG1
D-E	M118O	S141OG
F-G	A215O ^c	W182NE ^c
F-G	D212OD1	Y185OH
B-F	D212OD1	Y57OH
B-G	D205O ^c	Y57OH ^c
C-G	K216O ^{c,d}	T46O ^c
C-G	D85OD2 ^c	D212OD1 ^c
C-G	K216NZ ^{c,d}	D85OD2 ^c

^a α Helix A spans the membrane with amino acids 11 through 29; α helix B, 44 through 62; α helix C, 79 through 96; α helix D, 108 through 127; α helix E, 135 through 154; α helix F, 173 through 191; and α helix G, 204 through 223. ^bOnly hydrogen bonds within the hydrocarbon of the bilayer are tabulated. ^cHydrogen-bonded through a molecule of water. ^dLysine 216 is posttranslationally modified with a retinal.

α helices within the membrane to produce the tertiary structure of the protein. Once each membrane-spanning α helix has become inserted in the bilayer of phospholipid, however, the **hydrophobic effect** is no longer in operation. Consequently, the importance of hydrogen bonding and the importance of the hydrophobic effect in the formation of tertiary structure are reversed once that portion of the polypeptide is within the hydrocarbon of the bilayer of phospholipid. During the assembly of the α helices, hydrogen-bond donors and acceptors can be responsible for significant, favorable standard enthalpy of formation, yet no favorable change in standard free energy, other than that associated with the packing efficiency, occurs when two hydrophobic surfaces are juxtaposed. The situation, however, is similar to that observed in the folding of a water-soluble protein in that the hydrophobic effect in the one case immerses the α helix in the bilayer and in the other produces the initial hydrophobic collapse of the polypeptide. After these hydrophobically driven events, the establishment of the final tertiary structure relies not at all or much less, respectively, on the hydrophobic effect because it has already been expended.

Prolines are present in many of the α helices in membrane-spanning bundles.³⁸⁸ As it does in an α helix in a water-soluble protein, a proline always occupies a kink in a membrane-spanning α helix. There are also kinks in membrane-spanning α helices that do not incorporate a proline,³²⁶ but it has been noted that wherever there is such a kink in a membrane-spanning α helix there will be a high frequency of homologous proteins that have prolines at that position.³⁸⁸ In addition, it has been demonstrated that the introduction of a proline at an unkinked position in a membrane-spanning α helix usually produces a protein unable to assume its native structure,³⁸⁹ while changing a proline at a kink to an alanine has little effect on the protein.³⁸⁸ Evidently, a kink that has been established by evolution through natural selection in a membrane-spanning α helix prefers to have a proline at the position of the kink, but not exclusively, while established, unkinked segments of α helix are too rigidly held within the structure to tolerate the inevitable kink that results from inserting a proline.³⁸⁸ A proline just beyond one of the ends of a membrane-spanning α helix can promote the doubling back of the polypeptide to form the next membrane-spanning α helix.³⁹⁰

The question of whether or not there are **cystines** within the membrane-spanning segments of integral membrane-bound proteins is of chemical interest. First, the interior of a bilayer of phospholipid is devoid of glutathione or any other small mercaptan. Second, oxygen is more soluble in the hydrocarbon of a bilayer than it is in water. Third, the oxidation of two thiols to a disulfide performed by oxygen is a free radical reaction that should proceed normally within the hydrocarbon. Fourth, cystine is one of the most hydrophobic of the side

chains. In spite of these facts, there do not appear to be any cystines in the membrane-spanning segments of integral membrane-bound proteins. One of the first indications of this peculiar absence was the observation that acetylcholine receptor from *Torpedo californica*, although it has a total of 13 cysteines in its 20 membrane-spanning α helices, has no cystines in those membrane-spanning α helices.³⁹¹ Although there are a number of cystines in the extracytoplasmic portions of the crystallographic molecular models listed in Table

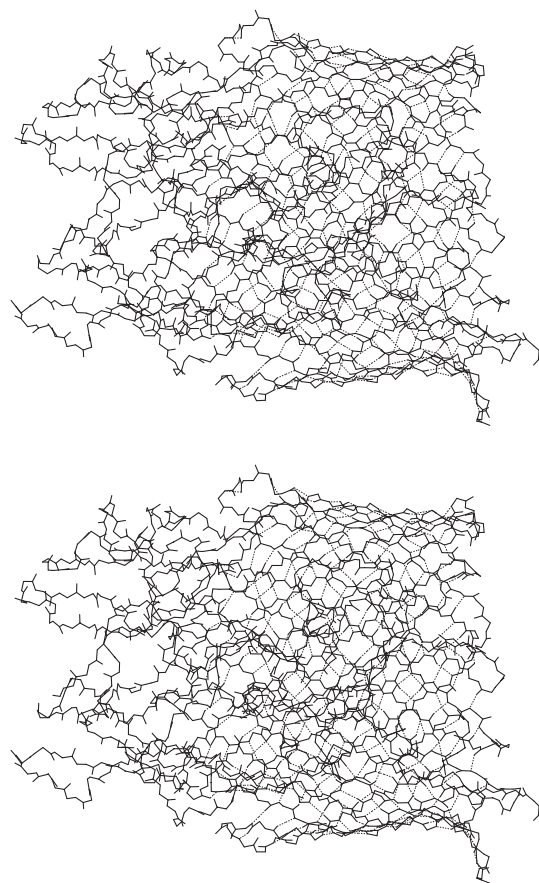


Figure 14-20: Skeletal drawing of the polypeptide backbone of the crystallographic molecular model of ferrichrome-iron receptor from *E. coli*.³⁵⁷ Membranes from a strain of *E. coli* that overexpresses ferrichrome-iron receptor were dissolved in 2% Triton X-100, and the ferrichrome-iron receptor was purified chromatographically from this solution. In the final chromatographic step, which was by molecular exclusion, the protein was transferred from 2% Triton X-100 to 1% *n*-octyl β -D-glucoside. It was crystallized from this solution by the use of poly(ethylene glycol) as a precipitant. The protein is a monomer, so the β sheet must be high enough to span the outer membrane all the way around. Only the hydrogen bonds in the portions of the β sheet that span the membrane are included. The extracellular surface of the protein is on the top; the surface facing the periplasm is on the bottom of the drawing. This drawing was produced with MolScript.^{77b}

14–6,^{331,333,335,337,351,391} there is none in any of the membrane-spanning portions of these proteins, with the exception of a cystine in the center of the large water-filled pore passing through maltoporin.³⁵³ This latter cystine, however, is not in contact with the hydrocarbon of the outer membrane.

The photosynthetic reaction center from *R. viridis* (Figure 14–17) has 11 **membrane-spanning α helices**. Each traverses the hydrocarbon of the bilayer of phospholipid in one unbroken α helix. The amino acid sequences of these 11 α helices³²² are –SLGVLSLFSGLMWFFTIGIWFYNA–, –LKEGGLWLIASFFMFVAVWSW WGRTYLRAQA–, –AWAFLSAIWLWMVLGFIRPILM–, –PFHGLSIAFLYGSALLFAMHGATILAV–, –MEGIHRWAIWM AVLVTLTGGIGILL–, –GFFGVATFFAALGIILAWSAVL–, –GGLWQIITICATGAFVSWALREVEICRKL–, –HIPFAFAIL AYLTLVLFPRPVM–, –PAHMIAISFFFTNALALALHGALVLS AA–, –GTLGIHRLGLLLSLAVFFSALCMII–, and –IAQLV WYAQWLVIWTVVLLYLRRDR–.^{392–394} In each of these amino acid sequences there is a region of at least 20 amino acids in length that contains no amino acids that are charged at neutral pH with the exception of the arginines at the carboxy-terminal ends of the third and the eighth α helices. The reason that these 11 α helices were able to be inserted into the bilayer of phospholipid is that they are composed of hydrophobic amino acids and the bilayer is an organic solvent into which the side chains of these amino acids have dissolved.

These 11 hydrophobic segments of amino acid sequence are similar to the five listed earlier for the single membrane-spanning α helices from anchored membrane-bound proteins but differ in flavor. The hydrogen-bond donors, tryptophan and tyrosine, are more uniformly distributed over the length of these segments; the neutral but hydrophilic hydrogen-bond donors and acceptors glutamine, asparagine, and histidine now occasionally appear; and the frequency with which glycine is encountered is greater. Each of these subtle changes indicates that these amino acid sequences are from a bundle of α helices gathered together as a protein rather than from individual α helices spanning the extremely hydrophobic environment of the membrane as isolated entities. Even in α helices completely buried in the center of one of these bundles, however, such as the three α helices –PRMNNMSFWLLPPSFLLLASSM–, –ASVDLTIFSLHLAGVSSILGAINFITTN–, and –LFVWSV MITAVLLLLSLPVLAAGITMLLTD– from bovine mitochondrial cytochrome-*c* oxidase,³³³ the preponderance of hydrophobic amino acids at the center of each persists.

Even though most integral membrane-bound proteins are dissolved in solutions of a nonionic detergent before they are crystallized (Table 14–6), they carry **molecules of phospholipid** from the original membranes along with them into the respective crystals, and if the maps of electron density are clear enough, these mole-

cules can be recognized and included in the crystallographic molecular model. For example, in a crystallographic molecular model of photosynthetic reaction center from *R. sphaeroides*, there is a molecule of diphosphatidylglycerol;³⁹⁵ in one for bovine mitochondrial cytochrome-*c* oxidase, there are five molecules of phosphatidylethanolamine and three molecules of phosphatidylglycerol;³³³ and in one for ubiquinol-cytochrome-*c* reductase from *S. cerevisiae*, there is a molecule of phosphatidylcholine, two of phosphatidylethanolamine, one of phosphatidylinositol, and one of diphosphatidylglycerol.³⁷¹ At least two diether lipids (see 14–8) are represented by their 1-(2,6,10,14-tetramethylhexadecan-16-yl)-2-(2,10,14-trimethylhexadecan-16-yl)glyceryl groups in the crystallographic molecular model of bacteriorhodopsin.³²⁶

The fatty acyl or isoprenyl chains of all of these molecules of phospholipid or diether lipid usually lie within the straight crevices running between the membrane-spanning α helices in an orientation roughly normal to the plane of the membrane, but in one case the fatty acyl groups penetrate sideways into spaces between the α helices,³⁷¹ and the head group of this particular phospholipid is 0.8 nm away from the plane in which head groups are normally located (Figure 14–2). The head groups of all of the other phospholipids in these crystallographic molecular models are in the expected positions, and they engage in many hydrogen bonds with side chains of the protein and molecules of water occupying fixed locations bridging those head groups and the protein.

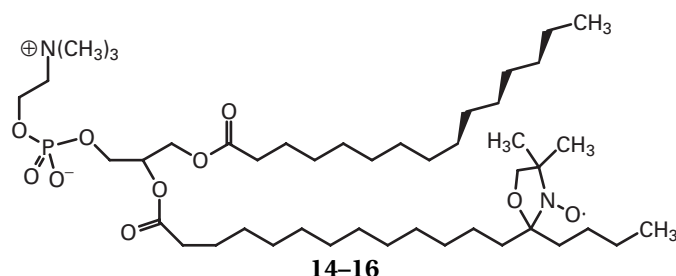
It is possible that these rigidly fixed molecules of phospholipid are structurally essential just as many of the molecules of water in the interior of a molecule of protein are structurally essential. If so, these observations may explain why integral membrane-bound proteins often lose their biological activity and even their tertiary structure when the ratio of detergent to protein becomes too large. They may also be examples of molecules of phospholipid within the boundary layer that are most severely immobilized, just as fixed locations for molecules of water on the surface of a crystallographic molecular model are the most severely immobilized molecules of the waters of hydration.

The sheath of hydrophobic side chains surrounding an integral membrane-bound protein is immersed in the bilayer of phospholipid and surrounded by the hydrocarbon of the amphipathic and neutral lipids. The lipids in this boundary layer are those molecules of lipid the behavior of which is affected at a given instant by the presence of the protein. Molecules of **lipid in the boundary layer** can be formally distinguished from those molecules of lipid that behave as if they were in an unadulterated bilayer of the same amphipathic lipids.

This distinction resembles in its ambiguity the distinction between water of hydration associated with a protein (Table 6–4) and water in the bulk solution. In the case of water of hydration, there is a gradual diminution

of the influence of the protein the farther a particular molecule of water is from its surface, but a water molecule several shells from the protein may still be influenced by it because of the nets of hydrogen bonds that ensnare both water and protein. Likewise, a molecule of amphipathic lipid somewhat distant from the protein may be marginally influenced by it when one or two of its methylenes strike against the surface of the protein as the linear hydrocarbon writhes within the liquid paraffin, but molecules of amphipathic lipid embracing the protein should be more severely affected. Networks of hydrogen bonds among the hydrophilic head groups of the phospholipids and sphingomyelins (Figure 14-6) may also spread the influence of the protein beyond its immediate vicinity. In this context, lipid in the boundary layer³⁹⁶ surrounding the protein and under its influence has been defined operationally just as water of hydration has been defined operationally. Just as in the case of waters of hydration, a single numerical value for moles of lipid in the boundary layer (mole of protein)⁻¹ is measured.

When 1-palmitoyl-2-stearoylphosphatidylcholine, to which a dimethyl cyclic nitroxyl radical is attached at the 14th carbon of the stearyl group,



is incorporated into bilayers of phosphatidylcholine containing an integral membrane-bound protein such as Ca^{2+} -transporting ATPase, the electron spin resonance spectrum that is observed (Figure 14-21A)³⁹⁷ can be decomposed into two spectra (Figure 14-21B,C) of which it is the sum.^{396,398} One of the component spectra is the same as that of nitroxylphosphatidylcholine **14-16** in pure vesicles of liquid phosphatidylcholine (Figure 14-21F), and one is that of nitroxylphosphatidylcholine **14-16** when its motion is restricted (Figure 14-21E). It was concluded³⁹⁹ that there are two sets of nitroxylphosphatidylcholines **14-16** present in these bilayers, one set constituted by molecules of **restricted mobility** located in the boundary layer immediately adjacent to the protein and the other constituted by molecules of **unrestricted mobility** in the bulk bilayer of phospholipid. The rates at which lipids exchange at positions within the boundary layer⁴⁰⁰ are around 10^7 s^{-1} so no one molecule of phospholipid remains for a significant amount of time within it.

As the ratio between phosphatidylcholine from eggs of *G. gallus* and Ca^{2+} -transporting ATPase in these

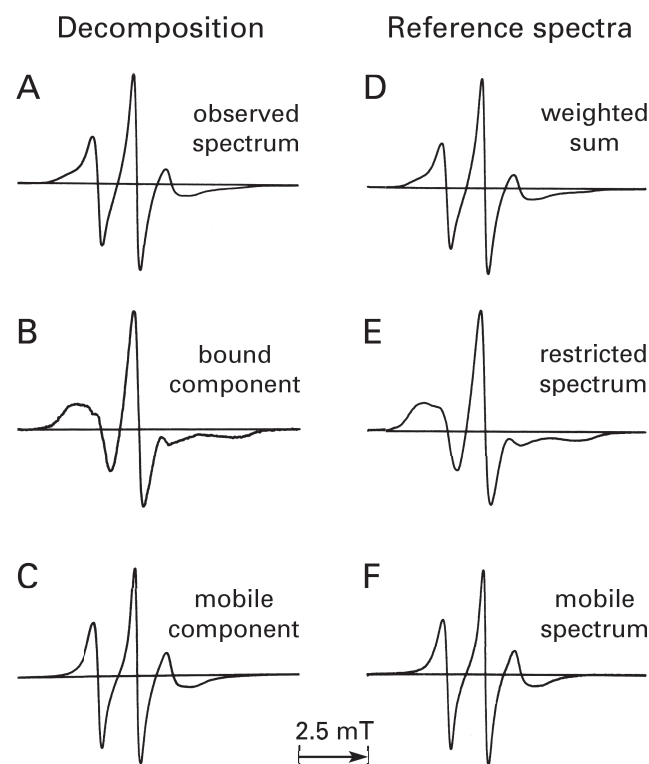


Figure 14-21: Decomposition of the electron spin resonance spectrum of nitroxylphosphatidylcholine **14-16** in vesicles of phospholipid in which Ca^{2+} -transporting ATPase from *O. cuniculus* has been incorporated.³⁹⁷ Ca^{2+} -Transporting ATPase, from which all indigenous phospholipid had been removed, was incorporated into vesicles of phosphatidylcholine from eggs of *G. gallus* into which nitroxylphosphatidylcholine **14-16** was also incorporated. In the final suspension of vesicles containing the protein, the molar ratios of Ca^{2+} -transporting ATPase ($110,500 \text{ g mol}^{-1}$) to phosphatidylcholine to phosphatidylcholine nitroxyl radical were 1:55:0.4. The nitroxyl radical was acting as a probe present in dilute concentration within a solvent of phosphatidylcholine. The observed electron spin resonance spectrum (A) of this probe in this environment could be decomposed into two component spectra. One component (C), which accounted for 54% of the spins producing spectrum A, had the same spectrum as the probe dissolved in the same preparation of phosphatidylcholine in the absence of protein (F). The other component (B), which accounted for 46% of the spins producing spectrum A, was assumed to represent boundary lipid. It had the same spectrum as that of the probe dissolved in a viscous bilayer composed of dipalmitoylphosphatidylcholine and palmitoyloleoylphosphatidylcholine at a ratio of 4:1 (E). The effect of the immobilization of the probe by the viscous bilayer of phospholipid resembles the effect of the protein on the probe (compare spectra B and E). A summation of reference spectra E and F, at a molar ratio of 0.46 to 0.54, produced theoretical spectrum D, which reproduces the observed spectrum A. All spectra are the amplitude of the first derivative of the adsorption of the microwave energy as a function of the strength of the magnetic field (tesla). Reprinted with permission from ref 397. Copyright 1984 American Chemical Society.

membranes is increased, the fraction of restricted nitroxylphosphatidylcholine **14-16** decreases. This decrease results from a competition between unlabeled molecules of phosphatidylcholine and molecules of nitroxylphos-

phatidylcholine **14–16** for positions in the boundary layer adjacent to the protein and the increase in the concentration of the former. From the numerical values of the fraction of restricted nitroxylphosphatidylcholine **14–16** as a function of the molar ratio between phosphatidylcholine and protein, the ratio between the affinities of nitroxylphosphatidylcholine **14–16** and unmodified phosphatidylcholine for positions adjacent to protein can be estimated, and the number of molecules of phosphatidylcholine occupying positions adjacent to the protein can be calculated.⁴⁰¹

In the case of Ca^{2+} -transporting ATPase at 25 °C,^{397,398} phosphatidylcholine and nitroxylphosphatidylcholine **14–16** have equal affinity for positions around the protein and the number of positions is 22 mol (mol of protein)⁻¹. In other words, there are, on the average, 22 molecules of phospholipid in the boundary layer around a molecule of Ca^{2+} -transporting ATPase, each of which can exchange with no bias for a molecule of nitroxylphosphatidylcholine **14–16**, and these 22 molecules of natural, unlabeled phosphatidylcholine occupy locations at which a molecule of nitroxylphosphatidylcholine **14–16** would be restricted in its motion by the protein. Similar measurements with spin-labeled cholesterol demonstrated that cholesterol was able to occupy all of the positions around Ca^{2+} -transporting ATPase within the boundary layer but with an affinity about two-thirds that of natural phosphatidylcholine.³⁹⁷

The **number of lipids in the boundary layer** has been estimated to be 94 ± 10 for each homodimer of cytochrome-*c* oxidase,³⁹⁹ 40 ± 7 for each heteropentamer of acetylcholine receptor,⁴⁰² 21 ± 3 for each monomer of bovine rhodopsin,⁴⁰³ and 10 for each folded polypeptide of myelin proteolipid protein.⁴⁰⁴ The numbers of lipids in the boundary layer is roughly equal to the number of molecules of lipid needed to cover the surface of the sheath for that protein with one layer of linear alkane aligned normal to the plane of the membrane.⁴⁰⁵

It is not necessarily the case that when a position in the boundary layer is occupied by a natural phospholipid rather than nitroxylphosphatidylcholine **14–16**, the motion of that natural phospholipid is noticeably affected. The side chains of the amino acids presented by the sheath surrounding the membrane-spanning bundle of α helices to the linear hydrocarbon of the phospholipid are themselves branched hydrocarbons that are free to rotate fluidly, and the hydrocarbon of phosphatidylcholine may experience little change as it enters or leaves these positions. When Ca^{2+} -transporting ATPase was incorporated into vesicles of dioleoylphosphatidylcholine in which either the 2nd carbons or the 9th and 10th carbons on the two fatty acids were labeled with deuterium rather than a nitroxyl radical, the effect of the protein on the motion of these lipids could be followed by deuterium nuclear magnetic resonance spectroscopy.⁴⁰⁶ The protein was present at a ratio of about 1 mol (100 mol of phospholipid)⁻¹. Statistically signifi-

cant **increases in the anisotropy** at the 9th and 10th carbons were detected when the protein was added to the bilayer but not at the 2nd carbon. The increase in anisotropy observed, even with the assumption that only 20–30% of the lipid was in the boundary layer, was only about 10–20% for the lipids in these positions. This small increase indicates that the protein does not force the hydrocarbon in this region to assume conformations much more irregular than those it would normally assume. The deuterium spin-lattice relaxation times, which are measures of the fluidity of the hydrocarbon, were barely altered by the presence of the protein, and no evidence for two separate sets of phospholipids remaining distinct over time intervals greater than 5 ms was observed.

Just as with soluble proteins, many integral membrane-bound proteins (Table 14–6) are homooligomers (Figures 14–16 and 14–19) or heterooligomers (Figures 14–17 and 14–18). Now that synthetically pure detergents are available, it is possible to dissolve a membrane, and if an oligomeric protein of interest is stable enough, the dissolution can produce a monodisperse solution of that membrane-bound oligomer. Each complex between a micelle of the detergent and an oligomer of that protein in such a solution has the same shape, as judged by its frictional coefficient, and the same molar mass, as judged by sedimentation equilibrium.^{407,408} Furthermore, it seems to be the case that **the oligomer found in solution** upon dissolving the membrane is often, but not always,²²⁴ the same one originally present in the membrane.²²²

While **quantitative cross-linking** gives reliable determinations of the number of subunits in water-soluble oligomeric proteins and for integral membrane-bound proteins in solutions of detergent,^{224,347} it fails to do so for membrane-bound oligomers that are still within the membrane. The problem is that the membrane-bound proteins within a native membrane are always at too high a concentration, and it is technically difficult to dilute them sufficiently to prevent them from cross-linking intermolecularly. Even when the membranes contain only one protein, quantitative cross-linking usually produces covalent polymers so large that their complexes with dodecyl sulfate do not even enter an acrylamide gel upon electrophoresis.²²⁵ If particular care is taken to reconstitute a membrane-bound protein at high ratios of phospholipid to protein so that each reconstituted vesicle is large and each contains only a few molecules of the protein, intermolecular cross-linking can be suppressed sufficiently that intramolecular cross-linking can give an accurate assessment of the oligomeric state of the protein.⁴⁰⁹ The ideal preparation for assessment of the number of subunits in the oligomer of an integral membrane-bound protein by quantitative cross-linking would be a suspension of vesicles in which each vesicle contained only one copy or no copies of the oligomer.

Oligomeric proteins constructed from identical subunits always incorporate rotational axes of symmetry into their structures. In a bilayer of amphipathic lipids, all of the identical subunits of either an oligomeric, anchored membrane-bound protein or an oligomeric, integral membrane-bound protein are inserted so that they point in the same direction. Because the same hydrophobic segments of their common amino acid sequence span the bilayer, all folded polypeptides of the same sequence float at the same depth in the membrane and have the same orientation. These inescapable requirements placed upon the common structure of the subunits of oligomeric membrane-bound proteins force any rotational axis of symmetry relating the individual subunits in the protein to be normal to the plane of the bilayer. Therefore, a membrane-bound homooligomeric protein can have only one rotational axis of symmetry, and that axis will be normal to the plane of the membrane. Consequently, all **membrane-bound oligomeric proteins** must have **cyclic symmetry**, and dihedral symmetry⁴¹⁰ is prohibited. The same argument can be made for membrane-bound heterooligomers formed from homologous, superposable subunits. These heterooligomers will contain rotational axes of pseudosymmetry normal to the plane of the membrane.

Closed structures with rotational axes of symmetry are even more exclusive necessities for oligomeric membrane-bound proteins than they are for soluble proteins. A screw axis of symmetry is incompatible with the vectorial two-dimensional distribution of identical or homologous subunits enforced by the bilayer, and helical polymeric fibers are not available structures. The only interfaces that could propagate a linear polymer of indefinite length in a membrane would have to form from complementary faces arrayed at precisely 180° across from each other on each subunit, or the row of subunits produced by them would eventually come around upon itself to form an unbroken or a broken ring. During evolution by natural selection, therefore, every time two complementary faces appear at random on the surface of one monomer of a membrane-bound protein such that those two faces can produce a series of interfaces joining several of the subunits in an oligomer, either an incomplete ring or a complete ring of an integral number of subunits will always form. A complete ring, because no pair of complementary faces remains unassociated, is a more stable structure than an incomplete ring. If the angle between the two complementary faces on a single subunit is an integral quotient of 360°, where the integer is greater than or equal to 3, complete rings containing that number of subunits will form. Because only one rotational axis of symmetry normal to the bilayer is available, magic numbers such as four or six, applicable to soluble oligomeric proteins having sets of perpendicular rotational axes of symmetry, are irrelevant to membrane-bound oligomeric proteins. In addition, unlike a soluble protein such as hemoglobin, a membrane-bound pro-

tein cannot be tetrameric under one set of conditions and dimeric under another.^{410,411}

It is not surprising that the oligomeric membrane-bound proteins the structures of which have been directly observed are all assembled around **single rotational axes of symmetry normal to the plane of the bilayer**. In this case, the distinction between anchored and integral membrane-bound proteins is irrelevant because both are constrained by the same requirements. Thus both bacteriorhodopsin,⁴¹² an integral membrane-bound protein, and the hemagglutinin of influenza virus,²⁸⁷ an anchored membrane-bound protein, have three identical subunits arrayed around a 3-fold rotational axis of symmetry normal to the plane of the membrane. Photosynthetic reaction center, an integral membrane-bound protein containing two different polypeptides with homologous amino acid sequences,^{392,393} has those two folded polypeptides arrayed around a 2-fold rotational axis of pseudosymmetry normal to the plane of the membrane (Figure 14-17).⁴¹³ Exo- α -sialidase of influenza virus, an anchored membrane-bound protein, has four identical subunits arrayed around a 4-fold rotational axis of symmetry.⁴¹⁴

Acetylcholine receptors from *T. californica* and *T. marmorata* are integral membrane-bound proteins. Each is an $\alpha_2\beta\gamma\delta$ heteropentamer^{415,416} constructed from four unique polypeptides,^{417,418} designated α , β , γ , and δ on the basis of their electrophoretic mobilities. All four of these polypeptides are glycoproteins,^{391,419} and all four are homologous in sequence.⁴²⁰⁻⁴²³ All four can be readily aligned, and in the six pairwise comparisons the percent identity averages around 40%.⁴²² These four distinct subunits were derived from a common ancestor and all four of them assume the same unique superposable tertiary structure upon folding.³⁴⁶ The five subunits, $\alpha_2\beta\gamma\delta$, in the native structure of acetylcholine receptor are arrayed around a 5-fold **rotational axis of pseudosymmetry** normal to the plane of the membrane (Figure 14-22).^{424,425}

Gap junction connexon, an integral membrane-bound protein, has six identical subunits arrayed around a 6-fold rotational axis of symmetry.⁴²⁶ In each gap junction between two cells, each connexon, which is a ring of six subunits in the plasma membrane of one of the cells, is associated by a 2-fold rotational axis of symmetry with another gap junction connexon in the membrane of the other cell to produce a dodecamer that is a dimer of hexamers with dihedral symmetry of point group 622 (D_6) with one hexamer in each plasma membrane. Within the plasma membrane of a given cell, the connexons are in crystalline arrays of indefinite surface area.⁴²⁷ A different hexamer of identical subunits arrayed around a 6-fold rotational axis of symmetry normal to the plane of the bilayer⁴²⁸ is the major constituent of the luminal plasma membrane of urinary bladder, and this protein is also present in the cell in a crystalline array.⁴²⁹

There is a group of proteins the crystallographic

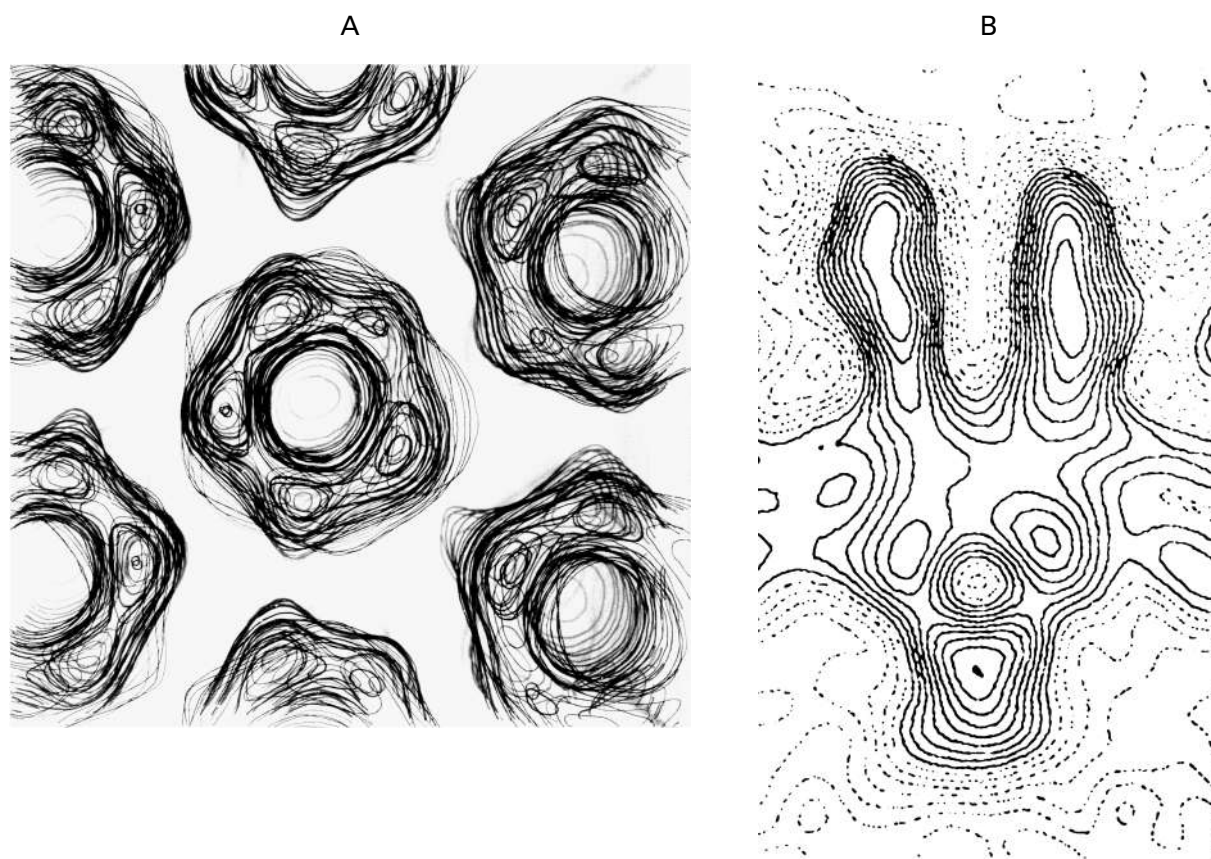


Figure 14-22: Map of electron scattering density of acetylcholine receptor embedded in a glass of amorphous ice.^{424,425} Membranes enriched in acetylcholine receptor were prepared from electric organs of *Torpedo marmorata* by differential centrifugation. The membranes were resuspended in 0.1 M tris(hydroxymethyl)aminomethane hydrochloride, pH 6.8, and allowed to stand at 10 °C for 1 month, at which time long (<1 μm) cylindrical tubes 70 nm in diameter had formed. These tubes were helical, crystalline arrays of molecules of acetylcholine receptor within tubular bilayers of the amphipathic lipid from the membranes. The asymmetric unit in the helical array was a dimer of identical acetylcholine receptors. The asymmetric units formed the rows of a 15-stranded left-handed helical array in one dimension and the rows of a 5-stranded right-handed helical array in the other dimension of the $(-15,5)$ surface lattice. These tubes were embedded in a thin layer of amorphous ice on a film of carbon on an electron microscopic grid. Digitized electron micrographs of these tubes were submitted to Fourier transformation. The layer lines of the resulting diffraction pattern were indexed, and variations in phase and amplitude along the layer lines were measured from these diffraction patterns. These functions were then submitted to Fourier-Bessel inversion to obtain a three-dimensional map of electron scattering density for the tube. (A) View perpendicular to the surface of the tube of this map of scattering density. The image was made by stacking about 20 successive sheets of clear plastic of the appropriate thickness, each with a cross section of the map traced upon it. The successive sections chosen were 0.5 nm apart. Reprinted with permission from *Nature*, ref 424. Copyright 1985 Macmillan Magazines Limited. (B) Cross section through the center of a molecule of acetylcholine receptor in a plane normal to the axis of the tube. The blocklike structure at the bottom of the image is thought to be a protein other than acetylcholine receptor. The bilayer of amphipathic lipid is to the right and left. There is a deep cylindrical depression on the upper, extracytoplasmic surface of the protein and a small shallow depression on the lower, cytoplasmic surface of the protein. The five subunits arrayed about a 5-fold rotational axis of pseudosymmetry produce a thick cylindrical pipe about 7 nm in diameter and 5 nm in height with a wall 2.5 nm thick extending out from the extracytoplasmic surface of the membrane. Reprinted with permission from ref 425. Copyright 1990 Rockefeller University Press.

molecular models of which at first glance seem to contradict the axiom that all rotational axes of symmetry in integral membrane-bound homooligomers must be normal to the plane of the membrane. Aquaporin serves as a paradigm for the proteins in this group. Aquaporin is an α_4 homotetramer, which as expected has cyclic symmetry of point group 4 (C_4). There is no doubt, however, that the subunit of the aquaporin tetramer (Figure 14-23)³⁴⁸ is the product of an internal duplication. This conclusion follows from the fact that the amino acid sequence from Valine 51 to Cysteine 88 in the first half of bovine aquaporin can be readily aligned (30% identity

with no gaps) with the amino acid sequence from Glycine 167 to Threonine 204 in the second half of the molecule. This alignment brings into register the amino acids forming the second and third membrane-incorporated α helices (Figure 14-23) from the first half of the folded polypeptide with those forming the membrane-incorporated sixth and seventh α helices, respectively, from the second half of the folded polypeptide. There are eight membrane-incorporated α helices in all, so these pairings are the ones expected from an internal duplication. Outside this central region of the protein, the two amino acid sequences cannot be aligned with statistical signifi-

cance, but when the crystallographic molecular model of the subunit is viewed from the proper angle (Figure 14–23), it can be seen that there is an obvious 2-fold rotational axis of pseudosymmetry that superposes the first half of the folded polypeptide onto the second half. This superposition is made all the more compelling by the fact that the peculiar structure of the third membrane-incorporated α helix, which passes only halfway across the membrane before the polypeptide doubles back, is exactly mimicked by that of the seventh membrane-incorporated α helix.

The puzzling aspect of this 2-fold rotational axis of pseudosymmetry within the subunit of aquaporin is that it is parallel to the plane of the membrane. Most soluble proteins or subunits of soluble proteins that are the products of duplications of the genes encoding their ancestors have 2-fold rotational axes of pseudosymmetry relating the duplicated halves (Figure 9–18). It has usually been assumed that these are the vestiges of 2-fold rotational axes of symmetry that related the two halves when they were identical subunits in a homodimer before the carboxy terminus of one subunit was joined to the amino terminus of the other subunit by the duplication of the ancestral gene. If this were the case in aquaporin, then 2-fold rotational axis of symmetry relating the two identical subunits of the homodimer that was the ancestral protein would have had to be parallel to the plane of the membrane.

Because this symmetry is not possible, it follows that there was no ancestral homodimer of aquaporin and that the gene that was duplicated was one encoding a monomeric integral membrane-bound protein. Because the amino terminus of that ancestral monomer was on the cytoplasmic side of the membrane and its carboxy terminus was on the extracytoplasmic side, once the fourth membrane-incorporated α helix of the new internally duplicated protein had been inserted into the membrane, the machinery responsible for inserting the ancestral monomer into the membrane came upon a segment of polypeptide that formed the first membrane-spanning α helix in the ancestral protein but was now on the wrong side of the membrane. It was simply incorporated in the opposite direction along with the other three α helices of the duplicated half. When the two halves of the internally duplicated protein cleaved to each other, as they were bound to do, the usual 2-fold rotational axis relating the two halves of a dimer was created. Because the two halves of the protein were inserted in the membrane in opposite directions, that 2-fold rotational axis of pseudosymmetry had to be parallel to the plane of the membrane.

There are a number of other integral membrane-bound proteins the subunits or monomers of which have 2-fold rotational axes of pseudosymmetry parallel to the plane of the membrane relating the two halves of what are assumed to be the products of internal duplications.^{430–434} Because the membrane-spanning portions of

these proteins are all bundles of membrane-spanning α helices, a situation that makes any superposition more likely, and because most of these proteins have evolved to the extent that obvious alignments of amino acid sequence can no longer be made, the conclusion that the two halves result from an internal duplication usually relies almost entirely on the order in which the superposed α helices occur in the sequence of the protein. The order in which the α helices paired by the superposition around the rotational axis of pseudosymmetry occur in the second half of the amino acid sequence is the same

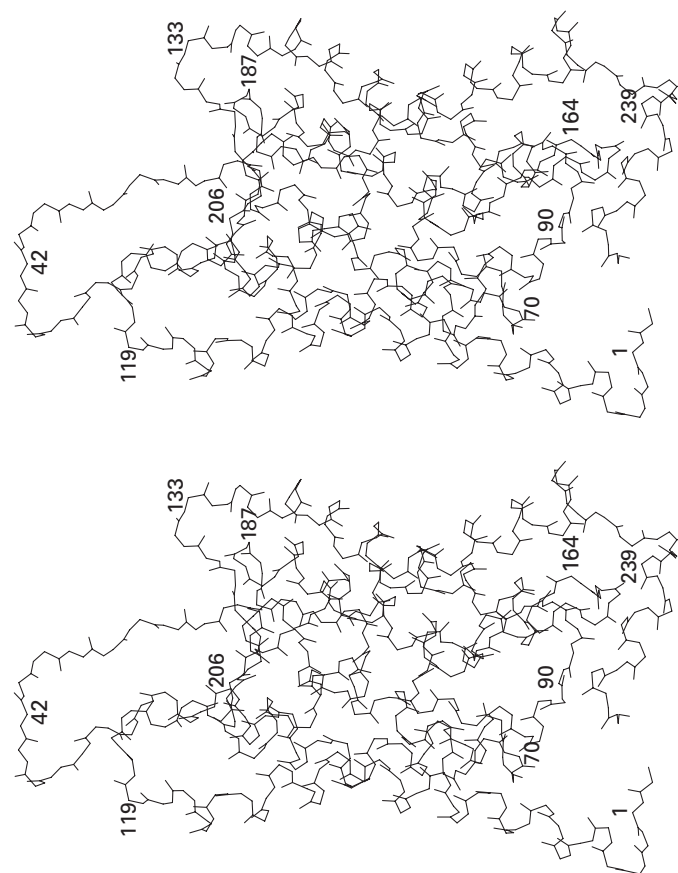


Figure 14–23: Skeletal drawing of the polypeptide backbone of the crystallographic molecular model (Table 14–6) of one of the subunits of isoform 1 of aquaporin purified from bovine erythrocytes.³⁴⁸ The view is down the 2-fold rotational axis of pseudosymmetry relating the two halves of the internal duplication in this protein. This drawing was produced with MolScript.⁷⁷⁸

as the order in which they occur in the first half of the amino acid sequence.

In each of these other proteins, as is the case with aquaporin, the junction between the carboxy terminus of the first half of the protein and the amino terminus of the second half of the protein is located on the opposite side of the membrane from the amino terminus of the first half. In all of these other cases, this topography results from the fact that the two halves have an odd number of membrane-spanning α helices. Consequently, following the gene duplication, the first segment of polypeptide encoding a membrane-spanning α helix at the amino terminus of the second half of each of these proteins found itself on the wrong side of the membrane during the insertion of the complete protein into the membrane and was simply incorporated in the opposite direction from the direction in which its twin at the amino terminus of the first half of the protein had been incorporated. The existence of these proteins reiterates the fact that larger proteins are created by the fusions and duplications of genes encoding their smaller ancestors.

The interfaces within the membrane among the subunits of an oligomeric membrane-bound protein should be stabilized from dissociation by different non-covalent forces from those stabilizing interfaces within soluble proteins.⁴³⁵ Because the faces forming an **interface between subunits** become surrounded by hydrocarbon upon dissociation rather than by water, the hydrophobic effect should be irrelevant. On the other hand, hydrogen bonding, which provides little favorable free energy to the formation of an interface within a water-soluble protein, should exert its full effect on stabilizing the interface of an oligomeric membrane-bound protein within the bilayer because when an interface dissociates, the hydrogen-bond donors and acceptors within it lose their acceptors and donors, respectively. These considerations were validated in a study of the effect of site-directed mutation on the interface between two identical membrane-spanning α helices within the interface forming the α_2 dimer of glycophorin.^{435–437} When the threonine within the interface was mutated to an alanine, the increase in its dissociation constant was much greater than when that threonine was mutated to a serine. Changes in hydrophobicity produced by other mutations, however, did not correlate with the respective changes in dissociation constant, but changes in the detailed stereochemical fit of one face to the other did correlate with changes in dissociation constant. These results suggest that both hydrogen bonding and stereochemical fit, but not hydrophobicity, determine the strength of an interface within the hydrocarbon of a membrane.

For those integral membrane-bound proteins that have not yet been crystallized in three dimensions, molecular models of lower resolution are often obtained from **image reconstruction** of two-dimensional crystalline arrays of these proteins still within the membrane. The protein most suited for such an approach so far has

been bacteriorhodopsin from *H. salinarium* because it is already in a two-dimensional crystalline array within the plasma membrane of the bacterium.⁴³⁸ The space group of this two-dimensional array is *P3*, and the asymmetric units related by the 3-fold rotational axes of symmetry within the unit cell are individual subunits of bacteriorhodopsin (Figure 14–14). Although bacteriorhodopsin is one of the few membrane-bound proteins that is naturally crystalline, most integral membrane-bound proteins can be induced to crystallize in two dimensions.⁴³⁹ Such crystals are then embedded in a **glass of amorphous ice**⁴⁴⁰ and are examined in a **cryo-electron microscope**, an electron microscope in which the embedded specimen is maintained at a temperature of 4 K while images are prepared.⁴⁴¹

A **two-dimensional crystalline array** of an integral membrane-bound protein within a bilayer of lipids (Figure 14–24)⁴³⁹ is a three-dimensional distribution of electron scattering density, $\theta(x,y,z)$, that is periodic in the two dimensions of the plane of the membrane. The electrons that it scatters in an electron microscope will form an electron diffraction pattern (Figure 14–25).⁴⁴² This diffraction pattern does not arise from reflections generated by sets of parallel planes running through a three-dimensional lattice (Figure 4–8) but from reflections generated by sets of parallel lines running through the two-dimensional lattice (Figure 4–4) of the projection of the three-dimensional array on a plane normal to the beam of electrons. Each reflection has an amplitude, an index, and a phase, but as always, the phases cannot be measured directly.

The indexed set of amplitudes and phases of the electron diffraction pattern from such an array are the amplitudes and phases of the Fourier transform of the projection of the three-dimensional distribution of electron scattering density of the array upon the plane normal to the axis of the beam (Equation 9–5). Therefore, they are the amplitudes and phases of a central section through the three-dimensional Fourier transform of the three-dimensional distribution of electron scattering density (Equation 9–4).

An electron micrograph of a two-dimensional crystalline array of a membrane-bound protein (Figure 14–24) is a projection of the three-dimensional electron scattering density $\theta(x,y,z)$ of the array upon a plane normal to the axis of the beam of electrons (Equation 9–6). From the digitized distribution of contrast on the electron micrograph, the amplitudes and phases of the central section of the Fourier transform of the three-dimensional electron scattering density can be calculated by a computer (Equation 9–5). Because the array is of a two-dimensional crystal, the central section through its Fourier transform is a lattice of spots. Each spot in the transform calculated by the computer from the digitized micrograph corresponds to one of the reflections in the electron diffraction pattern (Figure 14–25) and has the same phase and the same relative amplitude as it does.

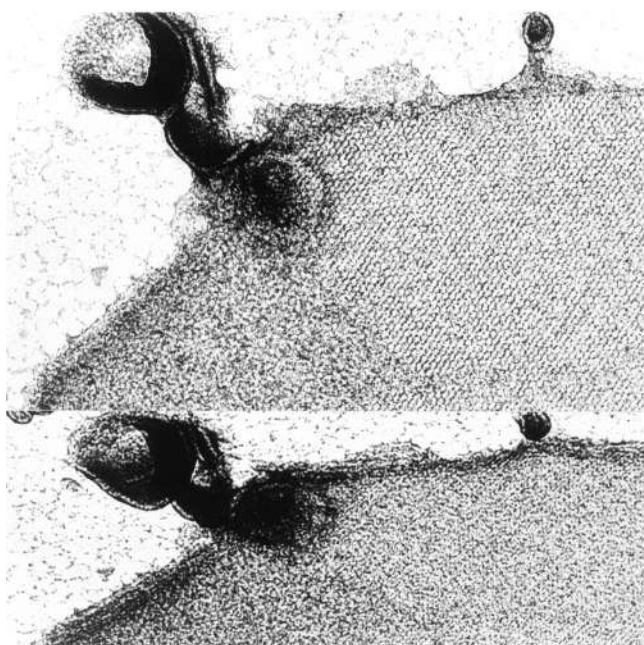


Figure 14-24: Electron micrograph of a two-dimensional crystalline array of bovine cytochrome-*c* oxidase.⁴³⁹ Mitochondria from bovine heart were sonicated, and the resulting fragments of membrane were separated by differential centrifugation. Fragments rich in cytochrome-*c* oxidase were extracted with Triton X-114 and then Triton X-100 to dissolve away other proteins. The purified particulate material was composed of fragments of membrane in which the only protein was cytochrome-*c* oxidase. These fragments of membrane were attached to carbon films and embedded in a glass of the negative stain uranyl acetate. Upon examination in the electron microscope, it was found that in many of the fragments the cytochrome-*c* oxidase had crystallized into two-dimensional arrays. The upper electron micrograph is of one of these arrays viewed normal to the electron beam. The lower electron micrograph is the same array tilted on a horizontal axis so that its plane was at an angle of 36° to the beam of electrons. Reprinted with permission from ref 439. Copyright 1977 Academic Press.

This permits the **phases** to be estimated from the micrograph and the **amplitudes** to be measured from the electron diffraction pattern.

The Fourier transform of the periodic three-dimensional distribution of electron density that is a three-dimensional crystal of a protein is a three-dimensional lattice of peaks in reciprocal space. Each peak has an amplitude and a phase. Each reflection in the diffraction of X-radiation from the crystal represents one of these peaks. The Fourier transform of the three-dimensional distribution of scattering density in a crystalline array of a membrane-bound protein, which is periodic in only two dimensions, is a lattice of parallel lines in reciprocal space. Each of these parallel **lattice lines** has an amplitude and a phase that vary periodically along its length. If the variations of the amplitudes and phases along each of these lattice lines could be measured, the three-dimensional distribution of electron scattering density in the unit cell of the crystalline array could be calculated by Fourier transformation of this set of functions.

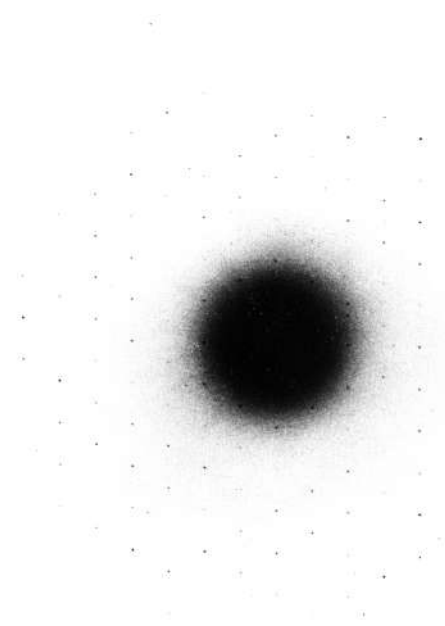


Figure 14-25: Electron diffraction pattern produced by a two-dimensional crystalline array of bacteriorhodopsin from *H. salinarium*.⁴¹² Fragments of membrane containing crystalline arrays of bacteriorhodopsin were attached to a film of carbon and embedded in a glass of glucose. The specimen was centered in the beam of electrons of an electron microscope and a photographic plate was used to record the reflections of the diffraction pattern. The reflections emerge from the specimen at characteristic angles determined by the lattice, and they are recorded on a piece of film at a known distance from the specimen. The dark central peak is the majority of the electrons that passed through the specimen undeflected. The sharp dots of varying intensity are the reflections themselves. Reprinted with permission from ref 412. Copyright 1975 Academic Press.

If a crystalline array is **tilted** in the electron beam (Figure 14-24, lower panel), a projection along an axis tilted relative to the axis normal to the plane of the array is recorded on the micrograph, and the amplitudes and phases of the electron diffraction pattern, which are now the amplitudes and phases of the Fourier transform of this new projection, have changed. Each of the micrographs and electron diffraction patterns in a series in which the specimen is systematically tilted represents a different central section through the lattice of lines in the Fourier transform of the three-dimensional array of scattering density that is periodic in two dimensions.⁴⁴² If enough of these central sections are gathered, the amplitudes and phases of the Fourier transform within certain ranges along the lattice lines can be gathered (Figure 14-26).³²⁷ The amplitudes can be gathered from either electron diffraction patterns or Fourier transforms of the digitized contrast on electron micrographs, but the phases can be obtained only from Fourier transforms of the digitized distributions of contrast on the electron micrographs. In such reconstructions, the details of the scattering density fade away at the two ends of the molecule above and below the bilayer of phospholipid

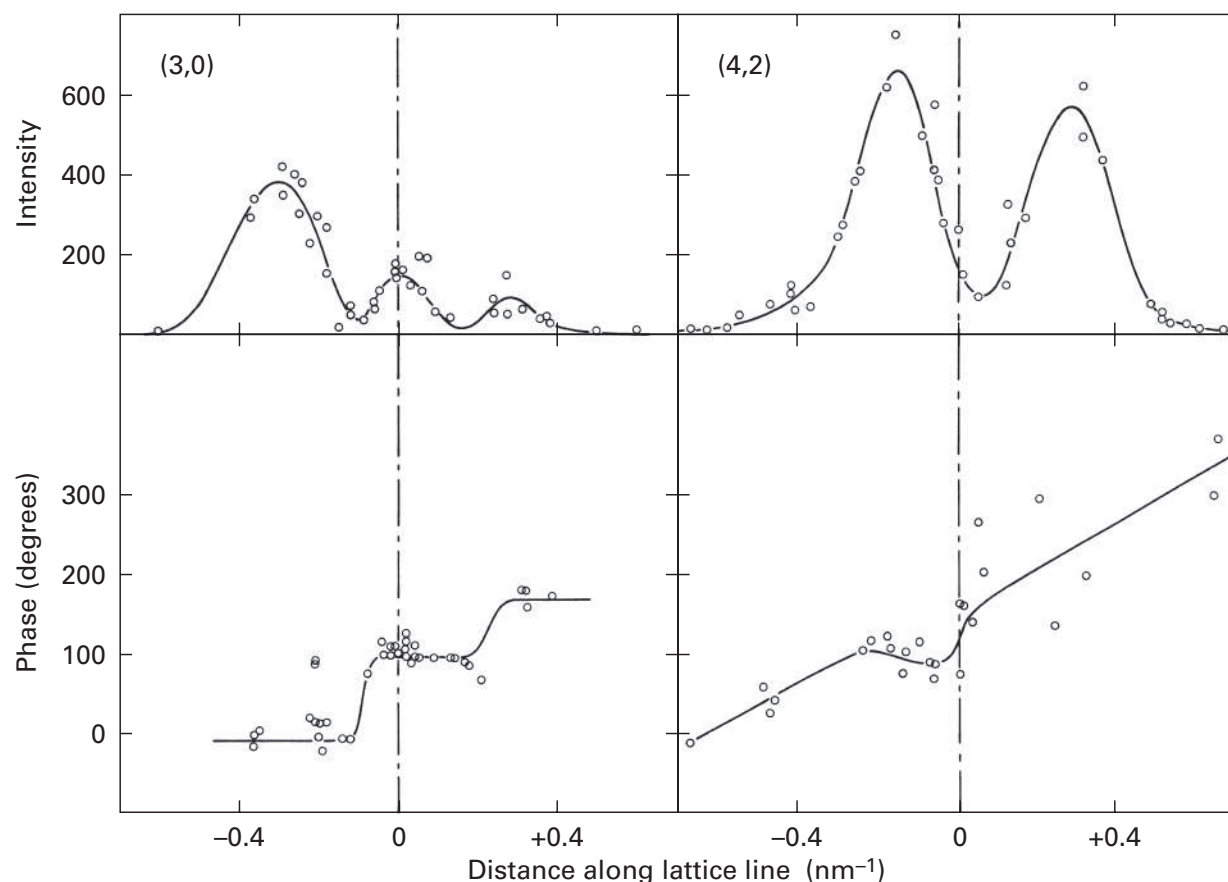


Figure 14-26: Variation of intensity and phase along two of the lattice lines [(3,0) and (4,2)] in the Fourier transform of the three-dimensional distribution of electron scattering density within the two-dimensional crystal of bacteriorhodopsin from *H. salinarium*.³²⁷ The intensities are the intensities of the reflections on electron diffraction patterns of the array (Figure 14-25) tilted at various angles. The phases were determined from Fourier transforms of the distribution of contrast on electron micrographs of the specimens tilted at the same angles. The lattice lines that occur in the Fourier transform of the electron micrograph are the same as the lattice lines upon which the reflections of the electron diffraction lie because the electron diffraction pattern is the same Fourier transform of the same array. Each data point on each graph represents a measurement from a different electron diffraction pattern or a different electron micrograph, respectively, from a specimen at a different angle of tilt. As the angle of tilt is varied, different positions along the lattice lines are sampled. The phase (degrees) or intensity (in arbitrary units) is presented as a function of the distance along the lattice line (nanometer⁻¹). Reprinted with permission from *Nature*, ref 327. Copyright 1975 Macmillan Magazines Ltd.

because the specimen can be tilted only between -60° and $+60^\circ$, and the amplitudes and phases on the lattice lines arising from the unseen portions of the protein are outside the regions that can be sampled with these tilts.

When the amplitudes of the electron diffraction patterns of crystalline arrays of bacteriorhodopsin embedded in a glass of glucose and tilted at various angles were combined with the phases from the digitized distributions of contrast in electron micrographs taken at the same angles of tilt, a map of the three-dimensional electron scattering density within the asymmetric unit in the unit cell of the two-dimensional crystal of the protein could be calculated.^{327,328,443} At low resolution, when only reflections from lattice lines with Bragg spacing out to 0.7 nm were included in the calculation,³²⁷ seven **rods of scattering density** aligned roughly perpendicular to the plane of the array within the membrane were observed.^{327,328} The seven rods are the seven α helices that span the membrane in bacteriorhodopsin (Figure 14-14).

When reflections from lattice lines with Bragg spacing out to 0.35 nm were included in the calculation of the map of electron scattering density,³²⁸ the same seven rods of electron scattering density were observed, but the rods became more detailed at their ends so that some of the connections between the rods could be observed and, more importantly, protrusions of electron scattering density appeared along the rods. These protrusions represent the side chains of the amino acids in the sequence of the protein. From the few connections observed and the pattern of the largest protrusions along each rod (representing the aromatic amino acids), a molecular model of the polypeptide, built with the known amino acid sequence, could be unambiguously placed into the map of electron scattering density. The map of electron density could be further improved by eliminating the contribution of diffuse electron scattering to the amplitudes of the electron diffraction, and the crystallographic molecular model was submitted to

refinement against the observed amplitudes of the electron diffraction.⁴⁴³ The final, refined molecular model duplicated the X-ray crystallographic molecular model of the protein that was subsequently reported³²⁵ in the arrangement of the α helices and their relative positions and orientations but lacked many of the atomic details of the latter molecular model. In particular, the details of the structure outside the membrane were ill-defined in the molecular model from electron diffraction, because the range of tilt that could be performed did not permit reflections from these regions to be gathered to sufficiently small Bragg spacing.

It has also been possible to obtain a map of scattering density of acetylcholine receptor from *Torpedo marmorata* from electron scattering and image reconstruction (Figure 14-22). A map of scattering density calculated from data sets with small Bragg spacing has sufficient detail within the bilayer to observe the protrusions of the side chains from the four membrane-spanning α helices of each of its five homologous subunits. From the pattern of these protrusions and the necessary structural homology of the five subunits, it was possible to insert the amino acid sequences of the various membrane-spanning segments into the map of electron scattering density.³⁴⁵ The structure outside the membrane was again, ill-defined, but serendipitously, there already was a crystallographic molecular model of a water-soluble protein homologous to the globular extracytoplasmic portions of acetylcholine receptor. This crystallographic molecular model⁴⁴⁴ could be positioned in the map of scattering density for the portions of acetylcholine receptor on the extracytoplasmic surface of the bilayer of phospholipids, and the amino acid sequences of acetylcholine receptor could be substituted into the folded polypeptide that had been positioned. The molecular model that resulted could be submitted to refinement against the amplitudes of the electron diffraction to obtain a crystallographic molecular model of acetylcholine receptor both within the membrane and on its extracytoplasmic surface.³⁴⁶

Maps of electron-scattering density have been calculated from electron diffraction and Fourier transformation of digitized images of tilted specimens in cryo-electron microscopes for light-harvesting chlorophyll *a/b*-protein complex (Bragg spacing ≥ 0.34 nm),⁴⁴⁵ isoform 1 of human aquaporin (Bragg spacing ≥ 0.38 nm),^{446,447} sodium/proton antiporter NhaA (Bragg spacing ≥ 0.7 nm),⁴⁴⁸ gap-junction channel (Bragg spacing ≥ 0.75 nm),⁴⁴⁹ Na⁺/K⁺-exchanging ATPase (Bragg spacing ≥ 0.95 nm),⁴⁵⁰ and Ca²⁺-transporting ATPase (Bragg spacing ≥ 1.4 nm).⁴⁵¹ In the first two instances, the maps of electron scattering density were detailed enough to identify the membrane-spanning α helices and position them in their proper orientations and relative positions; but in the case of isoform 1 of aquaporin, an X-ray crystallographic molecular model (Bragg spacing ≥ 0.22 nm) was reported shortly afterward.³⁴⁸

Nuclear magnetic resonance has also been used to obtain structural information about either peptides forming single membrane-spanning α helices or small integral membrane-bound proteins with one or two membrane-spanning α helices. If the peptide or the small integral membrane-bound protein can be inserted into small, uniform micelles of detergent to form a monodisperse solution⁴³⁷ or if it can be monodispersely dissolved in a mixture of miscible organic solvents in water⁴⁵² in such a way that its normal structure is retained, the usual two- and three-dimensional nuclear magnetic resonance spectra can be gathered from these solutions, and the chemical shifts of the nuclei of ¹hydrogens, ¹³carbons, and ¹⁵nitrogens in the peptide or protein can be assigned, just as though it were a peptide or protein that normally dissolves unassisted in water. The nuclear Overhauser effects among the nuclei of the ¹hydrogens can then be used to define the structures of these peptides or proteins.

It is also possible to obtain two- and three-dimensional nuclear magnetic resonance spectra⁴⁵³⁻⁴⁵⁵ from peptides or small proteins that span a membrane in a single α helix when they are inserted into oriented multibilayers of phospholipid such as those described in Figure 14-4. The chemical shifts of the nuclei of ¹hydrogens and ¹⁵nitrogens can be assigned, and information about the tilt of the α helices in the bilayers of phospholipid can be obtained from the fluctuations of the coupling constants between nuclei of the respective α ¹hydrogens and adjacent amido ¹⁵nitrogens.⁴⁵⁶

Electron spin resonance from probes attached to membrane-bound proteins has been used to obtain information about membrane-spanning α helices in larger integral membrane-bound proteins formed from bundles of such α helices. All of the naturally occurring cysteines in an integral membrane-bound protein are mutated to other amino acids. A segment of the amino acid sequence that is thought to be a membrane-spanning α helix in the native structure of the resulting cysteineless protein is identified from its hydropathy. Each of the consecutive amino acids in that segment is mutated in turn to a cysteine to produce a set of single point mutants. Each mutant is covalently modified at the respective cysteine with (1-oxyl-2,2,5,5-tetramethylpyrroline-3-methyl)methanethiosulfonate (see 12-11). The frequency with which molecular oxygen is able to collide with the nitroxyl radical in each of the covalently labeled native proteins is then estimated from the effect that changes in the concentration of molecular oxygen in the sample have on the rates of relaxation of the respective unpaired electron. Molecular oxygen, because it is paramagnetic, catalyzes the **relaxation of the unpaired electron** when it encounters the nitroxyl radical.

When the consecutively modified positions in the amino acid sequence of the protein are in a membrane-spanning α helix on the outer surface of the native structure of the integral membrane-bound protein, the

accessibility of the respective nitroxyl radicals to oxygen varies periodically along that α helix as it directs them periodically into the hydrocarbon of the phospholipid in which the molecular oxygen is dissolved or into the center of the protein in which it is not (Figure 14-27A).^{457,458}

When the modified positions in the amino acid sequence of the modified protein are in a loop between two membrane-spanning α helices, the rates of relaxation of the respective unpaired electrons are accelerated by chelated paramagnetic ions such as chromium oxalate (Figure 14-27B)⁴⁵⁷ or nickel *N,N'*-dicarboxymethyl-1,2-diaminoethane dissolved in the aqueous phase to which the loop is exposed.⁴⁵⁹ The increase in relaxation by oxygen is less pronounced for nitroxyl radicals in these locations because oxygen is less soluble in water than in hydrocarbon.

A segment of amino acid sequence in the protein is identified as a membrane-spanning α helix in the native integral membrane-bound protein if the excited state of nitroxyl radicals attached to its amino acids is not relaxed by the chelated paramagnetic ions⁴⁵⁹ but does display a pattern of exposure to molecular oxygen that fluctuates with a period of 3.6 aa (Figure 14-27). In the case of bacteriorhodopsin from *H. salinarium*, the prediction that the polypeptide between Threonine 128 and Tyrosine

131 would be a loop between two α helices while the polypeptide between Serine 132 and Threonine 142 would be within an α helix spanning the membrane was validated when a crystallographic molecular model of the protein (Figure 14-14) became available.³²⁸ That the periodicity observed in the catalysis of the relaxation of the unpaired electron coincided with periodicity of the exposure of the respective side chain to the hydrocarbon of the bilayer of phospholipid was also validated by the model.

Similar **periodicities of exposure** of covalently attached nitroxyl radicals have been observed for sets of consecutive mutants to cysteine in segments of amino acid sequence from lactose permease of *E. coli*,⁴⁶⁰ diphtheria toxin from corynebacterium β ,⁴⁶¹ and citrate transport protein from *S. cerevisiae*.⁴⁶² It is assumed that these segments from these integral membrane-bound proteins are membrane-spanning α helices even though their crystallographic molecular models are not yet available.

Fluctuations, also with a period of 3.6 aa, in levels of expression in oocytes of *X. laevis* and apparent dissociation constant for acetylcholine upon consecutive mutation to tryptophan of amino acids in a segment of the amino acid sequence of acetylcholine receptor from *T. californica* were presented as evidence that this segment spanned the membrane as an α helix.⁴⁶³ Likewise,

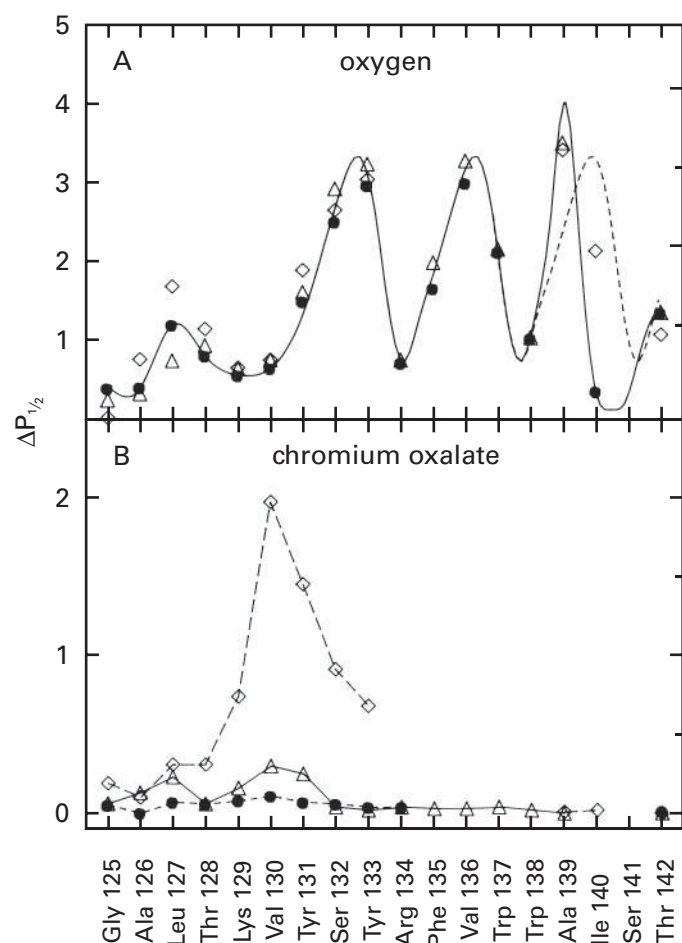


Figure 14-27: Periodic variation in exposure of side chains to the hydrocarbon of the bilayer of phospholipids along an α helix that spans the membrane in bacteriorhodopsin from *H. salinarium*.⁴⁵⁹ The gene for the protein was expressed in *E. coli*. All of the cysteines in the wild-type protein were replaced by site-directed mutation. A set of mutants of this cysteineless protein was then constructed and expressed. Each member of this set had one of the amino acids in the sequence from Glycine 125 to Threonine 142 replaced, respectively, by a cysteine, and the complete set contained each of the consecutive single point mutants. Each mutant protein was overexpressed in *E. coli*, and the product accumulated in the bacteria as a denatured precipitate. The respective polypeptides were purified⁴⁵⁸ and modified with (1-oxyl-2,2,5,5-tetramethylpyrrolidine-3-methyl)methanethiosulfonate at their single cysteines while in their unfolded states. The covalently modified, unfolded polypeptides were reconstituted to their native state by treatment with detergent, phospholipid, and retinal,⁴⁵⁸ and the refolded states of each protein were shown to be fully functional by kinetic and spectroscopic analysis, and by their ability to transport protons upon absorption of light. The spin-lattice relaxation time of the unpaired electron in each of the nitroxyl radicals was monitored indirectly by determining $P_{1/2}$, the continuous wave power at which the amplitude of the signal from the central absorption in the spectrum (Figure 12-35) was 50% of its absorption in the absence of saturation. The change in $P_{1/2}$ ($\Delta P_{1/2}$) for the respective nitroxyl radical at the noted position in the sequence in the presence, relative to the absence, of (A) molecular oxygen (O_2) and (B) chromium oxalate is presented. Measurements were made of native protein in the original reconstituted vesicles (●) or the vesicles dissolved in aqueous solutions 1% (Δ) or 10% (\diamond) in octyl glucoside. Compare the patterns of exposure to molecular oxygen (A) and chelated paramagnetic ion (B) to the crystallographic molecular model of the protein (Figure 14-14). Reprinted with permission from ref 457. Copyright 1990 American Association for the Advancement of Science.

consecutive site-directed mutation of a membrane-spanning α helix in lactose permease from *E. coli* has also suggested that the surface of the α helix facing the lipid is more tolerant to changes in the size of its side chains than the surface facing the interior of the protein, as long as the side chains remain hydrophobic.⁴⁶⁴ Consequently, when the frequency at which amino acids are substituted is examined as a function of their position in a set of aligned amino acid sequences of the same or a set of related integral membrane-bound proteins from different species of organisms, that frequency has also been observed to vary along segments that are membrane-spanning α helices with a period of about 3.6 aa,⁴⁶⁵ just as the exposure of a spin label to molecular oxygen varies (Figure 14–27). Those positions facing the interior of the protein are positions in which substitutions over time are less tolerated.

It is also possible to use **spin-labeled phospholipids** to determine whether or not a portion of membrane-bound protein is inserted into the bilayer of phospholipid and how deeply it sits within it. Phospholipids have been synthesized with a 3-oxa-1-oxyl-2,2,5,5-tetramethylpyrrolinyl group covalently attached to their head group and to carbon 5, 10, 12, and 14, respectively, of one of their fatty acyl chains (see 14–16). It has been well established that the fatty acyl groups bearing these nitroxyl radicals remain fully extended so each sits at a characteristic depth in the membrane, and the nitroxyl radical on the head group sits at the surface of the membrane. A 3-oxa-1-oxyl-2,2,5,5-tetramethylpyrrolinyl group will quench a nearby fluorescent chromophore attached to a protein. By comparing the degree of quenching of the fluorophore as a function of the mole fraction of the nitroxylphospholipid in the membrane for nitroxyl radicals covalently attached at different positions on the labeled phospholipid, the **depth at which the fluorophore sits in the bilayer** of phospholipid can be estimated.^{81,466} For example, it was estimated that a 3-(2',2'-dimethylnaphth-7'-yl)-3-oxopropyl group covalently attached as a fluorophore to the sulfur of a cysteine placed by site-directed mutation at position 81 in the amino acid sequence of cysteineless cholesterol oxidase from *Brevibacterium sterolicum* sits 0.8 ± 0.3 nm from the center of the bilayer when the native protein is inserted in the membrane.⁴⁶⁷

Although these spectroscopic methods have been applied to several integral membrane-bound proteins, the majority of the assignments of membrane-spanning α helices in proteins for which crystallographic molecular models are unavailable have been made from genetic and chemical observations.

The general problem of identifying within the amino acid sequence of an integral membrane-bound protein, for which a crystallographic molecular model is unavailable, those segments of greater than 20 aa in length that span the bilayer of phospholipid as α helices, rather than simply spanning the interior of the globular

protein on either side of the bilayer of phospholipid, is supposed to have both a computational solution and an experimental solution. The computational solution^{300,468–470} is based on one or the other of the scales of numerical values for the hydropathies of the amino acids. The values from one of these scales or differences between the values from two of the scales⁴⁷⁰ are averaged over the amino acid sequence of a given segment. If the mean numerical value of the average for the 21 aa in a given segment is greater than a certain magnitude, where hydrophobic amino acids have positive values and hydrophilic amino acids have negative values on the scale chosen, then there is a high probability that that segment spans the membrane as an α helix within the native structure of the folded polypeptide. For example, for the scale of Kyte and Doolittle, if the average numerical value of the hydropathy is greater than +1.6, the segment of 21 aa probably spans the membrane.^{300,470} In the case of the criterion of White and Wimley,⁴⁷⁰ it is the difference between the hydropathy displayed by an amino acid in a peptide during its partition into an isotropic organic phase (octanol) and the hydropathy displayed during partition onto the surface as opposed to the interior of a bilayer of phospholipid.

These **computational approaches to designating the membrane-spanning α helices** are based on the assumptions that the change in free energy for the insertion of an α helix into a 3.6 nm layer of hydrocarbon from an aqueous phase is directly related to the partition of model solutes for its amino acid side chains and peptide bonds between water and an isotropic phase of hydrocarbon,^{468,470} that scales of hydropathy regardless of their origin ultimately reflect the free energy of this partition; that the hydrocarbon of the bilayer of phospholipid is more nonpolar than the interior of any protein; and that a longer stretch of polypeptide, 24 amino acids, is required to span a bilayer than is required to span the interior of a molecule of protein.³⁰⁰

The 11 amino acid sequences known to span the membrane in photosynthetic reaction center from *R. viridis* (Figure 14–17) were designated as the only membrane-spanning α helices in its native structure by one of these computational algorithms^{392–394} before the crystallographic molecular model became available, and every hydrophobic segment of the amino acid sequence of this protein ultimately observed to span the membrane had been so designated. This result might be taken as an indication of the **reliability** of these predictions. In another instance, however, an entire set of assignments based on mean hydropathy seems to have failed. At least one of the computational algorithms designated five of the segments of the amino acid sequence of unspecific monooxygenase from mammalian liver as membrane-spanning α helices in addition to its amino-terminal anchor, which does span the membrane.⁴⁷¹ When, however, the amino-terminal anchor is removed from the protein and several other of its hydrophobic amino acids

are mutated to hydrophilic amino acids, the membrane-bound protein becomes a water-soluble protein that has been crystallized.^{294,472} From an examination of the resulting crystallographic molecular model, it is clear that, other than its amino-terminal anchor, mammalian unspecific monooxygenase contains no membrane-spanning α helices. In fact, the bacterial form of this monooxygenase^{294,473,474} is a normal, water-soluble protein the crystallographic molecular model of which is superposable on that of the mammalian protein.

In addition to such **overprediction**, these algorithms also miss some membrane-spanning α helices, particularly in large integral membrane-bound proteins in which several of the α helices that span the membrane pass entirely through the center of the protein without contacting the hydrocarbon of the bilayer of phospholipid. For example, of the 56 membrane-spanning α helices in the α_2 dimer of the crystallographic molecular model of bovine cytochrome-*c* oxidase (Table 14-6),³³³ 18 are not hydrophobic enough to have been designated as membrane-spanning, and 12 of those were passed over in an assignment of membrane-spanning α helices made before the crystallographic molecular model became available.³⁰⁰ Most of those that seem to be too hydrophilic pass through the center of the protein and have few or no contacts with the hydrocarbon of the bilayer of phospholipid.

Both the experience with unspecific monooxygenase, where hydrophobic segments that do not seem to span the bilayer of phospholipid were designated as hydrophobic enough to do so, and the experience with cytochrome-*c* oxidase, where hydrophobic segments that span the bilayer of phospholipid were not hydrophobic enough to be designated as doing so, suggest that there is no reliable method for making this designation by inspection alone. Some integral membrane-bound proteins may contain a set of membrane-spanning segments the hydrophobicities of which are so remarkable that they can be designated as traversing the bilayer of phospholipid without too much doubt. Other integral membrane-bound proteins, however, also contain another set of membrane-spanning segments that are hydrophobic but not so hydrophobic as to be distinguishable from those segments elsewhere within the amino sequence that merely span the globular portions of the protein on either side of the membrane. In many instances, it may be these less hydrophobic membrane-spanning segments, impossible to identify by inspection, that are most intimately involved with the function of the integral membrane-bound protein, particularly if it catalyzes the transport of a hydrophilic solute across the bilayer of phospholipid. This makes their identification even more desirable.

The success with which these various algorithms predict membrane-spanning sequences has been assessed by comparing their assignments to the actual membrane-spanning sequences in the available crystal-

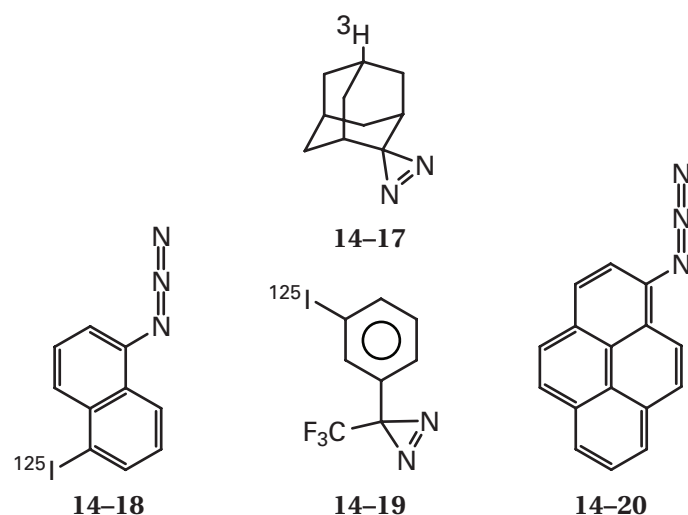
lographic molecular models of integral membrane-bound proteins.⁴⁷⁵ The **accuracy** with which they predict these direct observations varies from 91% to 99%. The most widely used⁴⁷⁶⁻⁴⁸⁴ algorithm for predicting the membrane-spanning α helices of a membrane-bound protein of unknown structure, that of Kyte and Doolittle, has an accuracy of only 93% and overpredicts membrane-spanning α helices by 13%. This overprediction arises from the fact that many α helices that simply span a globular portion of the protein outside the membrane are hydrophobic enough to be mistakenly assigned as spanning the membrane. The most accurate algorithm for predicting membrane-spanning α helices (99%) and for avoiding overprediction (<1%) was that of Wimley and White.⁴⁸⁵

The success with which these algorithms can designate membrane-spanning α helices was assessed with a set of proteins already known to be anchored membrane-bound proteins or integral membrane-bound proteins. It is probably the case that the rate of success is much lower when one of them is asked to determine whether or not a protein is membrane-bound in the absence of any information other than its sequence, a purpose for which these algorithms are often used. This ambiguity arises because many proteins that are not bound to membranes have long sequences buried in their interior that are hydrophobic enough to appear to be membrane-spanning α helices. Such decisions about whether or not a protein is membrane-bound should be viewed with caution.

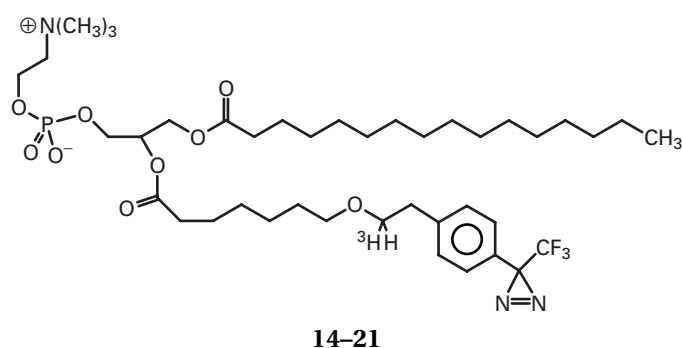
In spite of the success of the algorithm of Wimley and White, it is usually assumed that the segments identified computationally as membrane-spanning α helices should be considered only as candidates for spanning the membrane and that whether or not they do span the membrane should be validated experimentally.

The most direct experimental solution to the problem of identifying a membrane-spanning segment relies upon **covalent modification of the protein from within the liquid hydrocarbon** of the bilayer of phospholipid. Because only poorly nucleophilic amino acids are found within membrane-spanning segments, **nitrenes or carbenes** have been universally used as reagents for their selective modification. The precursor of a nitrene or carbene is incorporated into a hydrophobic molecule that partitions almost exclusively into the hydrocarbon of the bilayer of amphipathic lipids surrounding the membrane-spanning segments of the protein. The nitrene or carbene is generated from the precursor by photolysis, and it inserts into the membrane-spanning segments of the polypeptide, albeit in low yield. The intact polypeptides that are susceptible to the modification can be identified by electrophoresis in solutions of dodecyl sulfate,⁴⁸⁶⁻⁴⁸⁹ the regions of the polypeptides that have been modified can be identified by isolating and identifying peptides containing them,⁴⁹⁰⁻⁴⁹² and the particular amino acids modified can be identified by submitting the peptides to sequencing.⁴⁹³⁻⁴⁹⁶

The reagents that have been used are precursors of carbenes or nitrenes attached to two different types of **hydrophobic carriers**. 1-Tritiospiro[adamantane-4,3'-diazirine] (**14-17**),⁴⁸⁶ 5-[¹²⁵I]iodonaphthyl azide (**14-18**),⁴⁸⁸ 3-(trifluoromethyl)-3-(*m*-[¹²⁵I]iodophenyl)diazirine (**14-19**),⁴⁹⁷ and 1-azidopyrene (**14-20**)⁴⁹⁶ are examples of hydrophobic solutes that can diffuse freely through the liquid hydrocarbon of the bilayer of phospholipid:



Precursors of carbenes^{493,494,498,499} and nitrenes^{493,500} have also been incorporated covalently into phospholipids that can then be incorporated into the bilayers of phospholipids surrounding membrane-bound proteins. An example would be diazirinylphospholipid **14-21**:



In these derivatives of phospholipids, the precursor of the carbene or the nitrene can be incorporated into the fatty acyl chains, as in diazirinylphospholipid **14-21**, or it can be incorporated into the hydrophilic functional group esterified to the phosphate of the phospholipid.⁴⁹⁹ In the former case, amino acids within the hydrocarbon are the targets of the modification; and, in the latter case, amino acids at the two ends of the membrane-spanning segments.⁴⁹⁴

Originally it was thought that, by varying the position along the hydrocarbon of the fatty acid at which the carbene was located within a phospholipid, amino acids within the membrane-spanning segment located at dif-

ferent depths within the bilayer of phospholipid could be distinguished. Unfortunately, carbenes show a significant preference for insertion into nitrogen-hydrogen, oxygen-hydrogen, and sulfur-hydrogen bonds over carbon-hydrogen bonds, and this preference usually directs all of the carbenes in the liquid hydrocarbon, regardless of their mean depth in the bilayer of phospholipid, to the same one or two most susceptible amino acids in each membrane-spanning segment.^{493,494,496} Therefore, there is no obvious advantage to the derivatives of the phospholipids over the simpler hydrophobic precursors other than their elegance.

Often the incorporation observed with such hydrophobic reagents is consistent with the identification of membrane-spanning segments based on their mean hydropathy. Both glycophorin,⁴⁹⁴ which spans the membrane once, presumably with its only hydrophobic segment, and subunit IV of cytochrome-*c* oxidase,⁴⁹⁰ which also contains only one hydrophobic segment greater than 20 aa in length, which was later observed to span the membrane in the crystallographic molecular model,³³³ have been modified by either a carbene or a nitrene, respectively, incorporated into a phospholipid. The large majority of the incorporation on each case was located in a peptide 68 or 49 amino acids in length, respectively, that contained the hydrophobic segment of greater than 20 aa picked out by the computational algorithms. When (iodophenyl)diazirine (**14-19**) was used to modify bacteriorhodopsin, incorporation also was found to occur⁵⁰⁰ in a region of the amino acid sequence containing a segment that had been identified computationally as spanning the membrane and that was later shown to be in the center of a membrane-spanning α helix in the crystallographic molecular model.³²⁸

All five of the subunits of acetylcholine receptor (Figure 14-22) are homologous in amino acid sequence, and each contains four segments of amino acid sequence that were judged to be hydrophobic enough to span the membrane. Both 1-azidopyrene and 3-(trifluoromethyl)-3-(*m*-[¹²⁵I]iodophenyl)diazirine label all of the subunits in the native protein in its native membrane. A set of labeled peptides, identified by their sequences, has been isolated from digests of the labeled protein. Among the members of this set are contained the first, the third, and the fourth of the hydrophobic segments from one or the other of the subunits.⁴⁹⁶ In one of these peptides, two cysteines were identified as the sites of modification, further evidence for the electrophilicity of nitrenes and carbenes. In the molecular model derived from electron diffraction and image reconstruction,³⁴⁶ all 20 of the hydrophobic segments, four from each subunit of the protein, are membrane-spanning α helices.

When adamantyldiazirine (**14-17**) was used to modify canine Na⁺/K⁺-exchanging ATPase, however, and the long tryptic peptides of the intact polypeptide that were modified by the reagent were isolated and identified,^{491,501} it was found that substantial amounts of

adamantylidene had been incorporated into three tryptic peptides that did not contain hydrophobic segments designated computationally as membrane-spanning. These three peptides contained segments greater than 20 aa in length that were hydrophobic but not sufficiently hydrophobic to be picked out by their mean hydropathy. Their sequences are -VNFPVENL CFVGFISMIGPP-, -QIGMIQALGGFFTYFVILAE-, and -PTWWFCAFPYSLIFVYDEV-. The location of these three segments in the structure of Na⁺/K⁺-exchanging ATPase can be inferred from the location of the homologous sequences in the crystallographic molecular model of Ca²⁺-transporting ATPase from *O. cuniculus* (Figure 14–15).³⁵¹ The first is located in the globular cytoplasmic domain of the protein distant from the bilayer of phospholipid, but the latter two are in membrane-spanning α helices, in spite of their low mean hydropathy. Both of these membrane-spanning α helices are on the outside surface of the bundle of α helices within the bilayer of phospholipid (Figure 14–15). Three of the segments of the amino acid sequence of Na⁺/K⁺-exchanging ATPase designated as membrane-spanning by their high mean hydropathy were found within other tryptic peptides modified by adamantyldiazirine, and the homologues of these three segments do also span the membrane in the crystallographic molecular model of Ca²⁺-transporting ATPase.

The **topography** of a membrane-spanning protein is a complete designation of those segments of its polypeptide that span the membrane and of the sides of the membrane, cytoplasmic or extracytoplasmic, on which those segments of its polypeptide that are not within the hydrocarbon of the bilayer of phospholipid are located. The topography of a membrane-spanning protein can be defined by identifying in turn the location of one or more of the amino acids within each of the segments of its polypeptide outside the membrane. While assembling these individual topographic assignments, if one end of a segment of hydrophobic amino acids in the sequence of an integral membrane-bound protein can be shown to be located on its cytoplasmic surface and the other end on its extracytoplasmic surface, then it can be concluded that that segment spans the bilayer of phospholipid.

The identification of the side of a membrane, cytoplasmic or extracytoplasmic, upon which a particular amino acid or peptide from the polypeptide of an integral membrane-bound protein is located can be made with oriented, sealed structures and an impermeant reagent. The most reliable **oriented, sealed structures** are **intact cells**, such as erythrocytes,¹⁹¹ or intact organelles, such as undamaged mitochondria or lysosomes.⁵⁰² Erythrocytes are ideal for this purpose because sealed inside-out vesicles that present only the cytoplasmic surfaces of their membrane-bound proteins to the solution can be prepared from intact erythrocytes.¹¹⁹ Intact animal cells grown in tissue culture^{503,504} or spher-

oplasts of bacteria* have also been used in such experiments. **Mitochondria** as they are usually prepared contain both an outer membrane, which is the porous cellular membrane isolating these organelles from direct contact with the cytoplasm, and an inner membrane, which is the tight, impermeable boundary of the functional mitochondrion. The outer membrane can be removed⁵⁰⁵ to produce sealed, unwrapped mitochondria that present the cytoplasmic, extramitochondrial surfaces of their membrane-bound proteins to the external solution.⁵⁰⁶ Sealed, inside-out vesicles, which present the extracytoplasmic, intramitochondrial surfaces of their membrane-bound proteins to the solution, can be prepared from unwrapped mitochondria.⁵⁰⁷

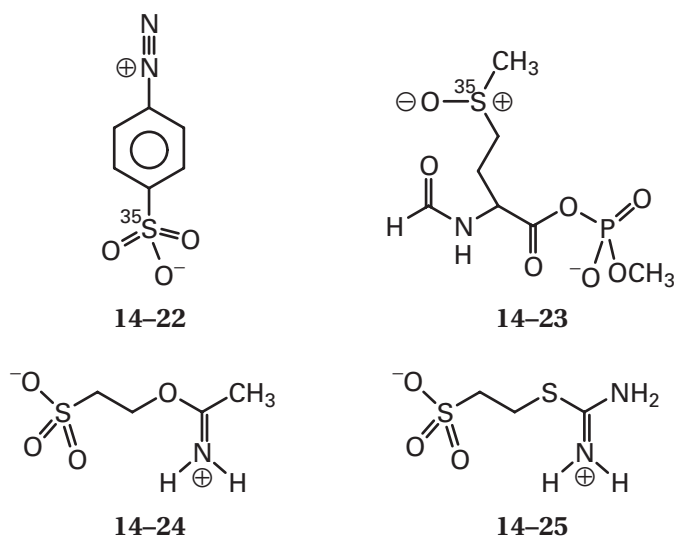
Sealed vesicles often form spontaneously from fragments of the constituent membranes during homogenization of a tissue. As these structures are adventitious, they are not necessarily sealed to all hydrophilic solutes. For example, vesicles of plasma membrane can be isolated from the electric organ of *T. californica* that are sealed to large solutes such as proteins,⁵⁰⁸ but only a minority of them are sealed to small solutes such as the cations of alkali metals.⁵⁰⁹ Occasionally, however, a suspension of homogeneously and tightly sealed vesicles, in which all of the proteins are oriented as they were when the membrane containing them was in the cell, can be prepared from a homogenate.^{510,511}

It is also possible to use purified membrane-bound proteins reconstituted into sealed vesicles of phospholipid. It is usually the case that during a reconstitution the membrane-bound protein inserts at random in either of the two possible orientations, cytoplasmic surface directed outward or extracytoplasmic surface directed outward. If only one of these two surfaces is susceptible to endopeptidolytic digestion when the protein is in its native structure, as is often the case, digestion of the reconstituted vesicles will nick only those molecules of protein exposing that surface, and intact polypeptides, derived exclusively from molecules of protein inserted in the opposite orientation, can be purified by electrophoresis or molecular exclusion chromatography performed in solutions of dodecyl sulfate.^{512,513}

Many impermeant reagents have been used to modify integral membrane-bound proteins in such sealed, impermeable structures. Because a bilayer of phospholipid contains a continuous sheet of hydrocarbon 3.6 nm wide, charged solutes or solutes with large numbers of donors and acceptors for hydrogen bonds cannot pass through it. An **impermeant reagent for covalent modification** is such a hydrophilic solute that also contains an electrophilic functional group appro-

* A spheroplast is a bacterial cell that has been stripped of its outer membrane.

priate for the modification of proteins. Diazotized *p*-[³⁵S]sulfanilic acid (**14-22**),⁵¹⁴ *N*-formyl-[³⁵S]sulfinylmethionyl methylphosphate (**14-23**),⁵¹⁵ isethionyl [¹⁴C]acetimidate (**14-24**),⁵¹⁶ 2-S-[¹⁴C]thiuroniummethanesulfonate (**14-25**),⁵¹⁷



and pyridoxal phosphate⁵¹⁸ and sodium borohydride (Figure 10-3) are impermeant reagents that have been used to modify only the surface of a protein presented to the external solution in a suspension of sealed membranes.

For example, both intact bovine mitochondria and sealed inside-out vesicles of bovine mitochondria were separately modified with pyridoxal phosphate and sodium borohydride, and labeled ADP, ATP carrier, an integral membrane-bound protein, was purified from each sample. Thermolytic peptides containing the labeled lysines were isolated from the protein and identified by their amino-terminal sequences. It was found that Lysine 146 of the protein was modified by pyridoxal phosphate and sodium borohydride in the protein from the sealed inside-out vesicles, while Lysines 95, 198, 205, 259, and 267 were modified by pyridoxal phosphate and sodium borohydride in the protein from intact mitochondria.⁵¹⁹ Lysine 146 was assigned to the extracytoplasmic, intramitochondrial surface of the protein; and the other lysines, to its cytoplasmic, intramitochondrial surface.

The tryptic peptide HLLVMKGAPER, the amino acid sequence containing Lysine 501 from ovine Na⁺/K⁺-exchanging ATPase, could be isolated from digests of the intact protein by immunoadsorption with immunoglobulins G directed against its carboxy terminus. Lysine 501 would not incorporate pyridoxal phosphate when the protein was in sealed vesicles that presented only the extracytoplasmic surface of the protein to the solution but readily incorporated pyridoxal phosphate when the vesicles were opened by adding the surfactant saponin.⁵²⁰ Saponin, by combining with the cholesterol

in the bilayer of phospholipid, is able to form large (8.0 nm) holes in natural membranes⁵²¹ without significantly altering the membrane-bound proteins. The results of these experiments demonstrated that Lysine 501 is located on the cytoplasmic surface of Na⁺/K⁺-exchanging ATPase, a topographical assignment later verified crystallographically.³⁵¹

One of the drawbacks of labeling integral membrane-bound proteins with such impermeant electrophiles is that the shorter segments between two candidates for spanning the membrane often lack a suitable **nucleophilic side chain**. For example, in the MotA protein from *E. coli*, the short segment connecting the first and second candidates for spanning the membrane contains only a tyrosine, and the short segment connecting the third and the fourth contains only a glutamate, two nucleophiles that can be difficult to label. Consequently, **cysteines** were placed in turn within these segments at position 24, 190, and 196 in the amino acid sequence of the protein, respectively, in three separate **site-directed mutations** of the cysteineless version of the protein. The modification of each of these cysteines by the impermeant fluorescent reagent fluorescein 5-maleimide proceeded at similar rates whether the mutant protein was in intact spheroplasts of the bacteria or osmotically disrupted spheroplasts. Six other cysteines placed in the much larger segments connecting the second and the third candidate and following the fourth candidate for spanning the membrane reacted slowly with the fluorescein 5-maleimide in the intact spheroplasts and rapidly in the disrupted spheroplasts. These results demonstrated that positions 24, 190, and 196 are located in extracytoplasmic segments and the rest of the connecting segments are cytoplasmic, and that all four of the candidates for spanning the membrane do so.⁵²²

Similar experiments have been performed on segments of the amino acid sequence of several other integral membrane-bound proteins but in these instances in order to assess variations from position to position in accessibility to the aqueous phase rather than topography. Each amino acid in a particular segment was mutated in turn to a cysteine, and the susceptibility of that cysteine to modification by a polar electrophile was assayed.⁵²³⁻⁵²⁵ Variations in accessibility provided information about the structure of the polypeptide within that segment in the native state of the protein.

Enzymes can also be used as impermeant reagents. For example, the impermeant enzyme protein-glutamine γ -glutamyltransferase (Equation 13-45) has been used to catalyze the modification of exposed glutamines on the surfaces of membrane-bound proteins with fluorescent primary amines.⁵²⁶ The enzyme lactoperoxidase (Equation 10-33) has also been used as an impermeant reagent.⁵²⁷ Although this enzyme probably produces a small, diffusible, activated form of iodine, perhaps IOH, that species is so reactive that it never makes it across a bilayer of biological phospholipids.⁵²⁸

Endopeptidases or immunoglobulins G are also impermeant reagents. Each of the **endopeptidases** pronase,⁵²⁹ chymotrypsin,³²⁰ and papain⁵³⁰ is able to cleave native, human band 3 anion transport protein ($n_{aa} = 911$),⁵³¹ an integral membrane-spanning protein in human erythrocytes, within the short segment of the polypeptide, -QDHPLQKTYNYNVLMPKPWQGGLP-, between Glutamine 545 and Proline 568. Papain cleaves after Glutamine 550, and chymotrypsin cleaves after Tyrosine 553.⁵³¹ These cleavages occur quantitatively when any one of the endopeptidases is added to the extracytoplasmic solution in which intact erythrocytes are suspended. These results demonstrate that this short segment of amino acids is fully exposed on the extracytoplasmic surface of the intact protein. Two different **monoclonal immunoglobulins G** raised against purified native acetylcholine receptor from the electric organ of *T. californica* recognize as an antigen the synthetic peptide KAEEYILKKPRSELMFEEQ, which is an amino acid sequence from the interior of one of the five homologous polypeptides composing the protein. Presumably, the natural epitopes on the intact protein are composed of sequences from this region. It could be shown that these monoclonal immunoglobulins G were bound only at the cytoplasmic surfaces of membranes containing this protein.⁵³² From this result, it was concluded that this sequence in the native structure of acetylcholine receptor is exposed on the cytoplasmic surface of the protein, a topographical assignment later validated by the crystallographic molecular model derived from electron diffraction and image reconstruction.³⁴⁶

Because oligosaccharides are added to all glycoproteins in the extracytoplasmic lumina of the Golgi membranes, any asparagine, serine, or threonine in an integral membrane-bound protein that is glycosylated must be located on its extracytoplasmic surface. The asymmetry of the biosynthesis of these oligosaccharides can also be used to make topographical assignments for segments in the amino acid sequence of a protein that are not normally glycosylated. If a sequence encoding glycosylation, for example, -NST-,⁵³³ is introduced by site-directed mutation into a segment of amino acid sequence between two candidates for spanning the membrane found in an integral protein located in the plasma membrane, the **glycosylation** of the new asparagine in that sequence demonstrates that the segment is located on the extracytoplasmic surface of the protein.^{534,535} Unfortunately, the steric requirements for access of the asparagine during its glycosylation are fairly stringent,⁵³³ so no conclusion can be made from a negative result. In addition, this approach has been observed to give misleading assignments.^{536,537}

It is also possible to **insert by genetic manipulation entire enzymes** into a segment in the polypeptide of an integral membrane-bound protein located between two hydrophobic segments and use the activity of that enzyme to identify the location of the modified segment. For exam-

ple, alkaline phosphatase has been inserted consecutively into the hydrophilic segments between the 12 candidates for spanning the membrane in lactose permease from *E. coli*⁵³⁸ and the six candidates for spanning the membrane in MalG protein from *E. coli*,⁵³⁹ and the intact cells bearing each of these constructs were assayed for alkaline phosphatase activity with the impermeant reactant *p*-nitrophenyl phosphate. In both cases, alkaline phosphatase inserted into every other hydrophilic segment was located extracytoplasmically, results demonstrating that every candidate in each protein actually does span the membrane. Both β -lactamase^{540,541} and chloramphenicol *O*-acetyltransferase⁵⁴² have also been used in this way. The principal difficulty with these approaches is that the insertion of an entire molecule of protein into a short segment between two membrane-spanning segments could disrupt the normal topography of the protein. One way to minimize this problem is to insert the enzyme at random and select for modified proteins that retain their full biological activity in addition to having the extraneous enzymes inserted into the desired segments.⁵⁴³

Another drawback of all of the approaches for assessing topography that involve large numbers of site-directed mutations, such as the insertion of cysteines at consecutive positions in a cysteineless version of the protein or the fusion of an integral membrane-bound protein with another protein, is that there must be an **efficient expression system** for that integral membrane-bound protein. This requirement usually confines their use to bacteria because there are few convenient, efficient expression systems for most integral membrane-bound proteins from animals that produce high yields of the protein. For example, methods that rely on the production of fusion proteins have been applied almost exclusively to bacteria.

This lack of efficient expression systems is a common problem with evaluating the results of any site-directed mutation of membrane-bound proteins from animals. Occasionally one of these proteins can be expressed in a functional form in *E. coli*,⁵⁴⁴ but usually they must be expressed in cultured cell lines derived from animal tissue. Although the product of a site-directed mutation can often be detected immunochemically,⁵⁴⁵ its expression in animal cells usually precludes the production of sufficient quantities of a mutant for purification and direct study. Rather, the structural changes resulting from a mutation are inferred from changes in the function of the protein in intact cells^{546,547} or crude preparations of membranes from the cells. An extreme example is the expression in oocytes of *X. laevis* of ionic channels from nervous tissue or transport proteins for metabolites such as glucose, which can be detected only by the robust fluxes of the substrates that they produce across the membrane.^{523,548-550} It is also possible to detect expressed channels for water in these oocytes again because of the high water permeability they create.⁵⁵¹

The extensive **topographical experiments** that have been performed on human band 3 anion transport protein from erythrocytes serve as an example of the cumulative application of these strategies to one integral membrane-bound protein. Band 3 anion transport protein is an integral membrane-bound protein in the plasma membrane of the erythrocyte that is responsible for the transport of anions such as chloride, bicarbonate, or phosphate⁵⁵² across the membrane, and it is the integral membrane-bound protein present in the highest concentration in this plasma membrane. The human protein is composed of a single polypeptide 911 amino acids long that bears covalently attached carbohydrate⁵⁵² and spans the bilayer of phospholipid in its native structure.¹⁹¹ Human band 3 anion transport protein has a detachable domain on the cytoplasmic side of the bilayer of phospholipid that can be released by trypsin from fragments of membrane⁵⁵³ and that constitutes the first 360 amino acids from the amino terminus of the folded polypeptide.⁵³¹ After cleavage from the membrane, the detached domain is freely water-soluble.

The domain 550 amino acids long, left in the membrane after the amino-terminal domain has been detached by endopeptidolytic digestion and washed away, is still able to transport anions as rapidly as does the intact protein.⁵⁵⁴ Its amino acid sequence contains at least 10 individual hydrophobic segments, 20 aa or more in length, that are candidates for spanning the membrane (Figure 14-28).^{300,555} The formal amino terminus of this embedded domain in the human protein,⁵⁵⁶ Glycine 361, must be on the cytoplasmic surface of the protein because cleavage by trypsin at the α amide of Glycine 361 to release the detachable domain occurs at that surface.

Lysine 430 in human band 3 anion transport protein can be modified from the extracytoplasmic surface by formylation followed by reduction with sodium borohydride, which, under the appropriate conditions, is impermeant to intact erythrocytes,⁵³⁰ so the hydrophobic segment between Glutamine 404 and Glycine 428 (segment *a* in Figure 14-28) spans the membrane. Tyrosine 486 is modified extensively by lactoperoxidase and [¹²⁵I]I⁻ when the protein is in inside-out vesicles made from erythrocytes but only weakly when it is in intact cells, a result that places Tyrosine 486 on the cytoplasmic surface of the membrane. Consequently, the hydrophobic segment between Methionine 435 and Phenylalanine 478 (segment *b* in Figure 14-28) spans the membrane.⁵²⁸ In the region between Lysine 542 and Proline 568 (between segments *d* and *e* in Figure 14-28) there is a loop of polypeptide in the native structure of the protein that is susceptible to cleavage by pronase,⁵²⁹ chymotrypsin,³²⁰ pepsin A,⁵²⁷ and papain⁵³⁰ but only from the extracytoplasmic surface of the membrane. Consequently, the polypeptide spans the membrane once between Tyrosine 486 and Lysine 542. There are two hydrophobic segments greater than 20 aa in length in this region (segments *c* and *d* in Figure 14-28).

Tyrosine 596 is not iodinated by lactoperoxidase and [¹²⁵I]I⁻ when the lactoperoxidase is present only at the extracytoplasmic side of the membrane, and Asparagine 593 does not bear an *N*-linked oligosaccharide, even though it is in the proper sequence to be so modified. These negative results suggest that this part of the amino acid sequence is on the cytoplasmic surface of the membrane and that the hydrophobic segment between Asparagine 569 and Arginine 589 (segment *e* in Figure 14-28) spans the membrane.⁵²⁷

Tyrosine 628 is iodinated in the presence of extracytoplasmic lactoperoxidase and [¹²⁵I]I⁻,⁵²⁷ the peptide bond between Threonine 629 and Glutamine 630 is cleaved by extracytoplasmic papain⁵³⁰ in intact erythrocytes, and Asparagine 642 is glycosylated,⁵⁵⁷ so the hydrophobic segment between Lysine 600 and Aspartate 621 (segment *f* in Figure 14-28) spans the membrane. Pyridoxal phosphate and Na[³H]BH₄ can modify Lysine 691 when band 3 anion transport protein is in inside-out

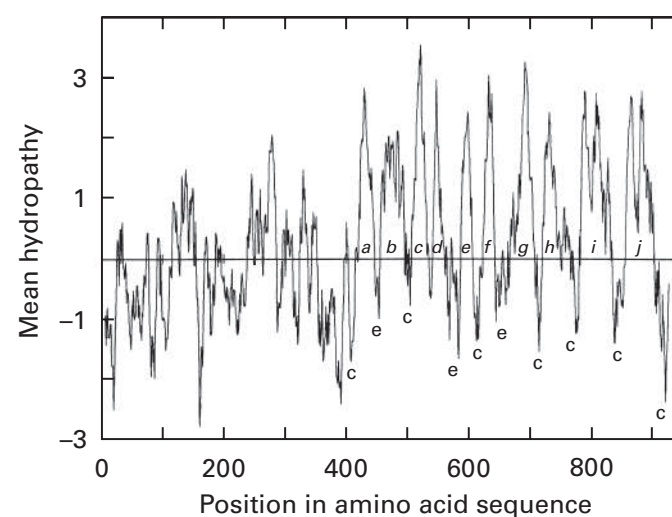


Figure 14-28: Plot of the distribution of hydropathy over the amino acid sequence of murine band 3 anion carrier ($n_{aa} = 929$).⁵⁵⁵ Each amino acid in the sequence of amino acids is assigned its numerical value of hydropathy in the scale of Kyte and Doolittle.³⁰⁰ A moving average with a span of 11 positions is calculated from this sequence of numbers. The numerical value of mean hydropathy assigned to each position in the amino acid sequence is the average of the segment of the 11 positions of which it is the central number. Positive values indicate hydrophobic locations; negative values, hydrophilic locations. Ten long hydrophobic segments are found within the last 530 amino acids of the sequence (*a-j*). Segments *b*, *i*, and *j* are long enough (>40 amino acids) to contain two membrane-spanning α helices. The amino-terminal cytoplasmic domain in the murine protein is 19 amino acids longer than the one in the human protein, so each potential membrane-spanning segment is 19 amino acids farther along in the sequence of the murine protein. The human protein and the murine protein, however, can be aligned with 92% identity and a gap of only one amino acid over the last 530 amino acids. The hydrophilic segments in the amino acid sequence of the murine protein that have been assigned to the cytoplasmic (c) surface or extracytoplasmic (e) surface of the human protein by chemical experiments are designated. Adapted with permission from *Nature*, ref 555. Copyright 1985 Macmillan Magazines Ltd.

vesicles but not when it is in intact erythrocytes, so the hydrophobic segment between Proline 660 and Glutamate 681 (segment *g* in Figure 14–28) spans the membrane.⁵⁵⁸ Trypsin can cleave the polypeptide in the native protein at Lysine 743, but only from the cytoplasmic surface,^{536,559} so the hydrophobic segment between Lysine 698 and Proline 722 (segment *h* in Figure 14–28) does not span the membrane. Aspartate 821 is modified with 1-ethyl-3-[3-(trimethylammonio)propyl]carbodiimide and [³⁵S]sulfanilic acid (Figure 10–5) when the protein is in inside-out vesicles but not when it is in intact erythrocytes,⁵⁶⁰ so the long hydrophobic segment between Arginine 760 and Arginine 808 (segment *i* in Figure 14–28) spans the membrane either twice or not at all. The carboxy terminus of anion carrier when it is in inside-out vesicles but not when it is in right-side-out vesicles can bind an immunoglobulin raised against its amino acid sequence,⁵⁶¹ so the long hydrophobic segment between Lysine 829 and Arginine 870 (segment *j* in Figure 14–28) spans the membrane either twice or not at all.

One measure of the success of such topographical assignments is to compare the conclusions reached experimentally with crystallographic molecular models of the protein that become available at a later date. For example, the topographies of two of the subunits of bovine cytochrome-*c* oxidase were examined before its crystallographic molecular model (Table 14–6) became available. Chymotrypsin cleaves the polypeptide of subunit III in a region between Tryptophan 34 and Phenylalanine 37 as well as in a region between Tryptophan 99 and Tryptophan 116 when it has access only to the cytoplasmic, extramitochondrial surface of cytochrome-*c* oxidase in reconstituted vesicles.⁵⁶² Glutamate 90, within the hydrophobic segment between Arginine 79 and Histidine 103 in subunit III, is modified by dicyclohexylcarbodiimide,⁵⁶³ which is a hydrophobic carbodiimide, and this result placed this hydrophobic segment in the bilayer of phospholipid. The peptide bond of Lysine 7 of subunit IV was susceptible to cleavage when intact bovine cytochrome-*c* oxidase was digested with trypsin from its extracytoplasmic, intramitochondrial surface in inside-out vesicles of mitochondria⁵⁶⁴ and only when sealed vesicles, in which the protein is oriented with its cytoplasmic, extramitochondrial surface outward, were opened with nonionic detergent.⁵⁶² The polypeptide of subunit IV was also susceptible to digestion by pronase in intact, unwrapped mitochondria,⁵⁶⁵ which expose only the cytoplasmic, extramitochondrial surface of cytochrome-*c* oxidase. Both of these results taken together demonstrated that subunit IV does span the membrane in cytochrome-*c* oxidase, with its amino terminus on the extracytoplasmic side. All of these conclusions were validated by the crystallographic molecular model.

Each of the four homologous subunits of acetylcholine receptor from *T. californica* contains a cysteine

connecting the cysteines homologous to Cysteine 128 and Cysteine 142 in the α subunit.³⁹¹ In each subunit, a glycosylated asparagine precedes the respective cysteine homologous to Cysteine 142 in the α subunit.³⁹¹ Lysine 165 from the β subunit is accessible to modification by pyridoxal phosphate and Na[³H]BH₄ on the extracytoplasmic surface of sealed right-side-out vesicles derived from plasma membranes from the electric organ of *T. californica*.⁵⁶⁶ All of these facts placed the hydrophilic portions containing the first 210 aa in each of the four subunits on the extracytoplasmic side of the membrane. It has already been noted that an immunoglobulin recognizing the sequence from Lysine 360 to Glutamine 378 in the γ subunit is recognized by an immunoglobulin at the cytoplasmic surface of the protein. Lysine 380 from the α subunit is accessible to modification by pyridoxal phosphate and Na[³H]BH₄ only when sealed right-side-out vesicles derived from plasma membranes of electric organ from *T. californica* are opened with saponin, but Lysine 486 from the γ subunit is accessible on the extracytoplasmic surface of the same vesicles.⁵¹¹ These results placed the former lysine on the cytoplasmic side and the latter on the extracytoplasmic side of the membrane. All of these results have been validated by the crystallographic molecular model of this protein derived from electron diffraction and image reconstruction.³⁴⁵

There are seven integral membrane-bound proteins that catalyze the active transport of inorganic cations across cellular membranes at the expense of the hydrolysis of MgATP. These are Na⁺/K⁺-exchanging ATPase (Na⁺/K⁺-ATPase) from animal plasma membranes, Ca²⁺-transporting ATPase (ER Ca²⁺-ATPase) from animal endoplasmic reticulum, calmodulin-regulated Ca²⁺-transporting ATPase from animal plasma membranes, H⁺/K⁺-exchanging ATPase from the luminal plasma membranes of gastric mucosa, K⁺-transporting ATPase from bacterial plasma membranes, H⁺-exchanging ATPase from fungal plasma membranes, and H⁺-exchanging ATPase from plant plasma membranes. Each of these seven proteins has a long polypeptide, designated the α polypeptide, which when in its native state is responsible for the catalysis of the respective active transport. All of the seven α polypeptides are homologous in sequence,^{567–571} and therefore each folds to create an α subunit that is superposable upon the native structure of all of the others.

Tryptic cleavages of the α subunit of canine Na⁺/K⁺-ATPase in the native membrane that occur at Arginine 262 and Lysine 30 can take place only when trypsin has access to the cytoplasmic surface of the protein.⁵⁷² Trypsin is able to cleave ER Ca²⁺-ATPase of *O. cuniculus* from the cytoplasmic surface of intact endoplasmic reticulum,⁵⁷³ at Arginine 198.⁵⁷⁴ Aspartate 369 is located in the active site of ovine Na⁺/K⁺-ATPase⁵⁷⁵ on its cytoplasmic surface. Lysine 766, Lysine 943, and Lysine 1012 in the α subunit of ovine Na⁺/K⁺-ATPase can be modified with pyridoxal phosphate and Na[³H]BH₄ when they are in sealed right-

side-out vesicles of plasma membrane only if those vesicles are opened with saponin,⁵⁷⁶⁻⁵⁷⁸ results that placed these amino acids on the cytoplasmic surface of the membrane. When Cu^{2+} is added to sealed right-side-out vesicles of plasma membrane, it catalyzes the oxidative cleavage of the α subunit of porcine Na^+/K^+ -ATPase in the presence of ascorbate and H_2O_2 within the segment between Tyrosine 895 and Lysine 905 and within the segment between Proline 965 and Threonine 979,⁵⁷⁹ results that placed these amino acids on the extracytoplasmic surface of the membrane. Immunoglobulins directed against the amino terminus and immunoglobulins directed against the carboxy terminus of H^+ -exporting ATPase from *Neurospora crassa* were bound only to the cytoplasmic surface of plasma membranes, and the amino terminus and carboxy terminus of the same protein were removed by trypsin only when it had access to the cytoplasmic surface of plasma membranes.⁵⁸⁰ The domain on human calmodulin-regulated Ca^{2+} -transporting ATPase to which calmodulin binds from the cytoplasmic surface of the membrane⁵⁸¹ is located on the carboxy terminus of the protein. All of these topographical observations have been validated by the crystallographic molecular model of ER Ca^{2+} -ATPase.³⁵¹

Once the membrane-spanning α helices in an integral membrane-bound protein have been identified, it is possible to determine how they are arranged within the membrane. For example, pairs of cysteines can be inserted systematically by site-directed mutation, one into each of two membrane-spanning segments in a cysteineless version of an integral membrane-bound protein, and those cysteines, if they are adjacent to each other in the native structure of the protein, can be cross-linked with a hydrophobic cross-linking reagent or turned into a cystine by oxidation. Any such **covalent cross-link between two membrane-spanning α helices** places them adjacent to each other in the bundle of α helices within the bilayer of phospholipid.⁵⁸²⁻⁵⁸⁵ Advantage can also be taken of the favorable free energy of formation of a hydrogen bond within the membrane. The replacement of a neutral amino acid within the membrane with a polar hydrogen-bond donor or acceptor will usually disrupt the structure of the protein, causing it to lose its function, but if a polar hydrogen-bond acceptor or donor, respectively, is then placed in an adjacent membrane-spanning α helix near enough to the polar donor or acceptor to form a hydrogen bond, the structure of the protein, and hence its function, can be rescued.⁵⁸⁶ The return of function demonstrates that the two α helices are adjacent to each other. Unfortunately, because of the need for an efficient system for expression and the need to score large numbers of mutants, these approaches have so far been confined to the analysis of integral membrane-bound proteins from bacteria.

As can be done with any other set of proteins, the amino acid sequences of integral membrane-bound proteins can be aligned, and statistically significant relation-

ships can be used to identify isoforms of the same protein in the same genome⁵⁸⁷⁻⁵⁹² or related proteins from different species of organisms. Often, similarities in the patterns of distribution of the hydrophobic segments that are candidates for spanning the membrane strengthen the conclusion that all of the aligned proteins share a common ancestor.⁵⁹³ One difficulty that arises in such **alignments of amino acid sequences**, however, is that, because the choice made by natural selection of the side chains to span the membrane is heavily biased in favor of the small set of the most hydrophobic, the amino acid sequences of unrelated membrane-spanning segments often seem to be more closely related to each other than unrelated amino acid sequences in general do.⁵⁹⁴ Nevertheless, in alignments of the amino acid sequences of integral membrane-bound proteins that are distantly related, the percentage of identity is about the same within membrane-spanning sequences as it is in sequences outside of the membrane.⁵⁹⁵

Almost all integral membrane-bound proteins remain in the membrane for their entire lives, but there are a set of proteins that begin their lives as water-soluble proteins, and if they are called upon, end their lives as integral membrane-bound proteins. The function of most of these proteins is to punch a hole in a membrane, either to short-circuit the normal gradients of metabolites and ions between the cytoplasm and the environment and thereby kill the cell or to permit another protein with which they are associated to thread its way through the hole and enter the foreign cell, usually as an act of subterfuge.

An example of a **protein that punches a hole in a membrane** is α -hemolysin from *S. aureus* (Table 14-6). The protein begins its life as a water-soluble monomer of 293 aa. To insert into the membrane of the cell it is to kill, each monomer puts forth a hairpin of β structure about 30 aa in length, and seven of these hairpins⁵⁹⁶ assemble side by side to form a cylindrical β barrel that after insertion spans the membrane of the doomed cell and forms the hole³⁶¹ through which the ions and metabolites pour. The large portion of each monomer on the exterior of the membrane associates with its six neighbors to form a thick torus of cyclic symmetry of point group 7 (C_7) on the outside of the cell that resembles the extracellular portion of acetylcholine receptor (Figure 14-22).

Colicin E1, which is secreted by *E. coli*, binds to proteins on the extracytoplasmic surface of cells other than *E. coli* that are to be killed. The portion of the protein responsible for forming the hole is a bundle of 10 α helices in the water-soluble form of the protein.⁵⁹⁷ Two of these α helices are completely buried in the center of the bundle. They are 16 and 17 amino acids in length and are composed entirely of hydrophobic amino acids. In addition, the amino acid sequences flanking these helices are also composed only of hydrophobic amino acids. When this portion of the protein has adsorbed to the surface of the bilayer of phospholipid of the mem-

brane of the cell that is to be killed, the hairpin of the central two hydrophobic α helices inserts into the hydrocarbon⁵⁹⁸ and associates with identical hairpins of α helices inserted into the membrane by other molecules of colicin E1 to form the mortal hole.

Cholera toxin from *V. cholerae*,⁵⁹⁹ heat-labile enterotoxin from *E. coli*,⁶⁰⁰ and verotoxin-1 from *E. coli*^{601,602} are homologous heterooligomers that create a **pore through the membrane through which a polypeptide is threaded**. Each contain a substructure that is a pentamer of identical subunits with cyclic symmetry of point group 5 (C_5). These pentameric rings are responsible for forming a pore through the membrane of the cell that is the target of the toxin. Another, unrelated subunit of the complex that forms the actual toxin is then threaded through the pore to intoxicate the cell. In the center of each of these pentameric rings is a symmetric ring of five identical α helices, one from each subunit; the five α helices are parallel to each other and to the 5-fold rotational axis of symmetry that superposes them. The length of these α helices varies from 11 to 20 aa, depending on the protein, but each is amphipathic.^{599,603} The ring of five subunits recognizes the oligosaccharide on a particular glycolipid in the membrane of the cell that it will eventually intoxicate and binds directly to the sugars.^{602,604} The pentamer then adsorbs to the surface of the bilayer of phospholipid⁶⁰⁵ and, presumably, the ring of α helices inserts into the membrane and the other subunit is threaded through the hydrophilic pore at its center. In the crystallographic molecular model of the intact, water-soluble complex of the components of the toxin, the carboxy terminus of this other subunit is already threaded into the pore, ready to enter the cell.⁶⁰⁰

In diphtheria toxin, another toxin that threads one of its domains into a cell through a pore formed by another of its domains, the domain that inserts a portion of its structure into the membrane to form the pore⁶⁰⁶ has a structure reminiscent of the pore-forming domain in colicin E1. It is a bundle of nine α helices, and the central core of this bundle is two α helices that are unusually hydrophobic,⁶⁰⁷ which are thought to form the pore itself.

Lipoproteins are complexes between specific proteins and heterogeneous mixtures of phospholipid, cholesterol, **triacylglycerol**, and **fatty acyl esters of cholesterol** that store the lipids they contain in repositories such as yolk or that transport the lipids they contain through extracellular fluids such as the serum of blood. The lipoprotein in the yolk of eggs is a complex between the protein vitellogenin and the lipids. The lipoproteins in serum are chylomicrons, very low density lipoprotein, low density lipoprotein, and high density lipoprotein.

In the mature lipoprotein in yolk, each molecule of the complex contains the posttranslationally modified version of one molecule of **vitellogenin**. In the yolks of

eggs from *G. gallus*, the vitellogenin (1897 aa) is posttranslationally cleaved into four fragments: lipovitellin I (1124 aa), phosvitin (252 aa), lipovitellin II (235 aa), and yolk glycoprotein 42 (284 aa).^{608,609} Each of these fragments has its own name because they were identified before intact vitellogenin was identified when they seemed to be separate proteins. The most peculiar and independent of these fragments is phosvitin, which contains 123 serines (50% of its amino acids), of which about 25 are phosphorylated. It is the vitellogenin lipoprotein in the yolks of eggs from *G. gallus* that is the source of the natural phosphatidylcholine that is purified from them.

The crystallographic molecular model of the mature lipoprotein formed between the lipid and these fragments of vitellogenin from *Ichthyomyzon unicuspis* contains a large globular cavity about 2.5 nm in diameter in which the lipid is located.⁶¹⁰ This **ball of lipid** is surrounded by three **slabs of β sheet** of 15, 9, and 6 strands in width, respectively, and one α helix.^{611,612} There are significant openings in this central cavity to the surrounding aqueous solution. The cavity contains about 30–35 molecules of lipid, about 20–25 of which are phospholipid, so the surfaces of the ball of lipid exposed to the solution are presumably paved by the head groups of these molecules of phospholipid, as are the surfaces of a bilayer of phospholipid. In the maps of electron density, several complete molecules of phospholipid could be observed as well as many fragments of linear alkane tucked into crevices formed by the pleats of the β sheets,⁶¹³ just as the linear alkane or the phospholipids in the crystallographic molecular models of integral membrane-bound proteins are tucked into crevices between the α helices. These complete molecules of phospholipid and fragments of linear alkane are not distributed as they would be in a spherical micelle, so the lipid must be irregularly packed into the cavity. The remaining approximately 10 molecules of lipid other than the phospholipid are triacylglycerols, cholesterol, and fatty acyl esters of cholesterol. None of the steroid has yet appeared in the maps of electron density and may be disordered within the interior of the ball of lipid.

The two major classes of lipoproteins in mammalian serum are low density lipoprotein (LDL) and high density lipoprotein (HDL). **Very low density lipoprotein** is a precursor from which lipid is stripped to produce low density lipoprotein. **Chylomicrons** seem to be hybrids of low and high density lipoprotein containing large amounts (84% by weight) of triacylglycerol.

Molecules of **low density lipoprotein** are spheres of lipid and protein that are remarkably uniform in size; their diameters are about 22 nm.⁶¹⁴ Unlike a molecule of the vitellogenin lipoprotein in yolk, which is a large protein in which a smaller ball of lipid is confined, low density lipoprotein is a large **sphere of lipid** (80% by weight) controlled by a much smaller amount of protein (20% by

weight). The sphere of lipid is composed on the average of about 2000 fatty acyl esters of cholesterol, 200 molecules of triacylglycerol, 1000 molecules of phospholipid, and 1000 molecules of cholesterol. The molar ratio of unesterified cholesterol to phospholipid is similar to that in the bilayer of a biological membrane. There is enough phospholipid and cholesterol to cover 70% of the sphere with a monolayer of the same dimensions as the monolayer of a biological membrane.⁶¹⁴ The remainder of the surface is covered by the protein.

Each molecule of low density lipoprotein contains one molecule of **apolipoprotein B100**.⁶¹⁵ Human apolipoprotein B100 is 4536 aa long. In addition to providing the remainder of the surface of a low density lipoprotein, apolipoprotein B100 sets the size of the sphere. That it does so was demonstrated by producing in cells in culture, by limited treatment with puromycin, a set of 13 different variants of low density lipoprotein, each formed from a fragment of apolipoprotein B100 of a different length from 1100 to 3600 aa. The lengths of the equators of these spherical variants of low density lipoprotein were directly proportional to the length of the fragment of apolipoprotein B100 they contained.⁶¹⁴ It was concluded that apolipoprotein B100 is a **belt around the waist** of a native low density lipoprotein; the length of the belt determines the size of the waist of the sphere of lipid.⁶¹⁵ Within the core of a molecule of low density lipoprotein, the lipid constituted by the fatty acyl ester of cholesterol and triacylglycerol may not be an isotropic fluid. Lamellar structures have been observed in this region in image reconstructions of molecules of low density lipoprotein embedded in amorphous ice.⁶¹⁶ These lamellae may or may not be present at physiological temperatures.

High density lipoprotein has similar molar ratios of fatty acyl esters of cholesterol, triacylglycerol, cholesterol, and phospholipid to those of low density lipoprotein and is also about 20% protein. Unlike the vitellogenin lipoprotein in yolk and low density lipoprotein, however, high density lipoproteins, even within the same individual, are a mixture of molecules of different size ranging in diameter from 4 to 10 nm. This is probably due to the fact that there are three different apolipoproteins, A1, A4, and A5, that are incorporated into the various high density lipoproteins. Each of these proteins is formed from **internal multiples** of a repeating segment 22 aa long. Each of these segments from any one of the three proteins can be aligned with any segment from any one of the three proteins with an average of 22% identity and no gaps. Human apolipoprotein A1 has 9 of these repeats; human apolipoprotein A4 has 13; and human apolipoprotein A5 has 13. A version of apolipoprotein A1 missing the first 43 aa of the 243 aa protein was expressed in *E. coli* and crystallized. The crystallographic molecular model of this variant of apolipoprotein A1 is a large ring formed from a consecutive series of α -helical segments, most of

which are 22 aa long.⁶¹⁷ The native molecules of high density lipoprotein may be enclosed by such rings of α -helical segments. The circular dichroic spectra of high density lipoprotein do suggest that the protein they contain is mostly α -helical.⁶¹⁸ Probes of lipid structure suggest that the phospholipid and cholesterol of high density lipoprotein are in a monolayer,⁶¹⁹ but the smallest high density lipoproteins are only about as wide as two molecules of phospholipid in a tail to tail orientation. This fact makes it hard to imagine how the lipid could be organized.

Suggested Reading

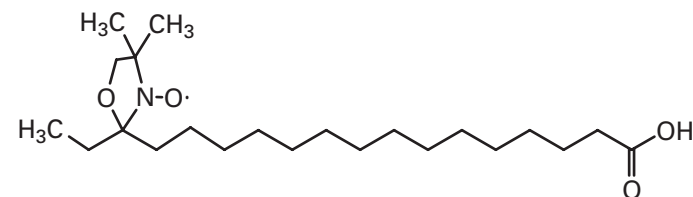
Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R., & Yoshikawa, S. (1996) The whole structure of the 13-subunit oxidized cytochrome-c oxidase at 2.8 Å, *Science* 272, 1136–1144.

Erickson, H.K. (1997) Cytoplasmic disposition of aspartate 821 in anion exchanger from human erythrocytes, *Biochemistry* 36, 9958–9967.

Problem 14-1: Pick out potential candidates for membrane-spanning α helices from the following amino acid sequence of an integral membrane-bound protein:

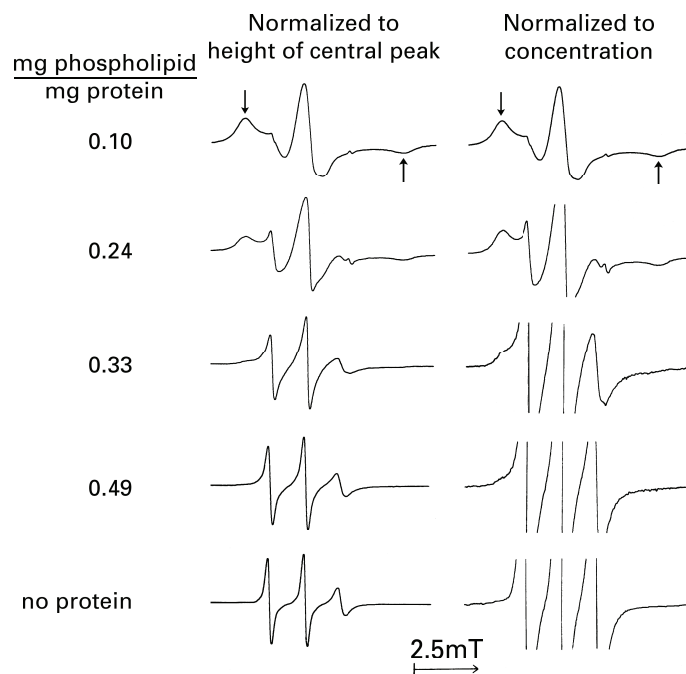
```
MNWTGLYTLLSGVNRHSTAIGRVWLSVIFIFRIMVLVVAES
VWGDEKSSFICNTLQPGSNVCYDQFFPISHVRLWSQQLILV
STPALLVAMHVAHQOHIEKKMLRLEGHGDPHLHLEEVKRHKVH
ISGTLWWTYVISVVFRLLEAVFMYVFYLLYPGYAMVRLVKC
DVYPCPNTVDCVFSRPTEKTVFTVFMLAASGICIIILNVAEVV
YLIIRACARRAQRSSNPPSRKSGSGFGHRLSPEYKQNEINKLL
SEQDGSCLKDILRRSPGTGAGLAEKSDRCSAC
```

Problem 14-2: Membranes containing only cytochrome-c oxidase can be isolated from mitochondria. By successive extractions with acetone, it is possible to deplete these membranes of their phospholipids and obtain preparations with different ratios of protein to phospholipid. The protein in these preparations retains its native conformation at all times. The following spin label was incorporated into the membranes containing different amounts of phospholipid:



2-ethyl-2-(14-carboxytetradecyl)-
4,4-dimethyloxazolidene *N*-oxyl radical

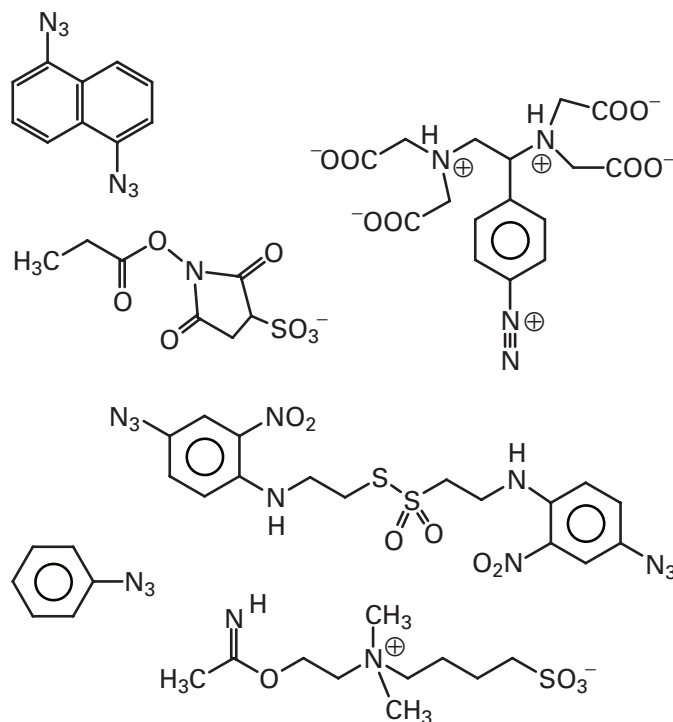
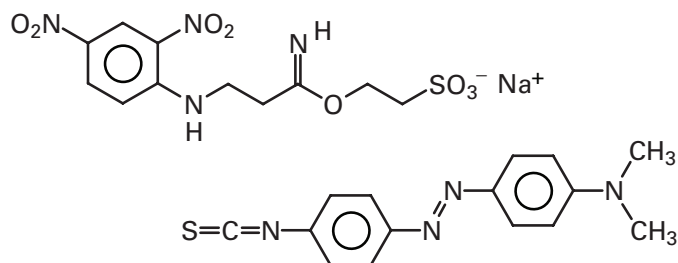
The ratio of protein to spin label was held constant. The ESR spectrum of each preparation was taken with the following result:



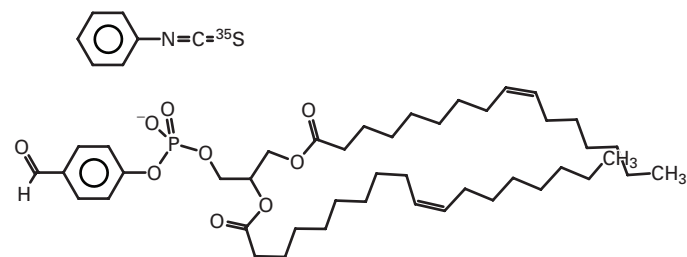
Electron spin resonance spectra of the spin label in buffered aqueous dispersions of membrane-bound cytochrome-*c* oxidase with various lipid contents. Ratio of spin label to protein remained constant. The lipid to protein ratio expressed as milligrams of lipid (milligram of protein)⁻¹ is indicated at the far left. Left, spectra normalized to the center-line height; right, the same spectra normalized to give equivalent values after two integrations. Therefore, the right column represents constant concentration of spin label. Reprinted with permission from ref 396. Copyright 1973 National Academy of Sciences.

The left-hand column gives an idea of spectral shape; the right-hand column, amplitude of the signal. The top spectrum is that of an immobilized probe; the bottom spectrum, of a mobile one. Explain these observations.

Problem 14-3: Label each of the following reagents as a hydrophobic reagent for modifying membrane-spanning segments of a protein or as an impermeant reagent. Indicate the reactive position in each reagent with an arrow, and circle the portion of the molecule that renders it hydrophobic or impermeant, respectively.



Problem 14-4: For what purpose were the following reagents synthesized and used to study membrane-bound proteins?

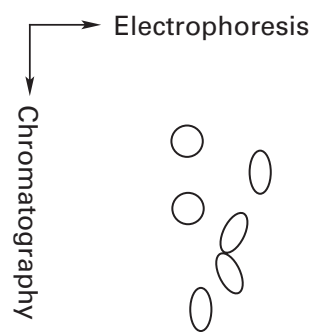


Phenyl [³⁵S]isothiocyanate modified the lysine in the segment -PNTALLSLVLMAGTFFFAMMLRKF- in an intact, native membrane-bound protein. Where is this lysine probably located relative to the bilayer of phospholipid?

Problem 14-5:

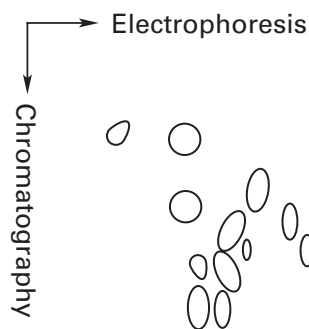
- The density of protein is 1.35 g cm⁻³. If one polypeptide of anion carrier were coiled so as to form a hard sphere, what would be its diameter? Compare this diameter to the width of a bilayer of phospholipid.
- N*-Formyl-[³⁵S]sulfinylmethionyl methylphosphate (**14-23**) reacts indiscriminately with lysines on the surfaces of protein molecules exposed to the solution in preparations of sealed membranes to form a derivative of the ε amino group that is radioactive. This reagent cannot pass through a membrane because of its polar character. Write the mechanism of this modification of lysine.

Intact erythrocytes were mixed with this reagent, and the reaction was allowed to proceed for 10 min. The cells were then washed three times with buffer. Band 3 anion transport protein was purified from these cells and was found to be radioactive. The polypeptide from this radioactive protein was cleaved with the endopeptidase thermolysin, and the digest was spread on a two-dimensional chromatogram. The chromatogram was placed over a sheet of photographic film and set aside for several days. The film was developed and radioactive peptides were located visually. The following is a diagrammatic representation of the spots observed on this film.



(C) Why was thermolysin used rather than trypsin?

This experiment was repeated with erythrocytes that had been broken open instead of intact erythrocytes. A representation of the spots observed on this film is shown in the following diagram.



You should convince yourself that each spot on the peptide maps corresponds to a unique lysine on a surface of the anion carrier. You should also understand that each erythrocyte contains 3×10^5 copies of band 3 anion transport protein in its membrane.

- (D) What two fundamental chemical properties of integral membrane-bound proteins are demonstrated by this experiment? How?
- (E) How does the surface area of the anion carrier exposed to the exterior of the cell compare to the surface area exposed to the cytoplasm?

Problem 14-6: Rhodopsin is a protein that is firmly embedded in the membranes of a vertebrate rod. If the sacs of membrane known as disks are purified from these rods, the only protein they contain in significant quantity is rhodopsin, an integral membrane-bound protein. Purified disks were dissolved in a solution of nonionic detergent and mixed with excess phospholipid in the same detergent. When the detergent was slowly removed, small unilamellar vesicles, 40–70 nm in diameter, form spontaneously. The rhodopsin molecules ended up embedded in the membranes of the vesicle. Spectral measurements demonstrated that the tertiary structure of the rhodopsin in the vesicles is the same as it was in the disk.

- (A) Papain, an endopeptidase, cleaves native rhodopsin at only one position in its entire sequence to yield two fragments from the original molecule of protein. In disk membranes, papain cleaves every rhodopsin molecule. In the reconstituted vesicles, it can cleave only 65% of the rhodopsin molecules. Draw a diagram of a reconstituted vesicle with bilayers of phospholipid and rhodopsin molecules and explain why 35% of the rhodopsin is resistant to cleavage.
- (B) Reconstituted vesicles were labeled with [125 I]iodide ion and lactoperoxidase, an enzyme that cannot pass through a bilayer of phospholipid. Those rhodopsin molecules that cannot be cleaved by papain are nevertheless labeled by lactoperoxidase. What does this experiment demonstrate about the rhodopsin molecule? How?

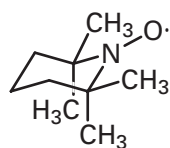
The Fluid Mosaic¹¹⁷

Every membrane in a living cell, regardless of its total surface area and shape, is an individual, intact, **isolated solution**. In each case, the solvent is a fluid bilayer of phospholipids and other amphipathic lipids of the appropriate surface area and shape, and the solutes are anchored membrane-bound proteins and integral membrane-bound proteins. The bilayer is a film of liquid paraffin and fused rings of hydrocarbon 3.6 nm in width sandwiched between thin hydrophilic lamellae 0.5–0.7 nm wide. The bilayer of phospholipid is isotropic in the two dimensions of the surface defined by its width. Because every membrane in a cell is a closed sac, this surface is finite, continuous, and unbounded. Each protein floats upon the sheet of the bilayer at an unvarying draught; its membrane-spanning α helices cannot move up or down in the bilayer owing to the hydrophobic effect. These proteins, however, unless pinned to structures outside the membrane, are free to diffuse in the two unbounded dimensions of the bilayer.

That the **solvent** from which biological membranes are composed is a **bilayer of amphipathic lipids** has

been demonstrated in many ways. The **diffraction of X-radiation** by biological membranes is mainly from the bilayer they contain. A myelinated nerve is a bundle of parallel axons each coated with a tubular spiral of myelin, which is the plasma membrane of a Schwann cell wrapped around and around the axon. Therefore, the plasma membranes of all the Schwann cells are cylindrically oriented. A myelinated nerve diffracts X-radiation. From the diffraction pattern, a radial distribution of electron density normal to the axis of an axon can be computed,⁶²⁰ and it is indistinguishable from that of multibilayers formed from an equimolar mixture of phosphatidylcholine and cholesterol (Figure 14-4D). Vesicles of biological membranes such as plasma membrane from erythrocytes, plasma membrane from the bacterium *M. laidlawii*, or endoplasmic reticulum from skeletal muscle diffract X-radiation in a circular pattern that can be thought of as the diffraction pattern of oriented bilayers of phospholipid (Figure 14-4A,C) spun around its center so that its equatorial and meridional reflections form circles. These circular diffraction patterns from biological membranes have the same periodicity as the diffraction patterns from vesicles containing only the lipids from the respective membranes,⁶²¹ and the amplitudes and periodicities of these diffraction patterns can be explained if they are assumed to arise from shells with distributions of electron density identical to those of bilayers of cholesterol and phosphatidylcholine (Figure 14-4C).⁶⁶

When **spin-labeled probes** are incorporated into biological membranes, they behave almost as if they were incorporated into bilayers of pure phospholipid. In oriented biological membranes such as flattened endoplasmic reticulum,⁶²² nerves,⁶²³ or oriented erythrocytes,⁶²³ fatty acids containing dimethyl nitroxyl radical **14-10** at various locations along the hydrocarbon incorporate with their long axes perpendicular to the surface of the membrane as they do in bilayers of phospholipid, and they display anisotropic motion resembling that displayed in oriented bilayers of pure amphipathic lipids. When 2,2,6,6-tetramethylpiperidine *N*-oxide



14-26

is incorporated into vesicles of endoplasmic reticulum from skeletal muscle, its spectrum is the same as its spectrum in vesicles of pure amphipathic lipid, and the absolute amplitude of its absorbance can be used to show that at least 85% of the amphipathic lipid in the endoplasmic reticulum is present as an unperturbed bilayer of phospholipid.⁶²⁴

Various microorganisms, such as fatty acid aux-

otrophs of *E. coli*, can be forced to incorporate high percentages of specific fatty acids into their plasma membranes. In preparations of these native membranes with a more homogeneous lipid composition, **phase transitions** between solid bilayers of phospholipid, with their hydrocarbons packed in hexagonal array, and liquid bilayers of phospholipid, with their hydrocarbons in the disordered fluid state, can be detected by X-ray diffraction.⁶²⁵ The phase transitions observed with such biological membranes are very similar to those observed when pure bilayers of the lipids extracted from these membranes undergo the same solid to liquid transformation.

These transitions can also be monitored by fluorescent probes. Fluorescent molecules such as *N*-phenyl-1-naphthylamine⁶²⁶ are hydrophobic enough to partition preferentially into the hydrocarbon of the bilayer of phospholipid and register the transition between solid and liquid by changes in the intensity of their emission of fluorescence. In vesicles of the amphipathic lipids purified from bacteria of *E. coli* the membranes of which were enriched in various fatty acids, in the intact native membranes purified from these cells, and in the whole cells themselves, the fluorescent probes detected the same phase transitions.⁶²⁷ Quantitative analysis of these results showed that at least 80% of the amphipathic lipid in the native membranes was in the form of a bilayer indistinguishable in its phase transitions from a bilayer of the purified lipids.⁶²⁶

It has already been noted that in a biological membrane each membrane-bound protein has its particular vectorial orientation. This orientation is maintained through the lifetime of a molecule of that protein by its inability to rotate even once 180° around any axis parallel to the surface of the membrane. For this to occur, the hydrophilic surfaces of the protein on the two sides of the membrane would have to pass through the 3.6 nm of hydrocarbon within the bilayer of phospholipid. This would require that the hydrogen bonds ensnaring these surfaces in the lattices of the liquid water (Figure 6-38) would all have to break simultaneously to permit the protein to capsiz. Apparently, this cannot be accomplished.

It is also the case that the phospholipids in a cellular membrane are asymmetrically distributed.⁶²⁸ These **asymmetric distributions of the phospholipids** have been demonstrated by submitting sealed, oriented biological membranes to digestion with phospholipases under nonlytic conditions^{629,630} or to modification with impermeant reagents,^{628,631} by extracting phospholipids from only one monolayer of a membrane with proteins that bind them preferentially,⁶³⁰ and by exchanging the phospholipids accessible on the outer monolayer of sealed membranes with radioactive phospholipids in other vesicles⁶³² in a reaction catalyzed by phospholipid transfer proteins.⁶³³ A phospholipid transfer protein carries a specific phospholipid tightly bound to itself,⁶³⁴ which it is able to exchange for another phospholipid of the same type at the external monolayer of a sealed

membrane. Spontaneous exchange of phospholipid between the outer monolayer of a membrane and vesicles of phospholipid in the absence of phospholipid transfer protein is also rapid enough to monitor asymmetry when the phospholipid exchanging is the dimyristoyl version.⁶³⁵ It is also possible to incorporate phospholipids labeled with 7-nitro-2,1,3-benzoxadiazol-4-yl groups on their fatty acyl substituents, allow the cells to equilibrate these labeled phospholipids across their plasma membranes, and then reduce the nitro groups to amino groups only on the outer monolayer of the resulting plasma membranes with impermeant dithionite.^{636,637} The reduction quenches the fluorescence only of those labeled phospholipids in the outer monolayer, and the resulting decrease in the fluorescence of the 7-nitro-2,1,3-benzoxadiazol-4-yl groups quantifies the asymmetry.

By one or the other of these procedures, the distributions of the various types of phospholipid across the bilayers of various biological membranes have been determined. In each case, the total moles of phospholipid in one monolayer always equals, within experimental error, the total moles in the other monolayer of the membrane, but the distribution of each type between the two monolayers is biased (Table 14-8). **Phosphatidylethanolamine and phosphatidylserine** are concentrated in the **cytoplasmic monolayers** of plasma membranes. Sphingomyelin is enriched in the extracytoplasmic monolayer of plasma membranes. Phosphatidylcholine, in animals, or phosphatidylglycerol, in bacteria, seems simply to make up the differences between the two monolayers.*

It is unclear whether or not cholesterol is asymmetrically distributed across biological membranes. Two independent measurements of cholesterol distribution in human erythrocytes found an equal ratio⁶⁴¹ or a ratio of about 2-fold in favor of the extracytoplasmic monolayer.⁶⁴² The membrane of influenza virus, derived directly from the plasma membrane of its host, has cholesterol evenly distributed between its two monolayers.⁹² Cholesterol, an aliphatic alcohol, should be able to pass readily through the bilayer.

The asymmetries in the distribution of the phospholipids are maintained by the enzyme **phospholipid-translocating ATPase**.^{643,644} This enzyme catalyzes the transport of phosphatidylethanolamine and phosphatidylserine from the extracytoplasmic monolayer of a membrane to the cytoplasmic monolayer^{645,646} and couples the transport to the hydrolysis of ATP.⁶⁴⁷ It is a

member of the same family of cation-transporting ATPases as Ca^{2+} -transporting ATPase (Figure 14-15).⁶⁴⁸ The active transport of the amino phospholipids to the cytoplasmic monolayer catalyzed by the enzyme drives phosphatidylcholine and sphingomyelin to the outer monolayer passively.^{643,645}

In sonicated vesicles of purified phospholipids, the rate at which a phospholipid, labeled in its head group with a tetramethyl cyclic nitroxyl radical, can pass from the external monolayer to the internal monolayer is slow. The time required for the distribution to come halfway to equilibrium at 30 °C was measured to be 6 h, but there was evidence that oxidation of the phospholipids was adventitiously accelerating the rate.⁶⁴⁹ When an exchange protein specific for phosphatidylcholine was used to exchange the phosphatidylcholine in small unilamellar vesicles of pure phosphatidylcholine, a very slowly exchanging component ($t_{1/2} \geq 10$ days at 37 °C) was observed that accounted for about 40% of the total phosphatidylcholine. This slow component was assigned to phosphatidylcholine on the inner monolayer that had to transfer to the outer monolayer before it could exchange.⁶³² In large unilamellar vesicles, however, the rate at which phosphatidylcholine can transfer between monolayers is more rapid.⁶³⁵ Equilibration occurs in less than 3 h at 37 °C in large unilamellar vesicles of pure dimyristoylphosphatidylcholine.

In biological membranes the **transfer of phospholipids between the two monolayers** of the bilayer seems to occur at about the same rate as that in large unilamellar vesicles of pure phospholipids or somewhat faster. [³²P]Phosphatidylcholine was observed to transfer from the extracytoplasmic monolayer to the cytoplasmic monolayer in an erythrocyte with a half-time for equilibration of 1–2 h at 37 °C.⁶³⁸ The same phospholipid labeled in its hydrophilic functional group with the same tetramethyl cyclic nitroxyl radical that equilibrated slowly in small vesicles of pure phospholipid could equilibrate in vesicles of membrane from the electric organ of *Electrophorus electricus* with a half-time of less than 10 min at 15 °C.⁶⁵⁰ In *Bacillus megaterium*, newly synthesized phosphatidylethanolamine in the cytoplasmic monolayer reaches equilibrium between the two monolayers within 30 min at 24 °C.⁶⁵¹ It is these passive equilibrations that must be constantly compensated for by the active transport of the aminophospholipids into the cytoplasmic monolayer of the plasma membrane catalyzed by phospholipid-translocating ATPase.

There is another enzyme, phospholipid scramblase,^{652,653} that catalyzes the passive transport of phospholipids between the two monolayers of a plasma membrane. The activity of this enzyme is controlled by levels of Ca^{2+} , so that the rapid equilibration of the phospholipids in the two monolayers catalyzed by this enzyme occurs only in particular circumstances.⁶³⁶ This control avoids the futile waste of MgATP that would occur if the enzyme were active continuously.

* There is an interesting inversion in the case of the membranes of unwrapped mitochondria in which phosphatidylethanolamine is concentrated in the extracytoplasmic, intramitochondrial monolayer and phosphatidylcholine is concentrated in the cytoplasmic, extramitochondrial monolayer, perhaps as a vestige of the ancestry of the mitochondrion, which is thought to have arisen from a prokaryotic symbiote.

Table 14–8: Asymmetric Distribution of Phospholipid and Sphingomyelin between the Two Sides of a Biological Membrane

	phospholipid ^a (% of total)													
	extracytoplasmic monolayer							cytoplasmic monolayer						
	PC ^b	PE	PS	SM	PG	DPG	PI	PC	PE	PS	SM	PG	DPG	PI
plasma membranes														
human erythrocyte ^{629,630}	20	5	<1	20			0.4	10	25	10	5			1.6
erythrocyte of <i>R. norvegicus</i> ⁶³⁸	30	5	1 ^c	10				15	20	15 ^c	<1			
human erythrocyte ⁶³¹	ND ^d	5	<1	ND				ND	25	15	ND			
<i>Bacillus megaterium</i> ⁶³¹		25			25 ^e				50			<5 ^e		
disks from rod outer segment ⁶³⁹	ND	10	5	ND				ND	30	5	ND			
unwrapped bovine mitochondria ⁶⁴⁰	10	20				15		30	10				5	

^aNumbers are mole percentages in each monolayer based on the total amount of phospholipid in the membrane. ^bPC, phosphatidylcholine; PE, phosphatidylethanolamine; SM, sphingomyelin; PG, phosphatidylglycerol; DPG, diphosphatidylglycerol; PI, phosphatidylinositol. ^cPhosphatidylserine plus phosphatidylinositol. ^dND, not determined. ^eBy difference.

As is the case with mixtures of pure phospholipid and cholesterol, the lipids in the bilayers of biological membranes can separate laterally into distinct phases. In particular, a separate phase that is enriched in cholesterol and sphingomyelin exists in plasma membranes of animal cells. This phase separates as patches of less fluid lipid that are surrounded by the rest of the lipid, which is more fluid. These patches are **rafts**.⁶⁵⁴ Rafts have a higher frequency of saturated fatty acyl groups on their lipids, a property characteristic of sphingomyelin that may explain its preferential inclusion. It is the higher concentration of **saturated fatty acyl groups** that causes the bilayer of a raft to be less fluid. Separated phases that resemble the rafts in natural membranes can be produced experimentally by adding cholesterol and phospholipids or sphingomyelins containing only saturated fatty acyl groups to multibilayers^{655,656} or monolayers⁶⁵⁷ composed of purified phospholipids with the normal composition of unsaturated fatty acids.

Rafts can be purified from the remainder of a biological membrane because they are much less soluble in the detergent Triton X-100 (14-12).⁶⁵⁸ When membranes from animal cells are dissolved with Triton X-100 at 4 °C, the rafts can be isolated as large aggregates. Rafts isolated in this way are enriched in lipid and depleted in protein relative to the rest of the membrane in which they are found.⁶⁵⁴ In addition to the **cholesterol and sphingomyelin**, rafts contain a high concentration of **glycosphingolipids**⁶⁵⁸ such as glucosylceramide, galactosylceramide, lactosylceramide, and globoside (Table 14-1). **Glycosylphosphatidylinositol-anchored proteins**,⁶⁵⁸⁻⁶⁶⁰ triply palmitoylated caveolin,⁶⁶¹ and doubly palmitoylated protein-tyrosine kinases⁶⁶² have been found to be preferentially associated with rafts. These proteins, like the glycolipids, insert only their fatty acyl groups into the membrane so their protein sits upon the raft. As a result, the raft itself is mostly assembled from lipid. The sizes of the rafts can be ascertained by measuring the sizes of the patches of these proteins sitting on them. These patches are about 300–500 nm in diameter.^{663,664}

Although rotational diffusion of a molecule of an integral membrane-bound protein about any axis parallel to the surface of the membrane does not occur and rotational diffusion of a molecule of phospholipid about any axis parallel to the surface of the membrane is slow, integral membrane-bound proteins, anchored membrane-bound proteins, and phospholipids all display one degree of rapid rotational diffusion about axes normal to the surface of the membrane and two degrees of translational diffusion along axes parallel to the surface of the membrane. These diffusional degrees of freedom are prescribed by the fact that a bilayer of amphipathic lipids is a two-dimensional solvent, and solutes dissolved in this solvent find themselves in a **two-dimensional solution**.

The translational diffusion of proteins over the two

dimensions of a plasma membrane was strikingly demonstrated by an experiment involving the **fusion of two cells**.⁶⁶⁵ Immunoglobulins G specific for antigens on the surface of murine (c11D) and human (VA-2) cells, respectively, were produced. These immunoglobulins were covalently modified with different fluorescent reagents, one that fluoresced green and one that fluoresced orange, respectively. The addition of the former immunoglobulins to mouse cells turned proteins in their plasma membranes green, and the addition of the latter immunoglobulins to human cells turned proteins in their plasma membranes orange. When a mouse cell was fused with a human cell, the hybrid was initially stained green at one side and orange at the other, but the two sets of antigenic proteins then diffused over the plasma membrane of the hybrid until, within 40 min, they were each uniformly distributed. This intermixing resulted from two-dimensional translational diffusion.

In a three-dimensional, isotropic solvent, such as an aqueous solution, the observed translational diffusion coefficient, D_T , is the proportionality constant (Equation 1-63) between the net flux, J , of a certain solute across a unit area, in a plane normal to its direction of net flux, and the gradient of its concentration, c , along the direction of net flux, x :

$$J = -D_T \left(\frac{\partial c}{\partial x} \right)_t \quad (14-2)$$

The units on J are moles centimeter⁻² second⁻¹, and on the gradient of concentration c they are (moles centimeter⁻³) centimeter⁻¹. Therefore, the units of D_T are centimeters² second⁻¹. The mean square displacement, $\overline{d^2}$, that the molecules of the solute will experience over a given time interval, t , is related to the diffusion coefficient by the equation

$$\overline{d^2} = 4 D_T t \quad (14-3)$$

In a three-dimensional solution, molecules of the solute will also experience rotational motion as well as translational motion, and the observed rotational diffusion coefficient, D_R , for this rotational motion can be defined, in analogy with Equation 14-3, as

$$D_R = \frac{\overline{\theta^2}}{2t} \quad (14-4)$$

where $\overline{\theta^2}$ is the mean square angular displacement experienced in time t . If θ is expressed in radians, the units on D_R are radians second⁻¹.

The theoretical relationship that is used to calculate a frictional coefficient for translational motion in three dimensions, f_{T3} , from the observed translational diffusion coefficient is

$$f_{T3} = \frac{k_B T}{D_T} \quad (14-5)$$

where k_B is Boltzmann's constant and T is the temperature. An analogous theoretical relationship that is used to calculate a frictional coefficient for rotational motion in three dimensions, f_{R3} , from the observed rotational diffusion coefficient can be written

$$f_{R3} = \frac{k_B T}{D_R} \quad (14-6)$$

If the molecules of the solute were hard spheres of radius r in a solvent of viscosity η (Equation 1-66), then

$$f_{T3} = 6\pi\eta r \quad (14-7)$$

and

$$f_{R3} = 8\pi\eta r^3 \quad (14-8)$$

Equations 14-7 and 14-8 are theoretical equations relating the frictional coefficient to the dimensions of the sphere.

Diffusion of a solute in a two-dimensional solvent, isotropic in those two dimensions, can be treated in parallel. The solvent is two-dimensional because the solute is confined to rotate only about an axis normal to the plane defined by the two dimensions and to translate in only the two dimensions. The observed **one-dimensional rotational diffusion coefficient** D_{R1} is defined by Equation 14-4 where the angular displacement is only around the axis normal to the plane. The observed **two-dimensional translational diffusion coefficient**, D_{T2} , is the proportionality constant (Equation 14-2) between the net flux of a substance across a unit width on a line normal to its direction of net flux, in moles centimeter⁻¹ second⁻¹, and the gradient of its concentration along the direction of net flux, in (moles centimeter⁻²) centimeter⁻¹. The units of D_{T2} are also centimeter² second⁻¹. The mean square displacement of one molecule is still governed by Equation 14-3. This relationship has been verified experimentally by following the lateral movements as a function of time of single, fluorescently labeled molecules of phospholipid in a bilayer of dioleoyl phosphatidylcholine.⁶⁶⁶ The frictional coefficient for one-dimensional rotational diffusion, f_{R1} , can be calculated from the observed rotational diffusion coefficient by a theoretical equation analogous to Equation 14-6, and the frictional coefficient for translation in a two-dimensional solvent, f_{T2} , can be calculated from the observed translational diffusion coefficient by a theoretical equation analogous to Equation 14-5, with the substitution of f_{R1} and f_{T2} for f_{R3} and f_{T3} , respectively.

When diffusion of a solute such as an integral membrane-bound protein in a two-dimensional solvent such as a bilayer of phospholipids is evaluated, it is usually assumed that the molecule can be treated as an equivalent right circular cylinder of radius r . The theoretical equations relating the **frictional coefficient** for its rotational diffusion, f_{R1} , to the dimensions of a right cylinder of any radius r in a two-dimensional solvent of width h about an axis normal to the surface of the solvent have an exact solution, which is

$$f_{R1} = 4\pi\eta_H r^2 h \quad (14-9)$$

For right cylinders of radius r that are the size of integral membrane-bound proteins (Figures 14-14 to 14-18) undergoing two-dimensional translational diffusion in a sheet of liquid paraffin with a width, h , of about 3-4 nm and a viscosity, η_H , significantly greater than that of the water that sandwiches the liquid paraffin on both sides with a viscosity, η_W , of 0.9 mPa s, at 298 K

$$f_{T2} \cong 4\pi\eta_H h \left(\ln \frac{\eta_H h}{\eta_W r} - \gamma_E \right)^{-1} \quad (14-10)$$

where γ_E is Euler's constant (0.5772).⁶⁶⁷ This theoretical relationship relating the frictional coefficient, f_{T2} , for its translational diffusion to the dimensions of a right cylinder is not exact because the equations for slow viscous flow used to derive Equation 14-7 in three dimensions have no exact solution in two dimensions.

Unlike Equations 14-7 and 14-8, Equations 14-10 and 14-9 cannot be used quantitatively to determine the equivalent of a Stokes' radius a (Equation 1-67) or the shape of a molecule of protein (Figure 12-1) or lipid in a bilayer of phospholipid. It is not possible to determine independently the value of η_H experienced by the diffusing solute because a bilayer of natural amphipathic lipids is not an isotropic sheet of liquid hydrocarbon (Figure 14-2). It is also not known what value should be used for h , the width of the bilayer. These equations, however, can be used qualitatively to show that the diffusion coefficients measured are in reasonable agreement with the sizes of the diffusing solutes, the width of a bilayer, and the expected viscosity of the liquid hydrocarbon.

The rotational diffusion constant for a naturally fluorescent molecule of protein such as bacteriorhodopsin or rhodopsin with their tightly bound retinals or a molecule of protein modified with a fixed fluorescent reagent is measured by monitoring the **decay in the anisotropy** of the fluorescence⁶⁶⁸ or phosphorescence⁶⁶⁹ of the chromophore following excitation with a flash of polarized light. The rotational diffusion coefficient of bacteriorhodopsin in bilayers of dimyristoylphosphatidylcholine is affected little by the concentration of the protein in the bilayer, and at 30 °C it is equal to $7 \times 10^4 \text{ s}^{-1}$. The

rotational diffusion coefficient of rhodopsin in densely packed disks of photoreceptors is $5 \times 10^4 \text{ s}^{-1}$ at 20°C .¹⁹² If the radius of an imaginary cylindrical bacteriorhodopsin, r , is taken as 2.0 nm and the width of the bilayer, h , as 4.5 nm, then the viscosity of the bilayer, η_{H} , sensed by the rotating protein (Equation 14–9), is 400 mPa s.

Large multilamellar vesicles of dimyristoylphosphatidylcholine, 25–50 μm across, can be prepared so that they contain various concentrations of bacteriorhodopsin.⁶⁷⁰ Because bacteriorhodopsin is fluorescent, the distribution of the protein over the membranes of a vesicle can be monitored in a microscope by the distribution of fluorescence. When a circular area 5 μm in diameter in the middle of a vesicle is submitted to intense irradiation by a laser, the retinal within the circle is photolytically bleached. After bleaching, the molecules of bacteriorhodopsin in the circular area are no longer fluorescent, but those surrounding the circle still are. As translational diffusion takes place, the circle slowly fills with fluorescent molecules entering from the perimeter, and the bleached circle gradually disappears. From such **recovery of fluorescence following photobleaching**,^{671,672} the translational diffusion coefficient D_{T_2} of the unbleached bacteriorhodopsin moving into the circle can be calculated.

Measurements were made of the two-dimensional translational diffusion coefficient for bacteriorhodopsin at several different temperatures and concentrations of the protein in the bilayers.⁶⁷⁰ The values varied between 0.1×10^{-8} and $4 \times 10^{-8} \text{ cm}^2 \text{ s}^{-1}$. As the mole fraction of bacteriorhodopsin in the bilayer was decreased from 1 mol (30 mol of phospholipid)⁻¹ to 1 mol (210 mol of phospholipid)⁻¹, the translational diffusion coefficient increased from 0.1×10^{-8} to $1.6 \times 10^{-8} \text{ cm}^2 \text{ s}^{-1}$ at 25°C . Because the diffusion coefficient was still increasing significantly at the lowest concentration at which measurements could be made, the translational diffusion coefficient at zero density, $D_{\text{T}_2}^0$, could not be determined accurately by extrapolation.

At the lowest concentrations of protein examined at 30°C , the translational diffusion coefficient of bacteriorhodopsin, $3.4 \times 10^{-8} \text{ cm}^2 \text{ s}^{-1}$, is that of a cylinder of protein with a radius r equal to 2.0 nm, in a bilayer of phospholipid with a width, h , of 4.5 nm, the **viscosity** of whose hydrocarbon, η_{H} , is 110 mPa s. The broadest dimension of the bundle of α helices in a single molecule of bacteriorhodopsin is about 4 nm (Figure 14–14), the width of a bilayer of natural phospholipid and cholesterol, including the hydrophilic functional groups, is 4–5 nm (Figure 14–4D), the viscosity of motor oil at 30°C is between 100 and 200 mPa s, and the viscosity of vegetable oil at 30°C is about 50 mPa s. Because realistic numerical values for these three parameters can be used to calculate, by Equations 14–9 and 14–10, a diffusion coefficient equal to the one that is measured, it appears that bacteriorhodopsin, at least when it is in bilayers of dimyristoylphosphatidylcholine, is diffusing freely and

predictably within the two-dimensional solvent formed by those bilayers. Its diffusion is a random walk driven by thermal energy through an isotropic, viscous medium just as is the diffusion of a soluble protein through an isotropic aqueous solution.

Translational diffusion coefficients have been measured for other integral membrane-bound proteins. Bacteriorhodopsin and rhodopsin are already fluorescent, but a purified membrane-bound protein can be covalently modified with a fluorescent electrophilic reagent and then incorporated into bilayers of phospholipid, and its two-dimensional translational diffusion coefficients can be measured⁶⁷³ by monitoring recovery of fluorescence following photobleaching. For anion carrier at a surface concentration of 1 mol (200 mol of phospholipid)⁻¹, the translational diffusion coefficient is $1.6 \times 10^{-8} \text{ cm}^2 \text{ s}^{-1}$ at 30°C .⁶⁷⁴ The translational diffusion coefficients of bovine rhodopsin, Ca^{2+} -transporting ATPase from endoplasmic reticulum of skeletal muscle, and acetylcholine receptor have all been determined by fluorescence photobleaching recovery at several temperatures in reconstituted membranes at high dilution [< 1 mol of protein (3000 mol of phospholipid)⁻¹].⁶⁷³ The values for the translational diffusion coefficients at 25°C are between 1.4×10^{-8} and $2 \times 10^{-8} \text{ cm}^2 \text{ s}^{-1}$ for all three proteins, and they are indistinguishable within the ranges of their standard deviations. From Equation 14–10, the viscosity calculated for the bilayer from these latter measurements is between 100 and 200 mPa s.

At high dilution at 25°C , the translational diffusion coefficients of all integral membrane-bound proteins are between 1×10^{-8} and $2 \times 10^{-8} \text{ cm}^2 \text{ s}^{-1}$. This means that after 1 s the average value of the square of the distance that a protein will be situated from the position it occupied initially will be $10 \mu\text{m}^2$. In a densely packed plasma membrane, however, the translational diffusion coefficients are significantly less. For example, for rhodopsin at 20°C , densely packed in the disks in photoreceptors, the translational diffusion coefficient⁶⁷⁵ is $0.3 \times 10^{-8} \text{ cm}^2 \text{ s}^{-1}$. A value of $0.02 \times 10^{-8} \text{ cm}^2 \text{ s}^{-1}$ at 37°C has been determined for randomly labeled proteins in the plasma membranes of L-6 cells.⁶⁷⁶ In the latter situation, the **mean square displacement** for one of these proteins after 1 s will be only $0.1 \mu\text{m}^2$. The diameter of a normal eukaryotic cell is about $20 \mu\text{m}$. Therefore, it should take about 100 min for a protein with this low value for its translational diffusion coefficient to spread over the plasma membrane of a cell of this size if the protein were added at only one point on its surface.

It is the **dense packing of the proteins** in a normal biological membrane that causes the diffusion coefficients of most proteins to be much less than they are when they are moving unhindered over a bilayer of phospholipids. The value of $0.02 \times 10^{-8} \text{ cm}^2 \text{ s}^{-1}$ at 37°C measured for the diffusion coefficient of proteins in a plasma membrane is in the same range as diffusion coef-

ficients of $0.06 \times 10^{-8} \text{ cm}^2 \text{ s}^{-1}$ at 25 °C for ubiquinol-cytochrome-*c* reductase in mitochondrial membranes⁶⁷⁷ and $0.006 \times 10^{-8} \text{ cm}^2 \text{ s}^{-1}$ for phycobilisomes in membranes of thylakoids.⁶⁷⁸ When the concentration of protein in the mitochondrial membranes was decreased systematically by incorporating endogenous phospholipid, the diffusion coefficient of ubiquinol-cytochrome-*c* reductase increased monotonically. Upon a 7-fold dilution, its diffusion coefficient had increased almost 20-fold.⁶⁷⁷

The translational diffusion coefficients of integral membrane-bound proteins vary dramatically with concentration but insignificantly with variations in the apparent radius, *r*, of the equivalent cylinder because of the logarithmic dependence of the frictional coefficient on this latter parameter (Equation 14–10). These are two additional reasons why translational diffusion coefficients cannot be used to provide any insight into the shapes of these proteins in the bilayer of phospholipid.

The **phospholipids** in a membrane also display translational diffusion. This can be followed by using phospholipids, such as phosphatidylethanolamine, that have been modified by fluorescent reagents, in the case of phosphatidylethanolamine at its primary amine. The **translational diffusion coefficients** for such a fluorescent lipid in bilayers of various phospholipids are between 4×10^{-8} and $9 \times 10^{-8} \text{ cm}^2 \text{ s}^{-1}$ at 25 °C.^{673,679,680} These values compare favorably with those calculated from spin-exchange among molecules of a nitroxylphosphatidylcholine in vesicles of various natural phospholipids, which are about $10 \times 10^{-8} \text{ cm}^2 \text{ s}^{-1}$ at 45 °C.⁶⁸¹ In bilayers of dimyristoylphosphatidylcholine and cholesterol at various ratios, the diffusion coefficients for a fluorescent phospholipid are between 1×10^{-8} and $3 \times 10^{-8} \text{ cm}^2 \text{ s}^{-1}$ at 25 °C.¹⁰³ The translational diffusion coefficients for lipids are not much greater than those for proteins.

Even though the diffusion coefficients for phospholipids are not much larger than those for proteins, Equation 14–10 does not describe their behavior. It fails to do so because a phospholipid does not span the membrane completely so the viscosity of the fluids at its two ends are dramatically different⁶⁸² and because, unlike a membrane-spanning protein, its cross-sectional area is the same as those of the phospholipids forming the bilayer.^{103,679,683}

Associated with the bilayer of a biological membrane is a microviscosity. The **microviscosity** is the viscosity experienced by a small hydrophobic solute dissolved in the liquid hydrocarbon while it is rotating isotropically. Because it is measured by following rotation, the microviscosity is the viscosity of the solvent in the immediate vicinity of the small solute. This microviscosity is generally estimated from the polarization retained in the fluorescence of a hydrophobic, fluorescent solute such as 2-methylanthracene⁶⁸⁴ after its excitation with a flash of polarized light. A hydrophobic

solute is used so that most of the molecules of that solute in the sample have been incorporated into the hydrocarbon of the bilayer of phospholipid. The more rapidly the solute is reorienting within the bilayer of phospholipid during the lifetime of the excited state, the greater will be the loss of its polarization. This loss of polarization can be calibrated by the behavior of the fluorescent solute in hydrocarbon solvents of known macroscopic viscosity.⁶⁸⁴

In bilayers made in the laboratory from purified natural phospholipid, microviscosities between 100 and 200 mPa s have been observed at 25 °C.^{685,686} The microviscosity determined in vesicles formed from only the lipids in a biological membrane is almost the same as that determined for the complete biological membranes from which the lipids were extracted,⁶⁸⁶ and addition of an integral membrane-bound protein to vesicles of pure phospholipid affects the microviscosity only slightly.⁶⁸⁷ These results suggest that the regions within the bilayer of phospholipid occupied by the fluorescent solutes used to monitor this property are mainly the **bulk lipid** between the molecules of protein. Because it is the rotational diffusion of the probe that senses the microviscosity rather than its translational diffusion, the presence of protein, which provides obstacles mainly to translation, has only a small effect. Nevertheless, the microviscosities determined are in the same range as the viscosities that seem to be controlling the translational and rotational diffusion of molecules of protein when they are dissolved at low concentrations in bilayers of phospholipids. The composition of the fatty acyl groups in the phospholipids forming biological membranes (Table 14–3) has been adjusted through evolution by natural selection to compensate for the different mean body temperatures of cold-blooded animals so that a similar microviscosity is maintained regardless of the mean temperature of the environment.⁶⁸⁸

The addition of **cholesterol** increases the observed microviscosity of bilayers of phospholipid by a factor of 5–10,⁶⁸⁵ an effect that seems to be considerably larger than that of cholesterol on the translational diffusion coefficient of phospholipids.¹⁰³ The addition of cholesterol to bilayers of phospholipid, however, increases the rotational diffusion coefficients of 1,6-diphenyl-1,3,5-hexatriene dissolved in them only by factors of 2.⁹⁶ These results suggest that the viscosity of the bilayer becomes more anisotropic upon addition of cholesterol.

Epidermal growth factor receptor is an integral membrane-bound protein that depends on its ability to diffuse translationally over a bilayer of phospholipid to accomplish its function. Epidermal growth factor is a polypeptide hormone that stimulates the growth of cells from a variety of tissues. Epidermal growth factor receptor is the protein in the plasma membrane of a cell to which epidermal growth factor binds to exert its effect on the cell. Human epidermal growth factor is a small (53 aa) soluble protein; human epidermal growth factor

receptor is a large (1186 aa) monomeric glycoprotein that spans the membrane.¹⁸² Epidermal growth factor receptor is a member of a group of structurally and functionally related receptors for growth factors on the cell surface characterized by an intrinsic activity for protein tyrosine kinase.¹⁸¹ The initial response in the cascade leading to the mitosis caused by the binding of epidermal growth factor to epidermal growth factor receptor is the activation of this protein-tyrosine kinase.

The amino acid sequence of human epidermal growth factor receptor¹⁸² can be divided into **two domains** of about equal size that are located on opposite sides of the plasma membrane. The extracytoplasmic domain of the protein (620 aa) contains the binding site for epidermal growth factor.¹⁸⁴ The cytoplasmic domain of the protein (540 aa) contains the active site for protein-tyrosine kinase.¹⁸³ Both of these domains have been produced independently, in their entirety, and they are well-behaved soluble proteins with the respective functions.^{183,184} Both have been crystallized, and crystallographic molecular models are available for each of them.⁶⁸⁹⁻⁶⁹¹ In the intact native protein the short segment between these two domains is composed of 23 hydrophobic amino acids, 15 of which are leucine, isoleucine, valine, methionine, or phenylalanine. There is no doubt that this segment spans the membrane, presumably in one α helix, connecting by this tether the extracytoplasmic domain and the cytoplasmic domain in the complete native protein. It has been shown that mutations, insertions, or deletions in this membrane-spanning segment have little effect on activation of the protein-tyrosine kinase,^{692,693} hence, its role in the activation of the protein-tyrosine kinase activity must not involve any severe structural requirements, short of spanning the bilayer of phospholipid and conjoining the two domains.

There have been a number of reports implicating the **dimerization** of epidermal growth factor receptor in the activation of its protein-tyrosine kinase. Moderate yields of covalent chemical cross-linking between monomers are observed but only after epidermal growth factor has been bound,^{694,695} bivalent immunoglobulins against epidermal growth factor receptor, but not Fab fragments from these immunoglobulins, can activate its tyrosine kinase in the absence of epidermal growth factor,⁶⁹⁶⁻⁶⁹⁸ and mutant forms of epidermal growth factor receptor can suppress the activation of wild-type epidermal growth factor receptor.

It has been possible to follow the dimerization of monomeric epidermal growth factor receptor as a function of time by quantitative cross-linking,⁶⁹⁹ just as the tetramerization of phosphoglycerate mutase was followed by quantitative cross-linking (Figure 13-17). After epidermal growth factor was added to epidermal growth factor receptor dissolved in a solution of the detergent Triton X-100 (14-12), the initially monomeric protein dimerized in a reaction that could be shown to be **kinet-**

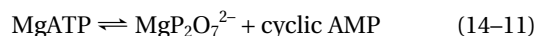
ically second-order in the concentration of protein. When the activation of protein-tyrosine kinase activity was followed in the same preparations, it was found that it was also a process second-order in the concentration of epidermal growth factor receptor and that the second-order rate constants for dimerization and the activation of protein-tyrosine kinase were identical. It follows that the rate-limiting step in the activation of the enzyme is dimerization of the protein. In the plasma membrane, this dimerization must result from collisions between monomers of epidermal growth factor receptor as they diffuse in two dimensions across the surface of the bilayer of phospholipid.

Because the protein spans the membrane in a single α helix, it seems unlikely that the order to dimerize is transmitted through this α helix across membrane, and consequently it should be the case that the binding of epidermal growth factor to the extracytoplasmic domain causes the extracytoplasmic domain to dimerize. In fact, the crystallographic molecular models of the extracytoplasmic domain of human epidermal growth factor receptor are symmetrical dimers of the complex between the protein and either epidermal growth factor⁶⁹⁰ or a related hormone.⁶⁸⁹ Even though the cytoplasmic domains of two intact molecules of epidermal growth factor receptor sterically inhibit the **dimerization of the extracytoplasmic domains**,⁵⁴⁵ during dimerization of the intact protein, dimerization of the cytoplasmic domains through an interface between them is also required for activation of the protein-tyrosine kinase. This conclusion follows from the facts that bivalent immunoglobulins G directed against the carboxy terminus of the protein activate its protein-tyrosine kinase to a level comparable to the activation by epidermal growth factor⁶⁹⁸ and that when epidermal growth factor is removed from the binding sites on dimerized protein, the intact protein dissociates into monomers greater than 40-fold more slowly than does a deletion mutant lacking the complete cytoplasmic domain.⁵⁴⁵

Therefore, dimerization of the extracytoplasmic domains produced by the binding of epidermal growth factor drags together the cytoplasmic domains, which are tethered through the membrane-spanning segment. The resulting juxtaposition of the cytoplasmic domains, which initially involves a steric problem because they are not positioned properly, nevertheless promotes, in turn, their dimerization through an interface that forms between them after they become properly aligned. The **dimerization of the cytoplasmic domains** leads to the activation of the protein-tyrosine kinase. In the crystallographic molecular model of the cytoplasmic domain, the asymmetric unit is a monomer, but because the space group of the crystal is *I*23, two monomers are related by an exact 2-fold rotational axis of symmetry, and the interface between them may be the same as the one in the dimeric cytoplasmic domains in the activated native protein.⁶⁹¹

Both observations of the kinetics of activation of epidermal growth factor receptor⁷⁰⁰ and of the binding of epidermal growth factor⁷⁰¹ and the crystallographic molecular models of the liganded, dimeric extracytoplasmic domains of epidermal growth factor receptor^{689,690} and the related fibroblast growth factor receptor⁷⁰² demonstrate that each monomer of the receptor binds a molecule of the respective hormone. The resulting dimer is rotationally symmetric, each subunit carrying its own molecule of hormone. In fact, the two symmetrically arrayed molecules of hormone are on completely opposite sides of the dimers in the crystallographic molecular models. **Growth hormone receptor**, however, which is unrelated to epidermal growth factor receptor but which also dimerizes upon binding of its hormone, forms a different kind of complex in which one molecule of hormone is bound by two molecules of the receptor.^{703,704} It is the formation of this **asymmetric complex** that dimerizes growth factor receptor and leads to its activation.⁷⁰⁵ One peculiarity of this process of activation, which is a consequence of both the fact that one molecule of hormone gathers together two molecules of receptor and the fact that molecules of receptor are not dimers before the hormone is added, is that high concentrations of growth hormone inhibit both the dimerization⁷⁰³ and the activation of growth hormone receptor.⁷⁰⁵ These facts require that monomers of growth hormone receptor be diffusing independently of each other through the plasma membrane before they are dimerized by binding to the two opposite sides of growth hormone.

There is another set of membrane-bound proteins that also relies on translational diffusion over the surface of the plasma membrane to fulfill its biological function. This is the **adenylate cyclase system**. The role of this set of proteins is also to respond to the presence of an agonist in the medium surrounding the cell. Binding of the agonist to the extracytoplasmic surface of a particular protein in the plasma membrane either activates or inhibits adenylate cyclase, which is the enzymatic activity of an active site at the cytoplasmic surface of a different protein in the same plasma membrane. This active site is responsible for the reaction



If the agonist increases the rate of production of cyclic AMP catalyzed by this enzyme, it is stimulatory; if it decreases the production, it is inhibitory.

The agonist initiates the process by binding to one of a set of receptors, each of which is a membrane-bound protein. A typical example of one of these receptors is **β -adrenergic receptor**. This protein has the site to which a **β -adrenergic agonist** such as epinephrine or norepinephrine binds as the first step in the stimulation of adenylate cyclase. β -Adrenergic receptor has been purified to homogeneity from microsomes of plasma membrane from hamster lung that have been dissolved in the

nonionic detergent digitonin.⁷⁰⁶ The ability of the receptor to bind agonist was used as an assay, and the purification involved affinity adsorption to a solid phase to which a β -adrenergic antagonist had been attached (Table 1-3). The purified receptor is a membrane-spanning glycoprotein. The polypeptide composing the protein from the hamster is 418 aa in length and contains seven hydrophobic segments, each greater than 20 aa in length, that are candidates to be α helices spanning the membrane.^{707,708}

β -Adrenergic receptor is a member of a **large family** of integral membrane-bound proteins responsible for responding to signals such as hormones, odorants, neurotransmitters, and light. In the human genome there are at least 950 genes encoding members of this family.⁷⁰⁹ Rhodopsin (Table 14-6) belongs to this family, and bacteriorhodopsin (Figure 14-14) is a homologous bacterial protein. Beginning 6 aa before its first membrane-spanning α helix and finishing 16 aa beyond the last, the amino acid sequence of human rhodopsin can be aligned with high statistical significance with that of human β 1-adrenergic receptor (21% identity; 2.6 gap percentage), so the structure of β 1-adrenergic receptor must be superposable upon that of rhodopsin⁷¹⁰ and hence upon that of bacteriorhodopsin (Figure 14-14). In particular, the seven hydrophobic segments in β 1-adrenergic receptor must be α helices that span the membrane. The segment (59 aa) of amino acid sequence on the extracytoplasmic side of the membrane amino-terminal to the first membrane-spanning α helix and the two segments (80 and 97 aa) on the cytoplasmic side of the membrane between the fifth and the sixth membrane-spanning α helices and carboxy-terminal to the seventh, respectively, are much longer in human β -adrenergic receptor than they are in bacteriorhodopsin and constitute extracytoplasmic and cytoplasmic domains involved in the binding of the hormone and the transmission of the information.

The **adenylate cyclase** itself is a much larger protein. The isoform of the human enzyme that responds to binding of an agonist to a β -adrenergic receptor is 1353 aa in length. As with all of the isoforms of the enzyme, the amino acid sequence of the adenylate cyclase contains 12 hydrophobic segments thought to span the plasma membrane as α helices. The pattern in which these α helices occur as well as alignments of segments of amino acid sequence suggest that the protein is a product of an internal duplication.⁷¹¹ In particular, two segments of amino acid sequence that contain no membrane-spanning α helices, each about 220 aa long, occur respectively in the middle of the protein, which is the carboxy terminus of the first half of the polypeptide, and at the carboxy terminus of the complete protein, which is the carboxy terminus of the second half of the polypeptide. The sequences of these two segments can be aligned,⁷¹¹ and their crystallographic molecular models are superposable.⁷¹² Each is preceded by six hydrophobic segments in close succession. Together, the two homol-

ogous segments that have no membrane-spanning α helices form a large cytoplasmic domain that is responsible for the catalysis of adenylate cyclase (Equation 14–11)⁷¹³ and for which crystallographic molecular models are available.^{711,712} In the various crystallographic molecular models, the two homologous cytoplasmic domains have superposable structures and are related to each other by a 2-fold rotational axis of pseudosymmetry.

The final component in the overall adenylate cyclase system is a guanosine nucleotide-binding protein or **G-protein**. There are several types of G-proteins present in the same membrane, one type mediating the stimulation of adenylate cyclase, another type mediating its inhibition, and other types with other roles in other systems.

The stimulation or inhibition of adenylate cyclase by an agonist requires the constant presence of **GTP** under physiological conditions⁷¹⁴ owing to the requirement that the stimulatory G-protein have GTP bound to it before the active site on adenylate cyclase can be stimulated to produce cyclic AMP. The stimulatory G-protein involved in this process binds GTP tightly⁷¹⁵ and under the appropriate circumstances catalyzes a slow hydrolysis of the GTP to GDP and inorganic phosphate.⁷¹⁶ Both of these properties are reminiscent of the binding and hydrolysis of GTP performed by tubulin. The rate of hydrolysis of GTP at the active site of a stimulatory G-protein is enhanced by the binding of agonist to β -adrenergic receptor,^{717,718} and this is expressed as a guanosine triphosphatase that is activated by the agonist.

Just as the slow hydrolysis of GTP in a microtubule is a timing device to give the growing end enough time to find its goal before it is eliminated for its failure to do so, the **slow hydrolysis of GTP** bound to the stimulatory G-protein is a timing device to terminate the activity of adenylate cyclase when the agonist is no longer present at the extracytoplasmic surface of the cell.^{716,719} When an agonist is abruptly removed from the binding site on β -adrenergic receptor by adding an antagonist to the solution in which the membranes are suspended, the adenylate cyclase activity decays slowly.⁷²⁰ The rate at which the adenylate cyclase activity decays is equal to the rate at which GTP is hydrolyzed within the active site of the G-protein that is coupled to β -adrenergic receptor and adenylate cyclase.⁷¹⁹ If this hydrolysis of GTP is blocked by cholera toxin, a specific inhibitor of this process, the adenylate cyclase no longer decays with time. The decrease in the rate of the GTPase activity as a function of the concentration of cholera toxin parallels the decrease in the fraction of the adenylate cyclase that is turned off.⁷¹⁶ The hydrolysis of the guanine nucleotide can also be prevented by the use of a chemical analogue of GTP such as guanosine 5'-O-(3-thiotriphosphate) that cannot be hydrolyzed by the G-protein. One of these analogues, as does cholera toxin, produces

adenylate cyclase activity that does not turn off when agonist is driven from the receptor.⁷¹⁶

After the GTP is hydrolyzed, GDP remains tightly bound to the G-protein, preventing the binding of GTP if the β -adrenergic receptor is unoccupied. The binding of agonist to the receptor, however, accelerates the dissociation of this bound GDP.⁷²¹

Therefore, the overall sequence of steps is the following. Guanosine triphosphate binds to the stimulatory G-protein activated by the binding of agonist to β -adrenergic receptor, and this complex between GTP and G-protein stimulates the production of cyclic AMP at the active site on adenylate cyclase as long as the GTP remains unhydrolyzed. Following hydrolysis, the tightly bound GDP prevents the G-protein from activating adenylate cyclase. When that GDP is released in a dissociation stimulated by β -adrenergic receptor to which agonist is bound, the empty site can again bind GTP.

Purified G-proteins are each constructed from **three different subunits**, α , β , and γ . The α polypeptides from all of the various G-proteins are highly homologous in sequence (averaging about 60% identity in pairwise comparisons),⁷²² but they differ significantly in length ($n_{aa} = 310\text{--}400$)⁷²² because of three regions in which long insertions can occur. The α polypeptide of the human stimulatory G-protein that is involved in the β -adrenergic system is 394 aa long;⁷²³ the β polypeptide, 340 aa; and the γ polypeptide, 70 aa. There is a crystallographic molecular model of an intact $\alpha\beta\gamma$ heterotrimeric G-protein.⁷²⁴ The α subunit has the tertiary structure of the proteins in an even larger family that bind guanosine nucleotides and control various cellular functions.⁷²⁵ The β subunit is a β propeller of seven blades (Figure 6–13).

The G-proteins involved in coupling receptors such as β -adrenergic receptor to their ultimate biological response are bound to membranes because they are **posttranslationally modified with lipid**.⁷²⁶ The α subunits are palmitoylated at a cysteine near their amino terminus (Cysteine 3 in the G-protein responsible for coupling β -adrenergic receptor to adenylate cyclase), and some are also myristoylated at their amino terminal. The γ subunits are S-geranylgeranylated (Figure 3–16) at a cysteine four amino acids from their carboxy termini. None of the subunits, however, has a membrane-spanning segment.⁷²⁴

Molecules of adenylate cyclase and molecules of β -adrenergic receptor diffuse about independently over the bilayer of phospholipids forming a membrane.⁷²⁷ There is no evidence that they associate with each other to form a specific complex. When increasing fractions of the β -adrenergic receptors in a membrane are destroyed by covalent modification, the final activities of adenylate cyclase achieved following addition of an agonist remain the same (Figure 14–29A).⁷²⁸ This observation is inconsistent with a strong, specific association between β -adrenergic receptors and adenylate cyclase because a permanently inactivated β -adrenergic receptor does not

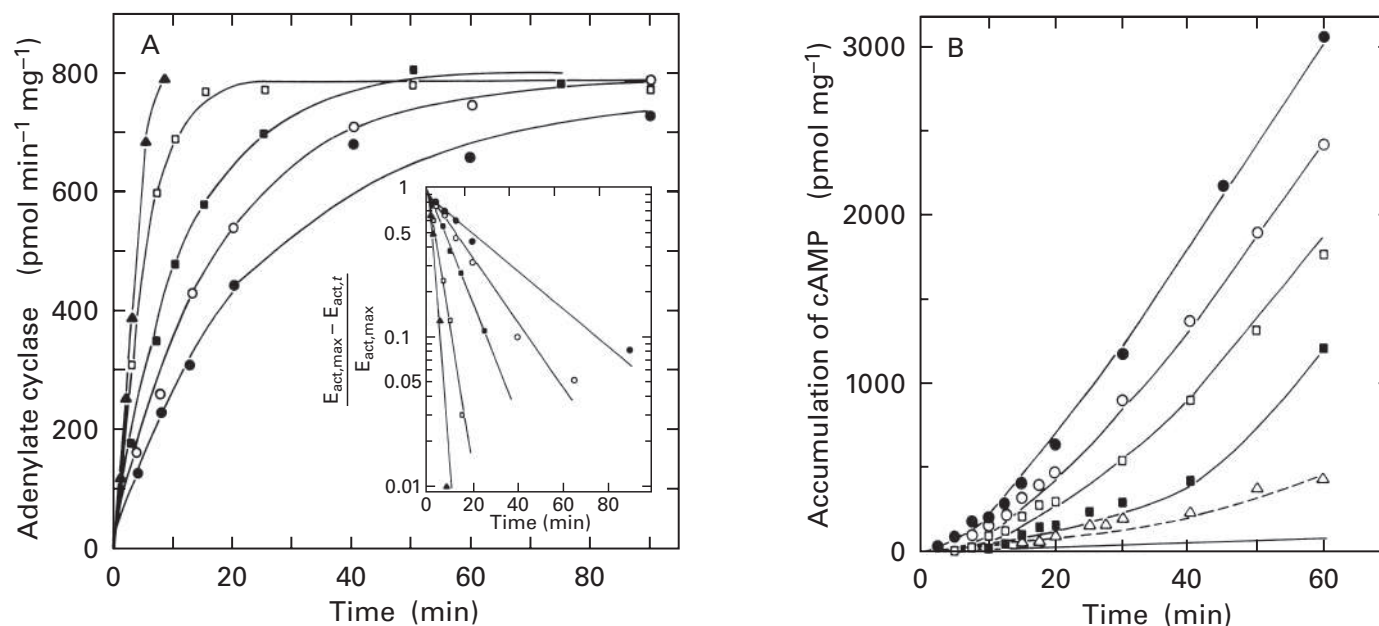


Figure 14-29: Rate of activation of adenylate cyclase as a function of the concentration of functional, occupied β -adrenergic receptor.⁷²⁸ (A) Samples of plasma membranes from erythrocytes of *Meleagris gallopavo* were exposed to various concentrations (\blacktriangle , none; \square , 17 μ M; \blacksquare , 43 μ M; \circ , 100 μ M; and \bullet , 230 μ M) of *N*-[2-hydroxy-3-(1-naphthoxy)propyl]-*N'*-(bromoacetyl)ethylenediamine, a specific, covalent, irreversible inhibitor of β -adrenergic receptor, to destroy the functional ability of various fractions of the receptor. Equivalent amounts of each sample (1.4 mg mL⁻¹) were mixed at 25 °C with saturating concentrations of the agonist epinephrine and guanosine β,γ -imidotriphosphate, an analogue of GTP that cannot be hydrolyzed by G-proteins. At various times, samples were removed, and the specific activity (picomoles of cyclic-3',5'-AMP minute⁻¹ milligram⁻¹) of adenylate cyclase was determined. As time progressed, the specific activity of the enzyme in the membranes increased monotonically until a maximum activity, $E_{act,max}$, was reached. The specific activity of adenylate cyclase is plotted as a function of time (minutes). Inset: Data in panel A are replotted on a semilogarithmic field to show that the approach to maximum activity is a first-order process. The amount of the competent adenylate cyclase that has not yet been activated at a given time t , $E_{inact,t}$, is equal to the amount of enzymatic activity at full activation, $E_{act,max}$, minus the adenylate cyclase activity observed at time t , $E_{act,t}$. The amount of adenylate cyclase activity that has not yet been activated is normalized by dividing by $E_{act,max}$. This normalized value is plotted as a function of time (minutes). (B) Equivalent samples of plasma membranes from erythrocytes of *M. gallopavo* (1.4 mg mL⁻¹) were mixed in several tubes with MgATP at 25 °C, and agonist-activated adenylate cyclase was initiated by adding guanosine β,γ -imidotriphosphate and various concentrations of the agonist epinephrine (\triangle , 0.5 μ M; \blacksquare , 1.0 μ M; \square , 3.0 μ M; \circ , 6 μ M; and \bullet , 15 μ M). At the indicated times, samples were removed from each tube and the total accumulations of cyclic-3',5'-AMP were assessed. The total accumulation of cyclic-3',5'-AMP [picomoles of cyclic-3',5'-AMP (milligram of protein)⁻¹] is plotted as a function of time (minutes). Panel B represents the integrated accumulation of cyclic-3',5'-AMP, and panel A shows the rate at which it is accumulating at any time. Reprinted with permission from ref 728. Copyright 1978 American Chemical Society.

produce a permanently unresponsive adenylate cyclase. Rather, it is the stimulatory G-protein that couples the binding of agonist on the β -adrenergic receptor to the activation of adenylate cyclase.

When the complete system for adenylate cyclase is depleted of G-protein by affinity adsorption, no **coupling between binding of an agonist and adenylate cyclase activity** can occur until it is added back.⁷²⁹ The membranes of Cyc-549 lymphoma cells contain both a β -adrenergic receptor and adenylate cyclase but lack a stimulatory G-protein and cannot display agonist-activated adenylate cyclase activity. When homogeneous stimulatory G-protein is added to these membranes, agonist-activated adenylate cyclase activity appears.⁷³⁰ Therefore, the G-protein must perform the coupling.

The complex between the β subunit and the γ subunit of the stimulatory G-protein binds tightly and

specifically to β -adrenergic receptor whether or not an agonist is bound.⁷³¹ The α subunit of the stimulatory G-protein also binds specifically and just as tightly to the β -adrenergic receptor whether or not it has guanosine 5'-O-(3-thiotriphosphate) bound to it. It forms an even tighter complex, however, with the β subunit and γ subunit of the stimulatory G-protein when it is unliganded with GTP or an analogue of GTP than when it is.⁷³¹ A complex between one β -adrenergic receptor and one G-protein has been identified in solutions derived from plasma membranes dissolved in nonionic detergent, but only when they are pretreated with β -adrenergic agonists.⁷³² In addition, the incorporation of purified homogeneous G-protein and purified homogeneous β -adrenergic receptor together into phospholipid vesicles causes the receptor to have a higher affinity for agonist, by a factor of more than 100, than it does in the absence of G-protein.⁷³³ This effect must result from the

formation of a **complex between β -adrenergic receptor and the respective G-protein** in these membranes, but it is eliminated by the addition of GTP, which binds to the G-protein.

The α subunit dissociates completely from the β subunit and γ subunit of stimulatory G-protein when GTP or a GTP analogue that cannot be hydrolyzed is bound to it.^{731,734,735} Consequently, the affinity of the α subunit of the stimulatory G-protein for the complex between β -adrenergic receptor and the β subunit and γ subunit of the stimulatory G-protein probably decreases after the β -adrenergic receptor has bound agonist and promoted the binding of GTP to the α subunit of the stimulatory G-protein.

From all of these observations, the **sequence of events at the β -adrenergic receptor** is thought to be the following.⁷³⁶ Before agonist binds there is a specific complex among β -adrenergic receptor, the β subunit and γ subunit of stimulatory G-protein,^{737,738} and the α subunit of stimulatory G-protein to which GDP is bound. Upon the binding of an agonist to β -adrenergic receptor, the dissociation of the GDP is stimulated as well as the association of GTP with the resulting empty α subunit of the stimulatory G-protein.^{739,740} The binding of GTP weakens the affinity of the α subunit of stimulatory G-protein for the remainder of this complex. This weakening increases the rate at which the α subunit exchanges between the complex and the bilayer of phospholipids in which it is in free solution. When it is dissociated from the complex in free solution within the bilayer of phospholipids, the complex between GTP and the α subunit is able to collide with a molecule of adenylate cyclase.

The α subunit of the stimulatory G-protein when it has guanosine 5'-O-(3-thiotriphosphate) bound to it forms a strong complex with the cytoplasmic domain of adenylate cyclase,⁷⁴¹ for which a crystallographic molecular model is available.⁷¹² The formation of this **complex between α subunit of the stimulatory G-protein and adenylate cyclase** stimulates the enzymatic activity of the cytoplasmic domain of adenylate cyclase by a factor of greater than 1000.⁷⁴¹ A tight complex between one intact molecule of adenylate cyclase and one molecule of G-protein has also been identified in solutions derived from plasma membranes dissolved with nonionic detergents.⁷⁴²

The system for adenylate cyclase in erythrocytes from *M. gallopavo* that is stimulated by β -adrenergic agonists takes several minutes to reach full enzymatic activity after a β -adrenergic agonist is added to a suspension of plasma membranes (Figure 14-29B).⁷²⁸ As the fraction of the β -adrenergic receptors occupied by agonist is increased, the duration of the lag preceding the expression of full enzymatic activity decreases (Figure 14-29B). The relationship between the rate at which full enzymatic activity is established, the final level of activity of the enzyme, and the fraction of the β -adrenergic

receptors occupied by agonist, which was determined by a direct measurement of bound agonist in separate experiments, is consistent⁷²⁸ with the **kinetic mechanism**



where $K_{d,A}$ is the dissociation constant for agonist, R is β -adrenergic receptor, A is agonist, E_{inact} is inactive adenylate cyclase, and E_{act} is active adenylate cyclase.

The meaning of the second step in the mechanism is that the rate at which adenylate cyclase is activated is directly proportional to the concentration of liganded β -adrenergic receptor, $A \cdot R$, and the concentration of inactive competent adenylate cyclase, E_{inact} , and consequently is first-order in each of these concentrations. When β -adrenergic receptor is saturated with agonist, all of the receptor is in the liganded form $A \cdot R$, and the reaction is governed solely by this second step (Equation 14-13). When the concentration of the complex between agonist and receptor at saturation $[A \cdot R]_{\text{sat}}$ is decreased systematically by decreasing the concentration of the competent β -adrenergic receptor by a specific covalent modification, the rate of production of active adenylate cyclase, E_{act} , containing the activated active site, decreases in direct proportion to the decrease in the concentration of occupied receptors at saturation (Figure 14-29A).⁷²⁸ Consequently, the rate of the reaction defined by the second step in the mechanism is **first-order in the concentration of occupied β -adrenergic receptor**. When β -adrenergic receptor is saturated with agonist, the rate of formation of active adenylate cyclase, E_{act} , is **first-order in the concentration of the unactivated adenylate cyclase**, E_{inact} (inset to Figure 14-29A).

The reaction governed by k_D (Equation 14-13) results from the **collision of two proteins** while they diffuse through the bilayer of phospholipids. When the microviscosity of the membrane, as judged by the residual polarization of a hydrophobic fluorescent molecule, is decreased by adding *cis*-vaccenic acid, the rate of production of active adenylate cyclase at saturating concentrations of agonist increases monotonically.⁷²⁸ When the viscosity of the plasma membrane is increased by removing some of its bulk phospholipid, the ability of the binding of agonist to β -adrenergic receptor to activate adenylate cyclase is significantly inhibited.⁷⁴³ It has also been observed that when the macroscopic viscosity of the membranes in the disks of retinal rods is decreased by halving the concentration of rhodopsin in them, the rate of phototransduction is accelerated 1.7-fold.⁷⁴⁴

Phototransduction results from the coupling of rhodopsin to cyclic GMP phosphodiesterase in a manner homologous to the coupling of β -adrenergic receptor to adenylate cyclase.

If, as is generally assumed, the two proteins that are colliding within the membrane to activate the enzymatic activity are adenylate cyclase and the complex between GTP and the α subunit of the stimulatory G-protein, it necessarily follows that the concentration of the complex between GTP and the α subunit of stimulatory G-protein in the bilayer of phospholipids must at all times be directly proportional to the concentration of liganded β -adrenergic receptor. To be so, the complex between GTP and the α subunit of stimulatory G-protein must be in constant communication with β -adrenergic receptor. Consequently, the α subunit and the complex between β -adrenergic receptor and the β subunit and γ subunit of stimulatory G-protein must be rapidly associating with and dissociating from each other in an equilibrium that maintains the required proportionality.^{728,745} The fact that one complex between agonist and β -adrenergic receptor is able to catalyze the exchange of GDP for GTP on more than 10 α subunits of stimulatory G-protein in the space of a few seconds^{736,737,746} suggests that the equilibration between these two proteins is much more rapid than the activation of adenylate cyclase (Figure 14-29). The fact that the α subunit of stimulatory G-protein is attached to the membrane only through its posttranslational lipid allows it to diffuse across the membrane more rapidly than if it were an integral membrane-bound protein, and this property increases the rate of its equilibration with β -adrenergic receptor.

A **membrane within a living cell** often differs in shape and extension from the same membrane purified from a homogenate of that cell. Small organelles such as mitochondria, chloroplasts, and lysosomes remain intact and are not visibly or functionally altered during gentle disruption of the cell and purification by centrifugation. The endoplasmic reticulum is constantly changing in its shape and contiguity even within the living cytoplasm, a property reflected in the fact that, upon homogenization, it readily disintegrates into small microsomes. The plasma membrane, however, in its natural state in an intact cell, is required to remain at all times a continuous enclosure surrounding the cytoplasm even while it must maintain a total surface area much greater than the membranes of any of the stable organelles it contains. When the cell is homogenized, the plasma membrane, as does the endoplasmic reticulum, also disintegrates into small microsomes, but much more reluctantly.

This reluctance seems to result from the fact that plasma membranes are skins stretched and pinned upon a frame. In bacteria and fungi, the frames are the outer membranes and cell walls on the extracytoplasmic surface, for when these integuments are digested away, a fragile, naked spheroplast remains that is easily disintegrated. In animal cells, however, there is often a frame on

the cytoplasmic side of the plasma membrane upon which the membrane is stretched and to which it is pinned. A limited number of molecules of integral membrane-bound proteins and lipids function as the pins. These pins connect the continuous bilayer of the plasma membrane to the frame at random points scattered over its surface but do not noticeably affect the physical properties of the bilayer of phospholipid. The stability provided by any one of these supports allows a plasma membrane to remain unbroken over its entire surface area even though it is a thin, fragile fluid film that disintegrates when it is removed from the frame.

The **cytoskeleton** is the frame upon which the plasma membrane of an animal cell is stretched and pinned. Although most of the proteins are free to diffuse, the proteins pinning the plasma membrane to the cytoskeleton do not. In the short run, the membrane is fixed at these points of attachment but fluid everywhere else. When the cytoskeleton collapses and, as a result, the pins cluster rather than remaining spread out, the unsupported plasma membrane in the abandoned regions slowly fragments into microsomes.⁷⁴⁷

The cytoskeleton of an **erythrocyte** is the best characterized. The proteins constituting the cytoskeleton of an erythrocyte are, however, found in most of the cells from other tissues and are thought to perform the same role in these other types of cells. When erythrocytes are added to a solution of nonionic detergent, the plasma membrane dissolves and leaves behind its cytoskeleton that has the shape of an erythrocyte but is a basket rather than a bag.⁷⁴⁸ The cytoskeleton exposed by this treatment contains mainly spectrin, actin,⁷⁴⁹ and protein 4.1.

Spectrin⁷⁵⁰ was originally identified by Rosenthal, Kregenow, and Moses⁷⁵¹ as a protein composing a fuzzy network on the cytoplasmic side of the plasma membrane of an erythrocyte. It was isolated by extraction of plasma membranes from erythrocytes at low ionic strength in the presence of a chelating agent for multivalent cations.⁷⁵¹ The protein from human erythrocytes is composed of an α polypeptide (2418 aa) and a β polypeptide (2136 aa)⁷⁵² and is an $\alpha\beta$ heterodimer or $(\alpha\beta)_2$ heterotetramer depending upon the conditions.⁷⁵³ The purified heterodimer has a high intrinsic viscosity ($[\eta] = 140 \text{ cm}^3 \text{ g}^{-1}$),⁷⁵⁴ indicating that it is elongated. The two polypeptides of the $\alpha\beta$ heterodimer are homologous in sequence, and each contains internally repeating domains (Figure 7-16).⁷⁵⁵ In electron micrographs, the $\alpha\beta$ heterodimer appears as a flexible two-stranded segment of rope about 100 nm long (Figure 14-30B).⁷⁵⁶ The two strands are not held together along their entire length and tend to splay. An $(\alpha\beta)_2$ heterotetramer is formed from two dimers associating end to end (Figure 14-30A).

The **actin** in the cytoskeleton is in the form of short thin filaments (Figure 9-1B) with a uniform length of 12-14 monomers.⁷⁵⁷⁻⁷⁵⁹ The protein dematin⁷⁶⁰ is associated with these actin filaments and controls the associations between them.⁷⁶¹

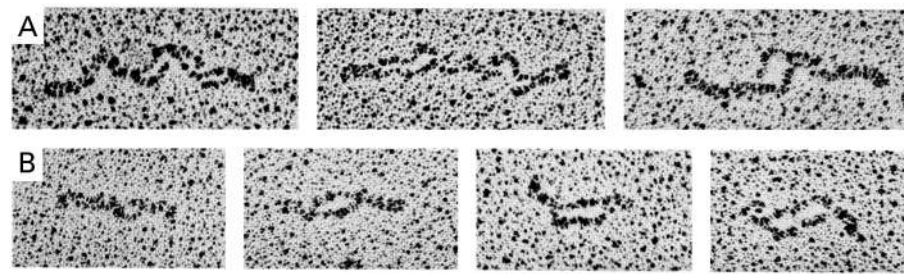


Figure 14-30: Electron micrographs of individual molecules of human spectrin as the $(\alpha\beta)_2$ heterotetramer (A) or $\alpha\beta$ heterodimer (B).⁷⁵⁶ Human erythrocytes were washed and lysed, and the resulting plasma membranes were washed to remove the hemoglobin. The purified plasma membranes were extracted with 0.1 mM EDTA, pH 8, at 0 °C for 40 h, and the membranes were then removed by centrifugation. The released spectrin was purified from the extract by molecular exclusion chromatography. Solutions containing the spectrin were brought to 70% glycerol and sprayed onto freshly cleaved mica. The surface of the mica was then sprayed at an angle of 9° to the surface with a mixture of platinum and carbon vaporized by electric discharge. The spray was applied while the sample was rapidly rotating about an axis normal to the surface. This causes the molecules to be surrounded by drifts of platinum, which is electron-dense. These drifts produce an outline of the molecules of protein. The film of platinum and carbon was then transferred from the mica to a grid for electron microscopy. The molecules are represented by the tortuous, elongated outlines. Magnification 170000×. Reprinted with permission from ref 756. Copyright 1979 Academic Press.

Human **protein 4.1** is a monomer with a polypeptide 864 aa in length. It is a globular protein that can bind to an $\alpha\beta$ heterodimer of spectrin near the other end of the rope from that which combines with another $\alpha\beta$ heterodimer to form the $(\alpha\beta)_2$ heterotetramer.^{762,763} When purified actin, protein 4.1, and spectrin are mixed together with ATP, they spontaneously form a macroscopic gel made up of heterotetramers of spectrin cross-linked by the short filaments of actin.⁷⁶⁴ This gel presumably is analogous to the meshwork seen in the cytoskeleton. Protein 4.1 promotes this association of spectrin and actin⁷⁶⁴⁻⁷⁶⁶ by binding simultaneously to a molecule of spectrin and a short filament of actin and linking them together.⁷⁶⁷

One set of the pins attaching the plasma membrane of an erythrocyte to the cytoskeleton is formed from two proteins, ankyrin and band 3 anion transport protein. Human erythrocytic ankyrin is a monomer constructed from one single polypeptide about 1880 aa in length.^{762,768} **Ankyrin** has a fairly high frictional ratio ($f/f_0 = 1.46$),⁷⁶⁸ and in electron micrographs it appears as a cluster of three to five globular domains.⁷⁶² In keeping with this structure, it has a detachable domain ($n_{aa} = 650$) that contains the binding site specific for spectrin.⁷⁶⁹ Ankyrin binds tightly to spectrin near the end of the rope that associates to form the $(\alpha\beta)_2$ heterotetramer, the opposite end from that to which protein 4.1 binds.⁷⁶² Intact ankyrin can also bind to the amino-terminal, detachable domain of band 3 anion transport protein in a simple bimolecular reaction ($K_d = 10^{-8}$ M) as well as to intact band 3 anion transport protein in solutions of nonionic detergents.⁷⁷⁰ A freely soluble heterodimer containing one polypeptide of ankyrin and one polypeptide from band 3 anion transport protein can be purified in solutions of nonionic detergent.⁷⁷¹

Because intact ankyrin binds tightly to both **band 3**

anion transport protein, which is an integral membrane-bound protein, and spectrin, which is incorporated into the cytoskeleton, it can link the cytoskeleton to the plasma membrane. As there are fewer molecules of ankyrin in an erythrocyte than molecules of band 3 anion transport protein, only a minority of the molecules of band 3 anion transport protein, presumably chosen at random, are linked to the cytoskeleton. Unlike the unattached molecules of band 3 anion transport protein, those that are attached to ankyrin and pin the cytoskeleton to the membrane are unable to diffuse translationally⁷⁷² or rotationally⁷⁶⁷ because they are attached rigidly to the cytoskeleton. Membrane protein band 4.2 stabilizes this interaction between ankyrin and band 3 anion transport protein.⁷⁷³

At the other end of an $\alpha\beta$ heterodimer of spectrin, the protein 4.1 that is creating the interaction of spectrin and actin is also linking the cytoskeleton to the membrane. Protein 4.1 binds strongly to phosphatidylserine on the inner surface of the bilayer of phospholipids⁷⁷⁴ and also to the cytoplasmic portion of glycophorin C⁷⁷⁵ and glycophorin A.⁷⁷⁶

In the short run, the molecules of band 3 anion transport protein, phosphatidylserine, glycophorin A, and glycophorin C are the stationary points around which flow the traffic of the proteins and lipids of the plasma membrane. In the long run, as the cell changes shape and size, these points of attachment also rearrange fluidly to accommodate the changes.

The pinning of the plasma membrane onto the cytoskeleton and the sculpting of the various membranes into the cellular organelles, like the assembly of microtubules, filaments of actin, and thick filaments or the mixing of a vast array of soluble proteins at high concentration to produce cytoplasm, seamlessly transforms protein chemistry into cell biology.

Suggested Reading

Peters, R., & Cherry, R.J. (1982) Lateral and rotational diffusion of bacteriorhodopsin in lipid bilayers: Experimental test of the Saffman–Delbrück equations, *Proc. Natl. Acad. Sci. U.S.A.* 79, 4317–4321.

Tolkovsky, A.M., & Levitski, A. (1978) Mode of coupling between the β -adrenergic receptor and adenylate cyclase in turkey erythrocytes, *Biochemistry* 17, 3795–3810.

Problem 14–7: Cytochrome b_5 is a protein that is firmly attached to the membranes of the endoplasmic reticulum. The amino acid sequence of the protein from *Rattus norvegicus* is

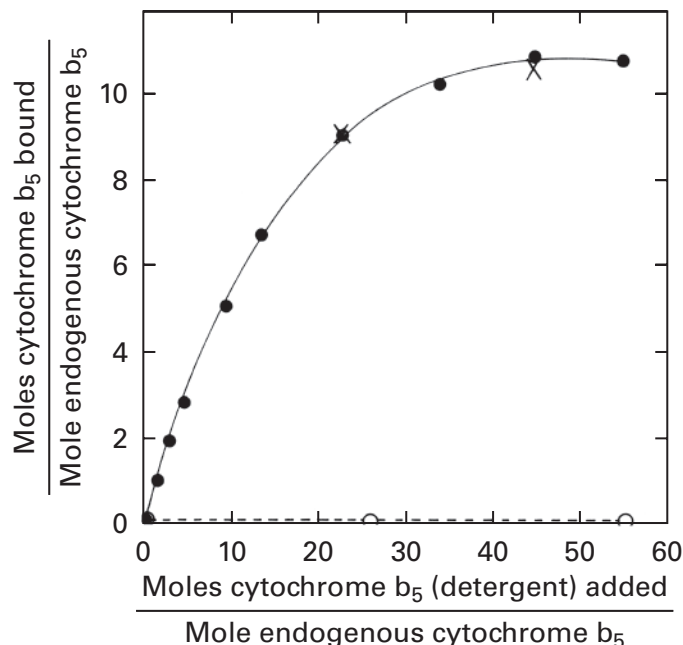
```
AEQSDKDVKYYTLEEIQKHKDSKSTWVILHHKVYDLTKFL
EEHPGGEEVLREQAGDATENFEDVGHSTDARELSKTYII
GELHPDDRSKIAKPSETLITTVESNSSWWTNWNVIPAISAL
VVALMYRLYMAED
```

When endoplasmic reticulum is treated with trypsin or pancreatic triacylglycerol lipase contaminated with trypsin, only the peptide bond following Lysine 90 is cleaved, and a protein containing the first 90 amino acids of cytochrome b_5 falls off the membrane. This soluble protein can be purified and crystallized, and its structure has been determined by X-ray crystallography. This protein will be referred to as cytochrome b_5 (trypsin) or cytochrome b_5 (lipase), respectively.

- Explain these observations in terms of the distribution of specific amino acids in the sequence of the protein.
- Draw a representation of the complete cytochrome b_5 molecule attached to the membrane.

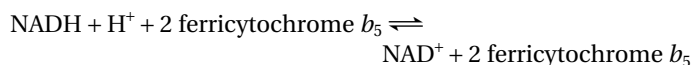
It is possible to release, by the use of a detergent, intact cytochrome b_5 from microsomes of endoplasmic reticulum and to purify this protein. This protein will be referred to as cytochrome b_5 (detergent).

- When various amounts of cytochrome b_5 (detergent) or cytochrome b_5 (lipase) are mixed with endoplasmic reticulum membranes and incubated for 18 h at 2 °C, and the membranes are then washed extensively, the detergent form attaches to the membranes while the lipase form does not (see the following figure). Explain this observation.

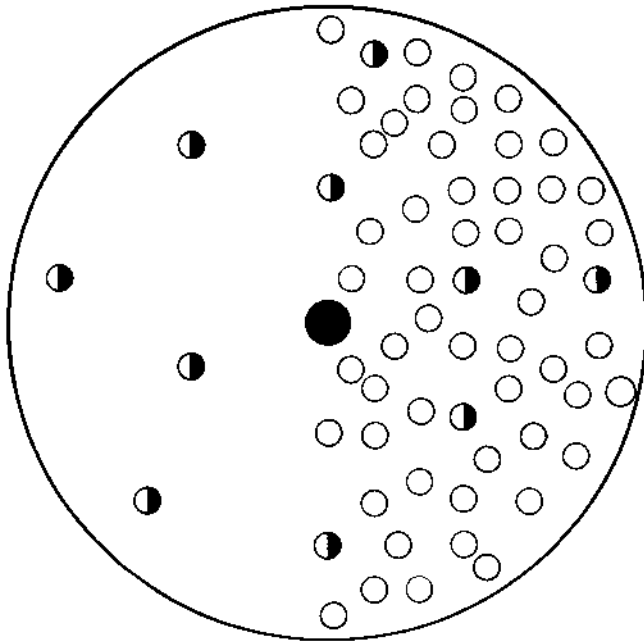


Binding of cytochrome b_5 (detergent) (●) or cytochrome b_5 (lipase) (○) to microsomes of endoplasmic reticulum from liver of *O. cuniculus*. The results are expressed as the moles of exogenous cytochrome b_5 bound to the microsomes for every mole of endogenous cytochrome b_5 in the microsomes as a function of the moles of exogenous cytochrome b_5 added for every mole of endogenous cytochrome b_5 present. The points marked with × were samples that were washed further in 0.5 M NaCl, pH 8.0, to remove any loosely bound cytochrome b_5 . Reprinted with permission from ref 275. Copyright 1972 *Journal of Biological Chemistry*.

Cytochrome- b_5 reductase is also attached to the endoplasmic reticulum. It catalyzes the following reaction:



The native endoplasmic reticulum membrane contains one molecule of cytochrome- b_5 reductase and 10 molecules of cytochrome b_5 for every $2.5 \times 10^3 \text{ nm}^2$ (see the following figure).



Schematic representation of the spatial relationships between phospholipid, cytochrome b_5 , and cytochrome- b_5 reductase on the surface of a microsomal vesicle. The surface areas used for these components were phospholipid, 0.63 nm^2 ; cytochrome b_5 , 4 nm^2 ; and cytochrome- b_5 reductase, 10 nm^2 . The entire area (diameter = 56 nm) would include 4000 molecules of phospholipid. The endogenous concentrations of the two proteins would place approximately 10 molecules of cytochrome b_5 (●) and approximately 1 molecule of reductase (●) in this area, assuming a random distribution on the outer surface of the membrane. The left sector indicates the molecular density with only endogenous cytochrome b_5 , and the right sector, the density with a 10-fold molar excess of cytochrome b_5 (detergent) (○). Reprinted with permission from ref 275. Copyright 1972 *Journal of Biological Chemistry*.

When all of these cytochrome b_5 molecules are in the oxidized form and the reaction is initiated by addition of reducing equivalents, the time required for the reduction of half of all the cytochrome b_5 molecules in the two types of membranes by the reductase is 0.47 s for the unenriched membranes and 0.13 s for the enriched membranes.

- (D) What ability must cytochrome b_5 possess in order to participate as a reactant in this reaction?
- (E) Why is the $t_{1/2}$ shorter in the case of the enriched membranes?
- (F) The concentration of cytochrome b_5 in the solution during the reduction experiments with the enriched membranes just described was about $2 \times 10^{-6} \text{ M}$, and the rate of the reaction catalyzed by the reductase was independent of the concentration of membranes suspended in the solution. In order to observe the same turnover rate, however, when cytochrome b_5 (trypsin) was the substrate, its concentration had to be $5 \times 10^{-5} \text{ M}$ and the rate

of the reaction catalyzed by the enzyme depended on the concentration of cytochrome b_5 (trypsin) in the solution. Explain these observations.

Problem 14-8: Calculate the translational diffusion coefficients at 20°C for integral membrane-bound proteins the apparent radii, α , of whose bundles of α helices are 1.0 , 2.0 , 4.0 , and 5.0 nm . Assume that the viscosity of the bilayer of phospholipid is 100 mPa s and the width of the bilayer of phospholipid is 5 nm .

Problem 14-9: The vertebrate rod is a cell in the retina responsible for registering light rays in the visual process. The end of the cell that performs this task is called the outer segment. It is a cylinder filled with disks, which are flattened circular, closed sacs pinched off from the plasma membrane.

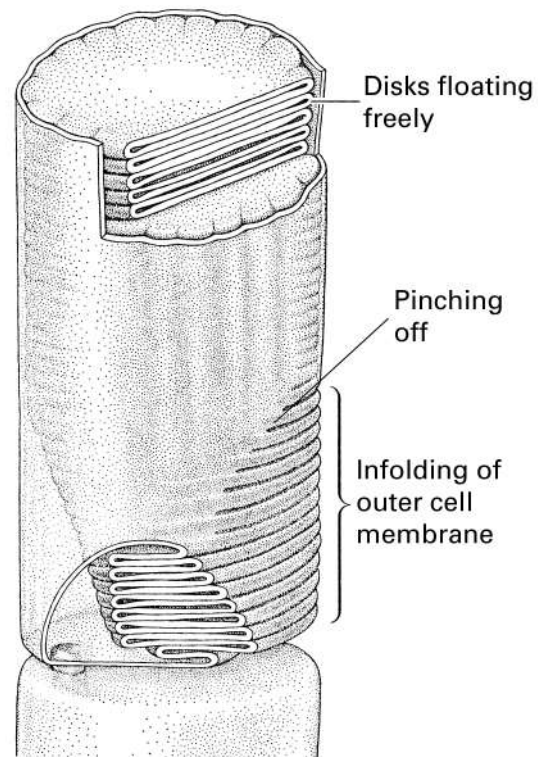


Diagram showing frog rod outer segment. The lamellar membranous structure of the rod outer segments consists of a stack of sacs, except near the base, where it consists of infoldings of the cell plasma membrane that are being pinched off. Reprinted with permission from ref 777. Copyright 1970 Scientific American Inc.

The disks are stacked in the rod as poker chips are stacked in a rack. In this way, greater than 90% of the membrane from which the disk is made lies normal to the cylindrical axis of the rod outer segment.

Rhodopsin (Table 14-6) is the only protein dissolved in the disk membrane.

- (A) In an outline of the cross section of a disk draw a schematic diagram of the molecular structure of the disk membrane. Include rhodopsin molecules, labeled with an arrow indicating direction of insertion, and phospholipids.

Rhodopsin is a protein to which is attached, by an imine linkage, one molecule of 11-*cis*-retinal. When 11-*cis*-retinal is exposed to intense light it isomerizes to all-*trans*-retinal. This process is known as bleaching and results in a color change from orange to clear.

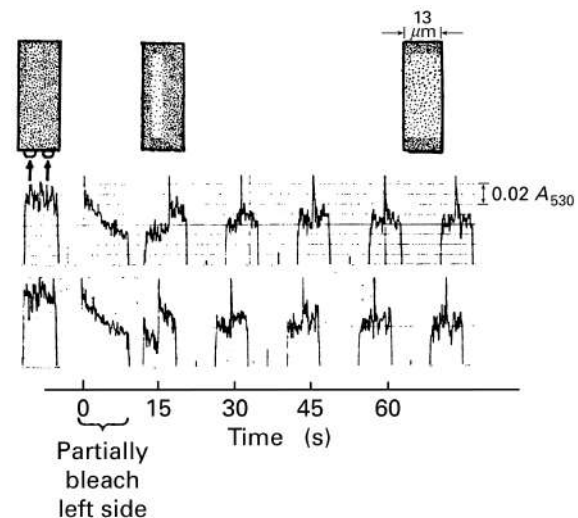
- (B) What does the following experiment demonstrate about the properties of a protein such as rhodopsin in a biological membrane? Reprinted with permission from *Nature*, ref 675. Copyright 1974 Macmillan Magazines Limited.

Rod outer segments were obtained by gently shaking retinas dissected under dim red light from the eyes of frog (*Rana catesbeiana*) and mudpuppy (*Necturus maculosus*) which had been dark-adapted for more than 10 h. The rods were shaken into a microchamber containing a standard Ringer solution and examined in a Shimadzu 50L microspectrophotometer (MSP) fitted with a high quantum efficiency photomultiplier (Hamamatsu type R375). Single rods which appeared intact and which lay flat on the bottom of the chamber were selected for observation, and all observations were completed within 30 min after the rods were isolated. The rhodopsin in isolated rods, once bleached, does not regenerate, hence dim red light was used for selection, focusing, and alignment.

The measuring beam of the MSP was limited by an aperture and a condensing lens to form a rectangle about $2 \times 20 \mu\text{m}$ in cross section. The long axis of the rectangle was aligned with the long axis of the rod and a simple motor-driven "alternator" optically shifted the rectangular measuring beam back and forth between the two sides of the rod. Thus the absorbance of the rhodopsin on each side of the rod could be compared directly. The wavelength of the measuring beam was set at the absorption peak of the visual pigment: 500 nm for frog, 530 nm for mudpuppy.

With suitable alignment and focusing the absorbance was essentially equal on both sides of the unbleached rods, as shown by the first pair of measurements on an unbleached rod at the beginning of each of the two recordings in the figure below. The alternator was then stopped momentarily and the intensity of the measuring beam was increased about 1000-fold to bleach some pigment on one side of the rod.

The exponential decrease in absorbance during the bleach was recorded, and then the intensity was dropped to the original level and the alternator turned on again. The figure below shows that immediately after the bleach the absorbance on the unbleached side was little changed, but there was a marked drop in absorbance on the bleached side. Within the next few seconds, however, the absorbance of the unbleached side decreased while that on the bleached side increased, and within less than 1 min the absorbance of the two sides became equal, reaching a final level midway between that of each side immediately after the bleach.



The diagrams of a rod depict the pigment distribution corresponding in time with the absorbance measurements shown below. The arrows indicate the location on the rod at which each absorbance measurement was made. Recordings made from two different rods are shown to give an indication of the repeatability of the measurements. In each experiment the chart recorder was run continuously, as shown by the time base. The alternator also ran continuously except during the bleach. The records thus consist of a repeated pattern in which absorbance measurements were made first on the left side, then the right side of the rod. Between each pair of measurements baseline measurements were also made to ensure that no drifts occurred (for clarity, these were omitted from the figure). The spikes on the traces were caused by switching transients in the alternator. The diameter of a disk is essentially equal to the width of the rod, and the width is measured in the MSP after completing each experiment.

References

1. Claude, A. (1941) *Cold Spring Harbor Symp. Quant. Biol.* 9, 263–271.
2. Hogeboom, G.H., Schneider, W.C., & Pallade, G.E. (1948) *J. Biol. Chem.* 172, 619–636.
3. Fleischer, S., & Kervina, M. (1974) *Methods Enzymol.* 31, 6–41.
4. Leighton, F., Poole, B., Beaufay, H., Baudhuin, P., Coffey, J.W., Fowler, S., & De Duve, C. (1968) *J. Cell Biol.* 37, 482–513.
5. Palade, G.E., & Siekevitz, P. (1956) *J. Cell Biol.* 2, 171–200.
6. Price, C.A. (1974) *Methods Enzymol.* 31, 501–519.
7. Schneider, W.C., Hogeboom, G.H., & Striebach, M.J. (1953) *Cancer Res.* 13, 617–663.
8. McKeel, D.W., & Jarett, L. (1970) *J. Cell Biol.* 44, 417–432.
9. Heidrich, H.G., Kinne, R., Kinne-Saffran, E., & Hannig, K. (1972) *J. Cell Biol.* 54, 232–245.
10. Booth, A.G., & Kenny, A.J. (1974) *Biochem. J.* 142, 575–581.
11. Lin, P.H., Selinfreund, R., Wakshull, E., & Wharton, W. (1987) *Biochemistry* 26, 731–736.
12. DePierre, J.W., & Karnovsky, M.L. (1973) *J. Cell Biol.* 56, 275–303.
13. Fleischer, B., & Zambrano, F. (1974) *J. Biol. Chem.* 249, 5995–6003.
14. Chang, K.J., Bennett, V., & Cuatrecasas, P. (1975) *J. Biol. Chem.* 250, 488–500.
15. Dallner, G. (1963) *Acta Pathol. Microbiol. Scand. Suppl.* 166, 1–94.
16. Nobel, P.S. (1974) *Methods Enzymol.* 31, 600–606.
17. Neville, D.M. (1960) *J. Biophys. Biochem. Cytol.* 8, 413–422.
18. Morgan, I.G., Wolfe, L.S., Mandel, P., & Gombos, G. (1971) *Biochim. Biophys. Acta* 241, 737–751.
19. Weinstein, D.B., Marsh, J.B., Glick, M.C., & Warren, L. (1969) *J. Biol. Chem.* 244, 4103–4111.
20. Atkinson, P.H., & Summers, D.F. (1971) *J. Biol. Chem.* 246, 5162–5175.
21. Wiley, W.R. (1974) *Methods Enzymol.* 31, 609–626.
22. Kaback, H.R. (1968) *J. Biol. Chem.* 243, 3711–3724.
23. Tanford, C. (1973) *The Hydrophobic Effect: Formation of Micelles and Biological Membranes*, Wiley, New York.
24. Farquhar, J.W. (1962) *J. Lipid Res.* 3, 21–30.
25. Zhang, D.L., Daniels, L., & Poulter, C.D. (1990) *J. Am. Chem. Soc.* 112, 1264–1265.
26. Komatsu, H., & Chong, P.L. (1998) *Biochemistry* 37, 107–115.
27. Nishihara, M., Morii, H., & Koga, Y. (1989) *Biochemistry* 28, 95–102.
28. Nishihara, M., Utagawa, M., Akutsu, H., & Koga, Y. (1992) *J. Biol. Chem.* 267, 12432–12435.
29. Orlandi, P.A., Jr., & Turco, S.J. (1987) *J. Biol. Chem.* 262, 10384–10391.
30. Turco, S.J., Hull, S.R., Orlandi, P.A., Jr., Shepherd, S.D., Homans, S.W., Dwek, R.A., & Rademacher, T.W. (1987) *Biochemistry* 26, 6233–6238.
31. Makaanu, C.K., Damian, R.T., Smith, D.F., & Cummings, R.D. (1992) *J. Biol. Chem.* 267, 2251–2257.
32. Rouser, G., Nelson, G.J., Fleischer, S., & Simon, G. (1968) in *Biological Membranes, Physical Fact and Function* (Chapman, D., Ed.) pp 5–69, Academic Press, London.
33. Erwin, J.A. (1973) *Lipids and Biomembranes of Eukaryotic Microorganisms*, Academic Press, New York.
34. Fiehn, W., Peter, J.B., Mead, J.F., & Gan-Elepano, M. (1971) *J. Biol. Chem.* 246, 5617–5620.
35. Kaback, H.R. (1971) *Methods Enzymol.* 22, 99–120.
36. Cronan, J.E., Jr., & Vagelos, P.R. (1972) *Biochim. Biophys. Acta* 265, 25–60.
37. Shaw, N., Smith, P.F., & Koostra, W.L. (1968) *Biochem. J.* 107, 329–333.
38. Qureshi, N., Honovich, J.P., Hara, H., Cotter, R.J., & Takayama, K. (1988) *J. Biol. Chem.* 263, 5502–5504.
39. Kamio, Y., & Nikaido, H. (1976) *Biochemistry* 15, 2561–2570.
40. Bangham, A.D., Standish, M.M., & Watkins, J.C. (1965) *J. Mol. Biol.* 13, 238–252.
41. Huang, C. (1969) *Biochemistry* 8, 344–352.
42. Hope, M.J., Bally, M.B., Webb, G., & Cullis, P.R. (1985) *Biochim. Biophys. Acta* 812, 55–65.
43. Moscho, A., Orwar, O., Chiu, D.T., Modi, B.P., & Zare, R.N. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 11443–11447.
44. Mueller, P., Rudin, D.O., Tien, H.T., & Wescott, W.C. (1962) *Nature* 194, 979–980.
45. Montal, M., & Mueller, P. (1972) *Proc. Natl. Acad. Sci. U.S.A.* 69, 3561–3566.
46. Levine, Y.K., & Wilkins, M.H. (1971) *Nat. New Biol.* 230, 69–72.
47. Wiener, M.C., & White, S.H. (1992) *Biophys. J.* 61, 434–447.
48. McIntosh, T.J., Magid, A.D., & Simon, S.A. (1987) *Biochemistry* 26, 7325–7332.
49. McIntosh, T.J., Magid, A.D., & Simon, S.A. (1989) *Biochemistry* 28, 17–25.
50. Reiss-Husson, F. (1967) *J. Mol. Biol.* 25, 363–382.
51. Small, D.M. (1967) *J. Lipid Res.* 8, 551–557.
52. Pearson, R.H., & Pascher, I. (1979) *Nature* 281, 499–501.
53. Hitchcock, P.B., Mason, R., Thomas, K.M., & Shipley, G.G. (1974) *Proc. Natl. Acad. Sci. U.S.A.* 71, 3036–3040.
54. Hauser, H., Pascher, I., & Sundell, S. (1988) *Biochemistry* 27, 9166–9174.
55. Thuren, T., Virtanen, J.A., & Kinnunen, P.K. (1987) *Biochemistry* 26, 5816–5819.
56. Hanai, T., Haydon, D.A., & Taylor, J. (1965) *J. Theor. Biol.* 9, 278–296.
57. McLaughlin, S. (1977) in *Current Topics in Membranes and Transport* (Bronner, F., & Kleinzeller, A., Eds.) Vol. 9, pp 71–144, Academic Press, New York.
58. Winiski, A.P., Eisenberg, M., Langner, M., & McLaughlin, S. (1988) *Biochemistry* 27, 386–392.
59. Thorgeirsson, T.E., Yu, Y.G., & Shin, Y.K. (1995) *Biochemistry* 34, 5518–5522.
60. Ben-Tal, N., Honig, B., Peitzsch, R.M., Denisov, G., & McLaughlin, S. (1996) *Biophys. J.* 71, 561–575.
61. Khouri, O., Sherrill, C., & Roise, D. (1996) *Biochemistry* 35, 14553–14560.
62. Ben-Tal, N., Honig, B., Miller, C., & McLaughlin, S. (1997) *Biophys. J.* 73, 1717–1727.

63. Gershfeld, N.L. (1989) *Biochemistry* 28, 4229–4232.
64. DeChavigny, A., Heacock, P.N., & Dowhan, W. (1991) *J. Biol. Chem.* 266, 5323–5332.
65. McElhaney, R.N., & Tourtellotte, M.E. (1969) *Science* 164, 433–434.
66. Engelman, D.M. (1971) *J. Mol. Biol.* 58, 153–165.
67. Pascher, I., Sundell, S., & Hauser, H. (1981) *J. Mol. Biol.* 153, 791–806.
68. Chapman, D., Williams, R.M., & Ladbroke, B.D. (1967) *Chem. Phys. Lipids* 1, 445–475.
69. Buldt, G., Gally, H.U., Seelig, A., Seelig, J., & Zaccari, G. (1978) *Nature* 271, 182–184.
70. Mattai, J., Sripada, P.K., & Shipley, G.G. (1987) *Biochemistry* 26, 3287–3297.
71. Mabrey, S., Mateo, P.L., & Sturtevant, J.M. (1978) *Biochemistry* 17, 2464–2468.
72. Shimshick, E.J., & McConnell, H.M. (1973) *Biochemistry* 12, 2351–2360.
73. Oldfield, E., & Chapman, D. (1972) *FEBS Lett.* 23, 285–297.
74. Chapman, D., & Urbina, J. (1974) *J. Biol. Chem.* 249, 2512–2521.
75. Keana, J.F.W., Keana, S.B., & Beetham, D. (1967) *J. Am. Chem. Soc.* 89, 3055–3056.
76. McConnell, H.M., & McFarland, B.G. (1970) *Q. Rev. Biophys.* 3, 91–136.
77. Stone, T.J., Buckman, T., Nordio, P.L., & McConnell, H.M. (1965) *Proc. Natl. Acad. Sci. U.S.A.* 54, 1010–1017.
78. Libertini, L.J., Waggoner, A.S., Jost, P.C., & Griffith, O.H. (1969) *Proc. Natl. Acad. Sci. U.S.A.* 64, 13–19.
79. Hubbell, W.L., & McConnell, H.M. (1971) *J. Am. Chem. Soc.* 93, 314–326.
80. McFarland, B.G., & McConnell, H.M. (1971) *Proc. Natl. Acad. Sci. U.S.A.* 68, 1274–1278.
81. Chattopadhyay, A., & London, E. (1987) *Biochemistry* 26, 39–45.
82. Seelig, A., & Seelig, J. (1974) *Biochemistry* 13, 4839–4845.
83. Seelig, J., & Niederberger, W. (1974) *J. Am. Chem. Soc.* 96, 2069–2072.
84. Seelig, J., & Browning, J.L. (1978) *FEBS Lett.* 92, 41–44.
85. Tardieu, A., Luzzati, V., & Reman, F.C. (1973) *J. Mol. Biol.* 75, 711–733.
86. Griffith, O.H., Dehlinger, P.J., & Van, S.P. (1974) *J. Membr. Biol.* 15, 159–192.
87. Sheetz, M.P., & Chan, S.I. (1972) *Biochemistry* 11, 4573–4581.
88. Dill, K.A., & Flory, P.J. (1981) *Proc. Natl. Acad. Sci. U.S.A.* 78, 676–680.
89. Papahadjopoulos, D., Vail, W.J., Jacobson, K., & Poste, G. (1975) *Biochim. Biophys. Acta* 394, 483–491.
90. White, S.H. (1977) *Ann. N.Y. Acad. Sci.* 303, 243–265.
91. Mendelsohn, R. (1972) *Biochim. Biophys. Acta* 290, 15–21.
92. Lenard, J., & Rothman, J.E. (1976) *Proc. Natl. Acad. Sci. U.S.A.* 73, 391–395.
93. Kroon, P.A., Kainosho, M., & Chan, S.I. (1975) *Nature* 256, 582–584.
94. Rothman, J.E., & Engelman, D.M. (1972) *Nat. New Biol.* 237, 42–44.
95. Weisz, K., Grobner, G., Mayer, C., Stohrer, J., & Kothe, G. (1992) *Biochemistry* 31, 1100–1112.
96. Straume, M., & Litman, B.J. (1987) *Biochemistry* 26, 5121–5126.
97. Oldfield, E., & Chapman, D. (1971) *Biochem. Biophys. Res. Commun.* 43, 610–616.
98. Lecuyer, H., & Dervichian, D.G. (1969) *J. Mol. Biol.* 45, 39–57.
99. Engelman, D.M., & Rothman, J.E. (1972) *J. Biol. Chem.* 247, 3694–3697.
100. Vist, M.R., & Davis, J.H. (1990) *Biochemistry* 29, 451–464.
101. DeKrieff, B., Demel, R.A., Slotboom, A.J., VanDeenan, L.L.M., & Rosenthal, A.F. (1973) *Biochim. Biophys. Acta* 307, 1–19.
102. Radhakrishnan, A., & McConnell, H.M. (2000) *Biochemistry* 39, 8119–8124.
103. Almeida, P.F., Vaz, W.L., & Thompson, T.E. (1992) *Biochemistry* 31, 6739–6747.
104. Hyslop, P.A., Morel, B., & Sauerheber, R.D. (1990) *Biochemistry* 29, 1025–1038.
105. Bangham, A.D., & Dawson, R.M.C. (1960) *Biochem. J.* 75, 133–138.
106. Fettiplace, R., Andrews, D.M., & Haydon, D.A. (1971) *J. Membr. Biol.* 5, 277–296.
107. Haydon, D.A., & Taylor, F.H. (1960) *Philos. Trans. R. Soc. London, A* 252, 225–248.
108. Haydon, D.A., & Taylor, J. (1963) *J. Theor. Biol.* 4, 281–296.
109. Tanford, C. (1974) *Proc. Natl. Acad. Sci. U.S.A.* 71, 1811–1815.
110. Rowe, E.S., & Cutrera, T.A. (1990) *Biochemistry* 29, 10398–10404.
111. Simon, S.A., & McIntosh, T.J. (1984) *Biochim. Biophys. Acta* 773, 169–172.
112. Dill, K.A., Koppel, D.E., Cantor, R.S., Dill, J.D., Bendedouch, D., & Chen, S.H. (1984) *Nature* 309, 42–45.
113. Cullis, P.R., & de Kruijff, B. (1979) *Biochim. Biophys. Acta* 559, 399–420.
114. Luzzati, V., & Husson, F. (1962) *J. Cell Biol.* 12, 207–219.
115. Stoeckenius, W. (1962) *J. Cell Biol.* 12, 221–229.
116. Lewis, R.N., Mannock, D.A., McElhaney, R.N., Turner, D.C., & Gruner, S.M. (1989) *Biochemistry* 28, 541–548.
117. Singer, S.J., & Nicolson, G.L. (1972) *Science* 175, 720–731.
118. Rosenberg, S.A., & Guidotti, G. (1969) *J. Biol. Chem.* 244, 5118–5124.
119. Steck, T.L., Weinstein, R.S., Straus, J.H., & Wallach, D.F. (1970) *Science* 168, 255–257.
120. Steck, T.L., & Yu, J. (1973) *J. Supramol. Struct.* 1, 220–232.
121. Tilney, L.G., & Detmers, P. (1975) *J. Cell Biol.* 66, 508–520.
122. Lu, Y., Bazzi, M.D., & Nelsestuen, G.L. (1995) *Biochemistry* 34, 10777–10785.
123. Bazzi, M.D., & Nelsestuen, G.L. (1987) *Biochemistry* 26, 115–122.
124. Mosior, M., & Epand, R.M. (1993) *Biochemistry* 32, 66–75.
125. Newton, A.C., & Keranen, L.M. (1994) *Biochemistry* 33, 6651–6658.
126. Johnson, J.E., Zimmerman, M.L., Daleke, D.L., & Newton, A.C. (1998) *Biochemistry* 37, 12020–12025.

127. Garcia, P., Gupta, R., Shah, S., Morris, A.J., Rudge, S.A., Scarlata, S., Petrova, V., McLaughlin, S., & Rebecchi, M.J. (1995) *Biochemistry* 34, 16228–16234.
128. McDonald, J.F., Evans, T.C., Jr., Emeagwali, D.B., Hariharan, M., Allewell, N.M., Pusey, M.L., Shah, A.M., & Nelsestuen, G.L. (1997) *Biochemistry* 36, 15589–15598.
129. Arnold, R.S., DePaoli-Roach, A.A., & Cornell, R.B. (1997) *Biochemistry* 36, 6149–6156.
130. Fricker, L.D., Das, B., & Angeletti, R.H. (1990) *J. Biol. Chem.* 265, 2476–2482.
131. Homa, F.L., Hollander, T., Lehman, D.J., Thomsen, D.R., & Elhammer, A.P. (1993) *J. Biol. Chem.* 268, 12609–12616.
132. Hagen, F.K., Van Wuyckhuysse, B., & Tabak, L.A. (1993) *J. Biol. Chem.* 268, 18960–18965.
133. Larsen, R.D., Rajan, V.P., Ruff, M.M., Kukowska-Latallo, J., Cummings, R.D., & Lowe, J.B. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 8227–8231.
134. Konishi, K., Van Doren, S.R., Kramer, D.M., Crofts, A.R., & Gennis, R.B. (1991) *J. Biol. Chem.* 266, 14270–14276.
135. Taljanidisz, J., Stewart, L., Smith, A.J., & Klinman, J.P. (1989) *Biochemistry* 28, 10054–10061.
136. Schleiff, E., & Turnbull, J.L. (1998) *Biochemistry* 37, 13052–13058.
137. Schleiff, E., & Turnbull, J.L. (1998) *Biochemistry* 37, 13043–13051.
138. Loeb-Hennard, C., & McIntyre, J.O. (2000) *Biochemistry* 39, 11928–11938.
139. Chin, D.J., Gil, G., Russell, D.W., Liscum, L., Luskey, K.L., Basu, S.K., Okayama, H., Berg, P., Goldstein, J.L., & Brown, M.S. (1984) *Nature* 308, 613–617.
140. Picot, D., Loll, P.J., & Garavito, R.M. (1994) *Nature* 367, 243–249.
141. Xu, Y., & Mitra, B. (1999) *Biochemistry* 38, 12367–12376.
142. Sukumar, N., Xu, Y., Gatti, D.L., Mitra, B., & Mathews, F.S. (2001) *Biochemistry* 40, 9870–9878.
143. Pratt, K.P., Shen, B.W., Takeshima, K., Davie, E.W., Fujikawa, K., & Stoddard, B.L. (1999) *Nature* 402, 439–442.
144. Macedo-Ribeiro, S., Bode, W., Huber, R., Quinn-Allen, M.A., Kim, S.W., Ortel, T.L., Bourenkov, G.P., Bartunik, H.D., Stubbs, M.T., Kane, W.H., & Fuentes-Prior, P. (1999) *Nature* 402, 434–439.
145. Ferguson, M.A., & Williams, A.F. (1988) *Annu. Rev. Biochem.* 57, 285–320.
146. Ramsden, J.J., & Schneider, P. (1993) *Biochemistry* 32, 523–529.
147. Micanovic, R., Kodukula, K., Gerber, L.D., & Udenfriend, S. (1990) *Proc. Natl. Acad. Sci. U.S.A.* 87, 7939–7943.
148. Ikezawa, H., Yamanegi, M., Taguchi, R., Miyashita, T., & Ohyabu, T. (1976) *Biochim. Biophys. Acta* 450, 154–164.
149. Stams, T., Nair, S.K., Okuyama, T., Waheed, A., Sly, W.S., & Christianson, D.W. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 13589–13594.
150. Sussman, J.L., Harel, M., Frolow, F., Oefner, C., Goldman, A., Toker, L., & Silman, I. (1991) *Science* 253, 872–879.
151. Ferguson, M.A., Homans, S.W., Dwek, R.A., & Rademacher, T.W. (1988) *Science* 239, 753–759.
152. Selvaraj, P., Rosse, W.F., Silber, R., & Springer, T.A. (1988) *Nature* 333, 565–567.
153. Huizinga, T.W., van der Schoot, C.E., Jost, C., Klaassen, R., Kleijer, M., von dem Borne, A.E., Roos, D., & Tetteroo, P.A. (1988) *Nature* 333, 667–669.
154. Stieger, S., Gentinetta, R., & Brodbeck, U. (1989) *Eur. J. Biochem.* 181, 633–642.
155. Toutant, J.P., Roberts, W.L., Murray, N.R., & Rosenberry, T.L. (1989) *Eur. J. Biochem.* 180, 503–508.
156. Duval, N., Krejci, E., Grassi, J., Coussen, F., Massoulie, J., & Bon, S. (1992) *EMBO J.* 11, 3255–3261.
157. Zhu, X.L., & Sly, W.S. (1990) *J. Biol. Chem.* 265, 8795–8801.
158. Wistrand, P.J., & Knuuttila, K.G. (1989) *Kidney Int.* 35, 851–859.
159. Stiernberg, J., Low, M.G., Flaherty, L., & Kincade, P.W. (1987) *J. Immunol.* 138, 3877–3884.
160. Waneck, G.L., Stein, M.E., & Flavell, R.A. (1988) *Science* 241, 697–699.
161. Scallon, B.J., Scigliano, E., Freedman, V.H., Miedel, M.C., Pan, Y.C., Unkeless, J.C., & Kochan, J.P. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 5079–5083.
162. Takeshima, K., Smith, C., Tait, J., & Fujikawa, K. (2003) *Thromb. Haemostasis* 89, 788–794.
163. Silvius, J.R., & l'Heureux, F. (1994) *Biochemistry* 33, 3014–3022.
164. Ghomashchi, F., Zhang, X., Liu, L., & Gelb, M.H. (1995) *Biochemistry* 34, 11910–11918.
165. Stoffyn, P., & Folch-Pi, J. (1971) *Biochem. Biophys. Res. Commun.* 44, 157–161.
166. Wilcox, C.A., & Olson, E.N. (1987) *Biochemistry* 26, 1029–1036.
167. Weimbs, T., & Stoffel, W. (1992) *Biochemistry* 31, 12289–12296.
168. Weyer, K.A., Schaefer, W., Lottspeich, F., & Michel, H. (1987) *Biochemistry* 26, 2909–2914.
169. Losonczy, J.A., Tian, F., & Prestegard, J.H. (2000) *Biochemistry* 39, 3804–3816.
170. Strittmatter, P., Kittler, J.M., Coghill, J.E., & Ozols, J. (1993) *J. Biol. Chem.* 268, 23168–23171.
171. Murray, D., Hermida-Matsumoto, L., Buser, C.A., Tsang, J., Sigal, C.T., Ben-Tal, N., Honig, B., Resh, M.D., & McLaughlin, S. (1998) *Biochemistry* 37, 2145–2159.
172. Dizhoor, A.M., Chen, C.K., Olshevskaya, E., Sinelnikova, V.V., Phillipov, P., & Hurley, J.B. (1993) *Science* 259, 829–832.
173. Hughes, R.E., Brzovic, P.S., Klevit, R.E., & Hurley, J.B. (1995) *Biochemistry* 34, 11410–11416.
174. Peitzsch, R.M., & McLaughlin, S. (1993) *Biochemistry* 32, 10436–10443.
175. Berg, O.G., Yu, B.Z., Rogers, J., & Jain, M.K. (1991) *Biochemistry* 30, 7283–7297.
176. Barnett, S.F., Ledder, L.M., Stirdivant, S.M., Ahern, J., Conroy, R.R., & Heimbrook, D.C. (1995) *Biochemistry* 34, 14254–14262.
177. Jain, M.K., Krause, C.D., Buckley, J.T., Bayburt, T., & Gelb, M.H. (1994) *Biochemistry* 33, 5011–5020.
178. Wendt, K.U., Lenhart, A., & Schulz, G.E. (1999) *J. Mol. Biol.* 286, 175–187.
179. Huang, H., Ball, J.M., Billheimer, J.T., & Schroeder, F. (1999) *Biochemistry* 38, 13231–13243.

180. Carpenter, G. (1987) *Annu. Rev. Biochem.* 56, 881–914.
181. Yarden, Y., & Ullrich, A. (1988) *Annu. Rev. Biochem.* 57, 443–478.
182. Ullrich, A., Coussens, L., Hayflick, J.S., Dull, T.J., Gray, A., Tam, A.W., Lee, J., Yarden, Y., Libermann, T.A., Schlessinger, J., Downward, J., Mayes, E.L., Whittle, N., Waterfield, M.D., & Seeburg, P.H. (1984) *Nature* 309, 418–425.
183. Wedegaertner, P.B., & Gill, G.N. (1989) *J. Biol. Chem.* 264, 11346–11353.
184. Lax, I., Mitra, A.K., Ravera, C., Hurwitz, D.R., Rubinstein, M., Ullrich, A., Stroud, R.M., & Schlessinger, J. (1991) *J. Biol. Chem.* 266, 13828–13833.
185. Sanders, C.R.N., Czerski, L., Vinogradova, O., Badola, P., Song, D., & Smith, S.O. (1996) *Biochemistry* 35, 8610–8618.
186. Otsu, K., Willard, H.F., Khanna, V.K., Zorzato, F., Green, N.M., & MacLennan, D.H. (1990) *J. Biol. Chem.* 265, 13472–13483.
187. Carroll, J., Shannon, R.J., Fearnley, I.M., Walker, J.E., & Hirst, J. (2002) *J. Biol. Chem.* 277, 50311–50317.
188. Tomita, M., & Marchesi, V.T. (1975) *Proc. Natl. Acad. Sci. U.S.A.* 72, 2964–2968.
189. Bretscher, M.S. (1971) *Nat. New Biol.* 231, 229–232.
190. Thomas, D.B., & Winzler, R.J. (1969) *J. Biol. Chem.* 244, 5943–5946.
191. Bretscher, M.S. (1971) *J. Mol. Biol.* 59, 351–357.
192. Cone, R.A. (1972) *Nat. New Biol.* 236, 39–43.
193. Torres, C.R., & Hart, G.W. (1984) *J. Biol. Chem.* 259, 3308–3317.
194. Nicolson, G.L., & Singer, S.J. (1971) *Proc. Natl. Acad. Sci. U.S.A.* 68, 942–945.
195. Nicolson, G.L., & Singer, S.J. (1974) *J. Cell Biol.* 60, 236–248.
196. Schindler, M., Hogan, M., Miller, R., & DeGaetano, D. (1987) *J. Biol. Chem.* 262, 1254–1260.
197. Glossmann, H., & Neville, D.M., Jr. (1971) *J. Biol. Chem.* 246, 6339–6346.
198. Fraser, P.D., Misawa, N., Linden, H., Yamano, S., Kobayashi, K., & Sandmann, G. (1992) *J. Biol. Chem.* 267, 19891–19895.
199. Bun-Ya, M., Nishimura, M., Harashima, S., & Oshima, Y. (1991) *Mol. Cell Biol.* 11, 3229–3238.
200. Tamai, Y., Toh-e, A., & Oshima, Y. (1985) *J. Bacteriol.* 164, 964–968.
201. Schmitt, R. (1968) *J. Bacteriol.* 96, 462–471.
202. Magagnin, S., Werner, A., Markovich, D., Sorribas, V., Stange, G., Biber, J., & Murer, H. (1993) *Proc. Natl. Acad. Sci. U.S.A.* 90, 5979–5983.
203. Julius, D., MacDermott, A.B., Axel, R., & Jessell, T.M. (1988) *Science* 241, 558–564.
204. Kiefer, H., Krieger, J., Olszewski, J.D., Von Heijne, G., Prestwich, G.D., & Breer, H. (1996) *Biochemistry* 35, 16077–16084.
205. Engel, I., Ottenhoff, T.H., & Klausner, R.D. (1992) *Science* 256, 1318–1321.
206. Zampighi, G., Kyte, J., & Freytag, W. (1984) *J. Cell Biol.* 98, 1851–1864.
207. Jorgensen, P.L. (1974) *Biochim. Biophys. Acta* 356, 36–52.
208. Makino, S., Reynolds, J.A., & Tanford, C. (1973) *J. Biol. Chem.* 248, 4926–4932.
209. Clarke, S. (1975) *J. Biol. Chem.* 250, 5459–5469.
210. Yu, B., Xie, J., Deng, S., & Hui, Y. (1999) *J. Am. Chem. Soc.* 121, 12196–12197.
211. Casey, J.R., & Reithmeier, R.A. (1993) *Biochemistry* 32, 1172–1179.
212. Ribeiro, A.A., & Dennis, E.A. (1974) *Biochim. Biophys. Acta* 332, 26–35.
213. Simons, K., Helenius, A., & Garoff, H. (1973) *J. Mol. Biol.* 80, 119–133.
214. Helenius, A., & Seoderlund, H. (1973) *Biochim. Biophys. Acta* 307, 287–300.
215. Lichtenberg, D., Robson, R.J., & Dennis, E.A. (1983) *Biochim. Biophys. Acta* 737, 285–304.
216. Paternostre, M.T., Roux, M., & Rigaud, J.L. (1988) *Biochemistry* 27, 2668–2677.
217. Kragh-Hansen, U., le Maire, M., Noel, J.P., Gulik-Krzywicki, T., & Moller, J.V. (1993) *Biochemistry* 32, 1648–1656.
218. Ueno, M. (1989) *Biochemistry* 28, 5631–5634.
219. Knol, J., Sjollem, K., & Poolman, B. (1998) *Biochemistry* 37, 16410–16415.
220. Roth, M., Lewit-Bentley, A., Michel, H., Deisenhofer, J., Huber, R., & Oesterhelt, D. (1989) *Nature* 340, 659–662.
221. Kyte, J. (1971) *J. Biol. Chem.* 246, 4157–4165.
222. Musatov, A., & Robinson, N.C. (1994) *Biochemistry* 33, 13005–13012.
223. Musatov, A., Ortega-Lopez, J., & Robinson, N.C. (2000) *Biochemistry* 39, 12996–13004.
224. Craig, W.S. (1982) *Biochemistry* 21, 2667–2674.
225. Craig, W.S. (1982) *Biochemistry* 21, 5707–5717.
226. Knuttel, K., Schneider, K., Schlegel, H.G., & Muller, A. (1989) *Eur. J. Biochem.* 179, 101–108.
227. Ashida, H., Yamamoto, K., Kumagai, H., & Tochikura, T. (1992) *Eur. J. Biochem.* 205, 729–735.
228. Corey, E.J., Cheng, H., Baker, C.H., Matsuda, S.P.T., Li, D., & Song, X. (1997) *J. Am. Chem. Soc.* 119, 1277–1288.
229. Yet, S.F., Moon, Y.K., & Sul, H.S. (1995) *Biochemistry* 34, 7303–7310.
230. Caron, M.G., & Lefkowitz, R.J. (1976) *J. Biol. Chem.* 251, 2374–2384.
231. Launay, J.M., Geoffroy, C., Mutel, V., Buckle, M., Cesura, A., Alouf, J.E., & Da Prada, M. (1992) *J. Biol. Chem.* 267, 11344–11351.
232. Hinkle, P.C., Kim, J.J., & Racker, E. (1972) *J. Biol. Chem.* 247, 1338–1339.
233. Goldin, S.M., & Tong, S.W. (1974) *J. Biol. Chem.* 249, 5907–5915.
234. Kasahara, M., & Hinkle, P.C. (1976) *Proc. Natl. Acad. Sci. U.S.A.* 73, 396–400.
235. Kaplan, R.S., Mayor, J.A., Johnston, N., & Oliveira, D.L. (1990) *J. Biol. Chem.* 265, 13379–13385.
236. Ruan, Z.S., Anantharam, V., Crawford, I.T., Ambudkar, S.V., Rhee, S.Y., Allison, M.J., & Maloney, P.C. (1992) *J. Biol. Chem.* 267, 10537–10543.
237. Indiveri, C., Tonazzi, A., & Palmieri, F. (1992) *Eur. J. Biochem.* 207, 449–454.
238. Huang, K.S., Bayley, H., Liao, M.J., London, E., & Khorana, H.G. (1981) *J. Biol. Chem.* 256, 3802–3809.
239. Mogi, T., Stern, L.J., Marti, T., Chao, B.H., & Khorana, H.G. (1988) *Proc. Natl. Acad. Sci. U.S.A.* 85, 4148–4152.
240. Kataoka, M., Shimizu, S., & Yamada, H. (1992) *Eur. J. Biochem.* 204, 799–806.

241. Haase, P., Deppenmeier, U., Blaut, M., & Gottschalk, G. (1992) *Eur. J. Biochem.* 203, 527–531.
242. Kasahara, M., & Hinkle, P.C. (1977) *J. Biol. Chem.* 252, 7384–7390.
243. Troll, H., Malchow, D., Muller-Taubenberger, A., Humbel, B., Lottspeich, F., Ecke, M., Gerisch, G., Schmid, A., & Benz, R. (1992) *J. Biol. Chem.* 267, 21072–21079.
244. Sakurai, N., & Sakurai, T. (1997) *Biochemistry* 36, 13809–13815.
245. van Hoek, A.N., Wiener, M.C., Verbavatz, J.M., Brown, D., Lipniunas, P.H., Townsend, R.R., & Verkman, A.S. (1995) *Biochemistry* 34, 2212–2219.
246. Siebers, A., Kollmann, R., Dirkes, G., & Altendorf, K. (1992) *J. Biol. Chem.* 267, 12717–12721.
247. Collinson, I.R., Runswick, M.J., Buchanan, S.K., Fearnley, I.M., Skehel, J.M., van Raaij, M.J., Griffiths, D.E., & Walker, J.E. (1994) *Biochemistry* 33, 7971–7978.
248. Middleton, R.E., Pheasant, D.J., & Miller, C. (1994) *Biochemistry* 33, 13189–13198.
249. Santacruz-Toloz, L., Perozo, E., & Papazian, D.M. (1994) *Biochemistry* 33, 1295–1299.
250. Hamada, H., & Tsuruo, T. (1988) *J. Biol. Chem.* 263, 1454–1458.
251. Zeidel, M.L., Ambudkar, S.V., Smith, B.L., & Agre, P. (1992) *Biochemistry* 31, 7436–7440.
252. Zeidel, M.L., Nielsen, S., Smith, B.L., Ambudkar, S.V., Maunsbach, A.B., & Agre, P. (1994) *Biochemistry* 33, 1606–1615.
253. Dean, R.M., Rivers, R.L., Zeidel, M.L., & Roberts, D.M. (1999) *Biochemistry* 38, 347–353.
254. Pourcher, T., Leclercq, S., Brandolin, G., & Leblanc, G. (1995) *Biochemistry* 34, 4412–4420.
255. Jezek, P., Orosz, D.E., & Garlid, K.D. (1990) *J. Biol. Chem.* 265, 19296–19302.
256. Perozo, E., & Hubbell, W.L. (1993) *Biochemistry* 32, 10471–10478.
257. Giangiacomo, K.M., Garcia-Calvo, M., Knaus, H.G., Mullmann, T.J., Garcia, M.L., & McManus, O. (1995) *Biochemistry* 34, 15849–15862.
258. Ferris, C.D., Cameron, A.M., Haganir, R.L., & Snyder, S.H. (1992) *Nature* 356, 350–352.
259. Awasthi, S., Singhal, S.S., Pikula, S., Piper, J.T., Srivastava, S.K., Torman, R.T., Bandorowicz-Pikula, J., Lin, J.T., Singh, S.V., Zimniak, P., & Awasthi, Y.C. (1998) *Biochemistry* 37, 5239–5248.
260. Hill, K., Model, K., Ryan, M.T., Dietmeier, K., Martin, F., Wagner, R., & Pfanner, N. (1998) *Nature* 395, 516–521.
261. Kaupp, U.B., Niidome, T., Tanabe, T., Terada, S., Bonigk, W., Stuhmer, W., Cook, N.J., Kangawa, K., Matsuo, H., Hirose, T., et al. (1989) *Nature* 342, 762–766.
262. Preston, G.M., Carroll, T.P., Guggino, W.B., & Agre, P. (1992) *Science* 256, 385–387.
263. Sanders, C.R.N., & Landis, G.C. (1995) *Biochemistry* 34, 4030–4040.
264. Sanders, C.R.N., & Schwonek, J.P. (1992) *Biochemistry* 31, 8898–8905.
265. Glover, K.J., Whiles, J.A., Wu, G., Yu, N., Deems, R., Struppe, J.O., Stark, R.E., Komives, E.A., & Vold, R.R. (2001) *Biophys. J.* 81, 2163–2171.
266. Tribet, C., Audebert, R., & Popot, J.L. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 15047–15050.
267. Price, E.M., Rice, D.A., & Lingrel, J.B. (1990) *J. Biol. Chem.* 265, 6638–6641.
268. Vilsen, B., Andersen, J.P., & MacLennan, D.H. (1991) *J. Biol. Chem.* 266, 18839–18845.
269. Walton, G.M., Chen, W.S., Rosenfeld, M.G., & Gill, G.N. (1990) *J. Biol. Chem.* 265, 1750–1754.
270. Lechleiter, J., Hellmiss, R., Duerson, K., Ennulat, D., David, N., Clapham, D., & Peralta, E. (1990) *EMBO J.* 9, 4381–4390.
271. Pei, G., Tiberi, M., Caron, M.G., & Lefkowitz, R.J. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 3633–3636.
272. Kong, C.T., Yet, S.F., & Lever, J.E. (1993) *J. Biol. Chem.* 268, 1509–1512.
273. Sunahara, R.K., Niznik, H.B., Weiner, D.M., Stormann, T.M., Brann, M.R., Kennedy, J.L., Gelernter, J.E., Rozmahel, R., Yang, Y.L., Israel, Y., et al. (1990) *Nature* 347, 80–83.
274. Spatz, L., & Strittmatter, P. (1971) *Proc. Natl. Acad. Sci. U.S.A.* 68, 1042–1046.
275. Strittmatter, P., Rogers, M.J., & Spatz, L. (1972) *J. Biol. Chem.* 247, 7188–7194.
276. Springer, T.A., Strominger, J.L., & Mann, D. (1974) *Proc. Natl. Acad. Sci. U.S.A.* 71, 1539–1543.
277. Springer, T.A., Kaufman, J.F., Terhorst, C., & Strominger, J.L. (1977) *Nature* 268, 213–218.
278. MacNair, D.C., & Kenny, A.J. (1979) *Biochem. J.* 179, 379–395.
279. Kenny, A.J., Booth, A.G., George, S.G., Ingram, J., Kershaw, D., Wood, E.J., & Young, A.R. (1976) *Biochem. J.* 157, 169–182.
280. Maroux, S., & Louvard, D. (1976) *Biochim. Biophys. Acta* 419, 189–195.
281. Brunner, J., Hauser, H., Braun, H., Wilson, K.J., Wacker, H., O'Neill, B., & Semenza, G. (1979) *J. Biol. Chem.* 254, 1821–1828.
282. Haselbeck, A. (1989) *Eur. J. Biochem.* 181, 663–668.
283. Nishimoto, M., Clark, J.E., & Masters, B.S. (1993) *Biochemistry* 32, 8863–8870.
284. Norman, R.L., Johnson, E.F., & Muller-Eberhard, U. (1978) *J. Biol. Chem.* 253, 8640–8647.
285. Haugen, D.A., van der Hoeven, T.A., & Coon, M.J. (1975) *J. Biol. Chem.* 250, 3567–3570.
286. Yagi, K., Arai, Y., Kato, N., Hirota, K., & Miura, Y. (1989) *Eur. J. Biochem.* 180, 509–513.
287. Skehel, J.J., & Waterfield, M.D. (1975) *Proc. Natl. Acad. Sci. U.S.A.* 72, 93–97.
288. Pattus, F., Verger, R., & Desnuelle, P. (1976) *Biochem. Biophys. Res. Commun.* 69, 718–723.
289. Wilson, I.A., Skehel, J.J., & Wiley, D.C. (1981) *Nature* 289, 366–373.
290. Mathews, F.S., Argos, P., & Levine, M. (1972) *Cold Spring Harbor Symp. Quant. Biol.* 36, 387–395.
291. Rey, F.A., Heinz, F.X., Mandl, C., Kunz, C., & Harrison, S.C. (1995) *Nature* 375, 291–298.
292. Kobe, B., Center, R.J., Kemp, B.E., & Pountourios, P. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 4319–4324.
293. Walter, M.R., Windsor, W.T., Nagabhushan, T.L., Lundell, D.J., Lunn, C.A., Zauodny, P.J., & Narula, S.K. (1995) *Nature* 376, 230–235.

294. Williams, P.A., Cosme, J., Sridhar, V., Johnson, E.F., & McRee, D.E. (2000) *Mol. Cell* 5, 121–131.
295. Ozols, J., & Gerard, C. (1977) *J. Biol. Chem.* 252, 8549–8553.
296. Malissen, M., Malissen, B., & Jordan, B.R. (1982) *Proc. Natl. Acad. Sci. U.S.A.* 79, 893–897.
297. Frank, G., Brunner, J., Hauser, H., Wacker, H., Semenza, G., & Zuber, H. (1978) *FEBS Lett.* 96, 183–188.
298. Verhoeven, M., Fang, R., Jou, W.M., Devos, R., Huylebroeck, D., Saman, E., & Fiers, W. (1980) *Nature* 286, 771–776.
299. Tomita, M., Furthmayr, H., & Marchesi, V.T. (1978) *Biochemistry* 17, 4756–4770.
300. Kyte, J., & Doolittle, R.F. (1982) *J. Mol. Biol.* 157, 105–132.
301. Schulte, T.H., & Marchesi, V.T. (1979) *Biochemistry* 18, 275–280.
302. Visser, L., Robinson, N.C., & Tanford, C. (1975) *Biochemistry* 14, 1194–1199.
303. Spiess, M., Brunner, J., & Semenza, G. (1982) *J. Biol. Chem.* 257, 2370–2377.
304. Landolt-Marticorena, C., Williams, K.A., Deber, C.M., & Reithmeier, R.A. (1993) *J. Mol. Biol.* 229, 602–608.
305. Bolen, E.J., & Holloway, P.W. (1990) *Biochemistry* 29, 9638–9643.
306. Ren, J., Lew, S., Wang, Z., & London, E. (1997) *Biochemistry* 36, 10213–10220.
307. Zhang, Y.P., Lewis, R.N., Hodges, R.S., & McElhaney, R.N. (1995) *Biochemistry* 34, 2362–2371.
308. Zhang, Y.P., Lewis, R.N., Hodges, R.S., & McElhaney, R.N. (1992) *Biochemistry* 31, 11572–11578.
309. de Planque, M.R., Greathouse, D.V., Koeppe, R.E.N., Schafer, H., Marsh, D., & Killian, J.A. (1998) *Biochemistry* 37, 9333–9345.
310. Lew, S., Ren, J., & London, E. (2000) *Biochemistry* 39, 9632–9640.
311. Zhou, F.X., Merianos, H.J., Brunger, A.T., & Engelman, D.M. (2001) *Proc. Natl. Acad. Sci. U.S.A.* 98, 2250–2255.
312. Zhou, F.X., Cocco, M.J., Russ, W.P., Brunger, A.T., & Engelman, D.M. (2000) *Nat. Struct. Biol.* 7, 154–160.
313. Choma, C., Gratkowski, H., Lear, J.D., & DeGrado, W.F. (2000) *Nat. Struct. Biol.* 7, 161–166.
314. Gratkowski, H., Lear, J.D., & DeGrado, W.F. (2001) *Proc. Natl. Acad. Sci. U.S.A.* 98, 880–885.
315. Braun, P., & von Heijne, G. (1999) *Biochemistry* 38, 9778–9782.
316. MacLennan, D.H. (1970) *J. Biol. Chem.* 245, 4508–4518.
317. Briggs, M., Kamp, P.F., Robinson, N.C., & Capaldi, R.A. (1975) *Biochemistry* 14, 5123–5128.
318. Karlin, A., McNamee, M.G., Weill, C.L., & Valderrama, R. (1976) in *Methods in Receptor Research* (Blecher, M., Ed.) pp 1–35, Marcel Dekker, New York.
319. Ludwig, B., Downer, N.W., & Capaldi, R.A. (1979) *Biochemistry* 18, 1401–1407.
320. Drickamer, L.K. (1976) *J. Biol. Chem.* 251, 5115–5123.
321. Deisenhofer, J., Epp, O., Sinning, I., & Michel, H. (1995) *J. Mol. Biol.* 246, 429–457.
322. Allen, J.P., Feher, G., Yeates, T.O., Komiya, H., & Rees, D.C. (1987) *Proc. Natl. Acad. Sci. U.S.A.* 84, 6162–6166.
323. Nogi, T., Fathir, I., Kobayashi, M., Nozawa, T., & Miki, K. (2000) *Proc. Natl. Acad. Sci. U.S.A.* 97, 13561–13566.
324. Michel, H. (1982) *J. Mol. Biol.* 158, 567–572.
325. Pebay-Peyroula, E., Rummel, G., Rosenbusch, J.P., & Landau, E.M. (1997) *Science* 277, 1676–1681.
326. Luecke, H., Schobert, B., Richter, H.T., Cartailier, J.P., & Lanyi, J.K. (1999) *J. Mol. Biol.* 291, 899–911.
327. Henderson, R., & Unwin, P.N. (1975) *Nature* 257, 28–32.
328. Henderson, R., Baldwin, J.M., Ceska, T.A., Zemlin, F., Beckmann, E., & Downing, K.H. (1990) *J. Mol. Biol.* 213, 899–929.
329. Landau, E.M., & Rosenbusch, J.P. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 14532–14535.
330. Kolbe, M., Besir, H., Essen, L.O., & Oesterheld, D. (2000) *Science* 288, 1390–1396.
331. Palczewski, K., Kumasaka, T., Hori, T., Behnke, C.A., Motoshima, H., Fox, B.A., Le Trong, I., Teller, D.C., Okada, T., Stenkamp, R.E., Yamamoto, M., & Miyano, M. (2000) *Science* 289, 739–745.
332. Okada, T., Le Trong, I., Fox, B.A., Behnke, C.A., Stenkamp, R.E., & Palczewski, K. (2000) *J. Struct. Biol.* 130, 73–80.
333. Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R., & Yoshikawa, S. (1996) *Science* 272, 1136–1144.
334. Yoshikawa, S., Tera, T., Takahashi, Y., Tsukihara, T., & Caughey, W.S. (1988) *Proc. Natl. Acad. Sci. U.S.A.* 85, 1354–1358.
335. Ostermeier, C., Harrenga, A., Ermler, U., & Michel, H. (1997) *Proc. Natl. Acad. Sci. U.S.A.* 94, 10547–10553.
336. Lee, J.W., Chan, M., Law, T.V., Kwon, H.J., & Jap, B.K. (1995) *J. Mol. Biol.* 252, 15–19.
337. Xia, D., Yu, C.A., Kim, H., Xia, J.Z., Kachurin, A.M., Zhang, L., Yu, L., & Deisenhofer, J. (1997) *Science* 277, 60–66.
338. Zhang, Z., Huang, L., Shulmeister, V.M., Chi, Y.I., Kim, K.K., Hung, L.W., Crofts, A.R., Berry, E.A., & Kim, S.H. (1998) *Nature* 392, 677–684.
339. Iwata, S., Lee, J.W., Okada, K., Lee, J.K., Iwata, M., Rasmussen, B., Link, T.A., Ramaswamy, S., & Jap, B.K. (1998) *Science* 281, 64–71.
340. Abramson, J., Riistama, S., Larsson, G., Jasaitis, A., Svensson-Ek, M., Laakkonen, L., Puustinen, A., Iwata, S., & Wikstrom, M. (2000) *Nat. Struct. Biol.* 7, 910–917.
341. Iverson, T.M., Luna-Chavez, C., Cecchini, G., & Rees, D.C. (1999) *Science* 284, 1961–1966.
342. Chang, G., & Roth, C.B. (2001) *Science* 293, 1793–1800.
343. Doyle, D.A., Morais Cabral, J., Pfuetzner, R.A., Kuo, A., Gulbis, J.M., Cohen, S.L., Chait, B.T., & MacKinnon, R. (1998) *Science* 280, 69–77.
344. Zhou, Y., Morais-Cabral, J.H., Kaufman, A., & MacKinnon, R. (2001) *Nature* 414, 43–48.
345. Miyazawa, A., Fujiyoshi, Y., & Unwin, N. (2003) *Nature* 423, 949–955.
346. Unwin, N. (2005) *J. Mol. Biol.* 346, 967–989.
347. Chang, G., Spencer, R.H., Lee, A.T., Barclay, M.T., & Rees, D.C. (1998) *Science* 282, 2220–2226.
348. Sui, H., Han, B.G., Lee, J.K., Walian, P., & Jap, B.K. (2001) *Nature* 414, 872–878.
349. Fu, D., Libson, A., Miercke, L.J., Weitzman, C., Nollert, P., Krucinski, J., & Stroud, R.M. (2000) *Science* 290, 481–486.
350. Toyoshima, C., & Mizutani, T. (2004) *Nature* 430, 529–535.

351. Toyoshima, C., Nakasako, M., Nomura, H., & Ogawa, H. (2000) *Nature* 405, 647–655.
352. Weiss, M.S., Abele, U., Weckesser, J., Welte, W., Schiltz, E., & Schulz, G.E. (1991) *Science* 254, 1627–1630.
353. Schirmer, T., Keller, T.A., Wang, Y.F., & Rosenbusch, J.P. (1995) *Science* 267, 512–514.
354. Stauffer, K.A., Page, M.G., Hardmeyer, A., Keller, T.A., & Pauptit, R.A. (1990) *J. Mol. Biol.* 211, 297–299.
355. Cowan, S.W., Schirmer, T., Rummel, G., Steiert, M., Ghosh, R., Pauptit, R.A., Jansonius, J.N., & Rosenbusch, J.P. (1992) *Nature* 358, 727–733.
356. Forst, D., Welte, W., Wacker, T., & Diederichs, K. (1998) *Nat. Struct. Biol.* 5, 37–46.
357. Buchanan, S.K., Smith, B.S., Venkatramani, L., Xia, D., Esser, L., Palnitkar, M., Chakraborty, R., van der Helm, D., & Deisenhofer, J. (1999) *Nat. Struct. Biol.* 6, 56–63.
358. Ferguson, A.D., Hofmann, E., Coulton, J.W., Diederichs, K., & Welte, W. (1998) *Science* 282, 2215–2220.
359. Pautsch, A., & Schulz, G.E. (1998) *Nat. Struct. Biol.* 5, 1013–1017.
360. Koronakis, V., Sharff, A., Koronakis, E., Luisi, B., & Hughes, C. (2000) *Nature* 405, 914–919.
361. Song, L., Hobaugh, M.R., Shustak, C., Cheley, S., Bayley, H., & Gouaux, J.E. (1996) *Science* 274, 1859–1866.
362. Lindblom, G., & Rilfors, L. (1989) *Biochim. Biophys. Acta* 988, 221–256.
363. Takeda, K., Sato, H., Hino, T., Kono, M., Fukuda, K., Sakurai, I., Okada, T., & Kouyama, T. (1998) *J. Mol. Biol.* 283, 463–474.
364. Takeda, K., Matsui, Y., Kamiya, N., Adachi, S., Okumura, H., & Kouyama, T. (2004) *J. Mol. Biol.* 341, 1023–1037.
365. Sui, H., Walian, P.J., Tang, G., Oh, A., & Jap, B.K. (2000) *Acta Crystallogr., D* 56, 1198–1200.
366. Luna-Chavez, C., Iverson, T.M., Rees, D.C., & Cecchini, G. (2000) *Protein Expression Purif.* 19, 188–196.
367. Sarkar, H.K., Thorens, B., Lodish, H.F., & Kaback, H.R. (1988) *Proc. Natl. Acad. Sci. U.S.A.* 85, 5463–5467.
368. Reeves, P.J., Thurmond, R.L., & Khorana, H.G. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 11487–11492.
369. Strugatsky, D., Gottschalk, K.E., Goldshleger, R., Bibi, E., & Karlish, S.J. (2003) *J. Biol. Chem.* 278, 46064–46073.
370. Schobert, B., Cupp-Vickery, J., Hornak, V., Smith, S., & Lanyi, J. (2002) *J. Mol. Biol.* 321, 715–726.
371. Lange, C., Nett, J.H., Trumpower, B.L., & Hunte, C. (2001) *EMBO J.* 20, 6591–6600.
372. Hunte, C., Koepke, J., Lange, C., Rossmann, T., & Michel, H. (2000) *Structure (London)* 8, 669–684.
373. Hung, L.W., Wang, I.X., Nikaido, K., Liu, P.Q., Ames, G.F., & Kim, S.H. (1998) *Nature* 396, 703–707.
374. Yeates, T.O., Komiya, H., Rees, D.C., Allen, J.P., & Feher, G. (1987) *Proc. Natl. Acad. Sci. U.S.A.* 84, 6438–6442.
375. Chamberlain, A.K., Lee, Y., Kim, S., & Bowie, J.U. (2004) *J. Mol. Biol.* 339, 471–479.
376. Monne, M., Nilsson, I., Johansson, M., Elmhed, N., & von Heijne, G. (1998) *J. Mol. Biol.* 284, 1177–1183.
377. Segrest, J.P., De Loof, H., Dohlman, J.G., Brouillette, C.G., & Anantharamaiah, G.M. (1990) *Proteins: Struct., Funct., Genet.* 8, 103–117.
378. Janin, J., & Wodak, S. (1978) *J. Mol. Biol.* 125, 357–386.
379. Ponder, J.W., & Richards, F.M. (1987) *J. Mol. Biol.* 193, 775–791.
380. Schrauber, H., Eisenhaber, F., & Argos, P. (1993) *J. Mol. Biol.* 230, 592–612.
381. Wilson, K.S., Butterworth, S., Dauter, Z., Lamzin, V.S., Walsh, M., Wodak, S., Pontius, J., Richelle, J., Vaguine, A., Sander, C., Hooft, R.W.W., Vriend, G., Thornton, J.M., Laskowski, R.A., MacArthur, M.W., Dodson, E.J., Murshudov, G., Oldfield, T.J., Kaptien, R., & Rullmann, J.A.C. (1998) *J. Mol. Biol.* 276, 417–436.
382. Ridder, A.N., Morein, S., Stam, J.G., Kuhn, A., de Kruijff, B., & Killian, J.A. (2000) *Biochemistry* 39, 6521–6528.
383. Von Heijne, G. (1986) *EMBO J.* 5, 3021–3027.
384. van Klompenburg, W., Nilsson, I., von Heijne, G., & de Kruijff, B. (1997) *EMBO J.* 16, 4261–4266.
385. Eilers, M., Shekar, S.C., Shieh, T., Smith, S.O., & Fleming, P.J. (2000) *Proc. Natl. Acad. Sci. U.S.A.* 97, 5796–5801.
386. Chamberlain, A.K., Faham, S., Yohannan, S., & Bowie, J.U. (2003) *Adv. Protein Chem.* 63, 19–46.
387. Bowie, J.U. (1997) *J. Mol. Biol.* 272, 780–789.
388. Yohannan, S., Faham, S., Yang, D., Whitelegge, J.P., & Bowie, J.U. (2004) *Proc. Natl. Acad. Sci. U.S.A.* 101, 959–963.
389. Yohannan, S., Yang, D., Faham, S., Boulting, G., Whitelegge, J., & Bowie, J.U. (2004) *J. Mol. Biol.* 341, 1–6.
390. Nilsson, I., & von Heijne, G. (1998) *J. Mol. Biol.* 284, 1185–1189.
391. Kellaris, K.V., Ware, D.K., Smith, S., & Kyte, J. (1989) *Biochemistry* 28, 3469–3482.
392. Williams, J.C., Steiner, L.A., Ogden, R.C., Simon, M.I., & Feher, G. (1983) *Proc. Natl. Acad. Sci. U.S.A.* 80, 6505–6509.
393. Williams, J.C., Steiner, L.A., Feher, G., & Simon, M.I. (1984) *Proc. Natl. Acad. Sci. U.S.A.* 81, 7303–7307.
394. Michel, H., Weyer, K.A., Gruenberg, H., & Lottspeich, F. (1985) *EMBO J.* 4, 1667–1672.
395. McAuley, K.E., Fyfe, P.K., Ridge, J.P., Isaacs, N.W., Cogdell, R.J., & Jones, M.R. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 14706–14711.
396. Jost, P.C., Griffith, O.H., Capaldi, R.A., & Vanderkooi, G. (1973) *Proc. Natl. Acad. Sci. U.S.A.* 70, 480–484.
397. Silviu, J.R., McMillen, D.A., Saley, N.D., Jost, P.C., & Griffith, O.H. (1984) *Biochemistry* 23, 538–547.
398. East, J.M., Melville, D., & Lee, A.G. (1985) *Biochemistry* 24, 2615–2623.
399. Griffith, O.H., McMillen, D.A., Keana, J.F., & Jost, P.C. (1986) *Biochemistry* 25, 574–584.
400. Horvath, L.L., Brophy, P.J., & Marsh, D. (1988) *Biochemistry* 27, 46–52.
401. Brothier, J.R., Griffith, O.H., Brothier, M.O., Jost, P.C., Silviu, J.R., & Hokin, L.E. (1981) *Biochemistry* 20, 5261–5267.
402. Ellena, J.F., Blazing, M.A., & McNamee, M.G. (1983) *Biochemistry* 22, 5523–5535.
403. Pates, R.D., & Marsh, D. (1987) *Biochemistry* 26, 29–39.
404. Sankaram, M.B., Brophy, P.J., & Marsh, D. (1991) *Biochemistry* 30, 5866–5873.
405. Kyte, J. (1995) *Structure in Protein Chemistry*, 1st ed., p 541, Garland Publishing Inc., New York.

406. Seelig, J., Tamm, L., Hymel, L., & Fleischer, S. (1981) *Biochemistry* 20, 3922–3932.
407. Hendriks, J., Warne, A., Gohlke, U., Haltia, T., Ludovici, C., Lubben, M., & Saraste, M. (1998) *Biochemistry* 37, 13102–13109.
408. Boulanger, P., le Maire, M., Bonhivers, M., Dubois, S., Desmadril, M., & Letellier, L. (1996) *Biochemistry* 35, 14216–14224.
409. Heegaard, C.W., le Maire, M., Gulik-Krzywicki, T., & Moller, J.V. (1990) *J. Biol. Chem.* 265, 12020–12028.
410. Zottola, R.J., Cloherty, E.K., Coderre, P.E., Hansen, A., Hebert, D.N., & Carruthers, A. (1995) *Biochemistry* 34, 9734–9747.
411. Kaul, R.K., Murthy, S.N., Reddy, A.G., Steck, T.L., & Kohler, H. (1983) *J. Biol. Chem.* 258, 7981–7990.
412. Unwin, P.N., & Henderson, R. (1975) *J. Mol. Biol.* 94, 425–440.
413. Deisenhofer, J., Epp, O., Miki, K., Huber, R., & Michel, H. (1986) *Nature* 318, 618–624.
414. Varghese, J.N., Laver, W.G., & Colman, P.M. (1983) *Nature* 303, 35–40.
415. Neubig, R.R., Krodel, E.K., Boyd, N.D., & Cohen, J.B. (1979) *Proc. Natl. Acad. Sci. U.S.A.* 76, 690–694.
416. Moore, H.P., Hartig, P.R., & Raftery, M.A. (1979) *Proc. Natl. Acad. Sci. U.S.A.* 76, 6265–6269.
417. Weill, C.L., McNamee, M.G., & Karlin, A. (1974) *Biochem. Biophys. Res. Commun.* 61, 997–1003.
418. Reed, K., Vandlen, R., Bode, J., Duguid, J., & Raftery, M.A. (1975) *Arch. Biochem. Biophys.* 167, 138–144.
419. Vandlen, R.L., Wu, W.C., Eisenach, J.C., & Raftery, M.A. (1979) *Biochemistry* 18, 1845–1854.
420. Raftery, M.A., Hunkapiller, M.W., Strader, C.D., & Hood, L.E. (1980) *Science* 208, 1454–1456.
421. Noda, M., Takahashi, H., Tanabe, T., Toyosato, M., Furutani, Y., Hirose, T., Asai, M., Inayama, S., Miyata, T., & Numa, S. (1982) *Nature* 299, 793–797.
422. Noda, M., Takahashi, H., Tanabe, T., Toyosato, M., Kikuyotani, S., Furutani, Y., Hirose, T., Takashima, H., Inayama, S., Miyata, T., & Numa, S. (1983) *Nature* 302, 528–532.
423. Claudio, T., Ballivet, M., Patrick, J., & Heinemann, S. (1983) *Proc. Natl. Acad. Sci. U.S.A.* 80, 1111–1115.
424. Brisson, A., & Unwin, P.N. (1985) *Nature* 315, 474–477.
425. Toyoshima, C., & Unwin, N. (1990) *J. Cell Biol.* 111, 2623–2635.
426. Unwin, P.N., & Zampighi, G. (1980) *Nature* 283, 545–549.
427. Robertson, J.D. (1963) *J. Cell Biol.* 19, 201–221.
428. Brisson, A., & Wade, R.H. (1983) *J. Mol. Biol.* 166, 21–36.
429. Vergara, J., Longley, W., & Robertson, J.D. (1969) *J. Mol. Biol.* 46, 593–596.
430. Zheng, L., Kostrewa, D., Berneche, S., Winkler, F.K., & Li, X.D. (2004) *Proc. Natl. Acad. Sci. U.S.A.* 101, 17090–17095.
431. Khademi, S., O'Connell, J.R., Remis, J., Robles-Colmenares, Y., Miercke, L.J., & Stroud, R.M. (2004) *Science* 305, 1587–1594.
432. Van den Berg, B., Clemons, W.M., Jr., Collinson, I., Modis, Y., Hartmann, E., Harrison, S.C., & Rapoport, T.A. (2004) *Nature* 427, 36–44.
433. Dutzler, R., Campbell, E.B., & MacKinnon, R. (2003) *Science* 300, 108–112.
434. Locher, K.P., Lee, A.T., & Rees, D.C. (2002) *Science* 296, 1091–1098.
435. MacKenzie, K.R., & Engelman, D.M. (1998) *Proc. Natl. Acad. Sci. U.S.A.* 95, 3583–3590.
436. Lemmon, M.A., Flanagan, J.M., Treutlein, H.R., Zhang, J., & Engelman, D.M. (1992) *Biochemistry* 31, 12719–12725.
437. MacKenzie, K.R., Prestegard, J.H., & Engelman, D.M. (1997) *Science* 276, 131–133.
438. Blaurock, A.E., & Stoeckenius, W. (1971) *Nat. New Biol.* 233, 152–155.
439. Henderson, R., Capaldi, R.A., & Leigh, J.S. (1977) *J. Mol. Biol.* 112, 631–648.
440. Unwin, P.N., & Ennis, P.D. (1984) *Nature* 307, 609–613.
441. Fujiyoshi, Y., Mizusaki, T., Morikawa, K., Yamagishi, H., Aoki, Y., Kihara, H., & Harada, Y. (1991) *Ultramicroscopy* 38, 241–251.
442. Crowther, R.A., DeRosier, D.J., & Klug, A. (1970) *Proc. R. Soc. London A* 317, 319–340.
443. Grigorieff, N., Ceska, T.A., Downing, K.H., Baldwin, J.M., & Henderson, R. (1996) *J. Mol. Biol.* 259, 393–421.
444. Brejc, K., van Dijk, W.J., Klaassen, R.V., Schuurmans, M., van Der Oost, J., Smit, A.B., & Sixma, T.K. (2001) *Nature* 411, 269–276.
445. Kuhlbrandt, W., Wang, D.N., & Fujiyoshi, Y. (1994) *Nature* 367, 614–621.
446. Murata, K., Mitsuoaka, K., Hirai, T., Walz, T., Agre, P., Heymann, J.B., Engel, A., & Fujiyoshi, Y. (2000) *Nature* 407, 599–605.
447. Ren, G., Cheng, A., Reddy, V., Melnyk, P., & Mitra, A.K. (2000) *J. Mol. Biol.* 301, 369–387.
448. Williams, K.A. (2000) *Nature* 403, 112–115.
449. Unger, V.M., Kumar, N.M., Gilula, N.B., & Yeager, M. (1999) *Science* 283, 1176–1180.
450. Hebert, H., Purhonen, P., Vorum, H., Thomsen, K., & Maunsbach, A.B. (2001) *J. Mol. Biol.* 314, 479–494.
451. Toyoshima, C., Sasabe, H., & Stokes, D.L. (1993) *Nature* 362, 467–471.
452. Girvin, M.E., Rastogi, V.K., Abildgaard, F., Markley, J.L., & Fillingame, R.H. (1998) *Biochemistry* 37, 8817–8824.
453. Ketchum, R.R., Hu, W., & Cross, T.A. (1993) *Science* 261, 1457–1460.
454. Marassi, F.M., Ramamoorthy, A., & Opella, S.J. (1997) *Proc. Natl. Acad. Sci. U.S.A.* 94, 8551–8556.
455. Opella, S.J., Marassi, F.M., Gesell, J.J., Valente, A.P., Kim, Y., Oblatt-Montal, M., & Montal, M. (1999) *Nat. Struct. Biol.* 6, 374–379.
456. Mesleh, M.F., Lee, S., Veglia, G., Thiriot, D.S., Marassi, F.M., & Opella, S.J. (2003) *J. Am. Chem. Soc.* 125, 8928–8935.
457. Altenbach, C., Marti, T., Khorana, H.G., & Hubbell, W.L. (1990) *Science* 248, 1088–1092.
458. Hackett, N.R., Stern, L.J., Chao, B.H., Kronis, K.A., & Khorana, H.G. (1987) *J. Biol. Chem.* 262, 9277–9284.
459. Altenbach, C., Greenhalgh, D.A., Khorana, H.G., & Hubbell, W.L. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 1667–1671.
460. Voss, J., He, M.M., Hubbell, W.L., & Kaback, H.R. (1996) *Biochemistry* 35, 12915–12918.
461. Oh, K.J., Zhan, H., Cui, C., Altenbach, C., Hubbell, W.L., & Collier, R.J. (1999) *Biochemistry* 38, 10336–10343.

462. Kaplan, R.S., Mayor, J.A., Kotaria, R., Walters, D.E., & McHaourab, H.S. (2000) *Biochemistry* 39, 9157–9163.
463. Tamamizu, S., Guzman, G.R., Santiago, J., Rojas, L.V., McNamee, M.G., & Lasalde-Dominicci, J.A. (2000) *Biochemistry* 39, 4666–4673.
464. Hinkle, P.C., Hinkle, P.V., & Kaback, H.R. (1990) *Biochemistry* 29, 10989–10994.
465. Komiya, H., Yeates, T.O., Rees, D.C., Allen, J.P., & Feher, G. (1988) *Proc. Natl. Acad. Sci. U.S.A.* 85, 9012–9016.
466. Abrams, F.S., Chattopadhyay, A., & London, E. (1992) *Biochemistry* 31, 5322–5327.
467. Chen, X., Wolfgang, D.E., & Sampson, N.S. (2000) *Biochemistry* 39, 13383–13389.
468. Engelman, D.M., Steitz, T.A., & Goldman, A. (1986) *Annu. Rev. Biophys. Biophys. Chem.* 15, 321–353.
469. Eisenberg, D., Schwarz, E., Komaromy, M., & Wall, R. (1984) *J. Mol. Biol.* 179, 125–142.
470. White, S.H., & Wimley, W.C. (1999) *Annu. Rev. Biophys. Biomol. Struct.* 28, 319–365.
471. Vergeres, G., Winterhalter, K.H., & Richter, C. (1989) *Biochemistry* 28, 3650–3655.
472. Wester, M.R., Johnson, E.F., Marques-Soares, C., Dansette, P.M., Mansuy, D., & Stout, C.D. (2003) *Biochemistry* 42, 6370–6379.
473. Haniu, M., Armes, L.G., Tanaka, M., Yasunobu, K.T., Shastry, B.S., Wagner, G.C., & Gunsalus, I.C. (1982) *Biochem. Biophys. Res. Commun.* 105, 889–894.
474. Gonzalez, F.J., Nebert, D.W., Hardwick, J.P., & Kasper, C.B. (1985) *J. Biol. Chem.* 260, 7435–7441.
475. Jayasinghe, S., Hristova, K., & White, S.H. (2001) *J. Mol. Biol.* 312, 927–934.
476. Weber, A., Menzlaff, E., Arbing, B., Gutensohn, M., Eckerskorn, C., & Flugge, U.I. (1995) *Biochemistry* 34, 2621–2627.
477. Green, G.N., Fang, H., Lin, R.J., Newton, G., Mather, M., Georgiou, C.D., & Gennis, R.B. (1988) *J. Biol. Chem.* 263, 13138–13143.
478. Yamaguchi, M., Hatefi, Y., Trach, K., & Hoch, J.A. (1988) *J. Biol. Chem.* 263, 2761–2767.
479. Erni, B., & Zanolari, B. (1986) *J. Biol. Chem.* 261, 16398–16403.
480. Scarpati, E.M., Wen, D., Broze, G.J., Jr., Miletich, J.P., Flandermeier, R.R., Siegel, N.R., & Sadler, J.E. (1987) *Biochemistry* 26, 5234–5238.
481. Grandy, D.K., Marchionni, M.A., Makam, H., Stofko, R.E., Alfano, M., Frothingham, L., Fischer, J.B., Burke-Howie, K.J., Bunzow, J.R., Server, A.C., et al. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 9762–9766.
482. Nicoll, D.A., Longoni, S., & Philipson, K.D. (1990) *Science* 250, 562–565.
483. Runswick, M.J., Walker, J.E., Bisaccia, F., Iacobazzi, V., & Palmieri, F. (1990) *Biochemistry* 29, 11033–11040.
484. Fushimi, K., Uchida, S., Hara, Y., Hirata, Y., Marumo, F., & Sasaki, S. (1993) *Nature* 361, 549–552.
485. Mackinnon, R. (2005) *Science* 307, 1425–1426.
486. Bayley, H., & Knowles, J.R. (1980) *Biochemistry* 19, 3883–3892.
487. Prochaska, L., Bisson, R., & Capaldi, R.A. (1980) *Biochemistry* 19, 3174–3179.
488. Bercovici, T., & Gitler, C. (1978) *Biochemistry* 17, 1484–1489.
489. White, B.H., & Cohen, J.B. (1988) *Biochemistry* 27, 8741–8751.
490. Malatesta, F., Darley-Usmar, V., de Jong, C., Prochaska, L.J., Bisson, R., Capaldi, R.A., Steffens, G.C., & Buse, G. (1983) *Biochemistry* 22, 4405–4411.
491. Nicholas, R.A. (1984) *Biochemistry* 23, 888–898.
492. Merrill, A.R., & Cramer, W.A. (1990) *Biochemistry* 29, 8529–8534.
493. Brunner, J., & Richards, F.M. (1980) *J. Biol. Chem.* 255, 3319–3329.
494. Ross, A.H., Radhakrishnan, R., Robson, R.J., & Khorana, H.G. (1982) *J. Biol. Chem.* 257, 4152–4161.
495. Bisson, R., & Montecucco, C. (1981) *Biochem. J.* 193, 757–763.
496. Blanton, M.P., & Cohen, J.B. (1992) *Biochemistry* 31, 3738–3750.
497. Brunner, J., & Semenza, G. (1981) *Biochemistry* 20, 7174–7182.
498. Brunner, J., Spiess, M., Aggeler, R., Huber, P., & Semenza, G. (1983) *Biochemistry* 22, 3812–3820.
499. Burnett, B.K., Robson, R.J., Takagaki, Y., Radhakrishnan, R., & Khorana, H.G. (1985) *Biochim. Biophys. Acta* 815, 57–67.
500. Brunner, J., Franzusoff, A.J., Luescher, B., Zugliani, C., & Semenza, G. (1985) *Biochemistry* 24, 5422–5430.
501. Kyte, J. (1987) in *Perspectives in Biological Energy Transduction* (Mukohata, Y., Morales, M.F., & Fleischer, S., Eds.) Vol. 11, pp 231–239, Academic Press, Tokyo.
502. Schneider, D.L., Burnside, J., Gorga, F.R., & Nettleton, C.J. (1978) *Biochem. J.* 176, 75–82.
503. Evans, R.M., & Fink, L.M. (1977) *Proc. Natl. Acad. Sci. U.S.A.* 74, 5341–5344.
504. Sharkey, R.G. (1983) *Biochim. Biophys. Acta* 730, 327–341.
505. Schnaitman, C., & Greenawalt, J.W. (1968) *J. Cell Biol.* 38, 158–175.
506. Clarke, S. (1976) *J. Biol. Chem.* 251, 1354–1363.
507. Hackenbrock, C.R., & Hammon, K.M. (1975) *J. Biol. Chem.* 250, 9185–9197.
508. St. John, P.A., Froehner, S.C., Goodenough, D.A., & Cohen, J.B. (1982) *J. Cell Biol.* 92, 333–342.
509. Karpen, J.W., & Hess, G.P. (1986) *Biochemistry* 25, 1777–1785.
510. Forbush, B.D. (1982) *J. Biol. Chem.* 257, 12678–12684.
511. Dwyer, B.P. (1991) *Biochemistry* 30, 4105–4112.
512. Fung, B.K., & Hubbell, W.L. (1978) *Biochemistry* 17, 4403–4410.
513. O'Connell, M.A. (1982) *Biochemistry* 21, 5984–5991.
514. Berg, H.C. (1969) *Biochim. Biophys. Acta* 183, 65–78.
515. Bretscher, M.S. (1971) *J. Mol. Biol.* 58, 775–781.
516. Whiteley, N.M., & Berg, H.C. (1974) *J. Mol. Biol.* 87, 541–561.
517. Hundle, B.S., & Richards, W.R. (1990) *Biochemistry* 29, 6172–6179.
518. Cabantchik, I.Z., Balshin, M., Breuer, W., & Rothstein, A. (1975) *J. Biol. Chem.* 250, 5130–5136.
519. Bogner, W., Aquila, H., & Klingenberg, M. (1982) *FEBS Lett.* 146, 259–261.
520. Kyte, J., Xu, K.Y., & Bayer, R. (1987) *Biochemistry* 26, 8350–8360.

521. Bangham, A.D., & Horne, R.W. (1962) *Nature* 196, 952–953.
522. Zhou, J., Fazzio, R.T., & Blair, D.F. (1995) *J. Mol. Biol.* 251, 237–242.
523. Akabas, M.H., Stauffer, D.A., Xu, M., & Karlin, A. (1992) *Science* 258, 307–310.
524. Zhang, H., & Karlin, A. (1998) *Biochemistry* 37, 7952–7964.
525. Javitch, J.A., Shi, L., Simpson, M.M., Chen, J., Chiappa, V., Visiers, I., Weinstein, H., & Ballesteros, J.A. (2000) *Biochemistry* 39, 12190–12199.
526. Dutton, A., & Singer, S.J. (1975) *Proc. Natl. Acad. Sci. U.S.A.* 72, 2568–2571.
527. Brock, C.J., Tanner, M.J., & Kempf, C. (1983) *Biochem. J.* 213, 577–586.
528. Kalo, M.S. (1996) *Biochemistry* 35, 999–1009.
529. Bender, W.W., Garan, H., & Berg, H.C. (1971) *J. Mol. Biol.* 58, 783–797.
530. Jennings, M.L., Adams-Lackey, M., & Denney, G.H. (1984) *J. Biol. Chem.* 259, 4652–4660.
531. Tanner, M.J., Martin, P.G., & High, S. (1988) *Biochem. J.* 256, 703–712.
532. LaRochelle, W.J., Wray, B.E., Sealock, R., & Froehner, S.C. (1985) *J. Cell Biol.* 100, 684–691.
533. Nilsson, I., Saaf, A., Whitley, P., Gafvelin, G., Waller, C., & von Heijne, G. (1998) *J. Mol. Biol.* 284, 1165–1175.
534. Ahmad, M., & Bussey, H. (1988) *Mol. Microbiol.* 2, 627–635.
535. Chavez, R.A., & Hall, Z.W. (1991) *J. Biol. Chem.* 266, 15532–15538.
536. Kuma, H., Shinde, A.A., Howren, T.R., & Jennings, M.L. (2002) *Biochemistry* 41, 3380–3388.
537. Popov, M., Tam, L.Y., Li, J., & Reithmeier, R.A. (1997) *J. Biol. Chem.* 272, 18325–18332.
538. Manoil, C. (1991) *Methods Cell Biol.* 34, 61–75.
539. Boyd, D., Traxler, B., & Beckwith, J. (1993) *J. Bacteriol.* 175, 553–556.
540. Broome-Smith, J.K., & Spratt, B.G. (1986) *Gene* 49, 341–349.
541. Wang, R.C., Seror, S.J., Blight, M., Pratt, J.M., Broome-Smith, J.K., & Holland, I.B. (1991) *J. Mol. Biol.* 217, 441–454.
542. Zelazny, A., & Bibi, E. (1996) *Biochemistry* 35, 10872–10878.
543. Yun, C.H., Van Doren, S.R., Crofts, A.R., & Gennis, R.B. (1991) *J. Biol. Chem.* 266, 10967–10973.
544. Breyer, R.M., Strosberg, A.D., & Guillet, J.G. (1990) *EMBO J.* 9, 2679–2684.
545. Tanner, K.G., & Kyte, J. (1999) *J. Biol. Chem.* 274, 35985–35990.
546. Canessa, C.M., Horisberger, J.D., Louvard, D., & Rossier, B.C. (1992) *EMBO J.* 11, 1681–1687.
547. Loo, T.W., & Clarke, D.M. (1994) *Biochemistry* 33, 14049–14057.
548. Kutsuwada, T., Kashiwabuchi, N., Mori, H., Sakimura, K., Kushiya, E., Araki, K., Meguro, H., Masaki, H., Kumanishi, T., Arakawa, M., et al. (1992) *Nature* 358, 36–41.
549. Arbuckle, M.I., Kane, S., Porter, L.M., Seatter, M.J., & Gould, G.W. (1996) *Biochemistry* 35, 16519–16527.
550. Lasalde, J.A., Tamamizu, S., Butler, D.H., Vibat, C.R., Hung, B., & McNamee, M.G. (1996) *Biochemistry* 35, 14139–14148.
551. Kuwahara, M., Shinbo, I., Sato, K., Terada, Y., Marumo, F., & Sasaki, S. (1999) *Biochemistry* 38, 16340–16346.
552. Ho, M.K., & Guidotti, G. (1975) *J. Biol. Chem.* 250, 675–683.
553. Steck, T.L., Ramos, B., & Strapazon, E. (1976) *Biochemistry* 15, 1153–1161.
554. Grinstein, S., Ship, S., & Rothstein, A. (1978) *Biochim. Biophys. Acta* 507, 294–304.
555. Kopito, R.R., & Lodish, H.F. (1985) *Nature* 316, 234–238.
556. Mawby, W.J., & Findlay, J.B. (1982) *Biochem. J.* 205, 465–475.
557. Jay, D.G. (1986) *Biochemistry* 25, 554–556.
558. Erickson, H.K., & Kyte, J. (1998) *Biochem. J.* 336, 443–449.
559. Jennings, M.L., Anderson, M.D., & Monaghan, R. (1986) *J. Biol. Chem.* 261, 9002–9010.
560. Erickson, H.K. (1997) *Biochemistry* 36, 9958–9967.
561. Lieberman, D.M., & Reithmeier, R.A. (1988) *J. Biol. Chem.* 263, 10022–10028.
562. Zhang, Y.Z., Georgevich, G., & Capaldi, R.A. (1984) *Biochemistry* 23, 5616–5621.
563. Prochaska, L.J., Bisson, R., Capaldi, R.A., Steffens, G.C., & Buse, G. (1981) *Biochim. Biophys. Acta* 637, 360–373.
564. Malatesta, F., & Capaldi, R. (1982) *Biochem. Biophys. Res. Commun.* 109, 1180–1185.
565. Jarausch, J., & Kadenbach, B. (1985) *Eur. J. Biochem.* 146, 219–225.
566. Ewalt, K.L. (1994) *Biochemistry* 33, 5077–5088.
567. Kyte, J. (1981) *Nature* 292, 201–204.
568. Serrano, R., Kielland-Brandt, M.C., & Fink, G.R. (1986) *Nature* 319, 689–693.
569. Walderhaug, M.O., Post, R.L., Saccomani, G., Leonard, R.T., & Briskin, D.P. (1985) *J. Biol. Chem.* 260, 3852–3859.
570. Niggli, V., Penniston, J.T., & Carafoli, E. (1979) *J. Biol. Chem.* 254, 9955–9958.
571. Filoteo, A.G., Gorski, J.P., & Penniston, J.T. (1987) *J. Biol. Chem.* 262, 6526–6530.
572. Castro, J., & Farley, R.A. (1979) *J. Biol. Chem.* 254, 2221–2228.
573. Stewart, P.S., & MacLennan, D.H. (1976) *J. Biol. Chem.* 251, 712–719.
574. MacLennan, D.H., Brandl, C.J., Korczak, B., & Green, N.M. (1985) *Nature* 316, 696–700.
575. Bastide, F., Meissner, G., Fleischer, S., & Post, R.L. (1973) *J. Biol. Chem.* 248, 8385–8391.
576. Anderberg, S.J. (1995) *Biochemistry* 34, 9508–9516.
577. Thibault, D. (1993) *Biochemistry* 32, 2813–2821.
578. Matsumoto, D.C.T. (1994) Ph.D. Thesis, Department of Chemistry, University of California at San Diego, La Jolla, CA.
579. Shimon, M.B., Goldshleger, R., & Karlsh, S.J. (1998) *J. Biol. Chem.* 273, 34190–34195.
580. Mandala, S.M., & Slayman, C.W. (1989) *J. Biol. Chem.* 264, 16276–16281.
581. James, P., Maeda, M., Fischer, R., Verma, A.K., Krebs, J., Penniston, J.T., & Carafoli, E. (1988) *J. Biol. Chem.* 263, 2905–2910.
582. Sun, J., & Kaback, H.R. (1997) *Biochemistry* 36, 11959–11965.

583. Wu, J., Hardy, D., & Kaback, H.R. (1999) *Biochemistry* 38, 1715–1720.
584. Whitley, P., Nilsson, L., & von Heijne, G. (1993) *Biochemistry* 32, 8534–8539.
585. Pakula, A.A., & Simon, M.I. (1992) *Proc. Natl. Acad. Sci. U.S.A.* 89, 4144–4148.
586. Frillingos, S., Sahin-Toth, M., Lengeler, J.W., & Kaback, H.R. (1995) *Biochemistry* 34, 9368–9373.
587. Krupinski, J., Coussen, F., Bakalyar, H.A., Tang, W.J., Feinstein, P.G., Orth, K., Slaughter, C., Reed, R.R., & Gilman, A.G. (1989) *Science* 244, 1558–1564.
588. Bakalyar, H.A., & Reed, R.R. (1990) *Science* 250, 1403–1406.
589. Feinstein, P.G., Schrader, K.A., Bakalyar, H.A., Tang, W.J., Krupinski, J., Gilman, A.G., & Reed, R.R. (1991) *Proc. Natl. Acad. Sci. U.S.A.* 88, 10173–10177.
590. Premont, R.T., Chen, J., Ma, H.W., Ponnappalli, M., & Iyengar, R. (1992) *Proc. Natl. Acad. Sci. U.S.A.* 89, 9809–9813.
591. Gao, B.N., & Gilman, A.G. (1991) *Proc. Natl. Acad. Sci. U.S.A.* 88, 10178–10182.
592. Katsushika, S., Chen, L., Kawabe, J., Nilakantan, R., Halnon, N.J., Homcy, C.J., & Ishikawa, Y. (1992) *Proc. Natl. Acad. Sci. U.S.A.* 89, 8774–8778.
593. Thorens, B., Sarkar, H.K., Kaback, H.R., & Lodish, H.F. (1988) *Cell* 55, 281–290.
594. Phelps, A., Schobert, C.T., & Wohlrab, H. (1991) *Biochemistry* 30, 248–252.
595. Hediger, M.A., Turk, E., & Wright, E.M. (1989) *Proc. Natl. Acad. Sci. U.S.A.* 86, 5748–5752.
596. Gouaux, J.E., Braha, O., Hobaugh, M.R., Song, L., Cheley, S., Shustak, C., & Bayley, H. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 12828–12831.
597. Parker, M.W., Postma, J.P., Pattus, F., Tucker, A.D., & Tsernoglou, D. (1992) *J. Mol. Biol.* 224, 639–657.
598. Shin, Y.K., Levinthal, C., Levinthal, F., & Hubbell, W.L. (1993) *Science* 259, 960–963.
599. Zhang, R.G., Westbrook, M.L., Westbrook, E.M., Scott, D.L., Otwinowski, Z., Maulik, P.R., Reed, R.A., & Shipley, G.G. (1995) *J. Mol. Biol.* 251, 550–562.
600. Sixma, T.K., Pronk, S.E., Kalk, K.H., Wartna, E.S., van Zanten, B.A., Witholt, B., & Hol, W.G. (1991) *Nature* 351, 371–377.
601. Stein, P.E., Boodhoo, A., Tyrrell, G.J., Brunton, J.L., & Read, R.J. (1992) *Nature* 355, 748–750.
602. Ling, H., Boodhoo, A., Hazes, B., Cummings, M.D., Armstrong, G.D., Brunton, J.L., & Read, R.J. (1998) *Biochemistry* 37, 1777–1788.
603. Sixma, T.K., Stein, P.E., Hol, W.G., & Read, R.J. (1993) *Biochemistry* 32, 191–198.
604. Sixma, T.K., Pronk, S.E., Kalk, K.H., van Zanten, B.A., Berghuis, A.M., & Hol, W.G. (1992) *Nature* 355, 561–564.
605. Ribi, H.O., Ludwig, D.S., Mercer, K.L., Schoolnik, G.K., & Kornberg, R.D. (1988) *Science* 239, 1272–1276.
606. Stenmark, H., McGill, S., Olsnes, S., & Sandvig, K. (1989) *EMBO J.* 8, 2849–2853.
607. Choe, S., Bennett, M.J., Fujii, G., Curmi, P.M., Kantardjieff, K.A., Collier, R.J., & Eisenberg, D. (1992) *Nature* 357, 216–222.
608. Deeley, R.G., Mullinix, D.P., Wetekam, W., Kronenberg, H.M., Meyers, M., Eldridge, J.D., & Goldberger, R.F. (1975) *J. Biol. Chem.* 250, 9060–9066.
609. Yamamura, J., Adachi, T., Aoki, N., Nakajima, H., Nakamura, R., & Matsuda, T. (1995) *Biochim. Biophys. Acta* 1244, 384–394.
610. Timmins, P.A., Poliks, B., & Banaszak, L. (1992) *Science* 257, 652–655.
611. Raag, R., Appelt, K., Xuong, N.H., & Banaszak, L. (1988) *J. Mol. Biol.* 200, 553–569.
612. Anderson, T.A., Levitt, D.G., & Banaszak, L.J. (1998) *Structure* 6, 895–909.
613. Thompson, J.R., & Banaszak, L.J. (2002) *Biochemistry* 41, 9398–9409.
614. Spring, D.J., Chen-Liu, L.W., Chatterton, J.E., Elovson, J., & Schumaker, V.N. (1992) *J. Biol. Chem.* 267, 14839–14845.
615. Chatterton, J.E., Phillips, M.L., Curtiss, L.K., Milne, R.W., Marcel, Y.L., & Schumaker, V.N. (1991) *J. Biol. Chem.* 266, 5955–5962.
616. Orlova, E.V., Sherman, M.B., Chiu, W., Mowri, H., Smith, L.C., & Gotto, A.M., Jr. (1999) *Proc. Natl. Acad. Sci. U.S.A.* 96, 8420–8425.
617. Borhani, D.W., Rogers, D.P., Engler, J.A., & Brouillette, C.G. (1997) *Proc. Natl. Acad. Sci. U.S.A.* 94, 12291–12296.
618. Wald, J.H., Krul, E.S., & Jonas, A. (1990) *J. Biol. Chem.* 265, 20037–20043.
619. Wald, J.H., Coormaghtigh, E., De Meutter, J., Ruysschaert, J.M., & Jonas, A. (1990) *J. Biol. Chem.* 265, 20044–20050.
620. Caspar, D.L.D., & Kirschner, D.A. (1971) *Nat. New Biol.* 231, 46–52.
621. Wilkins, M.H., Blaurock, A.E., & Engelman, D.M. (1971) *Nat. New Biol.* 230, 72–76.
622. Eletr, S., & Inesi, G. (1972) *Biochim. Biophys. Acta* 282, 174–179.
623. Hubbell, W.L., & McConnell, H.M. (1969) *Proc. Natl. Acad. Sci. U.S.A.* 64, 20–27.
624. McConnell, H.M., Wright, K.L., & McFarland, B.G. (1972) *Biochem. Biophys. Res. Commun.* 47, 273–281.
625. Esfahani, M., Limbrick, A.R., Knutton, S., Oka, T., & Wakil, S.J. (1971) *Proc. Natl. Acad. Sci. U.S.A.* 68, 3180–3184.
626. Treabue, H., & Overath, P. (1973) *Biochim. Biophys. Acta* 307, 491–512.
627. Overath, P., & Treabue, H. (1973) *Biochemistry* 12, 2625–2634.
628. Bretscher, M.S. (1972) *Nat. New Biol.* 236, 11–12.
629. Verkleij, A.J., Zwaal, R.F., Roelofsen, B., Comfurius, P., Kastelijn, D., & Deenen, L.L.V. (1973) *Biochim. Biophys. Acta* 323, 178–193.
630. Butikofer, P., Lin, Z.W., Chiu, D.T., Lubin, B., & Kuypers, F.A. (1990) *J. Biol. Chem.* 265, 16035–16038.
631. Gordesky, S.E., Marinetti, G.V., & Love, R. (1975) *J. Membr. Biol.* 20, 111–132.
632. Johnson, L.W., Hughes, M.E., & Zilversmit, D.B. (1975) *Biochim. Biophys. Acta* 375, 176–185.
633. Bloj, B., & Zilversmit, D.B. (1976) *Biochemistry* 15, 1277–1283.
634. Kamp, H.H., Wirtz, K.W.A., & VanDeenan, L.L.M. (1973) *Biochim. Biophys. Acta* 318, 313–325.

635. Wimley, W.C., & Thompson, T.E. (1990) *Biochemistry* 29, 1296–1303.
636. Williamson, P., Bevers, E.M., Smeets, E.F., Comfurius, P., Schlegel, R.A., & Zwaal, R.F. (1995) *Biochemistry* 34, 10448–10455.
637. McIntyre, J.C., & Sleight, R.G. (1991) *Biochemistry* 30, 11819–11827.
638. Renooij, W., Van Golde, L.M., Zwaal, R.F., & Van Deenen, L.L. (1976) *Eur. J. Biochem.* 61, 53–58.
639. Rothman, J.E., & Kennedy, E.P. (1977) *J. Mol. Biol.* 110, 603–618.
640. Krebs, J.J., Hauser, H., & Carafoli, E. (1979) *J. Biol. Chem.* 254, 5308–5316.
641. Blau, L., & Bittman, R. (1978) *J. Biol. Chem.* 253, 8366–8368.
642. Fisher, K.A. (1976) *Proc. Natl. Acad. Sci. U.S.A.* 73, 173–177.
643. Auland, M.E., Roufogalis, B.D., Devaux, P.F., & Zachowski, A. (1994) *Proc. Natl. Acad. Sci. U.S.A.* 91, 10938–10942.
644. Pomorski, T., Lombardi, R., Riezman, H., Devaux, P.F., van Meer, G., & Holthuis, J.C. (2003) *Mol. Biol. Cell* 14, 1240–1254.
645. Zachowski, A., Herrmann, A., Paraf, A., & Devaux, P.F. (1987) *Biochim. Biophys. Acta* 897, 197–200.
646. Bevers, E.M., Tilly, R.H., Senden, J.M., Comfurius, P., & Zwaal, R.F. (1989) *Biochemistry* 28, 2382–2387.
647. Zachowski, A., Henry, J.P., & Devaux, P.F. (1989) *Nature* 340, 75–76.
648. Tang, X., Halleck, M.S., Schlegel, R.A., & Williamson, P. (1996) *Science* 272, 1495–1497.
649. Kornberg, R.D., & McConnell, H.M. (1971) *Biochemistry* 10, 1111–1120.
650. McNamee, M.G., & McConnell, H.M. (1973) *Biochemistry* 12, 2951–2958.
651. Rothman, J.E., & Kennedy, E.P. (1977) *Proc. Natl. Acad. Sci. U.S.A.* 74, 1821–1825.
652. Basse, F., Stout, J.G., Sims, P.J., & Wiedmer, T. (1996) *J. Biol. Chem.* 271, 17205–17210.
653. Zhou, Q., Sims, P.J., & Wiedmer, T. (1998) *Biochemistry* 37, 2356–2360.
654. Simons, K., & Ikonen, E. (1997) *Nature* 387, 569–572.
655. Xu, X., & London, E. (2000) *Biochemistry* 39, 843–849.
656. Ahmed, S.N., Brown, D.A., & London, E. (1997) *Biochemistry* 36, 10944–10953.
657. Radhakrishnan, A., Anderson, T.G., & McConnell, H.M. (2000) *Proc. Natl. Acad. Sci. U.S.A.* 97, 12422–12427.
658. Brown, D.A., & Rose, J.K. (1992) *Cell* 68, 533–544.
659. Hanada, K., Nishijima, M., Akamatsu, Y., & Pagano, R.E. (1995) *J. Biol. Chem.* 270, 6254–6260.
660. Dietrich, C., Volovyk, Z.N., Levi, M., Thompson, N.L., & Jacobson, K. (2001) *Proc. Natl. Acad. Sci. U.S.A.* 98, 10642–10647.
661. Parton, R.G. (1996) *Curr. Opin. Cell Biol.* 8, 542–548.
662. Casey, P.J. (1995) *Science* 268, 221–225.
663. Vereb, G., Matko, J., Vamosi, G., Ibrahim, S.M., Magyar, E., Varga, S., Szollosi, J., Jenei, A., Gaspar, R., Jr., Waldmann, T.A., & Damjanovich, S. (2000) *Proc. Natl. Acad. Sci. U.S.A.* 97, 6013–6018.
664. Sheets, E.D., Lee, G.M., Simson, R., & Jacobson, K. (1997) *Biochemistry* 36, 12449–12458.
665. Frye, C.D., & Ediden, M. (1970) *J. Cell. Sci.* 7, 313–336.
666. Schmidt, T., Schutz, G.J., Baumgartner, W., Gruber, H.J., & Schindler, H. (1996) *Proc. Natl. Acad. Sci. U.S.A.* 93, 2926–2929.
667. Saffman, P.G., & Delbreuck, M. (1975) *Proc. Natl. Acad. Sci. U.S.A.* 72, 3111–3113.
668. Cherry, R.J., & Godfrey, R.E. (1981) *Biophys. J.* 36, 257–276.
669. Musier-Forsyth, K.M., & Hammes, G.G. (1990) *Biochemistry* 29, 3236–3241.
670. Peters, R., & Cherry, R.J. (1982) *Proc. Natl. Acad. Sci. U.S.A.* 79, 4317–4321.
671. Axelrod, D., Koppel, D.E., Schlessinger, J., Elson, E., & Webb, W.W. (1976) *Biophys. J.* 16, 1055–1069.
672. Schlessinger, J., Koppel, D.E., Axelrod, D., Jacobson, K., Webb, W.W., & Elson, E.L. (1976) *Proc. Natl. Acad. Sci. U.S.A.* 73, 2409–2413.
673. Vaz, W.L., Criado, M., Madeira, V.M., Schoellmann, G., & Jovin, T.M. (1982) *Biochemistry* 21, 5608–5612.
674. Chang, C.H., Takeuchi, H., Ito, T., Machida, K., & Ohnishi, S. (1981) *J. Biochem. (Tokyo)* 90, 997–1004.
675. Poo, M., & Cone, R.A. (1974) *Nature* 247, 438–441.
676. Schlessinger, J., Axelrod, D., Koppel, D.E., Webb, W.W., & Elson, E.L. (1977) *Science* 195, 307–309.
677. Chazotte, B., & Hackenbrock, C.R. (1988) *J. Biol. Chem.* 263, 14359–14367.
678. Mullineaux, C.W., Tobin, M.J., & Jones, G.R. (1997) *Nature* 390, 421–424.
679. Vaz, W.L., Clegg, R.M., & Hallmann, D. (1985) *Biochemistry* 24, 781–786.
680. Criado, M., Vaz, W.L., Barrantes, F.J., & Jovin, T.M. (1982) *Biochemistry* 21, 5750–5755.
681. Scandella, C.J., Devaux, P., & McConnell, H.M. (1972) *Proc. Natl. Acad. Sci. U.S.A.* 69, 2056–2060.
682. Hughes, B.D., Pailthorpe, B.A., White, L.R., & Sawyer, W.H. (1982) *Biophys. J.* 37, 673–676.
683. Galla, H.J., Hartmann, W., Theilen, U., & Sackmann, E. (1979) *J. Membr. Biol.* 48, 215–236.
684. Shinitzky, M., Dianoux, A.C., Gitler, C., & Weber, G. (1971) *Biochemistry* 10, 2106–2113.
685. Cogan, U., Shinitzky, M., Weber, G., & Nishida, T. (1973) *Biochemistry* 12, 521–528.
686. Shinitzky, M., & Inbar, M. (1974) *J. Mol. Biol.* 85, 603–615.
687. Kinoshita, K., Jr., Kawato, S., Ikegami, A., Yoshida, S., & Orii, Y. (1981) *Biochim. Biophys. Acta* 647, 7–17.
688. Farkas, T., Kitajka, K., Fodor, E., Csengeri, I., Lahdes, E., Yeo, Y.K., Krasznai, Z., & Halver, J.E. (2000) *Proc. Natl. Acad. Sci. U.S.A.* 97, 6362–6366.
689. Garrett, T.P., McKern, N.M., Lou, M., Elleman, T.C., Adams, T.E., Lovrecz, G.O., Zhu, H.J., Walker, F., Frenkel, M.J., Hoyne, P.A., Jorissen, R.N., Nice, E.C., Burgess, A.W., & Ward, C.W. (2002) *Cell* 110, 763–773.
690. Ogiso, H., Ishitani, R., Nureki, O., Fukai, S., Yamanaka, M., Kim, J.H., Saito, K., Sakamoto, A., Inoue, M., Shirouzu, M., & Yokoyama, S. (2002) *Cell* 110, 775–787.
691. Stamos, J., Sliwkowski, M.X., & Eigenbrot, C. (2002) *J. Biol. Chem.* 277, 46265–46272.
692. Kashles, O., Szapary, D., Bellot, F., Ullrich, A., Schlessinger, J., & Schmidt, A. (1988) *Proc. Natl. Acad. Sci. U.S.A.* 85, 9567–9571.

693. Carpenter, C.D., Ingraham, H.A., Cochet, C., Walton, G.M., Lazar, C.S., Sowadski, J.M., Rosenfeld, M.G., & Gill, G.N. (1991) *J. Biol. Chem.* 266, 5750–5755.
694. Cochet, C., Kashles, O., Chambaz, E.M., Borrello, I., King, C.R., & Schlessinger, J. (1988) *J. Biol. Chem.* 263, 3290–3295.
695. Fanger, B.O., Stephens, J.E., & Staros, J.V. (1989) *FASEB J.* 3, 71–75.
696. Yarden, Y., & Schlessinger, J. (1987) *Biochemistry* 26, 1434–1442.
697. Spaargaren, M., Defize, L.H., Boonstra, J., & de Laat, S.W. (1991) *J. Biol. Chem.* 266, 1733–1739.
698. Sherrill, J.M. (1997) *Biochemistry* 36, 5677–5684.
699. Canals, F. (1992) *Biochemistry* 31, 4493–4501.
700. Tanner, K.G. (1997) *Biochemistry* 36, 14889–14896.
701. Sherrill, J.M., & Kyte, J. (1996) *Biochemistry* 35, 5705–5718.
702. Plotnikov, A.N., Schlessinger, J., Hubbard, S.R., & Mohammadi, M. (1999) *Cell* 98, 641–650.
703. Cunningham, B.C., Ultsch, M., De Vos, A.M., Mulkerrin, M.G., Clauser, K.R., & Wells, J.A. (1991) *Science* 254, 821–825.
704. de Vos, A.M., Ultsch, M., & Kossiakoff, A.A. (1992) *Science* 255, 306–312.
705. Fuh, G., Cunningham, B.C., Fukunaga, R., Nagata, S., Goeddel, D.V., & Wells, J.A. (1992) *Science* 256, 1677–1680.
706. Benovic, J.L., Shorr, R.G., Caron, M.G., & Lefkowitz, R.J. (1984) *Biochemistry* 23, 4510–4518.
707. Dixon, R.A., Kobilka, B.K., Strader, D.J., Benovic, J.L., Dohlman, H.G., Frielle, T., Bolanowski, M.A., Bennett, C.D., Rands, E., Diehl, R.E., Mumford, R.A., Slater, E.E., Sigal, I.S., Caron, M.G., Lefkowitz, R.J., & Strader, C.D. (1986) *Nature* 321, 75–79.
708. Yarden, Y., Rodriguez, H., Wong, S.K., Brandt, D.R., May, D.C., Burnier, J., Harkins, R.N., Chen, E.Y., Ramachandran, J., Ullrich, A., et al. (1986) *Proc. Natl. Acad. Sci. U.S.A.* 83, 6795–6799.
709. Takeda, S., Kadowaki, S., Haga, T., Takaesu, H., & Mitaku, S. (2002) *FEBS Lett.* 520, 97–101.
710. Evers, A., & Klabunde, T. (2005) *J. Med. Chem.* 48, 1088–1097.
711. Zhang, G., Liu, Y., Ruoho, A.E., & Hurley, J.H. (1997) *Nature* 386, 247–253.
712. Tesmer, J.J., Sunahara, R.K., Gilman, A.G., & Sprang, S.R. (1997) *Science* 278, 1907–1916.
713. Tesmer, J.J., Sunahara, R.K., Johnson, R.A., Gosselin, G., Gilman, A.G., & Sprang, S.R. (1999) *Science* 285, 756–760.
714. Rodbell, M., Birnbaumer, L., Pohl, S.L., & Krans, H.M. (1971) *J. Biol. Chem.* 246, 1877–1882.
715. Pfeuffer, T. (1979) *FEBS Lett.* 101, 85–89.
716. Cassel, D., & Selinger, Z. (1977) *Proc. Natl. Acad. Sci. U.S.A.* 74, 3307–3311.
717. Cassel, D., & Selinger, Z. (1976) *Biochim. Biophys. Acta* 452, 538–551.
718. Pike, L.J., & Lefkowitz, R.J. (1980) *J. Biol. Chem.* 255, 6860–6867.
719. Cassel, D., Levkovitz, H., & Selinger, Z. (1977) *J. Cyclic Nucleotide Res.* 3, 393–406.
720. Cassel, D., Eckstein, F., Lowe, M., & Selinger, Z. (1979) *J. Biol. Chem.* 254, 9835–9838.
721. Cassel, D., & Selinger, Z. (1978) *Proc. Natl. Acad. Sci. U.S.A.* 75, 4155–4159.
722. Itoh, H., Kozasa, T., Nagata, S., Nakamura, S., Katada, T., Ui, M., Iwai, S., Ohtsuka, E., Kawasaki, H., Suzuki, K., & Kaziro, Y. (1986) *Proc. Natl. Acad. Sci. U.S.A.* 83, 3776–3780.
723. Robishaw, J.D., Russell, D.W., Harris, B.A., Smigel, M.D., & Gilman, A.G. (1986) *Proc. Natl. Acad. Sci. U.S.A.* 83, 1251–1255.
724. Wall, M.A., Coleman, D.E., Lee, E., Iniguez-Lluhi, J.A., Posner, B.A., Gilman, A.G., & Sprang, S.R. (1995) *Cell* 83, 1047–1058.
725. Sprang, S.R. (1997) *Annu. Rev. Biochem.* 66, 639–678.
726. Chen, C.A., & Manning, D.R. (2001) *Oncogene* 20, 1643–1652.
727. Orly, J., & Schramm, M. (1976) *Proc. Natl. Acad. Sci. U.S.A.* 73, 4410–4414.
728. Tolkovsky, A.M., & Levitzki, A. (1978) *Biochemistry* 17, 3795.
729. Pfeuffer, T. (1977) *J. Biol. Chem.* 252, 7224–7234.
730. Sternweis, P.C., Northup, J.K., Smigel, M.D., & Gilman, A.G. (1981) *J. Biol. Chem.* 256, 11517–11526.
731. Heithier, H., Frohlich, M., Dees, C., Baumann, M., Haring, M., Gierschik, P., Schiltz, E., Vaz, W.L., Hekman, M., & Helmreich, E.J. (1992) *Eur. J. Biochem.* 204, 1169–1181.
732. Limbird, L.E., Gill, D.M., & Lefkowitz, R.J. (1980) *Proc. Natl. Acad. Sci. U.S.A.* 77, 775–779.
733. Cerione, R.A., Codina, J., Benovic, J.L., Lefkowitz, R.J., Birnbaumer, L., & Caron, M.G. (1984) *Biochemistry* 23, 4519–4525.
734. Northup, J.K., Smigel, M.D., Sternweis, P.C., & Gilman, A.G. (1983) *J. Biol. Chem.* 258, 11369–11376.
735. Codina, J., Hildebrandt, J., Iyengar, R., Birnbaumer, L., Sekura, R.D., & Manclark, C.R. (1983) *Proc. Natl. Acad. Sci. U.S.A.* 80, 4276–4280.
736. Gilman, A.G. (1987) *Annu. Rev. Biochem.* 56, 615–649.
737. Hekman, M., Feder, D., Keenan, A.K., Gal, A., Klein, H.W., Pfeuffer, T., Levitzki, A., & Helmreich, E.J. (1984) *EMBO J.* 3, 3339–3345.
738. Fung, B.K. (1983) *J. Biol. Chem.* 258, 10495–10502.
739. Brandt, D.R., & Ross, E.M. (1986) *J. Biol. Chem.* 261, 1656–1664.
740. Tolkovsky, A.M., Braun, S., & Levitzki, A. (1982) *Proc. Natl. Acad. Sci. U.S.A.* 79, 213–217.
741. Sunahara, R.K., Dessauer, C.W., Whisnant, R.E., Kleuss, C., & Gilman, A.G. (1997) *J. Biol. Chem.* 272, 22265–22271.
742. Arad, H., Rosenbusch, J.P., & Levitzki, A. (1984) *Proc. Natl. Acad. Sci. U.S.A.* 81, 6579–6583.
743. Kazarov, A.R., Rozenkrants, A.A., & Sobolev, A.S. (1986) *Biokhimiya (Moscow)* 51, 355–363.
744. Calvert, P.D., Govardovskii, V.I., Krasnoperova, N., Anderson, R.E., Lem, J., & Makino, C.L. (2001) *Nature* 411, 90–94.
745. Tolkovsky, A.M., & Levitzki, A. (1978) *Biochemistry* 17, 3811–3817.
746. Pedersen, S.E., & Ross, E.M. (1982) *Proc. Natl. Acad. Sci. U.S.A.* 79, 7228–7232.
747. Elgsaeter, A., Shotton, D.M., & Branton, D. (1976) *Biochim. Biophys. Acta* 426, 101–122.
748. Steck, T.L. (1974) *J. Cell Biol.* 62, 1–19.

749. Sheetz, M.P., Painter, R.G., & Singer, S.J. (1976) *Biochemistry* 15, 4486–4492.
750. Marchesi, V.T., & Steers, E., Jr. (1968) *Science* 159, 203–204.
751. Rosenthal, A.S., Kregenow, F.M., & Moses, H.L. (1970) *Biochim. Biophys. Acta* 196, 254–262.
752. Gwynne, J.T., & Tanford, C. (1970) *J. Biol. Chem.* 245, 3269–3273.
753. Kam, Z., Josephs, R., Eisenberg, H., & Gratzer, W.B. (1977) *Biochemistry* 16, 5568–5572.
754. Clarke, M. (1971) *Biochem. Biophys. Res. Commun.* 45, 1063–1070.
755. Speicher, D.W., & Marchesi, V.T. (1984) *Nature* 311, 177–180.
756. Shotton, D.M., Burke, B.E., & Branton, D. (1979) *J. Mol. Biol.* 131, 303–329.
757. Brenner, S.L., & Korn, E.D. (1979) *J. Biol. Chem.* 254, 8620–8627.
758. Byers, T.J., & Branton, D. (1985) *Proc. Natl. Acad. Sci. U.S.A.* 82, 6153–6157.
759. Shen, B.W., Josephs, R., & Steck, T.L. (1986) *J. Cell Biol.* 102, 997–1006.
760. Rana, A.P., Ruff, P., Maalouf, G.J., Speicher, D.W., & Chishti, A.H. (1993) *Proc. Natl. Acad. Sci. U.S.A.* 90, 6651–6655.
761. Husain-Chishti, A., Faquin, W., Wu, C.C., & Branton, D. (1989) *J. Biol. Chem.* 264, 8985–8991.
762. Tyler, J.M., Hargreaves, W.R., & Branton, D. (1979) *Proc. Natl. Acad. Sci. U.S.A.* 76, 5192–5196.
763. Tyler, J.M., Reinhardt, B.N., & Branton, D. (1980) *J. Biol. Chem.* 255, 7034–7039.
764. Ungewickell, E., Bennett, P.M., Calvert, R., Ohanian, V., & Gratzer, W.B. (1979) *Nature* 280, 811–814.
765. Fowler, V., & Taylor, D.L. (1980) *J. Cell Biol.* 85, 361–376.
766. Ohanian, V., Wolfe, L.C., John, K.M., Pinder, J.C., Lux, S.E., & Gratzer, W.B. (1984) *Biochemistry* 23, 4416–4420.
767. Tsuji, A., Kawasaki, K., Ohnishi, S., Merkle, H., & Kusumi, A. (1988) *Biochemistry* 27, 7447–7452.
768. Bennett, V., & Stenbuck, P.J. (1980) *J. Biol. Chem.* 255, 2540–2548.
769. Bennett, V. (1978) *J. Biol. Chem.* 253, 2292–2299.
770. Bennett, V., & Stenbuck, P.J. (1980) *J. Biol. Chem.* 255, 6424–6432.
771. Bennett, V. (1982) *Biochim. Biophys. Acta* 689, 475–484.
772. Tsuji, A., & Ohnishi, S. (1986) *Biochemistry* 25, 6133–6139.
773. Korsgren, C., & Cohen, C.M. (1988) *J. Biol. Chem.* 263, 10212–10218.
774. Cohen, A.M., Liu, S.C., Lawler, J., Derick, L., & Palek, J. (1988) *Biochemistry* 27, 614–619.
775. Mueller, T.J., & Morrison, M. (1981) in *Erythrocyte Membranes 2: Recent Clinical and Experimental Advances* (Kruckeberg, W.C., Eaton, J.W., & Brewer, G. J., Eds.) pp 95–112, Alan R. Liss, New York.
776. Anderson, R.A., & Lovrien, R.E. (1984) *Nature* 307, 655–658.
777. Young, R.W. (1970) *Sci. Am.* 223, 80–91.
778. Kraulis, P.J. (1991) *J. Appl. Crystallogr.* 24, 946–950.

Index

A

- absorption of light, 592–613
 - elastic scattering, 593
 - electronic energy levels, 592–95
 - electronically excited state, 594
 - energy levels, 592–95
 - infrared light, 594
 - intersystem crossing, 595
 - Raman effect, 593
 - ultraviolet absorption spectra, 601
 - ultraviolet light, 594
 - vibrational energy levels, 592–95
 - visible light, 594
- absorption spectrum
 - fluorescence, 595
- α actinin
 - heterologous associations, 513, 516
- α -carbon diagram
 - crystallographic molecular model, 167
- acceptor
 - fluorescence resonance energy transfer, 604
- accessible surface area
 - hydration of a protein, 299
 - of amino acids, 276
 - of crystallographic molecular model, 273
 - of glyceraldehyde-3-phosphate dehydrogenase, 273
- acetate anion
 - acids and bases, 64
- acetate CoA-transferase
 - aligning amino acid sequence, 360
- acetic acid
 - titration curve, 67
- acetic anhydride, 546
- acetolactate synthase III
 - aligning amino acid sequence, 360
- acetone powder, 23
- N*-acetyl α -amides, 75
 - hydropathy, 245
 - of amino acids, 245
- acetylcholine receptor
 - α helix, 259
 - boundary layer of phospholipid, 786
 - covalent modification from within the bilayer, 797
 - crystallization, 772
 - cystines, 783
 - degradation by endopeptidases, 432
 - domains, 779
 - fluorescence resonance energy transfer, 608
 - image reconstruction, 793
 - membrane-spanning α helices, 772, 794
 - quaternary structure, 407, 451
 - rotational axis of
 - pseudosymmetry, 787–88
 - sequence of DNA, 106–7
 - sequencing of DNA, 101–2
 - topography of membrane-spanning proteins, 802
 - translational diffusion coefficient, 813
- acetylcholine-binding protein
 - point group, 469
- acetylcholinesterase
 - glycosylphosphatidylinositol-linked proteins, 765
- acetyl-CoA carboxylase
 - purification, 49
 - subunits, 437
- N*-acetyl-D-galactosamine
 - structure, 129
- N*-acetylgalactosaminidemucin- β -1,3-galactosyltransferase
 - purification, 29
- N*-acetylgalactosaminyltransferase
 - anchored membrane-bound proteins, 764
- N*-acetyl-D-glucosamine
 - structure, 129
- N*-acetylglucosamine kinase
 - purification, 29
- N*⁴-(β -*N*-acetylglucosaminyl)-L-asparaginase
 - posttranslational modification, 114
- acid
 - definition, 62
- acid dissociation constant
 - amino acids, 75–76
 - arginine, 75
 - asparagine, 75
 - aspartate, 75
 - cysteine, 75
 - effect of ionic interactions in
 - crystallographic molecular models, 300
 - glutamate, 75
 - glutamine, 75
 - histidine, 75
 - lysine, 75
 - nuclear magnetic resonance, 635–38
 - serine, 75
 - threonine, 75
 - tryptophan, 75
 - tyrosine, 75
- α_1 -acid glycoprotein
 - oligosaccharides on glycoproteins, 131
 - purification, 29
- acid hydrolysis
 - cleavage of polypeptide, 87
- acid peptidases, 49
- acid-base titration curve
 - nuclear magnetic resonance, 635
 - random coil, 660
 - ribonuclease, 33
- acid-base titration of a protein, 32
- acids and bases, 62–69
 - acetate anion, 64
 - aromaticity, 63
 - central atom, 62
 - conservation of charge, 66
 - conservation of mass, 66
 - delocalization, 63
 - hybridization, 63
 - induction, 63
 - nucleoside bases, 65
 - rehybridization, 63
 - titration curve, 66
 - water constant, 66
- acrosin inhibitor IIA
 - nuclear magnetic resonance, 626, 629, 631

- actin, 729–30
 - axes of symmetry, 452–53
 - covalent modification, 547
 - cross-linking, 549
 - cytoskeleton, 820
 - fluorescence resonance energy transfer, 609
 - image reconstruction, 503
 - peptide map, 432
 - peripheral membrane-bound proteins, 764
 - thin filament, 506
- activation volume
 - kinetics of folding, 703
- active sites, functional groups in
 - aligning crystallographic molecular models, 372
- active transport
 - phosphatidylethanolamine, 809
 - phosphatidylserine, 809
- acyl carrier protein
 - domains, 382
- acyl oxygen
 - electronic structure, 60
- acylating agents
 - reagents for covalent modification, 535
- acylation with fatty acid
 - anchored membrane-bound proteins, 765
- [acyl-carrier-protein] S-acyltransferase
 - domains, 381
- [acyl-carrier-protein] S-malonyltransferase
 - electrophoresis, 46
 - purification, 46
- acyl-CoA-binding protein
 - kinetics of folding, 697
- acylphosphatase
 - crystallization, 49
 - kinetics of folding, 702
 - proline isomerization, 701
- ADA regulatory protein
 - nuclear magnetic resonance, 632
- adenine
 - electronic structure, 65
- adenine–thymine pairs
 - flexibility, 321
- adenosine kinase
 - aligning crystallographic molecular models, 362–63
 - molecular taxonomy, 396
- S-adenosylmethionine decarboxylase
 - posttranslational modification, 114
- adenosylmethionine–8-amino-7-oxononate transaminase
 - molecular taxonomy, 396
- adenylate cyclase
 - diffusion, 817
 - domains, 817
 - purification, 29
- adenylate cyclase system, 816–20
 - adenylate cyclase, 816
 - β -adrenergic receptor, 816
 - collision of proteins, 819
 - complex between α subunit of the stimulatory G-protein and adenylate cyclase, 819
 - complex between β -adrenergic receptor and G-protein, 819
 - coupling, 818
 - G-protein, 817
 - kinetic mechanism, 819
 - kinetics, 818
- adenylate kinase
 - aligning crystallographic molecular models, 364
 - molecular taxonomy, 395
- adenylosuccinate lyase
 - interfaces, 480
- adenylyl-sulfate kinase
 - aligning crystallographic molecular models, 364
- ADP, ATP carrier
 - topography of membrane-spanning proteins, 799
- ADP-ribose diphosphatase
 - interfaces, 481
- α_1 -adrenergic receptor
 - binding of ligands, 47
- β -adrenergic agonist, 816
- β -adrenergic receptor
 - adenylate cyclase system, 816
 - diffusion, 817
 - integral membrane-bound protein, 816
 - purification, 29, 816
- adsorption chromatography
 - purification, 25
- affinity adsorption
 - agarose, 26
 - hydrophilic spacers, 27
 - purification, 26
- affinity elution
 - purification, 26
- agarose
 - affinity adsorption, 26
- aggreCAN
 - domains, 386
- aggregation
 - electrophoresis, 46
- α helices
 - angles between, 279
- crystallography, 165–67
 - intramolecular hydrogen bonds, 228
 - packing of side chains between, 279–85
- α helix
 - acetylcholine receptor, 259
 - amino-terminal end, 257
 - amphipathic, 259
 - capped, 257
 - carboxy-terminal end, 257
 - circular dichroism, 598
 - curved, 256
 - cytochrome *c*, 256
 - dihydrofolate reductase, 257
 - dipole, 258
 - dragline silk, 259
 - helical wheel, 258–59
 - hemoglobin, 259
 - left-handed, 256
 - membrane-spanning, 774, 776
 - nuclear magnetic resonance, 631
 - proline, 257
 - right-handed, 256
 - 2-fold rotational axis of symmetry, 280
 - secondary structure, 256–59
 - serine and threonine, 256
 - unsupported in water, 229
 - water, 256
- air and water interface
 - monolayer of lipids at, 761
- alanine
 - electronic structure, 76
- alanine dehydrogenase
 - convergent evolution, 373
- alanine-tRNA ligase
 - metalloproteins, 331
- alcohol dehydrogenase
 - β structure, 261
 - diffusion coefficient, 578
 - frictional coefficient, 578
 - frictional ratio, 578
 - interfaces, 480
 - molecular taxonomy, 395
 - multiple isomorphous replacement, 161
 - recurring structure, 373–74
 - sedimentation coefficient, 578
 - space groups, 463
 - water in crystallographic molecular models, 294
- aldehyde dehydrogenase
 - electrospray mass spectrometry, 417
- aldehyde:ferredoxin oxidoreductase
 - metalloproteins, 330

- aldehydes
 - reagents for covalent modification, 534
- algorithms for searching data banks
 - aligning amino acid sequences, 354
- aligning amino acid sequences, 346–61
 - acetate CoA-transferase, 360
 - acetolactate synthase III, 360
 - algorithms for searching data banks, 354
 - alignment score, 352
 - angiogenin, 361
 - antithrombin III, 360
 - argininosuccinate lyase, 361
 - azurin, 360
 - Ca²⁺-transporting ATPase, 364
 - carbonate dehydratase, 360
 - chymotrypsinogen, 360
 - computational alignment, 351
 - conservative replacement, 349
 - crystallin, 361
 - cytochrome *c*, 346–49, 351–52, 354–55, 360
 - cytochrome *f*, 360
 - 2-dehydro-3-deoxy-phosphogluconate aldolase, 360
 - dihydrolipoyllysine
 - acetyltransferase, 360
 - dihydrolipoyllysine
 - succinyltransferase, 360
 - dot matrix, 351–52
 - evolutionary distance, 355, 358
 - fibrinopeptides, 349
 - gap, 350–51
 - gap penalty, 353
 - gap percentage, 351
 - genetic code, 349
 - globins, 356
 - glucarate dehydratase, 361
 - H⁺/K⁺-exchanging ATPase, 364
 - H⁺-transporting two-sector ATPase, 360
 - haptoglobulin, 360
 - hemoglobin, 360
 - histocompatibility antigens, 360
 - homologues, 346
 - 4-hydroxy-2-oxoglutarate aldolase, 360
 - immunoglobulin, 360
 - invariant position, 348
 - jumbled amino acid sequences, 353
 - lactalbumin, 360
 - lamin A, 360
 - leghemoglobin, 360
 - lysozyme, 360
 - mandelate racemase, 361
 - matrix, 351
 - methylmalonyl-CoA
 - decarboxylase, 360
 - methylmalonyl-CoA mutase, 360
 - β 2-microglobulin, 353, 360
 - minimal mutational distance, 356
 - most parsimonious sequence of event, 356
 - myoglobin, 345, 360
 - Na⁺/K⁺-exchanging ATPase, 364
 - ovalbumin, 360
 - parvalbumin, 360
 - path of diagonal segments, 353
 - percentage of identity, 351
 - phylogenetic tree, 354–55
 - plastocyanin, 360
 - progressive alignment, 356–57
 - Protein Information Resource, 354
 - recent speciation, 357
 - record of intolerance, 348
 - ribonuclease, 361
 - searching data banks of amino acid sequences, 353–54
 - speciation of organisms, 354
 - statistical significance, 353
 - subtilisin, 360
 - succinate–propionate CoA-transferase, 360
 - Swiss-Prot sequence database, 354
 - tartronate-semialdehyde synthase, 360
 - tripeptidyl-peptidase II, 360
 - troponin *c*, 360
 - trypsinogen, 360
 - vacuolar H⁺-transporting two-sector ATPase, 360
 - vimentin, 360
 - weighting schemes, 351
- aligning crystallographic molecular models, 362–76
 - active sites, functional groups in, 372
 - adenosine kinase, 362–63
 - adenylate kinase, 364
 - adenylyl-sulfate kinase, 364
 - aspartate-semialdehyde dehydrogenase, 366
 - basic fibroblast growth factor, 366
 - chymotrypsin, 375
 - convergent evolution, 372
 - creatinase, 362–63
 - cytochromes *c*, 364–65
 - distantly related proteins, 366
 - elastase, 362–63
 - erythrocyruorin, 369
 - gaps, 364
 - globins, 369–72
 - 4- α -glucanotransferase, 362
 - glutathione synthase, 362
 - glyceraldehyde-3-phosphate dehydrogenase, 366
 - guanylate kinase, 364
 - hemoglobin, 369–72
 - hexokinase, 363
 - hydrophobic clusters, 369
 - interleukin1 β , 366
 - leghemoglobin, 369
 - lysozyme, 363
 - maltodextrin binding protein, 363
 - mandelate racemase, 366
 - methionyl aminopeptidase, 362–63
 - myoglobin, 369–72
 - P1 nuclease, 366
 - 6-phosphofructo-2-kinase, 364
 - phospholipase, 372
 - phospholipase C, 366
 - phosphopyruvate hydratase, 366
 - phosphoribosylamine–glycine ligase, 362
 - pyruvate kinase, 375
 - recurring structure, 373
 - ribokinase, 362
 - root mean square deviation, 362
 - structural alignment, 366
 - superoxide dismutase, 363
 - superposition, 362
 - thiamin pyridinylase, 363
 - tryptase, 362–63
- alignment score
 - aligning amino acid sequences, 352
- alkaline phosphatase
 - immunostaining, 566
 - topography of membrane-spanning proteins, 800
- alkaline phosphatase promoter
 - expressing DNA, 109
- alkanal monooxygenase (FMN-linked)
 - assembly of oligomers, 714
- alkyl glycosides
 - detergents, 769
- alkyl oligo(ethylene oxide) ethers
 - detergents, 768
- alkylhalidase
 - molecular taxonomy, 396
- alternative conformations
 - crystallography, 182
 - stereochemistry of side chains, 267
- alternative splicing
 - evolution of proteins, 350
- Alzheimer's disease, 508
- amide
 - electronic structure, 57–58
 - hydrogen bond in water, 217

- hydropathy, 242
 - infrared spectroscopy, 595
- amide I band, 595
 - secondary structure, 596
- amide II band, 595
- amide III band, 595
- amido proton exchange
 - random coil, 660
- amido protons
 - proton exchange, 640
- amine dehydrogenase
 - electron paramagnetic resonance, 647
- amino acid analysis, 11
- amino acid composition, 91
- amino acid residue
 - definition, 74
- amino acid sequence
 - bithorax complex, 106
 - filaggrin, 108
 - histidine–proline-rich glycoprotein, 106
 - refinement, 178
 - ribonuclease, 85
 - spider dragline silk, 106
 - vitellogenin, 106
- amino acids, 74–83
 - accessible surface areas of, 276
 - N*-acetyl- α -amides of, 245
 - acid dissociation constants of, 75–76
 - helical propensity, 259
 - helix-breaking, 259
 - helix-forming, 259
 - shape, 164
- amino terminus, 74
 - immunostaining, 566
 - posttranslational modification, 117
- 2-amino-4-(fluorophosphono) butanoic acid, 550
- aminocyclopropane carboxylate oxidase
 - electron nuclear double resonance, 650
 - electron paramagnetic resonance, 646
- aminodeoxychorismate synthase
 - domains, 380
- 5-aminolevulinate synthase
 - expressing DNA, 108
- 5-aminopentanamidase
 - assay, 20
- o*-aminotyrosine, 538
- ammonium sulfate precipitation
 - purification, 23
- amorphous ice
 - image reconstruction, 501, 790
- amphipathic α helix
 - anchored membrane-bound proteins, 764
- amphipathic lipid
 - definition, 745
- amplitude modulation
 - nuclear magnetic resonance, 619
- amplitude of the reflection
 - crystallography, 153
- amplitude of the structure factor
 - crystallography, 155
- amplitudes
 - image reconstruction, 790
- α -amylase
 - mass spectrometry, 92
 - sieving, 428
- α -amylase inhibitor HOE-467A
 - nuclear magnetic resonance, 631
 - proton exchange, 644
- β -amylase
 - sieving, 427
- anchored membrane-bound proteins, 764–65
 - acylation with fatty acid, 765
 - amino terminus, 764
 - amphipathic α helix, 764
 - carboxy terminus, 764
 - carboxypeptidase E, 764
 - coagulation factor V, 765
 - coagulation factor VIII, 765
 - cytochrome c_1 , 764
 - detachable domains, 773
 - domain, 764
 - dopamine- β -monooxygenase, 764
 - embedded anchor, 773
 - glycosylphosphatidylinositol-linked proteins, 765
 - 3-hydroxybutyrate dehydrogenase, 764
 - hydroxymethylglutaryl-CoA reductase, 764
 - interior segment, 765
 - isoprenylation, 765
 - (*S*)-mandelate dehydrogenase, 765
 - membrane-spanning α helices, 774
 - membrane-spanning anchor, 764
 - N*-acetylgalactosaminyltransferase, 764
 - receptor Tom 20, 764
 - signal sequences, 764
- angiogenin
 - aligning amino acid sequences, 361
- angular dependence
 - hydrogen bonds, 264–66
- anion exchange chromatography
 - sequencing oligosaccharides, 134
- anisotropic thermal parameters
 - crystallography, 176
- ankyrin
 - cytoskeleton, 821
 - heterologous associations, 516
 - peptide map, 434
 - peripheral membrane-bound protein, 821
- ankyrin domain, 386
- annexin
 - heterologous associations, 514
 - peripheral membrane-bound proteins, 764
- anomalous dispersion
 - multiple isomorphous replacement, 161
- anthranilate
 - phosphoribosyltransferase dodecyl sulfate gel electrophoresis, 432
- anthranilate synthase
 - dodecyl sulfate gel electrophoresis, 432
 - domains, 380
- antibodies
 - immunoglobulins, 555
- antibonding molecular orbital, 57
- antifreeze protein
 - domains, 384
- antigen
 - definition, 555
 - epitope on, 558
 - hapten attached to, 562
 - large tumor, 562
 - polyvalent, 564
 - synthetic peptide, 562–64
- antilone* pair, 69
- antisense strand
 - sequencing of DNA, 98
- antithrombin III
 - aligning amino acid sequences, 360
- apoferritin
 - frictional ratio, 426
 - molar mass, 418
 - multiple isomorphous replacement, 161
 - sieving, 424, 427
- apolipoprotein B100
 - length, 85
 - lipoproteins, 805
- apolipoprotein(a)
 - diffusion coefficient, 577–78
 - frictional coefficient, 577–78
 - frictional ratio, 577–78
 - sedimentation coefficient, 577–78
- apomyoglobin
 - kinetics of folding, 689, 693, 695, 697

- molten globule, 684
- apparent surface area of a protein sieving, 423
- approach to equilibrium kinetics of folding, 666
- approximation, 222–30
 - entropy of, 224
 - folding, 681
 - hydrogen bonds in crystallographic molecular models, 311
- aquaporin
 - crystallization, 772
 - crystallographic molecular model, 788
 - image reconstruction, 793
 - membrane-spanning α helices, 772
 - rotational axis of pseudosymmetry, 788
- L-arabinose
 - structure, 129
- arabinose-binding protein
 - molecular rotational axes of pseudosymmetry, 476
 - molecular taxonomy, 395
- AraC protein
 - domains, 378
- arachidonate 15-lipoxygenase
 - π helix in, 260
- arachidonic acid, 747
- arc* repressor
 - assembly of oligomers, 712
 - ionic interactions in crystallographic molecular models, 304
 - kinetics of folding, 703
 - nucleic acid, association of proteins with, 316
 - rotational axes of symmetry, 468
- archaeobacterial isoprenylether lipid phospholipid, 748
- architecture
 - molecular taxonomy, 396
- arginase
 - metalloproteins, 326
- arginine
 - acid dissociation constant, 75
 - association of proteins with nucleic acid, 315
 - covalent modification, 539
 - electronic structure, 80
 - hydropathy, 242
 - in interfaces, 478
 - water in crystallographic molecular models, 296
- argininosuccinate lyase
 - aligning amino acid sequences, 361
- arginyl endopeptidase
 - cleavage of polypeptide, 88
- armadillo domain, 386
- aromatic amino acids
 - hydropathy side chains, 275
 - stereochemistry of side chains, 269
- aromatic nitrogen heterocycles
 - electronic structure, 60–61
- aromatic ring
 - hydrogen bond, 208
- aromatic side chain
 - hydrogen bonds in crystallographic molecular models, 306
- aromaticity
 - acids and bases, 63
 - electronic structure, 60
- arrangement of subunits
 - cross-linking, 445
- arthroside
 - glycosphingolipid, 748
- aryl azides
 - reagents for covalent modification, 541
- aryl nitrenes
 - reagents for covalent modification, 542
- aryl-acylamidase
 - purification, 21
- asparagine
 - acid dissociation constant, 75
 - electronic structure, 79
 - hydropathy, 276
 - stereochemistry of side chains, 270
 - water in crystallographic molecular models, 296
- asparagine and adenine
 - hydrogen bonding, 317
- asparagine synthase
 - molecular taxonomy, 393
- aspartate
 - acid dissociation constant, 75
 - covalent modification, 539–41
 - stereochemistry of side chains, 270
- aspartate 1-decarboxylase
 - posttranslational modification, 114
- aspartate carbamoyltransferase
 - assembly of oligomers, 713, 714
 - circular permutation, 680
 - conformational change, 577
 - domains, 379
 - electron microscopy, 587
 - fluorescence resonance energy transfer, 608
 - frictional ratio, 577
 - heterologous interfaces, 508–10
 - heterooligomers, 508–10
- metalloproteins, 326, 332
- molar mass, 418, 419
- quaternary structure, 451
- sedimentation velocity, 576
- X-ray scattering, 584
- aspartate kinase
 - domains, 378
- aspartate kinase I-homoserine dehydrogenase I
 - domains, 381
 - kinetics of folding, 709
 - molar mass, 418, 420
 - sieving, 427
- aspartate transaminase
 - assembly of oligomers, 712
 - domains, 389
 - molecular taxonomy, 396
- aspartate-semialdehyde dehydrogenase
 - aligning crystallographic molecular models, 366
 - molecular taxonomy, 396
- aspartate-tRNA ligase
 - molecular taxonomy, 393
- aspartic acid
 - electronic structure, 79
 - water in crystallographic molecular models, 296
- aspartyl endopeptidase
 - crystallography, 182
 - domains, 384
- aspartyl imide, 115
- assay of proteins, 13–20
 - accuracy, 19
 - 5-aminopentanamidase, 20
 - biological, 19
 - chromatographic separation, 14
 - cis*-aconitase, 19
 - coenzymes in, 13
 - coenzyme A, 18
 - colorimetric, 18
 - continuous, 15
 - coupled, 16
 - cyclosporin synthase, 14
 - fumarate hydratase, 13, 19
 - galactonate dehydratase, 19
 - geranyltranstransferase, 14
 - glutamine-pyruvate transaminase, 19
 - glyceraldehyde-3-phosphate dehydrogenase (phosphorylating), 17
 - Hurler corrective factor, 19
 - 2-hydroxy-6-ketonona-2,4-diene-1,9-dioic acid 5,6-hydrolase, 18
 - 3-hydroxyacyl-CoA dehydrogenase, 17

- hydroxymethylglutaryl-CoA lyase, 17
- 2-hydroxyphytanoyl-CoA lyase, 13
- imidazoleglycerol-phosphate dehydratase, 17
- interference, 16
- maturation-promoting factor, 19
- medium-chain acyl-CoA dehydrogenase, 16
- membrane-bound proteins, 771
- methylamine-glutamate *N*-methyltransferase, 14
- 2-methyleneglutarate mutase, 16
- monophenol monooxygenase, 18
- myosin subfragment 1, 17
- (*S*)-pantolactone dehydrogenase, 18
- pH, 13
- phosphofructokinase, 19
- phosphomevalonate kinase, 17
- protocatechuate 3,4-dioxygenase, 16
- pyruvate carboxylase, 17
- radioactive reactant, 14
- receptors, 15
- reconstitution, 773
- ribose-phosphate diphosphokinase, 18
- selectivity, 19
- selenocysteine lyase, 19
- sensitivity, 19
- succinyldiaminopimelate transaminase, 20
- triacylglycerol lipase, 16
- tryptophan-tRNA ligase, 14
- assembly
 - definition, 659
- assembly map
 - ribosome, 716
- assembly of actin
 - F-actin capping protein, 730
 - fragmin, 730
 - gelsolin, 730
 - hydrolysis of ATP, 729
 - nebulin, 730
 - nucleation, 730
 - thin filament, 729
 - villin, 730
 - vinculin, 730
- assembly of fibrin
 - cross-linking, 721
 - fibrinogen, 717–20
 - irreversible, 721
 - kinetics, 720
 - protein-glutamine γ -glutamyltransferase, 721
 - protofibril, 720
- assembly of helical polymers, 717–33
- assembly of microtubules
 - apparent critical concentration, 727
 - bulk concentration, 724
 - catastrophic depolymerization, 728
 - centrosome, 723
 - colchicine, 729
 - critical concentration, 724–25
 - elongation, 723
 - hydrolysis of GTP, 727
 - kinetics, 724–26
 - microtubule-associated proteins, 730
 - minus end, 725
 - nucleation, 723
 - number concentration, 724
 - plus end, 725
 - protofilament, 721
 - seeds, 723
 - steady state, 727
 - tubulin, 722
- assembly of oligomers, 710–17
 - alkanal monooxygenase (FMN-linked), 714
 - arc* repressor, 712
 - aspartate carbamoyltransferase, 713–14
 - aspartate transaminase, 712
 - cross-linking for assay, 445
 - dihydrolipoyl dehydrogenase, 715
 - dihydrolipoyllysine-residue acetyltransferase, 715
 - dimer, 712
 - enzymatic activity, 713
 - fructose-bisphosphate aldolase, 713
 - fumarate hydratase, 713
 - heterooligomers, 713–17
 - HIV-1 retropepsin, 713
 - inorganic diphosphatase, 710
 - kinetics of, 711
 - loosely folded monomer, 713
 - malate dehydrogenase, 712
 - molten globule, 712
 - phosphoglycerate mutase, 710–11, 713
 - porphobilinogen synthase, 713
 - pyruvate dehydrogenase (acetyl-transferring), 715
 - pyruvate dehydrogenase complex, 715
 - quantitative cross-linking, 710–11, 713
 - ribosome, 715–17
 - steric crowding, 715
 - steroid Δ -isomerase, 712
 - tetramer, 710–11
- trimer, 713
- tryptophan synthase, 714
- assignments
 - nuclear magnetic resonance, 621–28
- association constant
 - intramolecular, 222
- association equilibrium constant
 - hydrogen bond, 210
- asymmetry of phospholipids, 808–10
 - phospholipid-translocating ATPase, 809
 - phospholipid scramblase, 809
 - diphosphatidylglycerol, 810
 - phosphatidylcholine, 810
 - phosphatidylethanolamine, 810
 - phosphatidylglycerol, 810
 - phosphatidylinositol, 810
 - phosphatidylserine, 810
- ATP diphosphatase
 - purification, 773
- ATP-dependent DNA helicase Rep
 - fluorescence resonance energy transfer, 610
- axes
 - crystallography, 151
- axes of symmetry, 451–56
 - actin, 452–53
 - fold of the symmetry, 452
 - helical polymer, 452
 - hexokinase, 454
 - malate dehydrogenase, 452–53
 - protocatechuate 3,4-dioxygenase, 454
 - rotational, 452
 - rotational axis of pseudosymmetry, 452
 - screw, 452
 - symmetry operations, 451
- axial ratio
 - collagen, 574–75
 - ellipsoid of revolution, 574
 - hydrodynamic particle, 574–75
- 1-azidopyrene
 - hydrophobic reagent for covalent modification, 797
- aziridine
 - reagent for covalent modification, 537
- azurin
 - aligning amino acid sequences, 360
 - metalloproteins, 331
- B**
 - β structure
 - nuclear magnetic resonance, 631
 - bacteriophage
 - cloning of DNA, 99

- T4 bacteriophage
helical surface lattice, 500
- bacteriorhodopsin
amplitude of electron diffraction, 792
 β barrel, 782
bound phospholipid, 784
covalent modification from within the bilayer, 797
crystallization, 772, 775
crystallographic molecular model, 777
electron diffraction, 791
electron spin resonance, 794
mean molar mass of an amino acid, 418
membrane-spanning helices, 772
phase of Fourier transform, 792
rotational axes of symmetry, 787
rotational diffusion coefficient, 812
translational diffusion coefficient, 813
- bad solvent
definition, 659
- Bam*HI site-specific
deoxyribonuclease
nucleic acid, association of proteins with, 321
- band 3 anion transport protein
cytoskeleton, 821
domains, 377
mean molar mass of an amino acid, 418
membrane-bound proteins, 767
topography of membrane-spanning proteins, 800–2, 806
vectorial insertion, 767
- bare zone
thick filament of myosin, 731–32
- barstar
kinetics of folding, 690, 695
- base
definition, 62
- base stacking
nucleic acid structure, 323
- basic fibroblast growth factor
aligning crystallographic molecular models, 366
- basic trypsin inhibitor
proton exchange, 642–44
- β barrel
 β structure, 260, 263
 α -hemolysin, 776
indole-3-glycerol-phosphate synthase, 262
integral membrane-bound proteins, 776
membrane-spanning, 776
packing of β structure, 285
porin OmpF, 782
red fluorescent protein, 287
retinol-binding protein, 287
- β bulge
 β structure, 260–62
- β -elimination
sequencing oligosaccharides, 133
- Bence-Jones protein
hydrogen bonds in crystallographic molecular models, 308
water in crystallographic molecular models, 294
- benzoate 4-monooxygenase
domains, 382
- benzoylformate decarboxylase
molecular taxonomy, 396
polyproline helix, 259
- benzyl bromide
reagent for covalent modification, 536
- β helix
 β structure, 261
2,3,4,5-tetrahydropyridine-2,6-dicarboxylate *N*-succinyl-transferase, 263
- bicelles
membrane-bound proteins, 773
- bicontinuous cubic phase
crystallization of integral membrane-bound proteins, 775
- bilayer of lipids, 745–63
cholesterol, 745, 759–60
cross-sectional area, 760
cross-sectional area of cholesterol, 759
distinct phases, 760
distribution of electron density, 760–61
mole fraction of cholesterol, 759
phospholipids, 745
width, 759
- bilayers of phospholipid
alkane proximal to the glyceryl groups, 758
conformation at the glyceryl backbone, 752
coordinated tilting of the hydrocarbon, 758
core of the hydrocarbon, 757
cross-sectional area, 754
crystallographic molecular models, 751–53
- deuterium nuclear magnetic resonance, 757
- electron density, 750
- electron spin resonance, 755–56
- electrostatic repulsion, 754
- heat of fusion, 755
- hydration, 751
- immiscible lipids, 755
- molecular motion, 756
- neutron diffraction, 750, 754
- neutron scattering density, 755
- nitroxyl fatty acid, 755–56
- order parameter, 756–57
- sliding fluctuations, 752
- solid and liquid, 754
- stereochemical paradox, 757–59
- steric effects of the hydration, 758
- steric repulsion, 751
- surface potential, 754
- width, 750
- X-ray diffraction, 750–51
- binding assays, 15
- biological assays, 19
- biological membranes
diffraction of X-radiation, 808
diffusion in, 811–13
phase transitions, 808
rafts in, 811
spin-labeled probes, 808
two-dimensional solution, 811
- biotin carboxylase
domains, 388
molecular taxonomy, 397
- biotin-dependent carboxylase
domains, 382
- N,N*-bis(2-hydroxyethyl)-2-aminoethanesulfonic acid
buffer, 68
- N,N*-bis(2-hydroxyethyl)glycine
buffer, 68
- 1,4-bis(2-sulfoethyl)piperazine
buffer, 68
- bisimidates, 441
- bithorax complex
amino acid sequence, 106
- BLAST, 354
- blunt ends
sequencing of DNA, 96
- β -*N*-acetylglucosamidase
sequencing oligosaccharides, 134
- bond angles
hydrogen bond, 206
- bond length
hydrogen bond, 205
- bonding
molecular orbital, 57

- bonding electrons
 - electronic structure, 56
- bonding molecular orbital
 - electronic structure, 56
- bound ions
 - molecular charge, 33
- boundary layer of phospholipid
 - acetylcholine receptor, 786
 - Ca²⁺-transporting ATPase, 785
 - cholesterol, 786
 - cytochrome-*c* oxidase, 786, 805
 - deuterium nuclear magnetic resonance spectroscopy, 786
 - electron spin resonance, 785
 - integral membrane-bound proteins, 784–86
 - nitroxylphosphatidylcholine, 785
 - rhodopsin, 786
- bovine pancreatic trypsin inhibitor
 - folding, 685
- bovine serum albumin
 - osmotic pressure, 419
- β -pleated sheet
 - secondary structure, 261–62
- β propeller
 - β structure, 260, 264
 - methanol dehydrogenase, 264
- Bragg spacing
 - crystallography, 158
- branching
 - oligosaccharides of glycoproteins, 128
- brominating agents
 - reagents for covalent modification, 539
- bromoperoxidase
 - tetrahedral symmetry, 487
- bromopyruvate
 - reagent for covalent modification, 549
- β sheets
 - packing of β structure, 285
- β structure, 285–87
 - alcohol dehydrogenase, 261
 - β barrel, 260, 263
 - β bulge, 260–62
 - β helix, 261
 - β -pleated sheet, 262
 - β propeller, 260, 264
- carbonate dehydratase, 261
- chymotrypsin, 261
- circular dichroism, 598
- concanavalin A, 261
- crystallography, 165–67
- fatty-acid-binding protein, 260
- gap, 260
- intramolecular hydrogen bonds, 228
- left-handed twist, 260
- micrococcal nuclease, 261
- secondary structure, 260–61
- btK kinase
 - domains, 386
- β turns
 - circular dichroism, 599
 - crambin, 265
 - crystallography, 165–67
 - definition, 261
 - intramolecular hydrogen bonds, 227
 - secondary structure, 261–64
 - type I, 263, 265
 - type II, 263
 - types of, 262
- buffers, 66
 - N,N*-bis(2-hydroxyethyl)-2-aminoethanesulfonic acid, 68
 - N,N*-bis(2-hydroxyethyl)glycine, 68
 - 1,4-bis(2-sulfoethyl)piperazine, 68
 - N*-[2-hydroxy-1,1-bis(hydroxymethyl)ethyl]glycine, 68
 - 1-(2-hydroxyethyl)-4-(3-sulfopropyl)piperazine, 68
 - N*-(2-sulfoethyl)cyclohexylamine, 68
 - N*-(2-sulfoethyl)morpholine, 68
 - N*-(3-sulfopropyl)-2-amino-1,3-dihydroxy-2-hydroxymethylpropane, 68
 - N*-(3-sulfopropyl)morpholine, 68
- bulk concentration
 - assembly of microtubules, 724
 - of polymer, 724
- bulk relative permittivity, 201
- κ bungarotoxin
 - interfaces, 480
 - point group, 466–67
- buoyant densities
 - cell fractionation, 743
- buoyant force
 - sedimentation velocity, 576
- buoyant mass
 - sedimentation velocity, 576
- buried hydrogen bonds
 - hydrogen bonds in
 - crystallographic molecular models, 306
 - proton exchange, 641
- buried ion pair
 - ionic interactions in
 - crystallographic molecular models, 303
- buried side chain, 273
- B* value
 - crystallography, 175
- C**
 - C₈E₅
 - detergent, 770
 - C₁₀E₆
 - detergent, 770
 - C₁₀E₈
 - detergent, 770
 - C₁₂E₆
 - detergent, 770
 - C₁₂E₈
 - detergent, 770
 - C₁₄E₈
 - detergent, 770
 - C₁₆E₈
 - detergent, 770
 - C2 domain, 386
 - CA protein
 - helical surface lattice, 500
 - CAD multienzyme complex
 - domains, 379, 390
 - E-cadherin
 - heterologous associations, 516
 - calcium
 - metalloproteins, 328–29
 - calculated amplitudes
 - crystallography, 172
 - calculated phases
 - crystallography, 173
 - caldesmon
 - frictional ratio, 577
 - calmodulin
 - evolution of proteins, 351
 - fluorescence resonance energy transfer, 608
 - nuclear magnetic resonance, 624
 - proton exchange, 645
 - calorimeter
 - thermodynamics of folding, 671
 - camel
 - immunoglobulin G, 559
 - carbamoyl-phosphate synthase
 - domains, 383
 - carbamoyl-phosphate synthase (ammonia)
 - domains, 379
 - carbamoyl-phosphate synthase (glutamine hydrolysing)
 - domains, 379
 - carbenes
 - insertion into nucleophiles, 543
 - intramolecular rearrangements, 543
 - reagents for covalent modification, 543
 - carbodiimides
 - reagents for covalent modification, 539–41

- carbonate dehydratase
 - aligning amino acid sequences, 360
 - β structure, 261
 - covalent modification, 547
 - dodecyl sulfate gel electrophoresis, 422
 - electrophoresis, 40
 - folding, 670
 - glycosylphosphatidylinositol-linked proteins, 765
 - molecular taxonomy, 393, 395
 - molten globule, 683
 - nuclear magnetic resonance, 635
- carbon–oxygen double bond
 - second hydrogen bond, 256
- carbonyl oxygen
 - electronic structure, 60
- carboxy terminus, 74
 - immunostaining, 566
 - posttranslational modification, 117
- carboxylesterase ESTA
 - interfaces, 478–80
- carboxylic acid
 - hydropathy, 242
- carboxymethyl cellulose
 - chromatography, 9
- 5-carboxymethyl-
 - 2-hydroxymuconate Δ -isomerase
 - molecular rotational axes of pseudosymmetry, 477
- carboxymethylenebutenolidase
 - molecular taxonomy, 396
- carboxypeptidase
 - molecular taxonomy, 395
- carboxypeptidase A
 - sequencing of polypeptides, 91
 - packing of α helices, 282
 - packing of side chains, 279
- carboxypeptidase B
 - sequencing of polypeptides, 91
- carboxypeptidase C
 - coiled coil of α helices, 283
 - folding, 679
 - stereochemistry of side chains, 271
- carboxypeptidase D
 - molecular taxonomy, 396
- carboxypeptidase E
 - anchored membrane-bound proteins, 764
- carrier frequency
 - nuclear magnetic resonance, 615
- cartoon
 - crystallographic molecular mode, 167
- cassettes
 - site-directed mutation, 111
- catabolite gene activator protein
 - association of proteins with nucleic acid, 320
- catalase
 - diffusion coefficient, 578
 - dodecyl sulfate gel electrophoresis, 422
 - domains, 388
 - electron nuclear double resonance, 650
 - frictional coefficient, 578
 - frictional ratio, 578
 - interfaces, 481
 - molar mass, 418
 - sedimentation coefficient, 578
 - sieving, 424
 - X-ray scattering, 583
- cathepsin D
 - purification, 29
- cathepsin K
 - molecular taxonomy, 393
- caveolin
 - in rafts, 811
- CD40 tumor necrosis factor receptor
 - heterologous associations, 514
- CDP-6-deoxy-L-threo- β -glycero-4-hexulose-3-dehydrase
 - electron paramagnetic resonance, 646
- cell fractionation, 743
 - buoyant densities, 743
 - chloroplasts, 744
 - differential centrifugation, 743
 - electron microscopy, 744
 - for purification of membrane-bound protein, 768
 - free-flow electrophoresis, 744
 - Golgi membranes, 744
 - isopycnic centrifugation, 744
 - lysosomes, 744
 - marker enzymes, 744
 - mitochondria, 744
 - peroxisomes, 744
 - plasma membrane, 744
 - precipitation, 744
 - rate sedimentation, 743
 - rough endoplasmic reticulum, 744
 - sedimentation coefficients, 743
- cell surface glycoprotein a-2
 - glycosylphosphatidylinositol-linked proteins, 765
- cell surface receptor CD2
 - kinetics of folding, 690
- cell wall, 743
- central atom
 - acids and bases, 62
- centrifugal potential
 - sedimentation equilibrium, 411
- centrosome
 - assembly of microtubules, 723
- ceramide
 - glycosphingolipid, 748
- cerebroside
 - glycosphingolipid, 748
- C^γ -endo conformation
 - proline, 270
- C^γ -exo conformation
 - proline, 270
- chaperone
 - folding, 705–8
- chaperone protein PapD
 - domains, 390
- chaperonin 60
 - cavity, 707
 - control of folding, 705–8
 - cross-linking, 444–45
 - hydrolysis of MgATP, 707
 - quaternary structure, 475
 - structure, 705
- charge-coupled device
 - crystallography, 155
- charged side chains
 - heterologous interfaces, 513
- chelation
 - ion, 203
- chemical method of sequencing
 - DNA, 103, 105
- chemical potential
 - osmotic pressure, 408
 - sedimentation equilibrium, 411
- chemical shift
 - nuclear magnetic resonance, 614
- chinese hamster ovary cells
 - expression of DNA, 110
- chitinase B
 - water in crystallographic molecular models, 294
- chloramphenicol O-acetyl transferase
 - ionic interactions in crystallographic molecular models, 302
- chloramphenicol O-acetyltransferase
 - fluorescence resonance energy transfer, 608
 - interfaces, 480
 - multiple isomorphous replacement, 160
 - point group, 468–69
 - space groups, 463
 - topography of membrane-spanning proteins, 800

- water in crystallographic molecular models, 294
 chloroacetamide, 546
N-chlorobenzenesulfonamide
 reagent for covalent modification, 538
 4-chlorobenzoyl-CoA dehalogenase
 interfaces, 481
p-chloromercuribenzoate
 reagent for covalent modification, 537
 chloroplasts, 743
 cell fractionation, 744
 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate
 detergent, 770
 cholera toxin
 punching a hole in a membrane, 804
 cholesterol
 bilayer of lipids, 745, 759–60
 effect on microviscosity, 814
 in boundary layer of phospholipid, 786
 cholesterol oxidase
 domains, 382
 spin-labeled phospholipids, 795
 water in crystallographic molecular models, 295
 choline *O*-acetyltransferase
 purification, 29
 choline-phosphate
 cytidyltransferase
 peripheral membrane-bound proteins, 764
 chorismate mutase
 convergent evolution, 372
 chromatogram
 definition, 4
 chromatography, 2–12
 adsorption, 11
 agarose, 8
 amino acid analysis, 11
 carboxymethyl cellulose, 9
 cellulose, 7
 chromatography by adsorption, 8
 column chromatography, 4
 countercurrent distribution, 5
 DEAE cellulose, 9
 definition, 3
 dextran, 8
 diameter of the particles, 6
 distribution of solute, 2
 Donnan formalism, 10
 elution volume, 4
 flow rate, 6
 gas-liquid chromatography, 4
 glycopeptides, 133
 gradient chromatography, 7
 high-pressure liquid chromatography, 6
 hydroxylapatite, 8
 included volume, 12
 interfacial denaturation, 3
 ion exchange, 8
 ionic double layer, 8
 irreversible adsorption, 3
 isocratic zonal chromatography, 7
 of membrane-bound proteins, 771
 mobile phase, 2
 molecular exclusion, 11
 paper chromatography, 4
 partition coefficient, 4
 peptide map, 433
 peptide separation, 90
 polymethacrylate, 8
 polystyrene, 7
 QAE cellulose, 9
 relative mobility, 4
 resolution, 4
 reverse-phase, 8
 saturation, 2–3
 selective adsorption, 3
 silica gel, 7
 stationary phase, 2
 theoretical plate, 4–5
 thin-layer, 4
 titration of charge, 10
 void volume, 4
 zonal chromatography, 4
 chylomicrons
 lipoproteins, 804
 chymotrypsin
 aligning crystallographic molecular models, 375
 β structure, 261
 cleavage of polypeptide, 88
 collisional quenching, 603
 covalent modification, 543
 free energy of folding, 674, 677
 hydration, 298
 hydrogen bonds in crystallographic molecular model, 306, 308
 chymotrypsin inhibitor
 free energy of folding, 677
 kinetics of folding, 702
 chymotrypsinogen
 aligning amino acid sequences, 360
 covalent modification, 547
 dodecyl sulfate gel electrophoresis, 422
 endopeptidolytic cleavage, 547
 frictional ratio, 426
 hydration, 298
 mean molar mass of an amino acid, 418
 molecular rotational axes of pseudosymmetry, 476
 sieving, 424
 thermodynamics of folding, 673
 X-ray scattering, 583
 circular dichroism, 597–601
 α helix, 598
 β structure, 598
 β turn, 599
 circular dichroic spectrum, 598
 circular polarizations, 597
 conformational change, 600
 cytochrome-*c* oxidase, 598
 cytochrome *c*₁, 598–99
 glyceraldehyde-3-phosphate dehydrogenase (phosphorylating), 600
 glycine hydroxymethyltransferase, 598
 kinetics of folding, 689
 molar ellipticity, 598
 molten globule, 684
 Na⁺/K⁺-transporting ATPase, 600
 optical rotation, 598
 peptides, 601
 phenylalanines, 598
 plane-polarized light, 597
 protein coat of Hepatitis B virus, 612
 random coil, 660
 random meander, 599
 subtilisin, 600
 tryptophans, 598
 tyrosines, 598
 circular permutation, 679–80
 aspartate carbamoyltransferase, 680
 dihydrofolate reductase, 680
 thiol:disulfide interchange protein dsbA, 680
 circular polarizations
 circular dichroism, 597
cis peptide bonds
 lectin IV, 252
 proline isomerization, 698
 secondary structure, 251–52
cis-aconitase
 assay, 19
 citraconic anhydride
 reagent for covalent modification, 536
 citrate (*si*) synthase
 crystallography, 171
 clathrates
 hydrophobic effect, 233
 clathrin, 498
 light scattering, 590

- CLC-0 chloride channel domains, 389
- cleavage of polypeptide
acid hydrolysis, 87
arginyl endopeptidase, 88
chymotrypsin, 88
cyanogen bromide, 87, 89
glutamyl endopeptidase, 88
hydroxylamine, 90
lysyl endopeptidase, 88
2-nitro-5-thiocyanatobenzoate, 87, 89
papain, 88
peptidyl-Asp metalloendo-peptidase, 88
thermolysin, 88
trypsin, 88
- cloning of DNA
bacteriophage, 99
colony, 99
complementary DNA, 98
DNA ligase (ATP), 96
DNA ligase (NAD⁺), 96
DNA-directed DNA polymerase, 97
DNA-directed RNA polymerase, 97
extensin, 100
hybridization of DNA, 100
library, 99
plaques, 99
plasmid, 99
polymerase chain reaction, 100
preparative electrophoresis, 101
probe, 100
RNA-directed DNA polymerase, 97
sequencing of DNA, 99
- closed structure
definition, 454
integral membrane-bound proteins, 787
- coagulation factor V
anchored membrane-bound protein, 765
- coagulation factor Va
electron microscopy, 585
- coagulation factor VIII
anchored membrane-bound protein, 765
- cobalt
metalloproteins, 330
- coenzymatic domain
domains, 382
- coenzyme
assay, 13
refinement, 180
- coenzyme A
assay, 18
- coenzyme-B
sulfoethylthiotransferase
posttranslational modification, 113
- cohesin domain, 386
hydropathy of side chains, 275
molecular taxonomy, 397
- coiled coil of α helices
carboxypeptidase C, 283
core of, 284
extracellular matrix protein COMP, 283
fibrinogen, 282, 717
general control protein GCN4, 283–84
heptad repeat, 282
hydrophobic amino acids in, 283
intermediate filaments, 506
keratin, 282
methyl-accepting chemotaxis protein, 284
myosin, 282, 730
packing of side chains, 279–85
synthetic peptides that form, 284
- coincident structure
molecular taxonomy, 396
- co-ion
definition, 8
- colchicine
assembly of microtubules, 729
- cold shock protein CspB
free energy of folding, 674
- cold shock-like protein
equilibrium constant for folding, 664
kinetics of folding, 664, 667, 702
- colicin E1
fluorescence, 603
punching a hole in a membrane, 803
- colicin E7 immunity protein
kinetics of folding, 694
- collagen
axial ratio, 574–75
frictional ratio, 575, 577
helical polymer, 503–6
4-hydroxyproline, 504
interstrand hydrogen bond, 504
intrinsic viscosity, 579
posttranslational modification, 122–23
proline isomerization, 701
triple helix, 504
viscosity, 579
- collagen type VI
domains, 386
- collagen type XII
electron microscopy, 585, 587
- collagen type XIV
peptide map, 434
- collagenase
crystallography, 181
- collisional quenching
chymotrypsin, 603
fluorescence, 602
fructose-bisphosphate aldolase, 603
immunoglobulin G, 603
- colonic mucin
oligosaccharides of glycoproteins, 128, 130
- colony
cloning of DNA, 99
- colorimetric assay, 18
- column chromatography, 4
- common ancestor
evolution of proteins, 346
- common fold
definition, 393
- complement fixation, 564
- complementarity-determining regions
immunoglobulins, 558
in immune complex, 559
- complementary DNA
cloning of DNA, 98
- complementary faces
quaternary structure, 455
- complex *N*-linked oligosaccharides
oligosaccharides on glycoproteins, 131
- complexes between dodecyl sulfate and polypeptides
sieving, 429
- compressibility, isothermal
of a molecule of protein, 278
packing of side chains, 278
thermodynamics of folding, 673
water, 192
- compression
electrophoresis of DNA, 106
- computational alignment
aligning amino acid sequences, 351
evaluation of, 368
- computed Fourier transform
image reconstruction, 501
- concanavalin A
 β structure, 261
interfaces, 480
packing of side chains, 281
posttranslational modification, 116
- concentration of protein
measurement, 21
osmotic pressure, 409
sedimentation equilibrium, 412

- concentration, units of, 196–99
 corrected volume fraction, 196
 equilibrium constant, 197
 partition coefficient, 198
 standard free energy of transfer, 198
 volume fraction, 196
- configurational entropy
 molten globule, 683
 thermodynamics of folding, 681–82
- configurational heat capacity
 water, 192
- conformational changes
 aspartate carbamoyltransferase, 577
 circular dichroism, 600
 during association of proteins with nucleic acid, 321
 fluorescence resonance energy transfer, 609
 immune complex, 561
- connections among nuclei
 nuclear magnetic resonance, 622
- conservation of charge
 acids and bases, 66
- conservation of mass
 acid and bases, 66
- conservative replacement
 aligning amino acid sequences, 349
 evolution of proteins, 349
- constraints
 refinement, 174–75
- continuous assay, 15
- continuous flow
 kinetics of folding, 694
- continuous wave nuclear magnetic spectrometers
 nuclear magnetic resonance, 614
- convergent evolution
 alanine dehydrogenase, 373
 aligning crystallographic molecular models, 372
 β -amino-acid oxidase, 373
 chorismate mutase, 372
 cytochrome P-450, 373
 3-dehydroquinone dehydratase, 373
 ferredoxin, 373
 L-lactate dehydrogenase, 373
 L-lactate dehydrogenase (cytochrome), 373
 nitric-oxide synthase, 373
 superoxide dismutase, 373
- coordinate system
 crystallography, 157
- copper
 metalloproteins, 331
- coproporphyrinogen oxidase
 purification, 26
- core electrons, 56
- core of the hydrocarbon
 bilayer of phospholipid, 757
- corrected volume fraction
 concentration, units of, 196
 definition, 196
- correlated spectrum
 nuclear magnetic resonance, 619
- coulomb effect, 57
- coulomb's law, 39
- countercurrent distribution, 5
- coupled assay, 16
- coupling constant
 nuclear magnetic resonance, 615
- covalency
 hydrogen bond, 215
- covalent bonds
 metal ion, 327
- covalent modification, 529–52
 acetic anhydride, 535
 acetylcholine receptor, 797
 actin, 547
 acylating agents, 535
 aldehydes, 534
 arginine, 539
 aryl azides, 541
 aryl nitrenes, 542
 aspartate, 539–41
 1-azidopyrene, 797
 aziridine, 537
 bacteriorhodopsin, 797
 benzyl bromide, 536
 brominating agents, 539
 bromopyruvate, 549
 carbenes, 543
 carbodiimides, 539–41
 carbonate dehydratase, 547
 N-chlorobenzenesulfonamide, 538
 chymotrypsin, 543
 chymotrypsinogen, 547
 citraconic anhydride, 536
 1,2-cyclohexanedione, 539
 cysteine, 530, 532, 536–37
 cytochrome b_{561} , 546, 550
 cytochrome-c oxidase, 797
 decomposition of reagent, 534
 2-dehydro-3-deoxy-6-phosphogluconate aldolase, 549
 diazonium salts, 538
 5-diazonium-1-hydrotetrazole, 538
 diazotized *p*-[^{35}S]sulfanilic acid, 799
 dicyclohexyl carbodiimide, 539
 diethyl pyrocarbonate, 536
 5-(dimethylamino)naphthalene-1-sulfonyl fluoride, 536
- diphenylethanedione, 539
- 5,5'-dithiobis(2-nitrobenzoate), 537
- DNA topoisomerase, 546
- DNA-directed RNA polymerase, 548
- electrophilic reagents, 529
- N*-(ethoxycarbonyl)-2-ethoxy-1,2-dihydroquinoline, 541
- N*-ethylmaleimide, 536–37
- N*-ethyl-5-phenylisoxazolium-3'-sulfonate, 541
- N*-ethyl-*N*'-[3-(dimethylamino)propyl]carbodiimide, 539
- ferredoxin-NADP⁺ reductase, 550
- fluorescent electrophiles, 606
- fluorosulfonic acids, 536
- N*-formyl-[^{35}S]sulfinylmethionyl methylphosphate, 799
- fructose-bisphosphate aldolase, 547
- fumarase, 536
- γ -glutamyltransferase, 546, 550
- glucose-6-phosphate isomerase, 545
- glutamate, 539–41
- glyceraldehyde-3-phosphate dehydrogenase (phosphorylating), 542
- glycophorin, 797
- histidine, 530, 531, 536
- N*-hydroxysuccinimide esters, 535
- ICl, 538
- iminothiolane, 549
- impermeant reagents for, 798
- iodoacetamide, 530
- 5-[^{125}I]iodonaphthyl azide, 797
- isethionyl[^{14}C] acetimidate, 799
- isocitrate lyase, 544
- isocyanates, 535
- isothiocyanates, 534
- kinetics, 531
- α lactalbumin, 545
- lactoperoxidase, 538
- lysine, 530–36
- membrane-spanning α helices, 796
- methionine, 530, 532, 536
- methyl acetimidate, 532–34
- myosin, 544
- Na⁺/K⁺-exchanging ATPase, 546, 797
- nitrene, 542
- 2-[(2-nitrophenyl) sulfonyl]-3-methyl-3'-bromoindolenine, 539
- oxidative cleavage, 544
- 4-(oxoacetyl)phenoxyacetic acid, 539

- papain, 546, 551
p-chloromercuribenzoate, 537
 pH effects, 531
 phosphoenolpyruvate
 carboxykinase (GTP), 543
 photolytic reactions, 541–44
p-nitrophenylethanedione, 539
 purpose of, 529–30, 546–47
 λ repressor, 547
 rhodopsin, 605
 ribonuclease, 550
 ribulose-bisphosphate
 carboxylase, 536, 546
 rose bengal, 536
 seminal ribonuclease, 545
 sigma factor rpoD, 548
 specificity of, 530–32
 succinic anhydride, 535
 sulfenyl halides, 538
 tetracycline repressor, 544
 3,4,5,6-tetrahydrophthalic
 anhydride, 536
 tetranitromethane, 538
 2-S-[¹⁴C]thiuroniummethane-
 sulfonate, 799
 trifluoroacetic anhydride, 535
 3-(trifluoromethyl)-3-(*m*-
 [¹²⁵I]iodophenyl)diazirine, 797
 2,4,6-trinitrobenzenesulfonate,
 536
 1-tritiospiro[adamantane-4,3'-
 diazirine], 797
 tryptophan, 538–39
 tyrosine, 536, 537–38
 UDP-*N*-acetylglucosamine
 1-carboxyvinyltransferase, 546
 vanadate, 544
 2-vinylpyridine, 537
 yield, 530
 crambin
 β turn, 265
 creatinase
 aligning crystallographic molecular
 models, 362–63
 creatine kinase
 fluorescence resonance energy
 transfer, 608
 CRINEPT
 nuclear magnetic resonance, 621
 critical concentration
 assembly of microtubules, 724–25
 critical micelle concentration
 detergent, 769
 Cro protein
 interface, 478
 cross-link
 posttranslational modification, 119
 cross-linking, 439–46, 548
 actin, 549
 arrangement of the subunits, 445
 assembly of fibrin, 721
 assembly of oligomers, 445
 chaperonin GroEL, 444–45
 count of the number of subunits,
 441
 cysteine, 549
 detection of heterologous
 associations, 519
 dimethyl suberimide, 440
 epidermal growth factor receptor,
 443
 glutaraldehyde, 443
 glycerol kinase, 442
 immunostaining, 566
 L-lactate dehydrogenase, 443–45
 ladders, 442
 lysines, 440
 membrane-spanning α helices, 803
 m-xylylene diisocyanate, 440
 myosin, 549
 Na⁺/K⁺-exchanging ATPase, 444–45
 quantitative cross-linking, 443–45
 ribonuclease, 548
 ribosome, 549
 stoichiometric ratio of subunits,
 445
 succinate-CoA ligase (ADP-
 forming), 445
 cross-linking reagent
 2-(*p*-nitrophenyl)-3-(3-carboxy-
 4-nitrophenyl)thio-1-propene,
 548
 cross-linking reagents, 441
 cross-sectional area
 bilayer of phospholipid, 754
 crystal packing
 crystallography, 170
 molecular rotational axes of
 symmetry, 465
 crystallin
 aligning amino acid sequences, 361
 evolution of proteins, 350
 folding, 668–69
 hydrogen bonds in crystallographic
 molecular models, 308
 crystalline array
 cytochrome-*c* oxidase, 791
 crystallization of proteins, 49–50
 acetylcholine receptor, 772
 acylphosphatase, 49
 aquaporin, 772
 bacteriorhodopsin, 772, 775
 cytochrome *o* ubiquinol oxidase,
 772
 crystallography, 50
 cytochrome-*c* oxidase, 772, 775
 endoplasmic reticulum Ca²⁺-
 transporting ATPase, 772
 ferrichrome-iron receptor, 772
 α -galactosidase, 49
 halorhodopsin, 772
 hanging drop, 50
 α -hemolysin, 772
 integral membrane-bound protein,
 775
 large-conductance mechano-
 sensitive channel, 772
 lipid A export ATP-binding protein,
 772
 maltoporin, 772
 nicotinate-nucleotide
 diphosphorylase, 49
 outer membrane protein A, 772
 outer membrane protein F, 772,
 775
 outer membrane protein TolC, 775
 outer membrane protein TolC, 772
 phosphoenolpyruvate
 carboxykinase, 49
 photosynthetic reaction center,
 772
 porin, 772
 potassium channel DcsA, 772
 protein MsbA, 772
 rhodopsin, 772
 succinate dehydrogenase, 772
 sucrose porin, 772
 ubiquinol-cytochrome-*c*
 reductase, 772, 775
 crystallization of integral membrane-
 bound proteins
 bicontinuous cubic phase, 775
 crystallographic asymmetric unit
 definition, 457
 space groups, 457
 crystallographic axis of symmetry
 definition, 461
 space groups, 461
 crystallographic molecular model,
 167–70
 α -carbon diagram, 167
 accessible surface area, 273
 aquaporin, 788
 bacteriorhodopsin, 777
 bilayer of phospholipid, 751–53
 cartoon, 167
 dilauroyl-*N,N*-dimethyl-
 phosphatidylethanolamine,
 753
 dilauroylphosphatidylethanol-
 amine, 753

- dilauroylphosphatidic acid, 753
 dimyristoylphosphatidylcholine, 753
 dimyristoylphosphatidylglycerol, 753
 endoplasmic reticulum Ca^{2+} -transporting ATPase, 778
 ferrichrome-iron receptor, 783
 lysozyme, 170
 myoglobin, 170
 penicillopepsin, 167–70
 photosynthetic reaction center, 780
 porin OmpF, 782
 potassium channel KcsA, 779
 random meander, 170
 skeletal representation, 167
 space-filling representation, 170
 ubiquinol-cytochrome-*c* reductase, 781
 crystallographic *R*-factor
 crystallography, 173
 definition, 173
 crystallography
 α helices, 165–67
 alternative conformations, 182
 amplitude of the reflection, 153
 amplitude of the structure factor, 155
 anisotropic thermal parameters, 176
 aspartyl endopeptidase, 182
 axes, 151
 Bragg spacings, 158
 β structure, 165–67
 β turns, 165–67
 B value, 175
 calculated amplitudes, 172
 calculated phases, 173
 charge-coupled device, 155
 citrate (*si*) synthase, 171
 collagenase, 181
 coordinate system, 157
 crystal packing, 170
 crystallization, 50
 crystallographic *R*-factor, 173
 cubic lattice, 151
 data set, 155
 deoxyribonuclease, 158, 182
 difference maps of electron density, 173
 diffraction, 149
 diffraction limit, 154
 dihydrofolate reductase, 180
 distribution of electron density, 156
 free *R*-factor, 176
 Friedel pair, 152
 fundamental unit cell, 151
 α -glucosidase, 183
 hexagonal lattice, 151
 hydrogen atoms, 182
 index, 151–53
 ions, 181
 lattice, 151
 layer line, 150
 lysozyme, 155
 molecular model, 163
 molecular replacement, 182
 monoclinic lattice, 151
 multiple isomorphous replacement, 158–61
 native structure, 171
 nitrite reductase, 181
 observed amplitudes, 173
 observed phases, 173
 oligosaccharides, 180
 orthorhombic lattice, 151
 phase of the reflection, 154
 photosynthetic reaction center, 180–181
 B-phycoerythrin, 181
 reflecting faces, 150
 resolution, 158
 rhombohedral lattice, 151
 ribonuclease T_1 , 184
 secondary structure, 165–67
 solvent flattening, 161
 structure factor, 155
 synchrotron, 156
 tetragonal lattice, 151
 tube of electron density, 162
 unit cell, 150
 unrefined map of electron density, 161
 CTP synthase
 domains, 380
 cubic expansion coefficient
 water, 193
 cubic lattice
 crystallography, 151
 cyanogen bromide
 affinity adsorption, 27
 cleavage of polypeptide, 87, 89
 cyclic-AMP dependent protein kinase
 domains, 382
 fluorescence resonance energy transfer, 608–609
 molecular taxonomy, 396
 radius of gyration, 581
 transitory heterologous associations, 513
 X-ray scattering, 582, 584
 cyclic point group, 466
 cyclic symmetry
 fluorescence resonance energy transfer, 608
 integral membrane-bound proteins, 787
 cyclin
 heterologous associations, 517
 yeast two-hybrid assay, 519
 cyclin-dependent kinase inhibitor 1
 yeast two-hybrid assay, 519
 cyclin-dependent protein kinase 2
 molecular taxonomy, 396
 cyclohexane monooxygenase
 growth on cyclohexane, 20
 1,2-cyclohexanedione
 reagent for covalent modification, 539
 cyclosporin synthase
 assay, 14
 cystathionine β -lyase
 molecular taxonomy, 396
 cystathionine β -synthase
 domains, 378
 cystathionine γ -synthase
 space groups, 464
 cysteic acid
 cysteine, 82
 cysteine
 acid dissociation constant, 75
 covalent modification, 530, 532, 536–37
 cross-linking, 549
 cysteic acid, 82
 disulfide, 81–82
 electronic structure, 80–82
 oxidation levels, 80–82
 sulfenic acid, 81–82
 sulfinic acid, 81–82
 sulfonate, 81–82
 cystines
 acetylcholine receptor, 783
 membrane-spanning α helices, 783
 posttranslational modification, 122
 ribonuclease, 124
 stereochemistry of side chains, 271
 thermodynamics of folding, 681
 thioredoxin, 125
 thrombomodulin, 125
 tris(2-carboxyethyl)phosphine, 125
 ultraviolet absorption spectra, 601
 cystines, formation of
 endoplasmic reticulum, 708
 glutathione, 708
 insulin-like growth factor, 708
 kinetics of folding, 708–9
 mixed disulfide, 708

- protein disulfide-isomerase, 125, 708–09
 reduction potential, 709
 ribonuclease, 708
 thiol:disulfide interchange protein, 708
 cytidine deaminase
 rotational axes of
 pseudosymmetry, 484
 cytochrome b_5
 diffusion, 822
 embedded anchor, 774
 purification, 773
 cytochrome b_{561}
 covalent modification, 546, 550
 cytochrome b_{562}
 water in crystallographic molecular models, 293
 cytochrome c
 α helix, 256
 aligning amino acid sequences, 346–49, 351–52, 354–55, 360
 aligning crystallographic molecular models, 364–65
 epitopes, 561
 fluorescence resonance energy transfer, 609
 folding, 685
 free energy of folding, 677
 frictional ratio, 426
 immune complex, 560
 kinetics of folding, 689, 692, 695, 697–99
 molten globule, 683–84
 nuclear magnetic resonance, 617
 proline isomerization, 702
 radius of gyration, 581
 sieving, 424, 428
 cytochrome- c oxidase
 bound phospholipid, 784
 boundary layer of phospholipid, 786, 805
 circular dichroism, 598
 covalent modification from within the bilayer, 797
 crystalline array, 791
 crystallization, 772, 775
 fluorescence resonance energy transfer, 609
 heterooligomer, 777
 membrane-spanning α helices, 772, 777, 784, 796
 passageway for cations, 778
 short subunits, 777
 topography of membrane-spanning proteins, 802
 cytochrome- c peroxidase
 peptide separation, 91
 cytochrome- c reductase
 translational diffusion coefficient, 814
 cytochrome c'
 kinetics of folding, 704
 cytochrome c_1
 anchored membrane-bound proteins, 764
 circular dichroism, 598–99
 cytochrome c_2
 kinetics of folding, 691
 nuclear magnetic resonance, 626
 cytochrome c_{551}
 nuclear magnetic resonance, 638
 cytochrome d ubiquinol oxidase
 resonance Raman spectrum, 596
 cytochrome f
 aligning amino acid sequences, 360
 water in crystallographic molecular models, 293–94
 cytochrome o ubiquinol oxidase
 crystallization, 772
 membrane-spanning helices, 772
 passageway for cations, 778
 cytochrome P-450
 convergent evolution, 373
 cytoplasm
 protein concentration, 1
 cytoplasmic surface
 of a membrane, 743
 cytosine
 electronic structure, 65
 cytoskeleton, 820–21
 actin, 820
 ankyrin, 821
 band 3 anion transport protein, 821
 erythrocyte, 820
 glycophorin, 821
 phosphatidylserine, 821
 protein 4.1, 821
 spectrin, 820
- D**
- D-5-deamino-5(S)-hydroxyneuraminic acid
 oligosaccharides of glycoproteins, 128
 D-alanine-D-alanine ligase
 domains, 388
 molecular taxonomy, 397
 D-amino-acid oxidase
 convergent evolution, 373
 data set
 crystallography, 155
 definition, 155
 databanks, searching
 coverage, 368
 error rate, 368
N-deacetylheparin *N*-sulfotransferase
 purification, 29
 dead time
 kinetics of folding, 688
 stopped-flow apparatus, 688
 DEAE cellulose
 chromatography, 9
 decay in the anisotropy
 rotational diffusion coefficient, 812
 decyl β -D-glucoside
 detergent, 770
 decyl β -D-maltoside
 detergent, 770
 2-dehydro-3-deoxy-6-phosphogluconate aldolase
 aligning amino acid sequence, 360
 covalent modification, 549
 molar mass, 418
 peptide map, 437–38
 purification, 29
 quaternary structure, 407
 dehydromerohistidine, 126
 3-dehydroquinate dehydratase
 convergent evolution, 373
 domains, 380
 tetrahedral symmetry, 487
 3-dehydroquinate synthase
 domains, 380
 deleterious mutation
 evolution of proteins, 348
 deletion
 evolution of proteins, 350
 delocalization
 acids and bases, 63
 electronic structure, 56
 denaturant
 definition, 660
 denatured state
 definition, 659
 denaturing proteins
 when sequencing polypeptides, 87
 δ -endotoxin CryIIIA
 packing of α helices, 285
 3-deoxy-7-phosphoheptulonate synthase
 purification, 25
 2'-deoxyadenosine, 95
 2'-deoxycytidine, 95
 2'-deoxyguanosine, 95
 deoxyhemoglobin
 molecular charge, 34

- deoxyribonuclease
 association of proteins with
 nucleic acid, 316
 crystallography, 158, 182
 endopeptidolytic cleavage, 547
 hydrogen bonds in
 crystallographic molecular
 models, 307–08
 multiple isomorphous
 replacement, 161
 Ramachandran plot, 255
 refinement, 176, 178
 secondary structure, 263
 space groups, 458
 water in crystallographic molecular
 models, 292
- dephospho-CoA kinase
 domains, 391
- deshielding
 hydrogen bond, 209
- desmin
 frictional ratio, 577
 sedimentation velocity, 577
- desmin filaments
 intermediate filaments, 506
- desmosine, 123
- detachable domains
 anchored membrane-bound
 proteins, 773
 domains, 376
 immunoglobulin G, 376, 378
- detergents
 alkyl glycosides, 769
 alkyl oligo(ethylene oxide) ethers,
 768
 C₈E₅, 770
 C₁₀E₆, 770
 C₁₀E₈, 770
 C₁₂E₆, 770
 C₁₂E₈, 770
 C₁₄E₈, 770
 C₁₆E₈, 770
 3-[(3-cholamidopropyl)dimethyl-
 lammonio]-1-propane-
 sulfonate, 770
 decyl β - β -glucoside, 770
 decyl β - β -maltoside, 770
 N,N-dimethyldecylamine *N*-oxide,
 770
 N,N-dimethyldodecylamine
 N-oxide, 769–70
 N,N-dimethyloctylamine *N*-oxide,
 770
 dodecyl β - β -glucoside, 770
 dodecyl β - β -maltoside, 770
 octyl β - β -glucoside, 770
 saponins, 769
- Tritons, 769
 Triton X-100, 770
- dethiobiotin synthase
 electrospray mass spectrometry, 417
- deuterium nuclear magnetic
 resonance spectroscopy
 bilayer of phospholipid, 757
 boundary layer of phospholipid,
 786
 4,6-di(bromomethyl)-3,7-dimethyl-
 1,5-diazabicyclo[3.3.0]octadiene-
 2,7-dione
 reagent for cross-linking, 549
- diacylglycerol kinase
 integral membrane-bound protein,
 766
- 2,2-dialkylglycine decarboxylase
 (pyruvate)
 point group, 470–71
- diameter of the particles
 chromatography, 6
- diaminopimelate epimerase
 domains, 383
- diazonium salts
 reagents for covalent modification,
 538
- 5-diazonium-1-hydrotetrazole
 reagent for covalent modification,
 538
- diazotized *p*-[³⁵S]sulfanilic acid
 impermeant reagent for covalent
 modification, 799
- dicarboxylate transporter
 immunostaining, 566
- dicarboxylic acids
 intramolecular hydrogen bonds,
 227
- dicyclohexyl carbodiimide
 reagent for covalent modification,
 539
- 2',3'-dideoxynucleotide
 sequencing of DNA, 104
- dielectric relaxation
 hydration of a protein, 297
 hydrophobic effect, 233
 water, 195
- diethyl pyrocarbonate, 550
 reagent for covalent modification,
 536
- difference in pK_a
 hydrogen bond, 210
- difference maps of electron density
 crystallography, 173
- difference maps of scattering density
 image reconstruction, 502
- differential centrifugation
 cell fractionation, 743
- differential scanning calorimetry
 domains, 388–89
- diffraction limit
 crystallography, 154
- diffraction of X-radiation
 biological membranes, 808
 crystallography, 149
- diffusion
 adenylate cyclase, 817
 β -adrenergic receptor, 817
 cytochrome *b*₅, 822
 frictional coefficient, 577–78
 frictional ratio, 577–78
 in biological membranes, 811–13
 rhodopsin, 823
 rotational diffusion coefficient, 811
 translational diffusion coefficient,
 811
- diffusion coefficient
 definition, 37
 measurement, 37
 molten globule, 684
 standard, 574
- dihedral angle
 nuclear magnetic resonance, 616
- dihedral angle χ_1
 stereochemistry of side chains, 267
- dihedral angle χ_2
 stereochemistry of side chains, 269
- dihedral angles ϕ and ψ
 secondary structure, 252–54
- dihedral point group
 definition, 470
- dihedral point group 222, 470–72
- dihedral point group 322, 472–73
- dihedral point group 422, 472
- dihedral point group 522, 472
- dihedral symmetry
 gap junction connexon, 787
- dihydrodipicolinate reductase
 proton exchange, 645
- dihydrodipicolinate synthase
 point group, 475
- dihydrofolate reductase
 α helix, 257
 circular permutation, 680
 crystallography, 180
 domains, 380
 hydration, 299
 kinetics of folding, 689, 691–692,
 694, 698, 704
 molecular taxonomy, 395
 nuclear magnetic resonance, 621,
 622, 625
 purification, 29
 water in crystallographic molecular
 model, 299

- dihydrofolate reductase–thymidylate synthase
 domains, 390
 dihydrolipoyl dehydrogenase
 assembly of oligomers, 715
 domains, 388
 dihydrolipoyllysine-residue acetyltransferase
 aligning amino acid sequence, 360
 assembly of oligomers, 715
 domains, 384, 390
 folding, 683
 hydrogen bonds in
 crystallographic molecular models, 309
 interfaces, 479
 ionic interactions, 300
 kinetics of folding, 702
 quaternary structure, 490
 space groups, 463
 dihydrolipoyllysine-residue (2-methylpropanoyl)transferase
 quaternary structure, 490
 dihydrolipoyllysine-residue succinyltransferase
 aligning amino acid sequence, 360
 mismatched symmetry, 511
 quaternary structure, 490
 space groups, 463
 dihydroneopterin aldolase
 interfaces, 480
 dihydroorotase
 domains, 379
 dilauroyl-*N,N*-dimethylphosphatidylethanolamine
 crystallographic molecular model, 753
 dilauroylphosphatidic acid
 crystallographic molecular model, 753
 dilauroylphosphatidylethanolamine
 crystallographic molecular model, 753
 dimer
 fundamental unit of quaternary structure, 474
 dimer of dimers
 point group 222, 471
 dimers of water
 water, 190
 dimethyl suberimidate
 cross-linking, 440
 6,7-dimethyl-8-ribityllumazine synthase
 icosahedral symmetry, 488
 interfaces, 488
 5-(dimethylamino)naphthalene-1-sulfonyl fluoride
 fluorescent reagent for covalent modification, 536
N,N-dimethyldecylamine *N*-oxide
 detergent, 770
N,N-dimethyldodecylamine *N*-oxide
 detergent, 769–770
N,N-dimethyloctylamine *N*-oxide
 detergent, 770
 dimyristoylphosphatidylcholine
 crystallographic molecular model, 753
 dimyristoylphosphatidylglycerol
 crystallographic molecular model, 753
 dipeptidyl-peptidase IV
 purification, 773
 diphenylethanedione
 reagent for covalent modification, 539
 diphosphatidylglycerol
 asymmetry of, 810
 phospholipid, 749
 diphtheria toxin
 punching a hole in a membrane, 804
 diphtheria toxin repressor
 dipolar interactions
 electron paramagnetic resonance, 649
 metalloproteins, 332–33
 disc electrophoresis, 42–44
 disordered water
 water in crystallographic molecular models, 290
 distance between dipoles
 fluorescence resonance energy transfer, 605
 distance distribution function
 X-ray scattering, 581
 distance measurements
 fluorescence resonance energy transfer, 607–9
 distribution coefficient
 molecular exclusion, 12
 distribution of electron density
 crystallography, 156
 distribution of scattering density
 image reconstruction, 501
 β -dystroglycan
 heterologous associations, 513
 disulfide
 cysteine, 81–82
 disulfide interchange, 124
 5,5'-dithiobis(2-nitrobenzoate)
 reagent for covalent modification, 537
 dithiothreitol
 use, 124
 divalent metal ions
 ion pairs, 203
 DNA-(apurinic or apyrimidinic site) lyase
 metalloproteins, 330
 nucleic acid, association of
 proteins with, 320
 E2 DNA-binding domain
 rotational axes of symmetry, 468
 DNA-directed DNA polymerase
 cloning of DNA, 97
 domains, 388
 fluorescence resonance energy transfer, 608–09
 posttranslational modification, 115
 purification, 30
 DNA-directed RNA polymerase
 cloning of DNA, 97
 covalent modification, 548
 quaternary structure, 407
 DNA ligase (ATP)
 cloning of DNA, 96
 DNA ligase (NAD⁺)
 cloning of DNA, 96
 DNA polymerase β
 nucleic acid, association of
 proteins with, 315
 DNA topoisomerase
 covalent modification, 546
 dodecyl β -D-glucoside
 detergent, 770
 dodecyl β -D-maltoside
 detergent, 770
 dodecyl sulfate gel electrophoresis
 carbonate dehydratase, 422
 catalase, 422
 catalogue of polypeptides, 431
 chymotrypsinogen, 422
 fructose-bisphosphate aldolase, 422
 fumarate hydratase, 422
 glutamate dehydrogenase, 422
 glyceraldehyde-3-phosphate dehydrogenase, 422
 L-lactate dehydrogenase, 422
 micelle of dodecyl sulfate, 421
 myoglobin, 422
 ovalbumin, 422
 phosphorylase, 422
 retardation coefficient, 422
 serum albumin, 422
 stacking, 422
 unfolded polypeptides, 421

- dolichyl-phosphate
 β -D-mannosyltransferase
 fluorescence resonance energy transfer, 609
 purification, 773
 domains, 376–91
 acetylcholine receptor, 779
 acyl carrier protein, 382
 [acyl-carrier-protein]
 S-acyltransferase, 381
 adenylate cyclase, 817
 aggreCAN, 386
 β -alanine– β -alanine ligase, 388
 aminodeoxychorismate synthase, 380
 anchored membrane-bound proteins, 764
 anion carrier, 377
 ankyrin domain, 386
 anthranilate synthase, 380
 antifreeze protein, 384
 AraC protein, 378
 armadillo domain, 386
 aspartate carbamoyltransferase, 379
 aspartate kinase, 378
 aspartate kinase–homoserine dehydrogenase, 381
 aspartate transaminase, 389
 aspartic endopeptidase, 384
 benzoate 4-monooxygenase, 382
 biotin-dependent carboxylase, 382, 388
 btk kinase, 386
 C2 domain, 386
 Ca²⁺-transporting ATPase, 779
 CAD multienzyme complex, 379, 390
 carbamoyl-phosphate synthase, 383
 carbamoyl-phosphate synthase (ammonia), 379
 carbamoyl-phosphate synthase (glutamine hydrolysing), 379
 catalase, 388
 chaperone protein PapD, 390
 cholesterol oxidase, 382
 CLC-0 chloride channel, 389
 coenzymatic domain, 382
 cohesin domain, 386
 collagen VI, 386
 CTP synthase, 380
 cyclic AMP-dependent protein kinase, 382
 cystathionine β -synthase, 378
 3-dehydroquinone dehydratase, 380
 3-dehydroquinone synthase, 380
 dephospho-CoA kinase, 391
 detachable domain, 376
 diaminopimelate epimerase, 383
 differential scanning calorimetry, 388–89
 dihydrofolate reductase, 380
 dihydrofolate reductase–thymidylate synthase, 390
 dihydrolipoyl dehydrogenase, 388
 dihydrolipoyllysine-residue acetyltransferase, 384, 390
 dihydroorotase, 379
 DNA-directed DNA polymerase I, 388
 dot matrix, 384
 dragline silk, 384
 EF hand, 386
 effect on folding, 680
 EGF domain, 386
 endopeptidolytic detachment, 377
 enoyl-[acyl-carrier-protein] reductase, 381
 enzymatic domain, 379–82
 epidermal growth factor receptor, 815
 erythronolide synthase, 381
 evolutionarily shifting domains, 388
 exon shuffling, 386
 extracellular matrix, 386
 Fab fragments, 376
 Factor XII, 386
 fatty-acid synthase, 381
 ferredoxin–NADP⁺ reductase, 376–77, 388
 fibrinogen, 388–89, 717
 fibronectin domain, 386
 functional domain, 382
 fundamental units of protein structure, 390
 galactose oxidase, 382
 gelsolin, 384
 gene fusion, 381
 gene multiplication, 384
 genetic detachment, 377
 glucose oxidase, 382
 glutaminase, 379
 glutamine–fructose-6-phosphate transaminase (isomerizing), 380
 glutamyl endopeptidase, 378
 glutathione reductase, 388–89
 glutathione synthase, 388
 glutathione-disulfide reductase, 382
 GMP synthase, 380
 granulin, 384
 hemagglutinin glycoprotein, 388
 hemocyanin, 384
 hemopexin domain, 386
 hexokinase, 384
 homoserine dehydrogenase, 378
 3-hydroxyacyl-[acyl-carrier-protein] dehydratase, 381
 imidazole glycerol phosphate synthase, 380, 383
 immunoglobulin domain, 386
 immunoglobulin G, 376, 378, 382, 390, 477, 555
 independently folding domain, 389
 independently shifting domains, 388
 indole-3-glycerol-phosphate synthase, 377, 384
 initiation factor IF3, 390
 internal duplication, 383
 internally repeating domain, 382
 kinetics of folding, 709
 kringle, 386, 390
 L-lactate dehydrogenase, 382–83
 laminin γ 1, 390
 leucine-rich repeat, 386
 lysozyme, 388
 mannose-6-phosphate receptor, 385
 methionyl aminopeptidase, 383
 modular domain, 385
 mosaic eukaryotic protein, 385
 multienzyme complex, 379–82
 NADH peroxidase, 388
 nebulin, 384
 operon, 381
 ovotransferrin, 384
 3-oxoacyl-[acyl-carrier-protein] reductase, 381
 3-oxoacyl-[acyl-carrier-protein] synthase, 381
 p120 GTPase activator, 386
 pantetheine-phosphate adenylyltransferase, 391
 peptidylamidoglycolate lyase, 377
 peptidylglycine monooxygenase, 377
 perlecan, 386
 6-phosphofructo-2-kinase/fructose-2,6-bisphosphate 2-phosphatase, 389
 phosphoglycerate kinase, 383
 phosphoinositide phospholipase C δ 1, 386–87
 phospholipase C γ , 386
 1-(5-phosphoribosyl)-5-[(5-phosphoribosylamino)methylidene amino] imidazole-4-carboxamide isomerase, 383
 phosphoribosylamine–glycine ligase, 388, 390

- phosphoribosylanthranilate isomerase, 377, 384
 phosphoribosylformylglycin-amidine cyclo-ligase, 390
 phosphoribosylformylglycin-amidine synthase, 380
 phosphoribosylglycinamide formyltransferase, 390
 3-phosphoshikimate 1-carboxy-vinyltransferase, 380, 382
 placental ribonuclease inhibitor, 384
 plasminogen, 388–89
 pleckstrin domain, 386
 prepromaginin, 384
 protein-tyrosine kinase ZAP-70, 390
 protein-tyrosine-phosphatase, 381
 pyruvate kinase, 382–83
 pyruvate oxidase, 385
 recurring domain, 382
 regulatory kinases, 386
 retinol-binding protein, 384
 RNA recognition motif, 386
 SAND domain, 386
 separately unfolding domains, 388–89
 serum albumin, 384
 sex-lethal protein, 390
 SH2 domain, 386
 SH3 domain, 386
 shikimate dehydrogenase, 380
 shikimate kinase, 380
 spectrin, 384–85, 390
 START domain, 386
 structural domain, 387
 sulfite oxidase, 377
 sulfite reductase, 382
 thermolysin, 389
 thioredoxin-disulfide reductase, 388
 thiosulfate sulfurtransferase, 383
 thrombospondin I, 386
 thymidylate synthase, 380
 titin, 385
 triose-phosphate isomerase, 382
 UDP-glucose 6-dehydrogenase, 384
 Donnan effect
 osmotic pressure, 410
 serum albumin, 419
 Donnan formalism
 chromatography, 10
 Donnan potential
 osmotic pressure, 411
 donor
 fluorescence resonance energy transfer, 604
 donors and acceptors of hydrogen bonds
 nucleic acid structure, 315
 dopamine- β -monooxygenase
 anchored membrane-bound proteins, 764
 dot matrix
 aligning amino acid sequences, 351–52
 domains, 384
 double-helical DNA
 local rotational axis, 467
 nucleic acid, association of proteins with, 314
 rotational axis of pseudosymmetry, 467
 structure of DNA, 95–96
 double-helical hairpin of RNA
 nucleic acid structure, 322
 double-mutant cycle
 hydrogen bonds in crystallographic molecular models, 310
 doubly wound, parallel β sheet
 recurring structure, 373
d π molecular orbitals
 phosphate, 83
 DQF
 nuclear magnetic resonance, 621
 dragline silk
 α helix, 259
 domains, 384
 dry weight, measurement of
 molar mass, 419
 dynamics
 nuclear magnetic resonance, 623
 dynein
 microtubules, 730
 dystrophin
 heterologous associations, 513
 immunoabsorbent, 566
E
 Edman degradation
 sequencing of polypeptides, 86
 EF hand, 386
 effective charge number
 electrophoresis, 39
 effective molarity
 intramolecular processes, 224
 effective sphere
 diffusion, 37
 efficiency of transfer
 fluorescence resonance energy transfer, 604
 EGF domain, 386
 elastase
 aligning crystallographic molecular models, 362–63
 elastic scattering
 absorption of light, 593
 electrolyte
 osmotic pressure, 411
 electron density
 bilayer of phospholipid, 750
 electron diffraction
 bacteriorhodopsin, 791
 electron microscopy
 aspartate carbamoyltransferase, 587
 cell fractionation, 744
 coagulation Factor Va, 585
 collagen type XII, 585, 587
 fibrinogen, 585, 587
 fibulin, 586
 image reconstruction, 501
 inversion-specific glycoprotein, 586
 α_2 -macroglobulin, 588
 myosin, 586
 negative stain, 585
 nidogen, 586
 phosphorylase kinase, 586–87
 ribosome, 588
 rotary shadowing, 585
 shape of a protein, 585–88
 viral protein coats, 588
 electron nuclear double resonance
 aminocyclopropane carboxylate oxidase, 650
 catalase, 650
 electron paramagnetic resonance, 649–50
 nitric-oxide synthase, 649–50
 photosynthetic reaction center, 650
 ribonucleoside-diphosphate reductase, 650
 electron paramagnetic resonance, 645–50
 amine dehydrogenase, 647
 aminocyclopropane carboxylate oxidase, 646
 bacteriorhodopsin, 794
 bilayer of phospholipid, 755–56
 boundary layer of phospholipid, 785
 CDP-6-deoxy-L-threo- β -glycero-4-hexulose-3-dehydrase, 646
 dipolar interactions, 649
 electron nuclear double resonance, 649–50
 formate C-acetyltransferase, 645, 647–48
 g factor, 648
 hyperfine splitting, 647

- membrane-spanning α helices, 793–95
- myoglobin, 648–49
- nitric-oxide synthase, 649–50
- nitrogenase, 646
- organic radical, 645
- 1-oxyl-2,2,5,5-tetramethylpyrrolin-3-yl group, 645
- ribonucleoside-diphosphate reductase, 645, 649
- spectrometer, 646
- spin quantum number, 646
- spin–spin coupling, 647
- electron paramagnetic resonance spectrometer, 646
- electron transfer flavoprotein peptide map, 435–36
- electronic energy levels
 - absorption of light, 592–95
- electronic structure, 55–61
 - acyl oxygen, 60
 - adenine, 65
 - alanine, 76
 - arginine, 80
 - aromatic nitrogen heterocycles, 60–61
 - aromaticity, 60
 - asparagine, 79
 - aspartic acid, 79
 - bonding electrons, 56
 - bonding molecular orbital, 56
 - carbonyl oxygen, 60
 - cysteine, 80–82
 - cytosine, 65
 - delocalization, 56
 - formal charge, 56
 - glutamic acid, 79
 - glutamine, 79
 - glycine, 76
 - guanidinium cation, 80
 - guanine, 65
 - histidine, 77
 - imidazole, 78
 - isoleucine, 76
 - leucine, 76
 - lone pairs of electrons, 56, 59
 - lysine, 80
 - methionine, 80
 - nucleoside bases, 65
 - phenylalanine, 76
 - phosphate, 83
 - π lone pair of electrons, 59
 - π molecular orbitals, 56
 - proline, 76
 - pyridine, 60
 - pyrrole, 61
 - serine, 77
 - σ bonds, 59
 - σ lone pair of electrons, 59
 - σ – π stereochemical representation, 56, 59
 - σ structure, 59
 - sulfate, 82
 - threonine, 77
 - tryptophan, 76
 - tyrosine, 77
 - uracil, 65
 - valence electrons, 56
 - valine, 76
- electronically excited state
 - absorption of light, 594
- electron-scattering density, map of image reconstruction, 793
- electrophilic reagents
 - covalent modification, 529
- electrophoresis, 36–45
 - [acyl-carrier-protein] S-malonyltransferase, 46
- aggregation, 46
- β lactoglobulin, 42
- carbonate dehydratase II, 40
- detection of heterologous associations, 515
- effective charge number, 39
- electrophoretic field, 38
- electrotransfer, 565
- equation governing, 40
- free electrophoretic mobility, 38, 41
- fructose-bisphosphate aldolase, 41
- hemoglobin, 41, 45
- Henry's function, 40
- immunoglobulin G, 42
- immunostaining, 566
- in 8M urea, 432
- ionic double layer, 38
- ionic strength, 39
- isocitrate dehydrogenase, 46
- moving boundary electrophoresis, 41
- myoglobin, 42
- number of amino acids, estimation of, 427
- ovalbumin, 38, 41–42
- ovomucoid, 42
- pepsin, 42
- phosphomevalonate kinase, 46
- polyacrylamide gel, 41
- relative mobility, 429
- retardation coefficients, 41, 426
- ribonuclease, 45
- running gel, 44
- sequencing of DNA, 101–3
- serum albumin, 42, 45
- sieving, 426–30
- stable moving boundaries, 42–43
- stacking, 43
- stacking gel, 43
- stain for enzymatic activity, 46
- terminal velocity, 37
- trypsin, 41, 45
- electrophoresis of DNA
 - compression, 106
- electrophoresis on gels of
 - polyacrylamide cast in solutions of dodecyl sulfate, 421–23
- electrophoretic field, 38
- electrospray mass spectrometer
 - mass spectrometry, 91
- electrospray mass spectrometry
 - aldehyde dehydrogenase, 417
 - dethiobiotin synthase, 417
 - L1 metallo- β lactamase, 417
 - molar mass, 416–17
 - rusticyanin, 417
 - ubiquinol-cytochrome-c reductase, 417
- electrostatic repulsion
 - bilayers of phospholipid, 754
- electrostatic work
 - ionic interactions in crystallographic molecular models, 301
- electrostriction, 197
- electrotransfer
 - electrophoresis, 565
- ellipsoid of revolution
 - axial ratio, 574
 - hydrodynamic particle, 574
- elongation
 - assembly of microtubules, 723
- elongation factor Ts
 - heterologous interfaces, 510
- elongation factor Tu
 - heterologous interfaces, 510
- elution volume
 - definition, 4
- embedded anchor
 - anchored membrane-bound proteins, 773
 - cytochrome b_5 , 774
 - glycophorin A, 774
 - HLA histocompatibility antigen, 774
 - sucrose α -glucosidase/oligo-1,6-glucosidase, 774
 - viral hemagglutinin, 774
- emission of light, 592–613
 - fluorescence, 594–95
 - phosphorescence, 595
- emission spectrum
 - fluorescence, 595
- 3'-end
 - sequencing DNA, 95

- 5'-end
sequencing DNA, 95
- end-labeled fluorescent fragments
sequencing of DNA, 104
- end-labeled fragments
sequencing of DNA, 103
- end-labeling
immunostaining, 567
- endo-1,4- β -galactosidase
sequencing oligosaccharides, 135
- endo-1,4- β -xylanase
nuclear magnetic resonance, 636
- endo- α -sialidase
sequencing oligosaccharides, 135
- endoglycosidases
sequencing oligosaccharides, 133
- β 3 endonexin
yeast two-hybrid assay, 519
- endopeptidase K
metalloproteins, 328–29
- endopeptidases
impermeant reagents, 800
posttranslational modification, 113
sequencing polypeptides, 87–88
topography of membrane-spanning proteins, 800
- endopeptidolytic analysis
of proton exchange, 642
- endopeptidolytic cleavage
chymotrypsinogen, 547
deoxyribonuclease, 547
- endopeptidolytic detachment
domains, 377
- endoplasmic reticulum, 743
cystines, formation of, 708
- endoplasmic reticulum
Ca²⁺-transporting ATPase
aligning amino acid sequences, 364
boundary layer of phospholipid, 785
crystallization, 772
crystallographic molecular model, 778
domains, 779
image reconstruction, 793
membrane-spanning α helices, 772
passageway for cations, 778
translational diffusion coefficient, 813
- endosomes, 743
- energy level
molecular orbital, 56
- energy levels
absorption of light, 592–95
- engrailed homeodomain
kinetics of folding, 702
- enoyl-[acyl-carrier-protein] reductase
domains, 381
- enrichment
definition, 21
- enthalpy change
hydrophobic effect, 233
- enthalpy of activation
proline isomerization, 699
- enthalpy of folding
thermodynamics of folding, 671
- enthalpy of formation
hydrogen bond, 210–11
- enthalpy of fusion
water, 190
- enthalpy of hydration
ion, 200–201
ion pair, 201
- enthalpy of vaporization
water, 190
- entropy of approximation
hydrogen bonds in crystallographic molecular models, 309
intramolecular processes, 224
- entropy of formation
hydrogen bond, 210
- entropy of hydration
ion, 202
- entropy of mixing
standard states, 196
- entropy of molecularity
intramolecular processes, 225
- entropy of rotational restraint
intramolecular processes, 224–26
- entropy of transfer
hydrophobic effect, 231
- enzymatic activity
assembly of oligomers, 713
- enzymatic domain
domains, 379–82
- enzymatic method of sequencing
DNA, 103–5
- epidermal growth factor
nuclear magnetic resonance, 636
- epidermal growth factor receptor
cross-linking, 443
dimerization, 814–15
domains, 815
integral membrane-bound protein, 766, 814
quantitative cross-linking, 815
- epitope tagging, 567
- epitopes
conformationally specific, 561–62
cytochrome *c*, 561
definition, 558
lysozyme, 561
micrococcal nuclease, 562
neuraminidase, 562
on antigen, 558
poliovirus, 561
rhinovirus, 561
sequence specific, 561–62
- Eps15 homology domain
heterologous associations, 514
- equilibrium, 414
- equilibrium constant
concentration, units of, 197
- equilibrium constant for folding, 662
cold shock-like protein, 664
measurement, 666
ribonuclease, 687
thermodynamics of folding, 664
- equivalence point
immunoprecipitate, 564
- erythrocrucorin
aligning crystallographic molecular models, 369
molecular axes of symmetry, 481
- erythrocyte
cytoskeleton, 820
- erythronate-4-phosphate
dehydrogenase
molecular taxonomy, 396
- erythronolide synthase
domains, 381
- erythropoietin
oligosaccharides of glycoproteins, 130
- N*-(ethoxycarbonyl)-2-ethoxy-1,2-dihydroquinoline
reagent for covalent modification, 541
- ethyl acetoacetate
tautomers, 70
- N*-ethyl[2,3-¹⁴C₂] maleimide, 550
- N*-ethyl-5-phenylisoxazolium-3'-sulfonate
reagent for covalent modification, 541
- N*-ethylmaleimide
reagent for covalent modification, 536–37
- N*-ethyl-*N'*-[3-(dimethylamino)propyl]carbodiimide
reagent for covalent modification, 539, 547
- ETS-domain protein Elk-1
association of proteins with nucleic acid, 316
- evolution
of interface, 455–56, 469
of quaternary structure, 455
- evolution of proteins
alternative splicing, 350

- appearance of new protein, 359
- calmodulin, 351
- conservative replacement, 349
- crystallin, 350
- deleterious mutation, 348
- deletion, 350
- fibrinogen, 359
- gene duplication, 358
- genetic drift, 348
- glycine, 349
- histone H4, 351
- hydropathy, 366
- insertion, 350
- introns, 350
- isoforms, 358
- L-lactate dehydrogenase, 359
- malate dehydrogenase, 359
- mutation probability, 349–50
- neutral replacement, 348
- orthologues, 358
- paralogues, 358
- point mutation, 350
- positive selection, 358
- protein phosphatase 2A, 351
- splicing of messenger RNA, 350
- start site, 350
- stop site, 350
- tolerance to replacement, 366
- α tubulin, 351
- ubiquitin, 351
- evolutionarily shifting domains, 388
- evolutionary distance
 - aligning amino acid sequences, 355, 358
- EX₁ limit
 - proton exchange, 643–44
- EX₂ limit
 - proton exchange, 643–44
- exact rotational axis of symmetry
 - space groups, 461
- excimer
 - fluorescence, 610
- excluded volume
 - thermodynamics of folding, 681
- exo- α -2,3-sialidase
 - sequencing oligosaccharides, 134
- exo- α -sialidase
 - rotational axes of symmetry, 787
 - sequencing oligosaccharides, 134
- exoglycosidases
 - sequencing oligosaccharides, 134
- exon shuffling
 - domains, 386
- exopeptidases
 - sequencing of polypeptides, 91
- exotoxin A
 - molecular charge, 34
- expressing DNA
 - alkaline phosphatase promoter, 109
 - 5-aminolevulinate synthase, 108
 - expression system, 108
 - expression vector, 108
 - Factor Xa, 109
 - fusion proteins, 109
 - β -galactosidase, 109
 - glutathione transferase, 109
 - inclusion bodies, 109
 - lacZ* promoter, 108
 - renin, 109
 - restriction site, 109
 - T3 promoter, 108
 - T7 promoter, 108
 - tacII* promoter, 109
 - trc* promoter, 109
 - ubiquitin, 109
- expression
 - glucose transporter, 776
 - Na⁺/K⁺-exchanging ATPase, 776
 - opsin, 776
 - unspecific monooxygenase, 776
- expression of DNA
 - chinese hamster ovary cells, 110
 - histidine tails, 110
 - insect cells, 109
 - mammalian cells, 109
 - murine L cells, 110
 - polyhedrosis virus, 109
 - yeast, 109
- expression system
 - expressing DNA, 108
- expression vector
 - expressing DNA, 108
- extended polymers
 - sieving, 428
- extracellular matrix
 - domains, 386
- extracellular matrix protein COMP
 - coiled coil of α helices, 283
- extracytoplasmic surface
 - of a membrane, 743
- extrapolation
 - free energy of folding, 673–74
- F**
- Fab fragments
 - domains, 376
 - immunoglobulin G, 556
 - univalence, 556
- F-actin capping protein
 - assembly of actin, 730
- Factor VIII
 - sequence of DNA, 106
- Factor IX
 - nuclear magnetic resonance, 629
- Factor Xa
 - expressing DNA, 109
- Factor XII
 - domains, 386
- Factor D
 - molecular taxonomy, 396
- family of domains
 - molecular taxonomy, 396
- FASTA, 354
 - evaluation of, 368
- fast-atom bombardment
 - mass spectrometry, 92
- fatty acid-binding protein
 - kinetics of folding, 694
- fatty-acid synthase
 - domains, 381
- fatty-acid-binding protein
 - β structure, 260
 - water in crystallographic molecular models, 294
- Fc fragment
 - immunoglobulin G, 556
- ferredoxin
 - convergent evolution, 373
 - space groups, 464
- ferredoxin–NADP⁺ reductase
 - covalent modification, 550
 - domains, 376–77, 388
 - molecular taxonomy, 393
 - water in crystallographic molecular models, 293
- ferrichrome-iron receptor
 - crystallization, 772
 - crystallographic molecular model, 783
 - hydrophobic sheath, 779
- ferritin
 - interfaces, 489–90
 - octahedral symmetry, 489–90
 - quaternary structure, 489–90
 - tetrahedral symmetry, 489–90
- fibrillar collagen
 - helical cable, 505
- fibrin
 - assembly of helical polymers, 717–21
 - rotational axes of symmetry, 719
- fibrin monomer
 - fibrinogen, 717
- fibrinogen
 - α -helical coiled coil, 282, 717
 - diffusion coefficient, 577–78
 - domains, 388–89, 717
 - electron microscopy, 585, 587
 - evolution of proteins, 359
 - fibrin monomer, 717
 - fibrinopeptides, 717

- frictional coefficient, 577–78, 588
- frictional ratio, 577–78
- $(\alpha\beta\gamma)_2$ heterohexamer, 717
- immunoelectron microscopy, 567
- mean molar mass of an amino acid, 418
- sedimentation coefficient, 577–78
- structure of, 717–20
- fibrinopeptides
 - aligning amino acid sequences, 349
 - fibrinogen, 717
- fibronectin
 - frictional ratio, 577
 - X-ray scattering, 583
- fibronectin domain, 386
- fibulin
 - electron microscopy, 586
- flaggrin
 - amino acid sequence, 108
- fixed orientation
 - fluorescence resonance energy transfer, 607
- flagellin
 - image reconstruction, 502
- flavodoxin
 - molecular taxonomy, 395
 - recurring structure, 373
- flow rate
 - chromatography, 6
- fluid mosaic, 807–24
- fluorescence, 601–10
 - absorption spectrum, 595
 - collisional quencher, 602
 - emission of light, 594–95
 - emission spectrum, 595
 - excimer, 610
 - kinetics of folding, 689
 - lifetime of, 602
 - lysozyme, 601, 603
 - molten globule, 684
 - phosphoglycerate kinase, 603
 - quantum yield, 602
 - quenching of, 602
 - ribonuclease, 601
 - ribonuclease T₁, 603
 - tryptophan, 601
- fluorescence resonance energy transfer, 603–10
 - acceptor, 604
 - acetylcholine receptor, 608
 - actin, 609
 - aspartate carbamoyltransferase, 608
 - ATP-dependent DNA helicase Rep, 610
 - calmodulin, 608
 - chloramphenicol O-acetyltransferase, 608
 - conformational changes, 609
 - creatine kinase, 608
 - cyclic AMP-dependent protein kinase, 608–9
 - cyclic-AMP dependent protein kinase, 608
 - cytochrome *c*, 609
 - cytochrome-*c* oxidase, 609
 - distance between dipoles, 605
 - distance measurements, 607–9
 - DNA-directed DNA polymerase, 608–9
 - dolichyl-phosphate β - ∂ -mannosyltransferase, 609
 - donor, 604
 - efficiency of transfer, 604
 - fixed orientation, 607
 - fluorescent electrophiles, 606
 - GTP-binding protein Cdc42, 609
 - high mobility group protein Z, 610
 - lifetime of the excited state, 605
 - lysozyme, 609
 - myosin, 609
 - Na⁺/K⁺-exchanging ATPase, 609
 - orientation factor, 606–7
 - orientational freedom, 607
 - overlap integral, 606
 - pancreatic trypsin inhibitor, 608
 - photo-lyase, 607
 - RecA protein, 609
 - rhodopsin, 605
 - Rho-GDP dissociation inhibitor, 609
 - steroid Δ -isomerase, 607
 - transcription factor AP-1, 609
 - troponin, 609
- fluorescent electrophiles
 - covalent modification, 606
 - fluorescence resonance energy transfer, 606
- fluorosulfonic acids
 - reagents for covalent modification, 536
- fold of the symmetry
 - axes of symmetry, 452
- folded state, 659
- folding
 - approximation, 681
 - bovine pancreatic trypsin inhibitor, 685
 - carbonate dehydratase, 670
 - carboxypeptidase C, 679
 - chaperone, 705–8
 - chaperonin 60, 705–8
 - crystallin, 668–69
 - cytochrome *c*, 685
 - definition, 659
 - dihydrolipoyllysine-residue acetyltransferase, 683
 - domains, effect on, 680
 - equilibrium constant for, 662
 - heat shock proteins 70, 705, 707
 - hydrogen-bonds, contributions to, 680–81
 - intermediate states, 668
 - α -lactalbumin, 669, 685
 - lysozyme, 678
 - micrococcal nuclease, 678–79
 - nucleation model, 685
 - 3-oxoacid CoA-transferase, 679
 - penicillin amidase, 679
 - pH, effect on, 662
 - phosphoribosylanthranilate isomerase, 668, 679
 - reversibility, 663
 - ribonuclease, 668, 678–79, 685
 - ribonuclease H, 678
 - ribonuclease T₁, 663
 - scanning calorimetry, 664
 - subtilisin, 679
 - temperature effect on, 662
 - tryptophan synthase, 685
 - WW domains, 683
- folding of a polypeptide
 - isomerization, 660
- footprinting, 548
- formal charge
 - electronic structure, 56
- formate C-acetyltransferase
 - electron paramagnetic resonance, 645, 647–48
- formate dehydrogenase
 - molecular rotational axes of symmetry, 465
- formate-tetrahydrofolate ligase
 - purification, 26
- N*-formyl-[³⁵S]sulfonilmethionyl methylphosphate
 - covalent modification, 799
 - impermeant reagent for covalent modification, 799
- 5-formyltetrahydrofolate cyclo-ligase
 - purification, 28–9
- forward scattering
 - scattering of electromagnetic radiation, 579
- Fourier transform nuclear magnetic resonance spectrometer
 - nuclear magnetic resonance, 614–15
- Fourier-Bessel transform
 - image reconstruction, 502
- fractionation factor
 - hydrogen bond, 209

- hydrogen bonds in crystallographic molecular models, 312
- fragment ions
 - mass spectrometry, 93
- fragmin
 - assembly of actin, 730
- frameshift
 - sequence of DNA, 106
- free electrophoretic mobility
 - definition, 38
 - dodecyl sulfate gel electrophoresis, 421
 - electrophoresis, 41
- free energies of association
 - interfaces, 471
- free energies of formation
 - hydrogen bonds in crystallographic molecular models, 310
- free energies of solvation
 - hydrophobic effect, 235
- free energy of folding
 - chymotrypsin, 674, 677
 - chymotrypsin inhibitor, 677
 - cold shock protein CspB, 674
 - cytochrome *c*, 677
 - extrapolation, 673–74
 - immunoglobulin G, 675
 - lysozyme, 677
 - myoglobin, 677
 - pH effect on, 675
 - protein L9, 675
 - proton exchange, 675
 - ribonuclease, 676
 - ribonuclease H, 675–677
 - ribonuclease T₁, 675, 677
 - site-directed mutation effect on, 675
 - thermodynamics of folding, 673
 - thioredoxin, 675, 677
- free energy of transfer
 - guanidinium, 660
 - hydropathy side chains, 274
 - hydrophobic effect, 231–38
 - tryptophan, 276
 - tyrosine, 276
 - urea, 660
- free induction decay
 - nuclear magnetic resonance, 615
- free *R*-factor
 - crystallography, 176
- free-flow electrophoresis
 - cell fractionation, 744
- French pressure cell, 20
- frequency labeling
 - nuclear magnetic resonance, 617
- frequency of maximum absorption
 - nuclear magnetic resonance, 614
- frictional coefficient, 577–78
 - alcohol dehydrogenase, 578
 - catalase, 578
 - diffusion, 577–78
 - diffusion coefficient, 37
 - fructose-bisphosphate aldolase, 578
 - β -galactosidase, 578
 - hydration of a protein, 299
 - hydrodynamic particle, 574
 - in membranes, 812
 - lysozyme, 578
 - manganese-stabilizing protein, 578
 - minimal, 574
 - prothrombin, 578
 - sedimentation velocity, 576–78
 - serum albumin, 578
 - string of spherical beads, 576
- frictional ratio, 574
 - alcohol dehydrogenase, 578
 - apoferritin, 426
 - apolipoprotein(a), 577–78
 - aspartate carbamoyltransferase, 577
 - caldesmon, 577
 - catalase, 578
 - chymotrypsinogen, 426
 - collagen, 575, 577
 - cytochrome *c*, 426
 - desmin, 577
 - diffusion, 577–78
 - fibrinogen, 577–78
 - fibronectin, 577
 - fructose-bisphosphate aldolase, 578
 - β -galactosidase, 578
 - glyceraldehyde-3-phosphate dehydrogenase (phosphorylating), 426
 - L-lactate dehydrogenase, 426
 - lysozyme, 578
 - manganese-stabilizing protein, 578
 - myoglobin, 426
 - ovalbumin, 426
 - plasminogen, 577
 - polynucleotide 3'-phosphatase/5'-kinase, 577
 - prothrombin, 578
 - sedimentation velocity, 577–78
 - serum albumin, 578
 - sieving, 425
 - urease, 426
 - vinculin, 577
- Friedel pair
 - crystallography, 152
 - definition, 152
- fructose 1,6-bisphosphatase
 - purification, 26
- fructose-bisphosphate aldolase
 - assembly of oligomers, 713
- collisional quenching, 603
- covalent modification, 547
- diffusion coefficient, 578
- dodecyl sulfate gel electrophoresis, 422
- electrophoresis, 41
- frictional coefficient, 578
- frictional ratio, 578
- heterooligomers, 508
- isoforms, 439–40
- molar mass, 418, 419
- molecular charge, 36
- quaternary structure, 439–40
- sedimentation coefficient, 578
- sieving, 424, 427–28
- L-fucose
 - structure, 129
- α -L-fucosidase
 - sequencing oligosaccharides, 134
- L-fucose-phosphate aldolase
 - point group, 469
- fumarate hydratase
 - assay, 13, 19
 - assembly of oligomers, 713
 - covalent modification, 536
 - dodecyl sulfate gel electrophoresis, 422
 - interfaces, 480
 - purification, 3
 - sieving, 424
- functional domain
 - domains, 382
- fundamental unit cell
 - crystallography, 151
- fusion proteins
 - expressing DNA, 109
- G**
- γ turn
 - secondary structure, 264
- galactonate dehydratase
 - assay, 19
- D-galactose
 - structure, 129
- galactose oxidase
 - domains, 382
 - map of electron density, 165
 - posttranslational modification, 122
- α -galactosidase
 - crystallization, 49
- β -galactosidase
 - diffusion coefficient, 578
 - expressing DNA, 109
 - frictional coefficient, 578
 - frictional ratio, 578
 - sedimentation coefficient, 578
 - sequencing oligosaccharides, 134

- sieving, 424
- yeast two-hybrid assay, 518
- galaside
 - glycosphingolipid, 748
- ganglioside
 - glycosphingolipid, 748
- gap junction connexon
 - dihedral symmetry, 787
 - rotational axes of symmetry, 787
- gap penalty
 - aligning amino acid sequences, 353
- gap percentage
 - aligning amino acid sequences, 351
- gap-junction channel
 - image reconstruction, 793
- gaps
 - aligning amino acid sequences, 350–51
 - aligning crystallographic molecular models, 364
- gas–liquid chromatography, 4
- gelsolin
 - assembly of actin, 730
 - domains, 384
 - heterologous associations, 516
- gene duplication
 - evolution of proteins, 358
- gene fusion
 - domains, 381
 - evolution of proteins, 373–74
- gene multiplication
 - domains, 384
- general control protein GCN4
 - coiled coil of α helices, 283–84
 - interfaces, 480
 - space groups, 464
- genetic code, 98
 - aligning amino acid sequences, 349
- genetic detachment
 - domains, 377
- genetic drift
 - evolution of proteins, 348
- genomic sequences
 - sequence of DNA, 108
- geodesic domes
 - icosahedral symmetry, 496
- geranylgeranylation
 - posttranslational modification, 117
- geranyltranstransferase
 - assay, 14
- g factor
 - electron paramagnetic resonance, 648
- γ -glutamyltransferase
 - covalent modification, 546, 550
- glial filaments
 - intermediate filaments, 506
- γ -linolenic acid, 747
- global rotational axis of symmetry
 - definition, 491
 - icosahedral symmetry, 491
 - quasi-equivalence, 491
- globins
 - aligning amino acid sequences, 356
 - aligning crystallographic molecular models, 369–72
- globoside
 - glycosphingolipid, 748
- glucan 1,4- α -glucosidase
 - molecular taxonomy, 396
- 4- α -glucanotransferase
 - aligning crystallographic molecular models, 362
 - interface in, 478
- glucarate dehydratase
 - aligning amino acid sequences, 361
- D-glucose
 - structure, 129
- glucose oxidase
 - domains, 382
 - interfaces, 480
- glucose transporter
 - expression, 776
- glucose-6-phosphate isomerase
 - covalent modification, 545
 - interfaces, 481
 - peptide map, 435, 437–38
- glucose-fructose oxidoreductase
 - interfaces, 480
- α -glucosidase
 - crystallography, 183
- D-glucuronic acid
 - structure, 129
- glutamate
 - acid dissociation constant, 75
 - covalent modification, 539–41
 - stereochemistry of side chains, 270
- glutamate dehydrogenase, 418
 - dodecyl sulfate gel electrophoresis, 422
- glutamate–ammonia ligase
 - molar mass, 420
 - posttranslational modification, 125
 - quaternary structure, 475
- glutamate–tRNA ligase
 - peptide map, 435
- glutamic acid
 - electronic structure, 79
 - water in crystallographic molecular models, 296
- glutaminase
 - domains, 379
- glutamine
 - acid dissociation constant, 75
 - electronic structure, 79
 - hydropathy, 276
 - stereochemistry of side chains, 270
 - water in crystallographic molecular models, 296
- glutamine amidotransferase
 - dodecyl sulfate gel electrophoresis, 432
- glutamine γ -glutamyltransferase
 - topography of membrane-spanning proteins, 799
- glutamine–fructose-6-phosphate transaminase (isomerizing)
 - domains, 380
- glutamine–pyruvate transaminase
 - assay, 19
- glutamyl endopeptidase
 - cleavage of polypeptide, 88
 - domains, 378
- glutamyl-tRNA reductase
 - purification, 31
- glutaraldehyde
 - cross-linking, 443
- glutaredoxin 2
 - nuclear magnetic resonance, 630
- glutathione
 - cystines, formation of, 708
 - function, 122
 - structure, 122
- glutathione-disulfide reductase
 - domains, 382, 388–89
 - molecular taxonomy, 394–95
 - point group, 467
 - stereochemistry of side chains, 267
- glutathione peroxidase
 - space groups, 463
- glutathione synthase
 - aligning crystallographic molecular models, 362
 - domains, 388
 - molecular taxonomy, 397
 - rotational axes of symmetry, 483
- glutathione transferase
 - detection of heterologous associations, 515
 - expressing DNA, 109
 - interfaces, 479
- glyceraldehyde-3-phosphate dehydrogenase (phosphorylating)
 - accessible surface area, 273
 - aligning crystallographic molecular models, 366
 - assay, 17
 - circular dichroism, 600
 - covalent modification, 542

- dodecyl sulfate gel electrophoresis, 422
- frictional ratio, 426
- molar mass, 418
- molecular rotational axes of symmetry, 464
- molecular taxonomy, 395–96
- purification, 24, 48
- sieving, 424
- space groups, 463
- glycerate dehydrogenase
 - molecular taxonomy, 396
 - quaternary structure, 483
- glycerol kinase
 - cross-linking, 442
- glycine
 - electronic structure, 76
 - evolution of proteins, 349
- glycine hydroxymethyltransferase
 - circular dichroism, 598
- glycoforms
 - oligosaccharides of glycoproteins, 130
- glycolipids
 - lipopolysaccharide, 749
 - vectorial insertion into membranes, 767
- glycopeptides
 - chromatography, 133
 - oligosaccharides on glycoproteins, 133
- glycophorin
 - covalent modification from within the bilayer, 797
 - cytoskeleton, 821
 - embedded anchor, 774
 - membrane-bound protein, 767
 - oligomeric interfaces, 790
- glycoprotein
 - definition, 127
- glycoproteins
 - vectorial insertion into membranes, 767
- N*-glycosidic linkage
 - oligosaccharides of glycoproteins, 128
- O*-glycosidic linkage
 - oligosaccharides of glycoproteins, 128
- glycosphingolipids, 748
 - arthroside, 748
 - cerebroside, 748
 - galaside, 748
 - ganglioside, 748
 - globoside, 748
 - in rafts, 811
 - isogloboside, 748
 - lactoside, 748
 - molluside, 748
 - mucoside, 748
 - neolactoside, 748
 - schistoside, 748
- glycosylase MutY
 - metalloproteins, 330
- glycosylation
 - immune complex, 558
 - topography of membrane-spanning proteins, 800
- glycosylphosphatidylinositol (GPI) anchor
 - posttranslational modification, 118
- glycosylphosphatidylinositol
 - diacylglycerol-lyase
 - glycosylphosphatidylinositol-linked proteins, 765
- glycosylphosphatidylinositol-linked proteins
 - acetylcholinesterase, 765
 - anchored membrane-bound proteins, 765
 - carbonate dehydratase, 765
 - cell surface glycoprotein a-2, 765
 - glycosylphosphatidylinositol diacylglycerol-lyase, 765
 - receptor for the Fc domain of immunoglobulinG, 765
 - variant surface glycoprotein, 765
- GMP synthase
 - domains, 380
- Golgi membranes, 743
 - cell fractionation, 744
 - synthesis of glycoproteins, 767
- good solvent
 - definition, 659
- Gouy-Chapman equation, 754
- G-protein
 - adenylate cyclase system, 817
 - hydrolysis of GTP, 817
 - subunits, 817
- gradient chromatography, 7
- gradient of concentration
 - sedimentation equilibrium, 411
- granulin
 - domains, 384
- granulocyte-colony-stimulating factor
 - molecular taxonomy, 399
- growth hormone
 - molecular taxonomy, 399
- growth hormone receptor
 - asymmetric complex, 816
 - dimerization, 816
 - integral membrane-bound protein, 816
- GTP
 - assembly of microtubules, 726–29
- GTP-binding protein Cdc42
 - fluorescence resonance energy transfer, 609
- guanidinium
 - denaturant, 660
 - free energies of transfer, 660
 - preferential solvation, 22, 661
- guanidinium cation
 - electronic structure, 80
- guanine
 - electronic structure, 65
- guanine and arginine
 - hydrogen bonding, 316
- guanine nucleotide-binding protein
 - heterologous associations, 517
- guanylate kinase
 - aligning crystallographic molecular models, 364
- Guinier plot
 - X-ray scattering, 581
- H**
 - H⁺/K⁺-exchanging ATPase
 - aligning amino acid sequences, 364
 - topography of membrane-spanning proteins, 802–3
 - H⁺-exchanging ATPase
 - topography of membrane-spanning proteins, 802–3
 - H⁺-transporting two-sector ATPase
 - aligning amino acid sequence, 360
 - halocyanin
 - resonance Raman spectrum, 596
 - halorhodopsin
 - crystallization, 772
 - membrane-spanning helices, 772
 - hanging drop
 - crystallization, 50
 - hapten, 563
 - antigen, 562
 - haptoglobulin
 - aligning amino acid sequence, 360
 - hardness
 - metal ion, 327
 - harsh treatments
 - produce heterogeneity, 49
 - head group
 - phospholipid, 747
 - heat capacity
 - hydrophobic effect, 231
 - kinetics of folding, 703
 - water, 191
 - heat capacity change of folding
 - thermodynamics of folding, 671

- heat of fusion
 - bilayer of phospholipid, 755
- heat shock protein 16.5
 - octahedral symmetry, 487–88
- heat shock protein 70
 - folding, 705, 707
- heat-labile enterotoxin
 - molecular rotational axes of symmetry, 465
 - point group, 469
- heavy atom
 - multiple isomorphous replacement, 158
- hedgehog protein
 - postranslational modification, 115
- helical cables
 - fibrillar collagen, 505
 - intermediate filament, 506
 - keratin, 508
 - of helical polymers, 502
- helical nets
 - packing of side chains, 280
- helical polymers, 499–508
 - axes of symmetry, 452
 - collagen, 503–6
 - helical cables of, 502
 - image reconstruction, 501–3
 - microtubule, 506
 - tRNA-intron endonuclease, 455
- helical surface lattice, 499
 - CA protein, 500
 - designation of, 499
 - flagellin, 499
 - microtubule, 722
 - radial angle, 500
 - T4 bacteriophage, 500
 - thick filament of myosin, 731–32
 - tobacco mosaic virus, 499–500
- helical wheel
 - α helix, 258–59
- helix
 - geometric parameters, 499
- 3_{10} helix
 - β turn, 262
- hemagglutinin glycoprotein
 - domains, 388
- heme
 - metalloproteins, 330
- heme-binding protein 23
 - interfaces, 480
- hemocyanin
 - domains, 384
 - postranslational modification, 122
 - resonance Raman spectrum, 596
 - rotational axes of pseudosymmetry, 485
- hemoglobin
 - α helix, 259
 - aligning amino acid sequences, 360
 - aligning crystallographic molecular models, 369–72
 - electrophoresis, 41, 45
 - heterooligomers, 508
 - hydration of a protein, 298
 - infrared spectrum, 595
 - interfaces, 471–72
 - mean molar mass of an amino acid, 418
 - molecular taxonomy, 394, 398
 - osmotic pressure, 420
 - peptide map, 432–33
 - quaternary structure, 407, 451
 - sieving, 428
 - stereochemistry of side chains, 267
- α -hemolysin
 - β barrel, 776
 - crystallization, 772
 - punching a hole in a membrane, 803
- hemopexin domain, 386
- Henry's function
 - electrophoresis, 40
- heptad repeat
 - coiled coil of α helices, 282
 - intermediate filaments, 506
- heterocycle
 - tautomers, 73
- heterocyclic side chain
 - hydropathy, 242
- $(\alpha\beta\gamma)_2$ heterohexamer
 - fibrinogen, 717
- heterologous association
 - definition, 508
- heterologous associations
 - α actinin, 513, 516
 - ankyrin, 516
 - annexin II, 514
 - CD40 tumor necrosis factor receptor, 514
 - cyclin, 517
 - detection, 515
 - β -dystroglycan, 513
 - dystrophin, 513
 - E-cadherin, 516
 - Eps15 homology domain, 514
 - gelsolin, 516
 - guanine nucleotide-binding protein, 517
 - heterooligomers, 508
 - histocompatibility antigen, 513, 517
 - importin, 514
 - integrin, 516
 - interfaces, 513
 - laminin, 516
- modular domains, 513
- myosin light chain kinase, 517
- nuclear import factor karyopherin α , 514
- nuclear localization signals, 514
- nucleolin, 516
- nucleoporin, 517
- PDZ domains, 514
- protein p11, 514
- protein-tyrosine phosphatase, 517
- proto-oncogene protein c-fos, 519
- proto-oncogene protein-tyrosine kinase ABL1, 517
- ribulose-bisphosphate carboxylase, 510
- SHC transforming protein, 517
- somatotropin, 519
- somatotropin receptor, 519
- synaptotagmin, 516
- T-cell receptor, 513
- α -thrombin, 513
- thrombomodulin, 513
- titin, 513
- transcription factor AP-1, 519
- transcription initiation factor TFIID, 517
- transitory, 513
- troponin C, 514
- troponin I, 514
- tumor necrosis factor receptor-associated factor 2, 514
- vitronectin, 516
- heterologous interface
 - definition, 508
- heterologous interfaces
 - aspartate carbamoyltransferase, 508–10
 - charged side chains, 513
 - elongation factor Ts, 510
 - elongation factor Tu, 510
 - heterooligomers, 508
 - immunoglobulin G, 513
 - interleukin-1, 513
 - interleukin-1 receptor, 513
 - karyopherin β 2, 513
 - protein G, 513
 - ribonuclease, 513
 - ribonuclease inhibitor, 513
 - ribulose-bisphosphate carboxylase, 510
 - synaptobrevin-II, 513
 - syntaxin-1A, 513
- heteromultimeric protein, 451
- heterooligomers, 508–19
 - aspartate carbamoyltransferase, 508–10
 - assembly of oligomers, 713–17

- fructose-bisphosphate aldolase, 508
- heterologous association, 508
- heterologous interface, 508
- histocompatibility antigen, 512
- homologous subunits, 508
- laminin, 513
- mismatched symmetry, 511
- modular domains in, 513
- multicatalytic endopeptidase complex, 508
- nidogen, 513
- nonstoichiometric ratios of subunits, 511
- pseudosymmetry, 508
- steric exclusion, 510
- hexagonal lattice
 - crystallography, 151
- hexamers of dimers
 - tetrahedral symmetry, 487
- hexokinase
 - aligning crystallographic molecular models, 363
 - axes of symmetry, 454
 - domains, 384
 - molecular taxonomy, 395
 - point group, 467
 - purification, 29
 - sieving, 427
- hierarchical classification
 - molecular taxonomy, 393
- high density lipoprotein
 - lipoproteins, 805
- high mobility group protein Z
 - fluorescence resonance energy transfer, 610
- highest occupied molecular orbital, 58
- high-mannose oligosaccharides
 - oligosaccharides on glycoproteins, 130
- high-pressure liquid
 - chromatography, 6
- high-resolution mass spectrum
 - posttranslational modification, 119
- histidine
 - acid dissociation constant, 75
 - covalent modification, 530–31, 536
 - electronic structure, 77
 - microscopic dissociation constants, 79
 - nuclear magnetic resonance, 635
 - tautomers, 78
 - titration curve, 79
- histidine ammonia-lyase
 - interfaces, 480
 - posttranslational modification, 126
- histidine decarboxylase
 - posttranslational modification, 114
- quaternary structure, 475
 - subunits, 436
- histidine tails
 - expression of DNA, 110
- histidinol-phosphate transaminase
 - binding of ligand, 47
- histocompatibility antigen
 - aligning amino acid sequences, 360
 - heterologous associations, 513, 517
 - heterooligomers, 512
 - hydrogen bonds in crystallographic molecular models, 306
 - packing of side chains, 279
- histone
 - association of proteins with nucleic acid, 316, 320
- histone H4
 - evolution of proteins, 351
- HIV-1 retropepsin
 - assembly of oligomers, 713
- HLA histocompatibility antigen
 - embedded anchor, 774
 - purification, 773
- HLA-linked B-cell antigen
 - purification, 773
- HMBC
 - nuclear magnetic resonance, 621
- HMQC
 - nuclear magnetic resonance, 621
- HNRNP arginine methyltransferase
 - quaternary structure, 476
- Hofmeister series, 22
- HOHAHA
 - nuclear magnetic resonance, 621
- homeodomain protein MAT α 2
 - association of proteins with nucleic acid, 320
- homogenization, 1, 20
- homologous subunits
 - heterooligomers, 508
- homologues
 - aligning amino acid sequences, 346
- homooligomeric proteins
 - frequency, 466
- homoserine dehydrogenase
 - domains, 378
- HSMQC
 - nuclear magnetic resonance, 621
- HSQC
 - nuclear magnetic resonance, 621
- Hurler corrective factor
 - assay, 19
- hybrid oligomers, 439
- hybridization
 - acids and bases, 63
- hybridization of DNA
 - cloning of DNA, 100
- hydrated effective sphere
 - definition, 574
- hydration, 577–78
 - bilayer of phospholipid, 751
 - definition, 189
 - X-ray scattering, 583
- hydration of a protein
 - accessible surface area, 299
 - chymotrypsin, 298
 - chymotrypsinogen, 298
 - dielectric relaxation, 297
 - dihydrofolate reductase, 299
 - frictional coefficient, 299
 - hemoglobin, 298
 - heterogeneity, 299
 - α -lactalbumin, 298
 - β -lactoglobulin, 298
 - lysozyme, 298–99
 - myoglobin, 298
 - oligomeric proteins, 577
 - ovalbumin, 298
 - pepsin, 298
 - preferential solvation, 297
 - quantification, 296–300
 - ribonuclease, 298–99
 - scattering at small angles, 299
 - self-diffusion of water, 297
 - serum albumin, 298–99
 - unfrozen water, 297
- hydrodynamic particle
 - definition, 573
 - ellipsoid of revolution, 574
 - frictional coefficient, 574
 - mass of, 573
 - volume of, 573
- hydrodynamic radius
 - definition, 574
- hydrogen
 - scattering length, 583
- hydrogen atoms
 - crystallography, 182
- hydrogen bonds, 204–22
 - amide in water, 217
 - angular dependence, 264–66
 - apparent equilibrium constant in water, 219
 - aromatic ring, 208
 - association equilibrium constant, 210
 - bifurcated, 206
 - bond angles, 206
 - bond length, 205
 - competition of water, 220
 - compressed, 214
 - covalency, 215
 - definition, 204
 - deshielding, 209

- difference in pK_a , 210
- distance between a donor and acceptor, 213
- effect of water, 216–20
- electrostatic attraction, 215
- enthalpy of formation, 210–11
- entropy of formation, 210
- fractionation factor, 209
- free energy of formation of low-barrier hydrogen bond, 215
- infrared spectrum, 209
- in interfaces, 478
- integral membrane-bound proteins, 782
- intramolecular, 207, 227
- low-barrier, 208–10
- lysozyme, 267
- membrane-spanning α helices, 775
- N*-methylacetamide, 217
- nuclear magnetic resonance, 209
- potential energy of, 213
- proton exchange, 641
- secondary structure, 264
- solvation, 208
- strength, 210
- stretching frequency, 209
- strong, short, 214
- strongest possible, 212
- sulfur, 207
- symmetric, 212
- water, 190
- water in crystallographic molecular models, 295
- wells of potential energy, 208
- zero-point energy, 208
- hydrogen bonds in crystallographic molecular models, 306–14
 - approximation, 311
 - aromatic side chain, 306
 - Bence-Jones protein, 308
 - buried hydrogen bonds, 306
 - chymotrypsin, 306, 308
 - clusters of hydrogen bonds, 306
 - contributions to folding, 680–81
 - crystallin, 308
 - deoxyribonuclease, 307–08
 - dihydrolipoyllysine-residue acetyl transferase, 309
 - donors and acceptors on the side chains, 306
 - double-mutant cycle, 310
 - entropy of approximation, 309
 - fractionation factor, 312
 - free energies of formation, 310
 - frequency, 306
 - histocompatibility antigen, 306
 - hydrogen-bond balance, 307
 - 4-hydroxybenzoate
 - 3-monooxygenase, 309
 - immune complex, 559–60
 - length of the hydrogen bond, 312
 - low-barrier hydrogen bonds, 312
 - lysozyme, 311
 - micrococcal nuclease, 312
 - myoglobin, 306, 309, 312
 - omit maps, 309
 - penicillopepsin, 309
 - phosphocarrier protein HPr, 312
 - protein-tyrosine kinase, 312
 - ribonuclease, 310, 313
 - ribulose-bisphosphate
 - carboxylase, 306–7, 306
 - stability of a protein, 311
 - stereochemistry, 306
 - steric effects, 309
 - streptococcal protein G, 312
 - sulfate-binding protein, 309
 - superoxide dismutase, 306
 - thermolysin, 311
 - transferrin, 312
 - troponin C, 312
 - trypsin, 309
 - tryptophan, 308
 - water, 308
 - hydrogen bonds in DNA, 230
 - hydrogen-bond balance
 - hydrogen bonds in
 - crystallographic molecular models, 307
 - hydrogen-bonded nearest neighbors
 - water, 193
 - hydrogen–carbon bonds
 - hydropathy side chains, 274
 - hydrophobic effect, 234
 - hydropathy, 241–46
 - accessibilities to water, 244
 - amide, 242
 - arginine, 242
 - asparagine, 276
 - carboxylic acid, 242
 - definition, 241
 - evolution of proteins, 366
 - glutamine, 276
 - heterocyclic side chain, 242
 - hydroxyl group, 241
 - lysine, 242
 - peptide bond, 242
 - scales of, 244
 - sulfur, 241
 - transfer between water and the gas, 241
 - hydropathy of side chains
 - aromatic amino acids, 275
 - cohesin domain, 275
 - free energies of transfer, 274
 - hydrogen–carbon bonds, 274
 - hydrophilic amino acids, 275
 - lysozyme, 275
 - polypeptide backbone, 276
 - ribonuclease T₁, 275
 - hydrophilic amino acids
 - hydropathy side chains, 275
 - hydrophobic clusters
 - aligning crystallographic molecular models, 369
 - hydrophobic collapse
 - kinetics of folding, 696
 - hydrophobic effect, 230–41
 - clathrates, 233
 - compensatory thermodynamic changes, 233
 - contributions to, 238
 - definition, 231
 - dielectric relaxation time, 233
 - enthalpy change, 233
 - entropy of transfer, 231
 - free energies of solvation, 235
 - free energy of transfer, 231–38
 - heat capacity, 231
 - hydrogen–carbon bonds, 234
 - in interfaces, 479
 - integral membrane-bound proteins, 783
 - neutron scattering, 233
 - partition coefficients between gas and water, 235–37
 - size of the cavity, 234
 - surface area, 238
 - thermodynamic properties of the water, 231
 - van der Waals forces, 235–37
 - 3-hydroxy fatty acids, 749
 - N*-[2-hydroxy-1,1-bis(hydroxymethyl)ethyl]glycine
 - buffer, 68
 - L-2-hydroxyisocaproate
 - dehydrogenase
 - molecular rotational axes of symmetry, 465
 - molecular taxonomy, 396
 - 4-hydroxy-2-oxoglutarate aldolase
 - aligning amino acid sequence, 360
 - 2-hydroxy-6-ketonona-2,4-diene-1,9-dioic acid 5,6-hydrolase
 - assay, 18
 - (S)-2-hydroxy-acid oxidase
 - space groups, 463
 - 3-hydroxyacyl-[acyl-carrier-protein] dehydratase
 - domains, 381

- 3-hydroxyacyl-CoA dehydrogenase assay, 17
- 4-hydroxybenzoate
3-monooxygenase
hydrogen bonds in
crystallographic molecular models, 309
- 3-hydroxybutyrate dehydrogenase
anchored membrane-bound proteins, 764
- 1-(2-hydroxyethyl)-4-(3-sulfoethyl) piperazine
buffer, 68
- 1-(2-hydroxyethyl)-4-(3-sulfopropyl) piperazine
buffer, 68
- hydroxyl group
hydropathy, 241
- hydroxylamine
cleavage of polypeptide, 90
- hydroxylapatite, 8
- hydroxymethylglutaryl-CoA lyase
assay, 17
- hydroxymethylglutaryl-CoA reductase
anchored membrane-bound proteins, 764
- 2-hydroxyphytanoyl-CoA lyase
assay, 13
- 4-hydroxyproline
collagen, 504
- N*-hydroxysuccinimide esters
reagents for covalent modification, 535
- hyperfine splitting
electron paramagnetic resonance, 647
- I**
- ice Ih
structure, 192
water, 190
- ICl
reagent for covalent modification, 538
- icosahedral point group 532, 488
- icosahedral symmetry
protein coat of satellite tobacco necrosis virus, 494–95
global rotational axis of symmetry, 491
hexagonal expansion, 497
protein coat of a virus, 488–98
protein coat of satellite panicum mosaic virus, 488–89
protein coat of southern bean mosaic virus, 494
- protein coat of tomato bushy stunt virus, 494
- quasi-equivalence, 491–98
- ideal gas law
osmotic pressure, 409
- image reconstruction
acetylcholine receptor, 793
actin, 503
amorphous ice, 501, 790
amplitude of electron diffraction, 792
amplitudes, 790
aquaporin, 793
Ca²⁺-transporting ATPase, 793
computational methods, 501
computed Fourier transform, 501
difference maps of scattering density, 502
distribution of scattering density, 501
electron microscope, 501
electron-scattering density, map of, 793
flagellin, 502
Fourier-Bessel transform, 502
gap-junction channel, 793
helical polymers, 501–3
lattice lines, 791
light-harvesting chlorophyll *a/b*-protein complex, 793
membrane-bound proteins, 790–93
Na⁺/K⁺-exchanging ATPase, 793
negative staining, 501
phase of Fourier transform, 792
phases, 790
refinement, 793
sodium/proton antiporter NhaA, 793
tilt, 791
tubulin, 502
two-dimensional crystalline array, 790
- imidazole
electronic structure, 78
- imidazole glycerol phosphate synthase
domains, 380, 383
- imidazoleglycerol-phosphate dehydratase
assay, 17
- iminothiolane
reagent for covalent modification, 549
- immune complexes
between immunoglobulin and antigen, 558
- complementarity-determining regions in, 559
- conformational changes, 561
- cytochrome *c*, 560
- glycosylation, 558
- hydrogen bonds, 559–60
- interfaces, 559
- lysozyme, 558–61
- micrococcal nuclease, 559–60
- viruses, 561
- water, 560
- immunity protein
light scattering, 591
- immunity protein Im9
nuclear magnetic resonance, 639
- immunization
to elicit immunoglobulins, 555
- immunoabsorbent, 563
definition, 566
dystrophin, 566
protein A, 566
purifying a peptide, 563
Shaker S4 K⁺ channel, 566
to follow covalent modification, 545
voltage-gated chloride channel, 566
- immunoabsorption, 566
membrane-bound proteins, 773
- immunoblotting, 565
- immunodiffusion, 47, 564
- immuno-electron microscopy, 567–68
fibrinogen, 567
 α 2-macroglobulin, 567
multicatalytic endopeptidase, 567
ribosome, 567–69
- immuno-electrophoresis, 47
- immunoglobulin A
structure, 557
- immunoglobulin D
oligosaccharides of glycoproteins, 127
- immunoglobulin domain, 386
- immunoglobulin ϵ receptor
steric exclusion, 510
- immunoglobulin G
bivalent fragment, 556
camel, 559
collisional quenching, 603
domains, 376, 378, 382, 390, 477, 555
electrophoresis, 42
Fab fragment, 556
Fc fragment, 556
free energy of folding, 675
frictional coefficient, 588
function, 555–57
heterologous interfaces, 513
hinges, 557

- impermeant reagents, 800
- infrared spectrum, 595
- mean molar mass of an amino acid, 418
- sieving, 424, 428
- structure, 555–57
- thermodynamics of folding, 682
- topography of membrane-spanning proteins, 800
- immunoglobulin G binding protein G
 - kinetics of folding, 694
 - nuclear magnetic resonance, 633
- immunoglobulin M
 - structure, 557
 - X-ray scattering, 584
- immunoglobulins
 - antibodies, 555
 - binding to antigens, 555
 - complementarity-determining regions, 558
 - function, 555
 - in serum, 555
 - molecular taxonomy, 397
 - monoclonal, 558
 - myeloma protein, 557
 - packing of β sheets, 285
 - packing of side chains, 281
 - polyclonal, 557
 - proline isomerization, 701
 - production by lymphocytes, 557
 - proton exchange, 645
 - purifying a peptide, 563
 - screen libraries, 567
 - immunoprecipitation, 564
 - detection of heterologous associations, 515
 - equivalence point, 564
- immunostaining
 - alkaline phosphatase, 566
 - amino terminus, 566
 - carboxy terminus, 566
 - cross-linking, 566
 - dicarboxylate transporter, 566
 - electrophoresis, 566
 - end-labeling, 567
 - NADH dehydrogenase (ubiquinone), 565–66
 - peroxidase, 566
- IMP dehydrogenase
 - point group, 469
 - quaternary structure, 475
- impermeant reagents
 - covalent modification, 798
 - topography of membrane-spanning proteins, 798
- impermeant solute
 - osmotic pressure, 408
- importin
 - heterologous associations, 514
- included volume
 - chromatography, 12
- inclusion bodies
 - expressing DNA, 109
- incremental scattering
 - light scattering, 415
- independently folding domain, 389
- independently shifting domains, 388
- index
 - crystallography, 151–53
- individual domain
 - molecular taxonomy, 393
- indole
 - tryptophan, 76
- indole-3-glycerol-phosphate synthase
 - domains, 377, 384
- induction
 - acids and bases, 63
- infrared absorption
 - water, 195
- infrared light
 - absorption of, 594
- infrared spectroscopy
 - selection rules, 595
- infrared spectrum
 - amide, 595
 - hemoglobin, 595
 - hydrogen bond, 209
 - immunoglobulin G, 595
 - of a protein, 595
 - ribonuclease, 595
 - secondary structure, 596
- inhibitors of peptidases
 - purification, 49
- initiation codon
 - sequence of DNA, 106
- initiation factor IF3
 - domains, 390
- inorganic diphosphatase
 - assembly of oligomers, 710
 - metalloproteins, 329, 332
 - molecular rotational axes of symmetry, 465
- insect cells
 - expression of DNA, 109
- insertion
 - evolution of proteins, 350
- insulin receptor
 - subunits, 436
- insulin-like growth factor
 - cystines, formation of, 708
- intact cells
 - topography of membrane-spanning proteins, 798
- integral membrane-bound proteins, 766
 - β -adrenergic receptor, 816
 - β barrel, 776
 - boundary layer of phospholipid, 784–86
 - closed structures, 787
 - crystallization, 775
 - cyclic symmetry, 787
 - diacylglycerol kinase, 766
 - epidermal growth factor receptor, 766
 - growth hormone receptor, 816
 - hydrogen bonds, 782
 - hydrophobic effect, 783
 - membrane-spanning α helices, 776
 - NADH dehydrogenase (ubiquinone), 766
 - oligomers, 786
 - packing of the α helices, 781
 - quantitative cross-linking, 786
 - rotational axes of
 - pseudosymmetry, 789
 - rotational axes of symmetry, 787
 - ryanodine receptor, 766
 - spin-labeled phospholipids, 795
 - water in the interior of, 781
- integrin
 - heterologous associations, 516
- intein
 - posttranslational modification, 115
- interdigitation of side chains
 - packing of side chains, 278
- interface
 - Cro protein, 478
 - definition, 455
 - evolution of, 455–56, 469
 - 4- α -glucanotransferase, 478
 - phosphopyruvate hydratase, 478
- interfaces
 - adenylosuccinate lyase, 480
 - ADP-ribose diphosphatase, 481
 - alcohol dehydrogenase, 480
 - arginine in, 478
 - carboxylesterase ESTA, 478–80
 - catalase, 481
 - chloramphenicol
 - O-acetyltransferase, 480
 - 4-chlorobenzoyl-CoA
 - dehalogenase, 481
 - concanavalin A, 480
 - dihydrolipoyllysine residue
 - acetyltransferase, 479
 - dihydroneopterin aldolase, 480
 - 6,7-dimethyl-8-ribityllumazine
 - synthase, 488
 - ferritin, 489–90

- free energies of association, 471
- fumarate hydratase II, 480
- general control protein GCN4, 480
- glucose oxidase, 480
- glucose-6-phosphate isomerase, 481
- glucose-fructose oxidoreductase, 480
- glutathione transferase, 479
- heme-binding protein 23, 480
- hemoglobin, 471–72
- heterologous associations, 513
- histidine ammonia-lyase, 480
- hydrogen bonds in, 478
- hydrophobic effect in, 479
- immune complex, 559
- interleukin-5, 481
- intimin receptor, 480
- isometric structures, 488
- κ bungarotoxin, 480
- lac* repressor, 480
- lectin, 480
- mannose-binding protein, 480
- methyl-accepting chemotaxis protein II, 480
- oligomeric integral membrane-bound proteins, 790
- oligomeric proteins, 478
- porin, 479
- protein coat of rhinovirus 14, 480
- protein coat of satellite panicum mosaic virus, 489
- quasi-equivalence, 492–94
- reversible dissociation, 471
- ribulose-phosphate 3-epimerase, 480
- structural swapping, 480–81
- structure of, 478
- superoxide dismutase, 480
- urate oxidase, 480
- variant surface glycoprotein, 480
- interfacial denaturation chromatography, 3
- interleukin
 - molecular taxonomy, 399
- interleukin 1
 - heterologous interfaces, 513
- interleukin 1 β
 - aligning crystallographic molecular models, 366
 - nuclear magnetic resonance, 628
 - water in crystallographic molecular models, 294
- interleukin 4
 - nuclear magnetic resonance, 629, 639
- interleukin 5
 - interfaces, 481
- interleukin 6
 - molten globule, 684
- interleukin 13
 - nuclear magnetic resonance, 627
- interleukin-1 receptor
 - heterologous interfaces, 513
- intermediate filament
 - helical cable, 506
- intermediate filaments, 506–8
 - coiled coils of α helices, 506
 - desmin filaments, 506
 - glial filaments, 506
 - heptad repeat, 506
 - keratin filaments, 506
 - neurofilaments, 506
 - tonofilaments, 506
 - vimentin filaments, 506
- intermediate states
 - folding, 668
 - molten globule, 683
- intermolecular aggregation
 - kinetics of folding, 705
- internal duplication
 - domains, 383
- internal duplications
 - molecular rotational axes of pseudosymmetry, 476
- internally repeating domain
 - domains, 382
- interstitial molecules of water, 194
- interstrand hydrogen bond
 - collagen, 504
- intersystem crossing, 542
 - absorption of light, 595
- intestinal fatty acid binding protein
 - kinetics of folding, 697
- intimin receptor
 - interfaces, 480
- intracellular membranes, 743
- intramolecular association constant, 222
- intramolecular hydrogen bonds
 - α helices, 228
 - β structure, 228
 - β turn, 227
 - dicarboxylic acids, 227
- intramolecular interference
 - scattering of electromagnetic radiation, 579
- intramolecular processes, 222–30
 - decrease in standard free energy of association, 226
 - effective molarity, 224
 - entropy of approximation, 224
 - entropy of molecularity, 225
 - entropy of rotational restraint, 224–26
- intramolecular proton transfer, 70
- intrinsic viscosity
 - collagen, 579
 - definition, 578
 - molten globule, 684
 - random coil, 660
- introns
 - evolution of proteins, 350
 - structure of DNA, 98
- invariant position
 - aligning amino acid sequences, 348
- inversion-specific glycoprotein
 - electron microscopy, 586
- iodoacetamide, 546
 - reagent for covalent modification, 530
 - specificity, 532
- iodoacetate, 550
- 5- $[^{125}\text{I}]$ iodonaphthyl azide
 - hydrophobic reagent for covalent modification, 797
- ion
 - chelation, 203
 - enthalpy of hydration, 200–201
 - entropy of hydration, 202
 - layer of hydration, 200
 - self-charging energy, 200
- ion exchange
 - media, 9
- ion exchange chromatography, 8
 - purification, 23
- ion pairs
 - carboxylate-ammonium, 202
 - divalent metal ions, 203
 - enthalpy of hydration, 201
 - ionic interactions in
 - crystallographic molecular models, 302
 - standard enthalpy of formation, 200
- ionic bonds
 - metal ion, 327
- ionic double layer
 - chromatography, 8
 - electrophoresis, 38
- ionic interactions, 199–204
- ionic interactions in crystallographic molecular models, 300–306
 - acid dissociation constants, 300
 - arc repressor, 304
 - buried acid–bases, 301
 - buried ion pair, 303
 - chloramphenicol *O*-acetyl transferase, 302
 - dihydrolipoyllysine-residue acetyltransferase, 300
 - electrostatic work, 301
 - frequency, 306

- ion pair, 302
- ionized hydrogen bond, 302
- α -lytic endopeptidase, 303
- neutron diffraction, 303
- pepsinogen, 303
- relative permittivity, 303
- site-directed mutation, 303
- subtilisin, 300
- tautomeric interactions, 300
- titration curve, 300
- xylose isomerase, 302
- ionic radius
 - metal ion, 327
- ionic strength, 203
 - definition, 39
 - electrophoresis, 39
- ionized hydrogen bond
 - ionic interactions in
 - crystallographic molecular models, 302
- ions
 - crystallography, 181
 - electrostatic repulsion, 203
- ion-trap mass spectrometer
 - mass spectrometry, 91
- iron
 - metalloproteins, 330
- iron-sulfur cluster
 - metalloproteins, 330
- irreversible adsorption
 - chromatography, 3
- isethionyl [^{14}C]acetimidate
 - impermeant reagent for covalent modification, 799
- isoaspartyl peptide bond, 115
- isocitrate dehydrogenase (NAD^+)
 - electrophoresis, 46
 - purification, 26, 29
- isocitrate lyase
 - covalent modification, 544
- isocratic zonal chromatography, 7
- isocyanates
 - reagents for covalent modification, 535
- isoelectric focusing, 47
- isoelectric point
 - definition, 33
- isoelectric precipitation
 - purification, 23
- isoforms
 - definition, 358
 - evolution of proteins, 358
 - malate dehydrogenase, 359
- isogloboside
 - glycosphingolipid, 748
- isoionic point
 - definition, 32
- isoleucine
 - electronic structure, 76
 - stereochemistry of side chains, 268
- isometric oligomeric proteins, 485–99
- isometric point groups, 486
- isometric structures
 - interfaces, 488
- isoprenylation
 - anchored membrane-bound proteins, 765
 - posttranslational modification, 117
- isopycnic centrifugation
 - cell fractionation, 744
- isothiocyanates
 - reagents for covalent modification, 534
- J**
- jumbled amino acid sequences
 - aligning amino acid sequences, 353
- K**
- K^+ -transporting ATPase
 - topography of membrane-spanning proteins, 802–3
- karyopherin $\beta 2$
 - heterologous interfaces, 513
- KcsA potassium channel
 - metalloproteins, 328
- keratin
 - coiled coil of α helices, 282
 - helical cable, 508
- keratin filaments
 - distribution in cell, 507
 - intermediate filaments, 506
- 3-keto fatty acids, 749
- keto-enol tautomers, 70
- α -ketoisocaproate oxygenase
 - purification, 25
- kinesin
 - microtubules, 730
 - molecular taxonomy, 393
- kinetic burst
 - kinetics of folding, 689
- kinetic dead-ends
 - kinetics of folding, 705
- kinetic mechanism
 - proton exchange, 643
- kinetics
 - assembly of microtubules, 724–26
 - covalent modification, 531
 - of assembly of oligomers, 711
- kinetics of folding, 688–710
 - activation volume, 703
 - acyl-CoA-binding protein, 697
 - apomyoglobin, 689, 693, 695, 697
 - approach to equilibrium, 666
- arc* repressor, 703
- aspartate kinase–homoserine dehydrogenase, 709
- barstar, 690, 695
- cell surface receptor CD2, 690
- chymotrypsin inhibitor 2A, 702
- circular dichroism, 689
- cold shock-like protein, 667, 702
- colicin E7 immunity protein, 694
- continuous flow, 694
- continuum of intermediate states, 704
- cystines, 708–9
- cytochrome *c*, 689, 692, 695, 697, 698–99
- cytochrome *c'*, 704
- cytochrome *c*₂, 691
- dead time, 688
- dihydrofolate reductase, 689, 691–92, 694, 698, 704
- dihydrolipoyllysine-residue acetyltransferase, 702
- dilution, 688
- domains, 709
- engrailed homeodomain, 702
- fatty acid-binding protein, 694
- fluorescence, 689
- heat capacity, 703
- human acylphosphatase, 702
- hydrophobic collapse, 696
- immunoglobulin G binding protein G, 694
- intermolecular aggregation, 705
- intestinal fatty acid binding protein, 697
- kinetic burst, 689
- kinetic dead-ends, 705
- β -lactoglobulin, 689, 691, 693–94, 698
- lysozyme, 689–91, 694, 698, 704, 710
- micrococcal nuclease, 689, 698
- molten globule, 703
- molten globules, 692–94
- multiple steps, 698
- myoglobin, 667
- δ -octopine dehydrogenase, 709
- parallel pathways, 704
- pH effect on, 703
- phosphoglycerate kinase, 690
- proline isomerization, 698–702
- protein A, 702
- protein G, 697
- protein L, 696
- protein L9, 703
- protein S6, 702
- proton exchange, 691–92, 698

- rapid mixing chamber, 688
- rate constant for folding, 667
- λ repressor, 702–03
- ribonuclease, 695
- ribonuclease H, 688–90, 693–94, 697
- scattering of X-radiation, 691
- site-directed mutation, 703
- stopped-flow apparatus, 688
- temperature jump, 695
- ubiquitin, 702
- viscosity effect on, 697, 703
- WW domain, 703
- kringle
 - domains, 386, 390
- L**
- α -lactalbumin
 - aligning amino acid sequences, 360
 - covalent modification, 545
 - folding, 669, 685
 - hydration of a protein, 298
 - metalloproteins, 329
 - molten globule, 683–84
 - thermodynamics of folding, 682
- β -lactamase
 - topography of membrane-spanning proteins, 800
 - water in crystallographic molecular models, 293
- D-lactate dehydrogenase
 - molecular taxonomy, 396
- L-lactate dehydrogenase
 - convergent evolution, 373
 - cross-linking, 443–45
 - dodecyl sulfate gel electrophoresis, 422
 - domains, 382–83
 - evolution of proteins, 359
 - frictional ratio, 426
 - isoforms of, 359
 - mean molar mass of an amino acid, 418
 - molar mass, 418
 - molecular taxonomy, 393–95
 - osmotic pressure, 419
 - packing of side chains, 288
 - purification, 29
 - quaternary structure, 407
 - recurring structure, 373–74
 - sieving, 424, 427
 - space groups, 463
- L-lactate dehydrogenase (cytochrome)
 - convergent evolution, 373
 - point group, 469–70
- β -lactoglobulin
 - electrophoresis, 42
- hydration of a protein, 298
- kinetics of folding, 689, 691, 693–94, 698
- molar mass, 418
- molecular charge, 33
- osmotic pressure, 419
- sieving, 428
- thermodynamics of folding, 671
- lactoperoxidase
 - reagent for covalent modification, 538
 - sieving, 424
 - topography of membrane-spanning proteins, 799
- lactose
 - preferential solvation, 31
- lactose permease
 - membrane-spanning α helices, 795
 - topography of membrane-spanning proteins, 800
- lactoside
 - glycosphingolipid, 748
- lacZ* promoter
 - expressing DNA, 108
- ladder
 - sequencing of DNA, 102
- lamin A
 - aligning amino acid sequence, 360
- laminar flow
 - viscosity, 578
- laminin
 - heterologous associations, 516
 - heterooligomers, 513
- laminin $\gamma 1$
 - domains, 390
- large tumor antigen, 562
- large-conductance mechanosensitive channel
 - crystallization, 772
 - membrane-spanning helices, 772
 - overexpression, 776
- Larmor frequency
 - nuclear magnetic resonance, 613
- lattice
 - crystallography, 151
- lattice lines
 - image reconstruction, 791
- layer line
 - crystallography, 150
- layer of hydration
 - ion, 200
- layers of hydration
 - repulsion, 202
- lectin
 - interfaces, 480
 - molecular rotational axes of symmetry, 465
 - space groups, 460
- lectin IV
 - cis* peptide bond, 252
 - metalloproteins, 330
- left-handed twist
 - β structure, 260
- leghemoglobin
 - aligning amino acid sequence, 360
 - aligning crystallographic molecular models, 369
- length of a hydrogen bond
 - hydrogen bonds in crystallographic molecular models, 312
- length of a polypeptide
 - definition, 407
- leucine
 - electronic structure, 76
- leucine-rich repeat, 386
- leucyl aminopeptidase
 - sequencing of polypeptides, 91
- Lewis acids, 326
- Lewis bases, 326
- Lewis structure, 56
- library
 - cloning of DNA, 99
- licheninase
 - molecular taxonomy, 398
- lifetime
 - of fluorescence, 602
- lifetime of the excited state
 - fluorescence resonance energy transfer, 605
- ligands
 - metalloproteins, 327
- light scattering
 - incremental scattering, 415
 - molar mass, 414–16
 - ovalbumin, 420
 - polarized light, 415
 - Rayleigh's ratio, 416
 - refractive index, 415
 - serum albumin, 416
 - virial coefficient, 416
 - Zimm plot, 416
- light-harvesting chlorophyll *a/b*-protein complex
 - image reconstruction, 793
- limiting viscosity number, 579
- linoleic acid, 747
- α -linolenic acid, 747
- lipid A export ATP-binding protein
 - crystallization, 772
 - membrane-spanning helices, 772
- lipopolysaccharide
 - glycolipid, 749
- lipoproteins, 804–5

- apolipoprotein B100, 805
- belt around the waist, 805
- chylomicrons, 804
- high density lipoprotein, 805
- lipovitellin, 804
- low density lipoprotein, 804
- phosvitin, 804
- very low density lipoprotein, 804
- vitellogenin, 804
- lipovitellin
 - lipoproteins, 804
- liquid water
 - water, 190
- local 3-fold rotational axis of
 - pseudosymmetry
 - quasi-equivalence, 491
- local conformational changes
 - ribonuclease H, 678
- local minimum
 - refinement, 176
- local rotational axis
 - double-helical DNA, 467
- local rotational axis of symmetry
 - definition, 491
 - palindromic sequence, 467
- lone pair of electrons
 - electronic structure, 56, 59
- low density lipoprotein
 - lipoproteins, 804
- low-affinity immunoglobulin γ Fc
 - region receptor
 - nonstoichiometric ratio of subunits, 512
- low-barrier hydrogen bond, 208–10
 - free energy of formation, 215
 - hydrogen bonds in
 - crystallographic molecular models, 312
- lowest unoccupied molecular orbital, 58
- lymphocytes
 - production of immunoglobulins, 557
- lysine
 - acid dissociation constant, 75
 - covalent modification, 530–36
 - cross-linking, 440
 - electronic structure, 80
 - hydropathy, 242
 - nucleic acid, association of proteins with, 315
 - water in crystallographic molecular models, 296
- lysosomes, 743
 - cell fractionation, 744
- lysozyme
 - aligning amino acid sequences, 360
- aligning crystallographic molecular models, 363
- crystallographic molecular model, 170
- crystallography, 155
- diffusion coefficient, 578
- domains, 388
- epitopes, 561
- fluorescence, 601, 603
- fluorescence resonance energy transfer, 609
- folding, 678
- free energy of folding, 677, 687
- frictional coefficient, 578
- frictional ratio, 578
- hydration of a protein, 298–99
- hydrogen bonds, 267
- hydrogen bonds in crystallographic molecular models, 311
- hydropathy side chains, 275
- immune complex, 558–61
- ionic interactions, 300
- kinetics of folding, 689–91, 694, 698, 704, 710
- mean molar mass of an amino acid, 418
- molar mass, 418
- molecular taxonomy, 393
- packing of side chains, 289
- preferential solvation, 661
- proton exchange, 645
- sedimentation coefficient, 578
- sieving, 424
- thermodynamics of folding, 682
- unfolding, 661
- virial coefficients, 419
- water in crystallographic molecular model, 299
- water in crystallographic molecular models, 296
- lysyl endopeptidase
 - cleavage of polypeptide, 88
- α -lytic endopeptidase
 - ionic interactions in
 - crystallographic molecular models, 303
 - stereochemistry of side chains, 268
 - water in crystallographic molecular models, 294
- M**
- α_2 -macroglobulin
 - electron microscopy, 588
 - immunoelectron microscopy, 567
- macroscopic acid dissociation
 - constant
 - definition, 71
- MADS-box protein MCM1
 - association of proteins with nucleic acid, 320
- magnesium
 - metalloproteins, 329, 332
- magnetic flux density
 - nuclear magnetic resonance, 613
- magnetogyric ratio
 - nuclear magnetic resonance, 613–14
- major cold-shock protein
 - nuclear magnetic resonance, 630
- major groove
 - nucleic acid structure, 315
- malate dehydrogenase
 - assembly of oligomers, 712
 - axes of symmetry, 452–53
 - evolution of proteins, 359
 - isoforms, 359
 - molecular rotational axes of symmetry, 465
 - sieving, 424, 428
 - space groups, 461
- malate dehydrogenase (oxaloacetate-decarboxylating) (NADP⁺)
 - purification, 29
- malate synthase
 - purification, 30
- maltodextrin binding protein
 - aligning crystallographic molecular models, 363
- maltoporin
 - crystallization, 772
- maltose-binding protein
 - molecular taxonomy, 396
 - multiple isomorphous replacement, 160
- mammalian cells
 - expression of DNA, 109
- (S)-mandelate dehydrogenase
 - anchored membrane-bound proteins, 765
- mandelate racemase
 - aligning amino acid sequences, 361
 - aligning crystallographic molecular models, 366
- manganese
 - metalloproteins, 330
- manganese-stabilizing protein
 - diffusion coefficient, 578
 - frictional coefficient, 578
 - frictional ratio, 578
 - sedimentation coefficient, 578
- mannose-6-phosphate receptor
 - domains, 385
- mannose-binding protein
 - interfaces, 480

- map of electron density
 - calculation of, 157
 - galactose oxidase, 165
- MAP protein kinase ERK2
 - molecular taxonomy, 396
- maps of electron density, 149–62
- marker enzymes
 - cell fractionation, 744
- mass
 - of hydrodynamic particle, 573
- mass spectrometer, 550
- mass spectrometry
 - α -amylase, 93
 - electrospray mass spectrometer, 91
 - fast-atom bombardment, 92
 - fragment ions, 93
 - ion-trap mass spectrometer, 91
 - matrix-assisted-laser-desorption ionization, 92
 - peptide map, 433
 - posttranslational modification, 119
 - proton exchange, 642
 - quadrupole mass spectrometer, 91
 - sequencing of polypeptides, 91–93
 - sequencing oligosaccharides, 136
 - tandem mass spectrometer, 93
 - time-of-flight mass spectrometer, 91
 - thioredoxin, 92
- matrix-assisted-laser-desorption ionization
 - mass spectrometry, 92
- maturation-promoting factor
 - assay, 19
- maturity-onset diabetes, 508
- mean molar mass of an amino acid
 - anion exchanger, 418
 - bacteriorhodopsin, 418
 - chymotrypsinogen, 418
 - fibrinogen, 418
 - hemoglobin, 418
 - immunoglobulin G, 418
 - L-lactate dehydrogenase, 418
 - lysozyme, 418
 - myosin, 418
 - Na⁺/K⁺-exchanging ATPase, 418
 - parvalbumin, 418
 - phosphorylase, 418
 - protein coat from R17 virus, 418
 - serum albumin, 418
- mean net molecular charge number
 - definition, 32
- mean net proton charge number
 - definition, 32
- medium-chain acyl-CoA dehydrogenase
 - assay, 16
- melting point
 - water, 190
- membrane alanyl aminopeptidase
 - purification, 773
- membrane in MalG protein
 - topography of membrane-spanning proteins, 800
- membrane-bound proteins, 763–824
 - anchored, 764–65
 - assay, 771
 - band 3 anion transport protein, 767
 - chromatography, 771
 - enzymes, 766
 - glycophorin, 767
 - identification of genes for, 767
 - image reconstruction, 790–93
 - immunoabsorption, 773
 - integral, 766
 - myristoylation, 765
 - overexpression, 776
 - peripheral, 763–64
 - photosynthetic reaction center, 765
 - prostaglandin-endoperoxide synthase, 766
 - proteolipid protein, 765
 - purification, 768–73
 - reconstitution, 771
 - site-directed mutation, 773
 - snorkeling, 780
 - sterol carrier protein, 766
 - vectorial insertion, 767
- membranes
 - cytoplasmic surface, 743
 - diffusion in, 812
 - extracytoplasmic surface, 743
 - frictional coefficient in, 812
 - microsomes, 743
 - microviscosity of, 814
 - punching holes in, 803–4
 - rotational diffusion coefficient in, 812
 - solvent is bilayer of lipids, 807
 - translational diffusion coefficient in, 812
 - viscosity of, 813
- membrane-spanning α helices
 - acetylcholine receptor, 772, 794
 - anchored membrane-bound proteins, 774
 - aquaporin, 772
 - bacteriorhodopsin, 772
 - computational assignment, 795–96
 - covalent modification, 796
 - cross-linking, 803
 - cytochrome-*c* oxidase, 772, 784, 796
 - cytochrome *o* ubiquinol oxidase, 772
 - cystines, 783
 - electron spin resonance, 793–95
 - halorhodopsin, 772
 - hydrogen bonds, 775
 - integral membrane-bound proteins, 776
 - lactose permease, 795
 - large-conductance mechano-sensitive channel, 772
 - lipid A export ATP-binding protein, 772
 - nuclear magnetic resonance, 793
 - periodicities of exposure, 794
 - photosynthetic reaction center, 772, 784, 795
 - potassium channel DcsA, 772
 - prolines, 783
 - protein MsbA, 772
 - rhodopsin, 772
 - succinate dehydrogenase, 772
 - synthetic hydrophobic peptide, 774
 - tryptophan, 774
 - tyrosine, 774
 - unspecific monooxygenase, 796
- membrane-spanning segment, 766
 - anchored membrane-bound protein, 774
- messenger RNA, 98
- metal ion
 - covalent bonds, 327
 - hardness, 327
 - ionic bonds, 327
 - ionic radius, 327
 - softness, 327
- metallic cations
 - metalloproteins, 326
- L1 metallo- β lactamase
- metallopeptidases, 49
- metalloproteins, 326–32
 - alanine-tRNA ligase, 331
 - aldehyde:ferredoxin oxidoreductase, 330
 - arginase, 326
 - aspartate carbamoyltransferase, 326, 332
 - α -thrombin, 328
 - azurin, 331
 - calcium, 328–29
 - cobalt, 330
 - copper, 331
 - diphtheria toxin repressor, 332–33
 - DNA-(apurinic or apyrimidinic site) lyase, 330
 - endopeptidase K, 328–29
 - glycosylase MutY, 330
 - heme, 330
 - inorganic diphosphatase, 329, 332

- iron, 330
- iron-sulfur cluster, 330
- KcsA potassium channel, 328
- α -lactalbumin, 329
- lectin IV, 330
- ligands, 327
- magnesium, 329, 332
- manganese, 330
- metallic cations, 326
- molybdenum, 330
- myoglobin, 326
- nickel, 330
- nitrate reductase, 330
- nitrile hydratase, 330
- pentacoordinate zinc, 331
- phosphoribosylaminoimidazole-carboxamide formyltransferase, 328
- plastocyanin, 331
- potassium, 328
- sodium, 328
- sulfenic acid, 330
- sulfinic acid, 330
- thermitase, 329
- tungsten, 330
- urease, 330
- UTP-hexose-1-phosphate uridylyltransferase, 330–31
- vanadium, 330
- zinc, 331–32
- zinc finger, 326, 331
- zinc-binding protein TroA, 331–32
- methanol dehydrogenase
 - β propeller, 264
- methionine
 - covalent modification, 530, 532, 536
 - electronic structure, 80
 - stereochemistry of side chains, 270
 - sulfone, 81–82
 - sulfoxide, 81–82
- methionine adenosyltransferase
 - rotational axis of symmetry, 480
- methionyl aminopeptidase
 - aligning crystallographic molecular models, 362–63
 - domains, 383
 - molecular rotational axes of pseudosymmetry, 476
- methyl acetimidate, 547
 - reagent for covalent modification, 532–34
- methylamine-glutamate
 - N*-methyltransferase assay, 14
- methyl-accepting chemotaxis protein
 - coiled coil of α helices, 284
 - interfaces, 480
- methylcrotonyl-CoA carboxylase
 - purification, 25
- 2-methyleneglutarate mutase
 - assay, 16
- methyl group
 - stereochemistry of side chains, 270
- methyl group of thymine
 - nucleic acid, association of proteins with, 318
- methylation
 - posttranslational modification, 115
 - sequencing oligosaccharides, 135
- methylmalonyl-CoA
 - carboxytransferase peptide map, 435
 - subunits, 435–36
- methylmalonyl-CoA decarboxylase
 - aligning amino acid sequences, 360
- methylmalonyl-CoA mutase
 - aligning amino acid sequence, 360
 - binding of ligands, 47
- micelle of dodecyl sulfate
 - dodecyl sulfate gel electrophoresis, 421
- micrococcal nuclease
 - β structure, 261
 - epitopes, 562
 - folding, 678–79
 - hydrogen bonds in crystallographic molecular models, 312
 - immune complex, 559–60
 - kinetics of folding, 689, 698
 - proton exchange, 645
 - purification, 26
- β_2 -microglobulin
 - aligning amino acid sequences, 353, 360
- microheterogeneity
 - oligosaccharides of glycoproteins, 129
- microscopic acid dissociation
 - definition, 62
- microscopic acid dissociation
 - constants, 62
 - histidine, 79
- microsin
 - posttranslational modification, 114
- microsomes
 - membranes, 743
- microtubules, 721–29
 - dynein, 730
 - helical polymer, 506
 - helical surface lattice, 722
 - kinesin, 730
 - polarity, 723
 - seam, 723
 - structure, 722
- microtubule-associated proteins
 - assembly of microtubules, 730
- microviscosity
 - cholesterol effect on, 814
 - of membranes, 814
- mini thick filaments of myosin, 732
- minimal mutational distance
 - aligning amino acid sequences, 356
- minor groove
 - nucleic acid structure, 315
 - nucleic acid, association of proteins with, 318
- mismatch repair protein MutS
 - nucleic acid, association of proteins with, 321
- mismatched symmetry
 - dihydrolipoyl dehydrogenase, 511
 - dihydrolipoyllysine-residue succinyltransferase, 511
 - heterooligomers, 511
 - oxoglutarate dehydrogenase (succinyl-transferring), 511
 - 2-oxoglutarate dehydrogenase complex, 511
- mitochondria, 743
 - cell fractionation, 744
 - topography of membrane-spanning proteins, 798
- mitochondrial H^+ -transporting two-sector ATPase
 - space groups, 461
- mixed disulfide
 - cystines, formation of, 708
- mobile phase
 - definition, 2
- model compound for an amino acid, 74
- modular domains
 - domains, 385
 - heterologous associations, 513
 - in heterooligomers, 513
- molar ellipticity
 - circular dichroism, 598
- molar mass, 408–21
 - apoferritin, 418
 - aspartate carbamoyltransferase, 418–419
 - aspartate kinase I-homoserine dehydrogenase I, 418, 420
 - catalase, 418
 - chymotrypsinogen, 418
 - definition, 408
 - 2-dehydro-3-deoxyphosphogluconate aldolase, 418
 - dry weight, measurement of, 419
 - electrospray mass spectrometry, 416–17

- fructose-bisphosphate aldolase, 418–19
- glutamate dehydrogenase, 418
- glutamate–ammonia ligase, 420
- glyceraldehyde-3-phosphate dehydrogenase, 418
- β -lactoglobulin, 418
- L-lactate dehydrogenase, 418
- light scattering, 414–16
- lysozyme, 418
- osmotic pressure, 408–11
- pepsin, 418
- phosphorylase, 418
- ribonuclease, 418–19
- sedimentation equilibrium, 411–14
- serum albumin, 412, 418
- molar volume, 197
- water, 192
- molecular asymmetric unit
 - definition, 472
 - molecular rotational axes of symmetry, 472
- molecular axes of symmetry
 - definition, 461
 - erythrocrucorin, 481
 - space groups, 461
- molecular charge, 32–36
 - β -lactoglobulin, 33
 - bound ions, 33
 - deoxyhemoglobin, 34
 - exotoxin A, 34
 - fructose-bisphosphate aldolase, 36
 - loosely bound ions, 34
 - plasminogen activator inhibitor 1, 34
 - ribonuclease, 35
 - tryptophanase, 34
- molecular dynamics
 - refinement, 177
- molecular exclusion chromatography
 - chromatography, 11
 - distribution coefficient, 12
 - hydrophilic media, 11
 - included volume, 12
 - purification, 24
 - sieving, 423
- molecular mass
 - definition, 408
- molecular model, 162–72
 - crystallography, 163
 - nuclear magnetic resonance, 631
- molecular orbital
 - antibonding, 57
 - bonding, 57
 - energy level, 56
 - node, 56
 - nonbonding, 57
 - phase, 56
- molecular replacement
 - crystallography, 182
- molecular rotational axes of
 - pseudosymmetry
 - arabinose binding protein, 476
 - 5-carboxymethyl-2-hydroxy-muconate Δ -isomerase, 477
 - chymotrypsinogen, 476
 - internal duplications, 476
 - methionyl aminopeptidase, 476
 - 4-oxalocrotonate tautomerase, 477
 - phaseolin, 477
 - pyruvate oxidase, 477
 - sulfite reductase, 476
 - thiosulfate sulfurtransferase, 476–77
- molecular rotational axes of
 - symmetry
 - crystal packing, 465
 - formate dehydrogenase, 465
 - glyceraldehyde-3-phosphate dehydrogenase (phosphorylating), 464
 - heat-labile enterotoxin, 465
 - L-2-hydroxyisocaproate dehydrogenase, 465
 - inorganic diphosphatase, 465
 - lectin, 465
 - malate dehydrogenase, 465
 - molecular asymmetric unit, 472
 - ribulose-bisphosphate carboxylase, 465
 - self-rotation function, 465
 - superposed α carbons, 465
 - triose-phosphate isomerase, 465
- molecular surface
 - definition, 277
- molecular taxonomy, 392–99
 - adenosine kinase, 396
 - adenosylmethionine–8-amino-7-oxononate transaminase, 396
 - adenylate kinase, 395
 - D-alanine–D-alanine ligase, 397
 - alcohol dehydrogenase, 395
 - alkylhalidase, 396
 - arabinose-binding protein, 395
 - architecture, 396
 - asparagine synthase, 393
 - aspartate transaminase, 396
 - aspartate-semialdehyde dehydrogenase, 396
 - aspartate–tRNA ligase, 393
 - benzoylformate decarboxylase, 396
 - biotin carboxylase, 397
 - carbonate dehydratase, 393, 395
 - carboxymethylenebutenolidase, 396
 - carboxypeptidase, 395
 - carboxypeptidase D, 396
 - cathepsin K, 393
 - cohesin domain, 397
 - coincident structure, 396
 - common fold, 393
 - cyclic-AMP dependent protein kinase, 396
 - cyclin-dependent protein kinase 2, 396
 - cystathionine β -lyase, 396
 - dihydrofolate reductase, 395
 - erythronate-4-phosphate dehydrogenase, 396
 - Factor D, 396
 - family of domains, 396
 - ferredoxin–NADP⁺ reductase, 393
 - flavodoxin, 395
 - glucan 1,4- α -glucosidase, 396
 - glutathione synthase, 397
 - glutathione-disulfide reductase, 394–95
 - glyceraldehyde 3-phosphate dehydrogenase, 395–96
 - glycerate dehydrogenase, 396
 - granulocyte-colony-stimulating factor, 399
 - growth hormone, 399
 - hemoglobin, 394, 398
 - hexokinase, 395
 - hierarchical classification, 393
 - L-2-hydroxyisocaproate dehydrogenase, 396
 - immunoglobulin, 397
 - interleukin, 399
 - kinesin, 393
 - D-lactate dehydrogenase, 396
 - L-lactate dehydrogenase, 393–95
 - licheninase, 398
 - lysozyme, 393
 - maltose-binding protein, 396
 - MAP protein kinase ERK2, 396
 - myohemerythrin, 394, 398
 - myosin, 393
 - newer proteins, 392
 - ornithine decarboxylase, 396
 - papain, 393–94, 398
 - phosphofructokinase, 395
 - phosphoglycerate dehydrogenase, 396
 - phosphoglycerate kinase, 395
 - phosphoglycerate mutase, 395
 - phosphopyruvate hydratase, 397
 - phosphoribosylamine–glycine ligase, 397
 - phosphorylase, 395

- phosphoserine transaminase, 396
 primordial proteins, 392
 protein coat of satellite tobacco
 necrosis virus, 393
 protein coat of tomato bushy stunt
 virus, 397
 pyruvate decarboxylase, 396
 pyruvate kinase, 394–95, 397
 ribokinase, 396
 ribonucleoside-diphosphate
 reductase, 397
 speciation of proteins, 393
 species of domains, 393
 subtilisin, 395
 succinate-CoA ligase, 395
 superfamily, 396
 synapsin Ia, 397
 thiamine pyridinylase, 396
 thioredoxin, 395
 thioredoxin-disulfide reductase, 393
 thiosulfate sulfurtransferase, 395
 triosephosphate isomerase, 394
 trypsin, 396
 tumor necrosis factor, 393
 tyrosine phenol-lyase, 396
 molecularity, 222–30
 entropy of, 225
 molluside
 glycosphingolipid, 748
 molten globules, 683–84
 α -lactalbumin, 683–84
 apomyoglobin, 684
 assembly of oligomers, 712
 carbonate dehydratase, 683
 circular dichroism, 684
 configurational entropy, 683
 cytochrome *c*, 683–84
 definition, 683
 diffusion coefficient, 684
 fluorescence, 684
 interleukin 6, 684
 intermediate states, 683
 intrinsic viscosity, 684
 kinetics of folding, 692–94, 703
 neutron scattering, 684
 nuclear magnetic resonance
 spectrum, 684
 proton exchange, 684
 rotational relaxation time, 684
 stable intermediates, 683
 ultrasound, 684
 molybdenum
 metalloproteins, 330
 monoclinic lattice
 crystallography, 151
 monoclonal immunoglobulin, 558
 monodisperse solution
 definition, 408
 monolayer of lipids
 air and water interface, 761
 surface area at zero pressure, 761
 surface pressure, 761
 monophenol monooxygenase
 assay, 18
 posttranslational modification, 122
 mosaic eukaryotic protein
 domains, 385
 MotA protein
 topography of membrane-
 spanning proteins, 799
 moving boundary electrophoresis, 41
 MQF
 nuclear magnetic resonance, 621
 mucin MUC2
 length, 85
 mucins, 132
 mucoside
 glycosphingolipid, 748
 multicatalytic endopeptidase
 complex
 heterooligomers, 508
 immunoelectron microscopy, 567
 multienzyme complex
 domains, 379–82
 multifunctional endopeptidase
 quaternary structure, 475
 multimeric protein
 definition, 407
 multiple isomorphous replacement,
 155
 alcohol dehydrogenase, 161
 anomalous dispersion, 161
 apoferritin, 161
 chloramphenicol *O*-acetyltrans-
 ferase, 160
 crystallography, 158–61
 deoxyribonuclease, 161
 heavy atom, 158
 maltose binding protein, 160
 trimethylamine-*N*-oxide reductase
 (cytochrome *c*), 160
 vector equations, 159
 murine L cells
 expression of DNA, 110
 mutagenic oligonucleotide
 site-directed mutation, 110
 mutation probability
 definition, 349
 evolution of proteins, 349–50
 myeloma protein
 immunoglobulins, 557
 myoglobin
 aligning amino acid sequences,
 345, 360
 aligning crystallographic molecular
 models, 369–72
 crystallographic molecular model,
 170
 dodecyl sulfate gel electrophoresis,
 422
 electron paramagnetic resonance,
 648–49
 electrophoresis, 42
 free energy of folding, 677
 frictional ratio, 426
 hydration of a protein, 298
 hydrogen bonds in
 crystallographic molecular
 models, 306, 309, 312
 kinetics of folding, 667
 metalloproteins, 326
 nuclear magnetic resonance, 635
 proton exchange, 642
 radius of gyration, 581
 sieving, 424
 thermodynamics of folding, 673
 X-ray scattering, 583
 myohemerythrin
 molecular taxonomy, 394, 398
myo-inositol, 747
 myosin, 730–32
 coiled coil of α helices, 282, 730
 covalent modification, 544
 cross-linking, 549
 electron microscopy, 586
 fluorescence resonance energy
 transfer, 609
 frictional coefficient, 589
 globular heads, 730
 light scattering, 590
 mean molar mass of an amino
 acid, 418
 molecular taxonomy, 393
 structure, 730
 [myosin-light-chain] kinase
 heterologous associations, 517
 proton exchange, 645
 myosin subfragment 1
 assay, 17
 myristoylation
 membrane-bound proteins, 765
N
 Na⁺/K⁺-exchanging ATPase
 aligning amino acid sequences, 364
 circular dichroism, 600
 covalent modification, 546
 covalent modification from within
 the bilayer, 797
 cross-linking, 444–45
 expression, 776

- fluorescence resonance energy transfer, 609
 image reconstruction, 793
 immunochemistry, 568
 mean molar mass of an amino acid, 418
 peptide map, 433
 purification, 768
 topography of membrane-spanning proteins, 799, 802
 NADH dehydrogenase (ubiquinone) immunostaining, 565–66
 integral membrane-bound protein, 766
 NADH peroxidase domains, 388
 N→O acyl migration posttranslational modification, 114
 native state definition, 659
 native structure crystallography, 171
 natural selection quaternary structure, 455
 nebulin assembly of actin, 730
 domains, 384
 negative staining electron microscopy, 585
 image reconstruction, 501
 neolactoside glycosphingolipid, 748
 net magnetization nuclear magnetic resonance, 613
 D-neuraminic acid oligosaccharides of glycoproteins, 128
 neuraminidase epitope, 562
 neurofilaments intermediate filaments, 506
 neutral replacement evolution of proteins, 348
 neutron diffraction bilayer of phospholipid, 750, 754
 ionic interactions in crystallographic molecular models, 303
 proton exchange, 645
 neutron scattering, 583–84 hydrophobic effect, 233
 molten globule, 684
 ribosome, 583–84
 neutron scattering density bilayer of phospholipid, 755
 newer proteins molecular taxonomy, 392
 nickel metalloproteins, 330
 nicotinate-nucleotide diphosphorylase crystallization, 49
 nidogen electron microscopy, 586
 heterooligomers, 513
 nitrate reductase metalloproteins, 330
 water in crystallographic molecular models, 293
 nitrene reagent for covalent modification, 542
 singlet, 542
 triplet, 542
 nitric-oxide synthase convergent evolution, 373
 electron nuclear double resonance, 649–50
 electron paramagnetic resonance, 649–50
 rotational axis of symmetry, 480
 nitrile hydratase metalloproteins, 330
 nitrite reductase crystallography, 181
 X-ray scattering, 583
 2-nitro-5-thiocyanatobenzoate cleavage of polypeptides, 87, 89
 nitrogenase electron paramagnetic resonance, 646
 peptide map, 435
 2-[(2-nitrophenyl) sulfenyl]-3-methyl-3'-bromoindolenine reagent for covalent modification, 539
 2-(*p*-nitrophenyl)-3-(3-carboxy-4-nitrophenyl)thio-1-propene cross-linking reagent, 548
p-nitrophenylethanedione reagent for covalent modification, 539
 nitrotyrosine ultraviolet absorption spectra, 601
 nitroxyl fatty acid bilayer of phospholipid, 755–56
 nitroxylphosphatidylcholine boundary layer of phospholipid, 785
N-methylacetamide hydrogen bond, 217
 node molecular orbital, 56
 nonbonding molecular orbital, 57
 nonionic detergents for purification of membrane-bound proteins, 768–71
 nonstoichiometric ratio of subunits low-affinity immunoglobulin γ Fc region receptor, 512
 pyruvate dehydrogenase complex, 512
 nonstoichiometric ratios of subunits heterooligomers, 511
n-tetradecanoyl amide posttranslational modification, 117
 nuclear import factor karyopherin α heterologous associations, 514
 nuclear localization signals heterologous associations, 514
 nuclear magnetic resonance, 613–40
 α helix, 631
 α -amylase inhibitor HOE-467A, 631
 acid dissociation constants, 635–38
 acid-base titration curve, 635
 acrosin inhibitor IIA, 626, 629, 631
 ADA regulatory protein, 632
 amplitude modulation, 619
 assignments, 621–28
 β structure, 631
 calmodulin, 624
 carbonate dehydratase, 635
 carrier frequency, 615
 chemical shift, 614
 connected 1 hydrogens, 630
 connections among nuclei, 622
 continuous wave nuclear magnetic spectrometers, 614
 correlated spectrum, 619
 coupling constant, 615
 CRINEPT, 621
 cytochrome *c*, 617
 cytochrome c_2 , 626
 cytochrome c_{551} , 638
 dihedral angle, 616
 dihydrofolate reductase, 621–22, 625
 DQF, 621
 dynamics, 623
 endo-1,4- β -xylanase, 636
 epidermal growth factor, 636
 factor IX, 629
 Fourier transform nuclear magnetic resonance spectrometer, 614–15
 free induction decay, 615
 frequency labeling, 617
 frequency of maximum absorption, 614

- glutaredoxin 2, 630
 histidine, 635
 HMBC, 621
 HMQC, 621
 HOHAHA, 621
 HSMQC, 621
 HSQC, 621
 hydrogen bond, 209
 immunity protein Im9, 639
 immunoglobulin G binding protein G, 633
 interleukin 1 β , 628
 interleukin 4, 629, 639
 interleukin 13, 627
 Larmor frequency, 613
 magnetic flux density, 613
 magnetogyric ratio, 613, 614
 major cold-shock protein, 630
 membrane-spanning α helices, 793
 molecular model, 631
 molten globule, 684
 MQF, 621
 myoglobin, 635
 net magnetization, 613
 nuclear Overhauser effect, 616
 nuclear Overhauser enhanced spectrum, 625–31
 nuclear spin, 613
 off-diagonal cross-peaks, 619
 order parameter, 623
 pancreatic trypsin inhibitor, 620
 phosphoglycerate mutase, 630
 proton exchange, 642
 PS, 621
 pyruvate dehydrogenase complex, 638
 random meander, 632
 rate of relaxation, 614
 refinement, 632
 regulatory protein GAL4, 633
 relaxation, 614
 resonance, 613
 ribonuclease, 635, 687
 ribonuclease H, 636–37
 ribosomal protein S17, 629
 ring current, 617
 saturation, 614
 SBC, 621
 sequencing oligosaccharides, 136
 spin diffusion, 617
 spin quantum number, 613–614
 spin states, 613
 spin–spin coupling, 615
 subtilisin, 635
 three-dimensional spectroscopy, 617–38
 threonine, 636
 time of mixing, 626
 TOCSY, 621
 transfer of saturation, 616
 transforming growth factor β 1, 627
 TROSY, 621
 tryptophan, 636
 two-dimensional spectroscopy, 617–38
 water, 632
 nuclear magnetic resonance
 molecular model, 631
 nuclear Overhauser effect
 nuclear magnetic resonance, 616
 nuclear Overhauser enhanced spectrum
 nuclear magnetic resonance, 625–31
 nuclear spin
 nuclear magnetic resonance, 613
 P1 nuclease
 aligning crystallographic molecular models, 366
 nucleation
 assembly of actin, 730
 assembly of microtubules, 723
 nucleic acid structure
 base stacking, 323
 bulges, 323
 donors and acceptors of hydrogen bonds, 315
 double-helical hairpin of RNA, 322
 major groove, 315
 minor groove, 315
 pairs of bases, 314
 phosphoryl oxygens, 314
 tertiary structure, 323
 tetraloop, 322
 transfer RNA, 322–23
 nucleic acid, association of proteins
 with, 314–25
 arginine, 315
 *Bam*HI site-specific deoxyribonuclease, 321
 catabolite gene activator protein, 320
 conformational changes, 321
 deoxyribonuclease, 316
 DNA-(apurinic or apyrimidinic site) lyase, 320
 DNA polymerase β , 315
 double helix of DNA, 314
 ETS-domain protein Elk-1, 316
 histones, 316, 320
 homeodomain protein MAT α 2, 320
 lysine, 315
 MADS-box protein MCM1, 320
 methyl group of thymine, 318
 minor groove, 318
 mismatch repair protein MutS, 321
 packing, 319
 phosphodiesterases, 315
 positively charged amino acids, 315
 protein gp 45, 321
 π systems of side chains, 322
 purine repressor, 320
 regulatory protein Cro, 315–16
 replication protein A, 322
 replication terminator, 316
 arc repressor, 316
 λ repressor, 316
 met repressor, 316
 trp repressor, 319, 321
 repressor protein CI, 321
 ribonucleoproteins, 323
 ribosome, 324
 ring of protein, 321
 shape of the surface of the DNA, 319
 single-stranded DNA, 321–22
 single-stranded DNA binding protein, 322
 site-specific deoxyribonuclease, 318
 TATA-box-binding protein, 320
 telomere end-binding protein, 322
 topoisomerase I, 316, 321
 transcription factor IIIA, 324
 transcription factor AP-1, 316–17
 transcription factor AREA, 319
 transcription factor Rob, 317
 U1 small nuclear ribonucleoprotein, 324
 water, 317
 zinc finger, 324–25
 zinc finger protein GLI1, 324
 nucleolin
 heterologous associations, 516
 nucleoporin
 heterologous associations, 517
 nucleoside 5'-monophosphates
 sequencing DNA, 95
 nucleoside bases
 acids and bases, 65
 electronic structure, 65
 nucleoside-diphosphate kinase
 point group, 474
 nucleotide
 definition, 95
 nucleus, 743
 number concentration
 assembly of microtubules, 724
 of polymer, 724
 number of amino acids in a protein
 sieving, 424
 number of amino acids, estimation of
 electrophoresis, 427

O

- observed amplitudes
 - crystallography, 173
- observed phases
 - crystallography, 173
- octahedral point group 432, 487–88
- octahedral symmetry
 - heat shock protein 16.5, 487–88
- D-octopine dehydrogenase
 - kinetics of folding, 709
- octyl β -D-glucoside
 - detergent, 770
- off-diagonal cross-peaks
 - nuclear magnetic resonance, 619
- oleic acid, 745
- oligomer
 - thermodynamics of folding, 670
- oligomeric integral membrane-bound proteins
 - interfaces, 790
- oligomeric interfaces
 - glycophorin, 790
- oligomeric protein
 - definition, 407, 455
- oligomeric proteins, 466–85
 - hydration, 577
 - integral membrane-bound proteins, 786
 - interfaces, 478
 - point groups, 466
- oligosaccharides of glycoproteins, 126–38
 - α 1-acid glycoprotein, 131
 - branching, 128
 - colonic mucin, 128, 130
 - complex *N*-linked
 - oligosaccharides, 131
 - crystallography, 180
 - β -5-deamino-5(S)-hydroxy-neuraminic acid, 128
 - definition, 127
 - glycoforms, 130
 - glycopeptides, 133
 - glycosidic linkage, 128
 - N*-glycosidic linkage, 128
 - O*-glycosidic linkage, 128
 - high-mannose oligosaccharides, 130
 - immunoglobulin D, 127
 - microheterogeneity, 129
 - β -neuraminic acid, 128
 - O*-linked oligosaccharides, 131
 - phytohemagglutinin, 137
 - proteoglycans, 132–33
 - sequence of monosaccharides, 129, 133
 - sialic acids, 128
 - thyroglobulin, 138
- O*-linked oligosaccharides
 - oligosaccharides on glycoproteins, 131
- omit maps of difference electron density
 - hydrogen bonds in crystallographic molecular models, 309
 - refinement, 178
- open reading frame
 - sequence of DNA, 106
- open structure
 - definition, 455
- operon
 - domains, 381
- opsin
 - expression, 776
- optical constant
 - scattering of electromagnetic radiation, 579
- optical rotation
 - circular dichroism, 598
- optical rotatory dispersion, 598
- orbitals, 55
- order parameter
 - bilayer of phospholipid, 756–57
 - nuclear magnetic resonance, 623
- organic radical
 - electron paramagnetic resonance, 645
- orientation factor
 - fluorescence resonance energy transfer, 606
 - fluorescence resonance energy transfer, 607
- orientational freedom
 - fluorescence resonance energy transfer, 607
- oriented bilayers
 - phospholipid, 750
- oriented helical polymeric proteins
 - X-ray diffraction, 502
- oriented, sealed vesicles
 - topography of membrane-spanning proteins, 798
- ornithine carbamoyltransferase
 - tetrahedral symmetry, 487
- ornithine decarboxylase
 - molecular taxonomy, 396
 - purification, 29
- orthogonal β sheets
 - packing of β structure, 285
- orthogonality, 61
- orthologues
 - definition, 358
 - evolution of proteins, 358
- orthorhombic lattice
 - crystallography, 151
- osmotic pressure
 - bovine serum albumin, 419
 - chemical potential, 408
 - concentration of the protein, 409
 - Donnan effect, 410
 - Donnan potential, 411
 - electrolyte, 411
 - hemoglobin, 420
 - ideal gas law, 409
 - impermeant solute, 408
 - L-lactate dehydrogenase, 419
 - β -lactoglobulin, 419
 - molar mass, 408–11
 - semipermeable membrane, 408
 - serum albumin, 410
 - virial coefficients, 409
- outer membrane, 743
 - bacteria, 749
- outer membrane protein A
 - crystallization, 772
- outer membrane protein F
 - crystallization, 772, 775
 - overexpression, 776
- outer membrane protein TolC
 - crystallization, 772, 775
- ovalbumin
 - aligning amino acid sequences, 360
 - dodecyl sulfate gel electrophoresis, 422
 - electrophoresis, 38, 41–42
 - frictional ratio, 426
 - hydration of a protein, 298
 - light scattering, 420
 - sieving, 424, 427–28
 - X-ray scattering, 583
- overexpression
 - large-conductance mechano-sensitive channel, 776
 - membrane-bound proteins, 776
 - outer membrane protein F, 776
- overlap integral
 - fluorescence resonance energy transfer, 606
- ovomucoid
 - electrophoresis, 42
- ovomucoid inhibitor
 - stereochemistry of side chains, 268
- ovotransferrin
 - domains, 384
 - sieving, 427
- 4-oxalocrotonate tautomerase
 - molecular rotational axes of pseudosymmetry, 477
- oxazolines
 - posttranslational modification, 114
- oxidation levels
 - cysteine, 80–82

- oxidative cleavage
 - covalent modification, 544
- 4-(oxoacetyl)phenoxyacetic acid
 - reagent for covalent modification, 539
- 3-oxoacid CoA-transferase
 - folding, 679
- 3-oxoacyl-[acyl-carrier-protein]
 - reductase
 - domains, 381
- 3-oxoacyl-[acyl-carrier-protein]
 - synthase
 - domains, 381
- oxoglutarate dehydrogenase
 - (succinyl-transferring)
 - mismatched symmetry, 511
- 2-oxoglutarate dehydrogenase
 - complex
 - mismatched symmetry, 511
- oxygen
 - quenching fluorescence, 603
- 1-oxyl-2,2,5,5-tetramethylpyrrolin-3-yl group
 - electron paramagnetic resonance, 645
- P**
- p120 GTPase activator
 - domains, 386
- packing
 - in space groups, 457
 - nucleic acid, association of
 - proteins with, 319
- packing of α helices
 - carboxypeptidase A, 282
 - δ -endotoxin CryIIIA, 285
 - integral membrane-bound
 - proteins, 781
 - photosynthetic reaction center, 782
 - ribonucleoside-diphosphate
 - reductase, 285
- packing of β sheets
 - immunoglobulin, 285
 - penicillopepsin, 287
- packing of β structure
 - β barrel, 285
 - β sheets, 285
 - orthogonal β sheets, 285
 - ribonucleoside-diphosphate
 - reductase, 286
- packing of side chains, 277–90
 - α helices, 279–85
 - carboxypeptidase A, 279
 - cavities, 289
 - coiled coil of α helices, 279–85
 - compressibility, 278
 - concanavalin A, 281
 - elasticity, 289
 - helical nets, 280
 - histocompatibility antigen, 279
 - immunoglobulin, 281
 - interdigitation of side chains, 278
 - L-lactate dehydrogenase, 288
 - lysozyme, 289
 - minimization of molecular
 - volume, 278
 - plastocyanin, 289
 - superoxide dismutase, 281
 - volume of a molecule of protein, 278
- pairs of bases
 - nucleic acid structure, 314
- palindromic sequence
 - local rotational axis of symmetry, 467
- palmitic acid, 745
- palmitoleic acid, 745
- pancreatic trypsin inhibitor
 - fluorescence resonance energy
 - transfer, 608
 - nuclear magnetic resonance, 620
- pantetheine-phosphate
 - adenylyltransferase
 - domains, 391
- (S)-pantolactone dehydrogenase
 - assay, 18
- papain
 - cleavage of polypeptide, 88
 - covalent modification, 546, 551
 - molecular taxonomy, 393–94, 398
- paper chromatography, 4
- parallel pathways
 - kinetics of folding, 704
- paralogues
 - definition, 358
 - evolution of proteins, 358
- paramagnetic ion
 - electron paramagnetic resonance, 645
- partial molar volume
 - calculation of, 197
 - definition, 197
- partial specific volume
 - sedimentation equilibrium, 412
- partition coefficient
 - chromatography, 4
 - concentration, units of, 198
 - definition, 198
- partition coefficients between gas
 - and water
 - hydrophobic effect, 235–37
- parvalbumin
 - aligning amino acid sequences, 360
 - mean molar mass of an amino
 - acid, 418
- PDZ domains
 - heterologous associations, 514
- penicillin amidase
 - folding, 679
- penicillopepsin
 - crystallographic molecular model, 167–70
 - hydrogen bonds in
 - crystallographic molecular
 - models, 309
 - packing of β sheets, 287
 - stereochemistry of side chains, 268
 - water in crystallographic molecular
 - models, 293, 295
- pentacoordinate zinc
 - metalloproteins, 331
- pepsin
 - electrophoresis, 42
 - hydration of a protein, 298
 - molar mass, 418
 - sieving, 424
- pepsinogen
 - ionic interactions in
 - crystallographic molecular
 - models, 303
- peptidases
 - produce heterogeneity, 48
- peptide bonds, 74
- hydropathy, 242
- planarity, 251
- secondary structure, 251
- peptide maps, 432–35
 - actin, 432
 - ankyrin, 434
 - chromatography, 433
 - collagen type XIV, 434
 - definition, 432
 - 2-dehydro-3-deoxy-phospho-
 - gluconate aldolase, 437–38
 - electron transfer flavoprotein, 435–36
 - glucose-6-phosphate isomerase, 435, 437–38
 - glutamate-tRNA ligase, 435
 - hemoglobin, 432–33
 - mass spectrometry, 433
 - methylmalonyl-CoA
 - carboxytransferase, 435
 - Na⁺/K⁺-exchanging ATPase, 433
 - nitrogenase, 435
 - phosphoglycerate dehydrogenase, 434
 - tryptic digests, 432
 - two or more polypeptides, 434
 - tyrosines, 433
- peptide separation
 - chromatography, 90

- cytochrome *c* peroxidase, 91
- phosphoglycerate kinase, 90
- peptides
 - circular dichroism, 601
- peptidylamidoglycolate lyase
 - domains, 377
- peptidyl-Asp metalloendopeptidase
 - cleavage of polypeptide, 88
- peptidylglycine monooxygenase
 - domains, 377
- peptidylprolyl isomerases
 - proline isomerization, 701
- percentage of identity
 - aligning amino acid sequences, 351
- periodic acid
 - sequencing oligosaccharides, 134, 136
- peripheral membrane-bound
 - proteins, 763–64
 - actin, 764
 - ankyrin, 821
 - annexin, 764
 - choline-phosphate cytidyltransferase, 764
 - phospholipase C, 764
 - protein kinase C, 764
 - protein Z, 764
 - prothrombin, 764
 - spectrin, 764
- perlecan
 - domains, 386
- peroxidase
 - immunostaining, 566
- peroxiredoxin
 - point group, 472, 475
- peroxisomes, 743
 - cell fractionation, 744
- pH
 - effect on assay, 13
 - effect on covalent modification, 531
 - effect on folding, 662
 - effect on free energy of folding, 675
 - effect on kinetics of folding, 703
 - effect on proton exchange, 644
- phage display
 - detection of heterologous associations, 515, 518
- phase
 - molecular orbital, 56
- phase of the reflection
 - crystallography, 154
- phase transitions
 - biological membranes, 808
- phaseolin
 - molecular rotational axes of pseudosymmetry, 477
 - tetrahedral symmetry, 487
- phases
 - image reconstruction, 790
- phenylalanine
 - circular dichroism, 598
 - electronic structure, 76
 - ultraviolet absorption spectra, 601
- phosphate
 - dp* π molecular orbitals, 83
 - electronic structure, 83
- phosphatidic acid
 - phospholipid, 747
- phosphatidylcholine
 - asymmetry of, 810
 - phospholipid, 745
- phosphatidylethanolamine
 - active transport, 809
 - asymmetry of, 810
 - phospholipid, 745
- phosphatidylglycerol
 - asymmetry of, 810
 - phospholipid, 749
- phosphatidylinositol
 - asymmetry of, 810
 - phospholipid, 747
- phosphatidylserine
 - active transport, 809
 - asymmetry of, 810
 - cytoskeleton, 821
 - phospholipid, 747
- phosphocarrier protein HPr
 - free energy of folding, 677
 - hydrogen bonds in crystallographic molecular models, 312
- phosphodiester backbone
 - association of proteins with nucleic acid, 315
 - nucleic acid structure, 314
- phosphoenolpyruvate carboxykinase (GTP)
 - covalent modification, 543
 - crystallization, 49
- 6-phosphofructo-2-kinase
 - aligning crystallographic molecular models, 364
- 6-phosphofructo-2-kinase/fructose-2,6-bisphosphate 2-phosphatase
 - domains, 389
- phosphofructokinase
 - assay, 19
 - molecular taxonomy, 395
- phosphoglycerate dehydrogenase
 - molecular taxonomy, 396
 - peptide map, 434
 - purification, 49
- phosphoglycerate kinase
 - domains, 383
- fluorescence, 603
- kinetics of folding, 690
- molecular taxonomy, 395
- peptide separation, 90
- purification, 24
- recurring structure, 373
- phosphoglycerate mutase
 - assembly of oligomers, 710–11, 713
 - molecular taxonomy, 395
 - nuclear magnetic resonance, 630
 - purification, 24
- phosphoinositide phospholipase C δ 1
 - domains, 386–87
- phospholipase
 - aligning crystallographic molecular models, 372
- phospholipase C
 - aligning crystallographic molecular models, 366
 - peripheral membrane-bound proteins, 764
- phospholipase C γ
 - domains, 386
- phospholipid scramblase
 - assymetry of phospholipids, 809
- phospholipids
 - asymmetric distribution of, 808–10
 - archaeobacterial isopranylether lipid, 748
 - bilayer of lipids, 745
 - diphosphatidylglycerol, 749
 - ether linkage, 747
 - flip flop, 809
 - sn*-glycerol 3-phosphate, 745
 - head group, 747
 - oriented bilayers, 750
 - phosphatidic acid, 747
 - phosphatidylcholine, 745
 - phosphatidylethanolamine, 745
 - phosphatidylglycerol, 749
 - phosphatidylinositol, 747
 - phosphatidylserine, 747
 - plasmalogen, 747
 - saturated fatty acids, 745
 - sphingomyelin, 748
 - translational diffusion coefficient, 814
 - unsaturated fatty acids, 745
- phospholipid-translocating ATPase
 - asymmetry of phospholipid, 809
- phosphomevalonate kinase
 - assay, 17
 - electrophoresis, 46
- phosphopyruvate hydratase
 - aligning crystallographic molecular models, 366
 - interface, 478
 - molecular taxonomy, 397

- phosphorescence
 emission of light, 595
 1-(5-phosphoribosyl)-5-[(5-phospho-
 ribosylamino)methylideneamino]
 imidazole-4-carboxamide
 isomerase
 domains, 383
 phosphoribosylamine–glycine ligase
 aligning crystallographic molecular
 models, 362
 domains, 388, 390
 molecular taxonomy, 397
 phosphoribosylaminoimidazole-
 carboxamide formyltransferase
 metalloproteins, 328
 phosphoribosylanthranilate
 isomerase
 domains, 377, 384
 folding, 668, 679
 phosphoribosylformylglycinamide
 cyclo-ligase
 domains, 390
 phosphoribosylformylglycinamide
 synthase
 domains, 380
 phosphoribosylglycinamide
 formyltransferase
 domains, 390
 phosphoribulokinase
 point group, 472, 474
 phosphoryl oxygens
 nucleic acid structure, 314
 phosphorylase
 dodecyl sulfate gel electrophoresis,
 422
 mean molar mass of an amino
 acid, 418
 molar mass, 418
 molecular taxonomy, 395
 recurring structure, 373
 phosphorylase *b*
 space groups, 463
 phosphorylase kinase
 electron microscopy, 586–87
 phosphoserine, 82
 phosphoserine transaminase
 molecular taxonomy, 396
 3-phosphoshikimate 1-carboxy-
 vinyltransferase
 domains, 380, 382
 phosphothreonine, 82
 phosphotyrosine, 82
 phosvitin
 lipoproteins, 804
 photo-lyase
 fluorescence resonance energy
 transfer, 607
 photolytic reactions
 covalent modification, 541–44
 photosynthetic reaction center
 bound phospholipid, 784
 crystallization, 772
 crystallographic molecular model,
 780
 crystallography, 180–81
 electron nuclear double resonance,
 650
 hydrophobic sheath, 780
 membrane-bound proteins, 765
 membrane-spanning α helices,
 772, 784, 795
 packing of the α helices, 782
 rotational axis of pseudosymmetry,
 777, 787
 phthalate-dioxygenase reductase
 water in crystallographic molecular
 models, 293
 B-phycoerythrin
 crystallography, 181
 phylogenetic tree
 aligning amino acid sequences,
 354–55
 phytohemagglutinin
 oligosaccharides of glycoprotein,
 137
 π character of a lone pair, 63
 π helix
 arachidonate 15-lipoxygenase, 260
 secondary structure, 260
 π lone pair of electrons
 electronic structure, 59
 π molecular orbitals
 electronic structure, 56
 placental ribonuclease inhibitor
 domains, 384
 plane-polarized light
 circular dichroism, 597
 plaques
 cloning of DNA, 99
 plasma membrane
 cell fractionation, 744
 definition, 743
 eubacteria, 749
 fungi, 748
 lipid composition, 748
 plasmalogen
 phospholipid, 747
 plasmid
 cloning of DNA, 99
 plasminogen
 diffusion coefficient, 577–78
 domains, 388–89
 frictional coefficient, 577–78
 frictional ratio, 577
 purification, 29
 sedimentation coefficient, 577–78
 plasminogen activator inhibitor 1
 molecular charge, 34
 plastocyanin
 aligning amino acid sequences, 360
 metalloproteins, 331
 packing of side chains, 289
 pleckstrin domain, 386
 point group
 acetylcholine-binding protein, 469
 κ bungarotoxin, 466–67
 chloramphenicol *O*-acetyltrans-
 ferase, 468–69
 cyclic, 466
 definition, 466
 2,2-dialkylglycine decarboxylase
 (pyruvate), 470–71
 dihydrodipicolinate synthase, 475
 L-fucose-phosphate aldolase, 469
 glutathione-disulfide reductase, 467
 heat-labile enterotoxin, 469
 hexokinase, 467
 IMP dehydrogenase, 469
 L-lactate dehydrogenase
 (cytochrome), 469–70
 nucleoside-diphosphate kinase, 474
 oligomeric proteins, 466
 peroxiredoxin, 472, 475
 phosphoribulokinase, 472, 474
 regular polyhedra, 486
 replicative DNA helicase, 469
 ribulose-phosphate 3-epimerase,
 472–73, 475
 L-ribulose-phosphate 4-epimerase,
 469
 serum amyloid P component, 469
 small nuclear ribonucleoprotein,
 470
 sulfate adenylyltransferase, 474
 superoxide dismutase, 475
 transcriptional activator NTRC1,
 470
 transitional endoplasmic reticulum
 ATPase, 469
 point group 11, 470
 point group 2, 466
 point group 222, 470–72
 dimer of dimers, 471
 point group 23, 486–87
 point group 3, 468–69
 point group 322, 472–73
 point group 4, 469
 point group 422, 472
 point group 432, 487–88
 point group 5, 469
 point group 522, 472

- point group 532, 488
- point group 6, 469
- point group 7, 469
- point mutation
 - evolution of proteins, 350
- polarized light
 - light scattering, 415
- poliovirus
 - epitopes, 561
- poly(ethylene glycol) precipitation
 - purification, 23
- polyacrylamide gel
 - electrophoresis, 41
- polyamide backbone
 - posttranslational modification, 114
- polyclonal immunoglobulin, 557
- polyhedrosis virus
 - expression of DNA, 109
- polymerase chain reaction
 - cloning of DNA, 100
- polymeric protein
 - definition, 407, 455
- polynucleotide
 - sequencing of DNA, 96
- polynucleotide 5'-hydroxyl-kinase
 - sequencing of DNA, 96
- polynucleotide 3'-phosphatase/5'-kinase
 - frictional ratio, 577
- polypeptide
 - definition, 74
- polypeptide backbone
 - hydropathy, 276
- polyproline helix
 - benzoylformate decarboxylase, 259
 - secondary structure, 259
- polysaccharides
 - storage, 127
 - structural, 126
- polyvalent antigen, 564
- porin
 - crystallization, 772
 - interfaces, 479
 - space groups, 461, 463
- porin OmpF
 - crystallographic molecular model, 782
 - water-filled channel, 778
- porphine, 61
- porphobilinogen synthase
 - assembly of oligomers, 713
- porphyrin, 61
- positive selection
 - evolution of proteins, 358
- positively charged amino acids
 - nucleic acid, association of proteins with, 315
- posttranslational modification, 113–26
 - N^4 -(β -*N*-acetylglucosaminy)-L-asparaginase, 114
 - S-adenosylmethionine decarboxylase, 114
 - amino terminus, 117
 - aspartate 1-decarboxylase, 114
 - carboxy terminus, 117
 - coenzyme-b sulfoethylthiotransferase, 113
 - concanavalin A, 116
 - cross-link, 119
 - cystine, 122
 - DNA polymerase, 115
 - endopeptidases, 113
 - galactose oxidase, 122
 - geranylgeranylation, 117
 - glutamate-ammonia ligase, 125
 - glycosylphosphatidylinositol (GPI) anchor, 118
 - hedgehog protein, 115
 - hemocyanin, 122
 - high-resolution mass spectrum, 119
 - histidine ammonia-lyase, 126
 - histidine decarboxylase, 114
 - intein, 115
 - isoprenylation, 117
 - mass spectrometry, 119
 - methylation, 115
 - microsin, 114
 - monophenol monooxygenase, 122
 - N \rightarrow O acyl migration, 114
 - n*-tetradecanoyl amide, 117
 - oxazolines, 114
 - polyamide backbone, 114
 - proinsulin, 114
 - pyroglutamate, 117
 - N*-2-pyruvylation, 117
 - RecA protein, 115
 - red fluorescent protein, 126
 - self cleavage, 114
 - signal sequences, 114
 - thiazolines, 114
 - vacuolar adenosinetriphosphatase, 115
- potassium
 - metalloproteins, 328
- potassium channel KcsA
 - crystallization, 772
 - crystallographic molecular model, 779
 - membrane-spanning helices, 772
- potassium thiocyanate
 - preferential solvation, 22
- potential energy
 - of hydrogen bond, 213
 - refinement, 175
- precipitation
 - cell fractionation, 744
- precipitation of proteins
 - purification, 22
- preferential solvation, 22
 - guanidinium, 22, 661
 - hydration of a protein, 297
 - lactose, 31
 - potassium thiocyanate, 22
 - sulfate, 22
 - urea, 661
- preparative electrophoresis
 - cloning of DNA, 101
- prepromagainin
 - domains, 384
- primer
 - polymerase, 97
- primordial proteins
 - molecular taxonomy, 392
- probe
 - cloning of DNA, 100
- procollagen-proline dioxygenase
 - purification, 29
- progressive alignment
 - aligning amino acid sequences, 356–57
- proinsulin
 - posttranslational modification, 114
- proline
 - α helix, 257
 - electronic structure, 76
 - C'*-*endo* conformations, 270
 - C'*-*exo* conformations, 270
 - membrane-spanning α helices, 783
 - stereochemistry of side chains, 270
- proline isomerization
 - cis* peptide bond, 698
 - collagen, 701
 - cytochrome *c*, 702
 - enthalpy of activation, 699
 - immunoglobulin, 701
 - kinetics of folding, 698–702
 - nativelike conformations, 700
 - peptidylprolyl isomerases, 701
 - rate constants, 699, 701
 - ribonuclease, 699–701
 - ribonuclease T₁, 700–01
 - slowly folding isomers, 699
- T3 promoter
 - expressing DNA, 108
- T7 promoter
 - expressing DNA, 108
- tacII* promoter
 - expressing DNA, 109
- promoter-specific transcription
 - factor Sp1
 - purification, 28

- prostaglandin-endoperoxide synthase
 - membrane-bound proteins, 766
- protection factor
 - proton exchange, 644
- protein
 - infrared spectrum, 595
- protein 4.1
 - cytoskeleton, 821
- protein A
 - immunoabsorbent, 566
 - kinetics of folding, 702
- protein CDC20
 - yeast two-hybrid assay, 519
- protein coat of a virus, 488–98
- protein coat of adenovirus
 - $T = 25$ icosahedral symmetry, 498
- protein coat of bacteriophage ϕ X174
 - common ancestor, 495
- protein coat of bacteriophage MS2, 496
 - structural swapping, 480
- protein coat of bacteriophage P22
 - $T = 7$ icosahedral symmetry, 497
- protein coat of beanpod mottle virus
 - common ancestor, 495
 - quasi-equivalence, 492
- protein coat of black beetle nodavirus
 - common ancestor, 495
 - quasi-equivalence, 492
- protein coat of Bluetongue virus
 - $T = 13$ icosahedral symmetry, 498
- protein coat of canine parvovirus
 - common ancestor, 495
- protein coat of cowpea mosaic virus
 - common ancestor, 495
 - quasi-equivalence, 492
- protein coat of foot-and-mouth disease virus
 - common ancestor, 495
 - quasi-equivalence, 494
- protein coat of Hepatitis B virus, 610
 - circular dichroism, 612
 - sedimentation coefficient, 611
 - sedimentation equilibrium, 610
- protein coat of herpes simplex virus
 - $T = 16$ icosahedral symmetry, 498
- protein coat of Mengo virus
 - common ancestor, 495
 - quasi-equivalence, 494
- protein coat of Nudaurelia ω Capensis virus
 - $T = 4$ icosahedral symmetry, 496
- protein coat of poliovirus
 - common ancestor, 495
 - quasi-equivalence, 494
- protein coat of polyoma virus
 - $T = 7$ icosahedral symmetry, 497
- protein coat of primate calcivirus
 - quasi-equivalence, 492
- protein coat from R17 virus
 - mean molar mass of an amino acid, 418
- protein coat of reovirus
 - $T = 13$ icosahedral symmetry, 498
- protein coat of rhinovirus
 - common ancestor, 495
 - interfaces, 480
 - quasi-equivalence, 494
- protein coat of satellite panicum mosaic virus
 - common ancestor, 495
 - icosahedral symmetry, 488–89
 - interfaces, 489
- protein coat of satellite tobacco necrosis virus
 - common ancestor, 495
 - icosahedral symmetry, 494–95
 - molecular taxonomy, 393
- protein coat of simian virus 40
 - $T = 7$ icosahedral symmetry, 497
- protein coat of Sindbis virus
 - $T = 4$ icosahedral symmetry, 496
- protein coat of southern bean mosaic virus
 - common ancestor, 495
 - icosahedral symmetry, 494
 - quasi-equivalent interfaces, 492–94
- protein coat of tomato bushy stunt virus
 - common ancestor, 495
 - icosahedral symmetry, 494
 - molecular taxonomy, 397
 - octahedral symmetry, 492
 - quasi-equivalence, 492
- protein coat of turnip yellow mosaic virus
 - quasi-equivalence, 492
- protein disulfide-isomerase, 710
 - cystine
 - cystines, formation of, 125, 708–09
 - reduction potential, 709
- protein G
 - heterologous interfaces, 513
 - kinetics of folding, 697
- protein geranylgeranyltransferase
 - purification, 29
- protein gp 45
 - nucleic acid, association of proteins with, 321
- protein HPr
 - space groups, 465
- Protein Information Resource
 - aligning amino acid sequences, 354
- protein kinase C
 - peripheral membrane-bound proteins, 764
- protein kinase N
 - purification, 26
- protein L
 - kinetics of folding, 696
- protein L9
 - free energy of folding, 675
 - kinetics of folding, 703
- protein MsbA
 - crystallization, 772
 - membrane-spanning helices, 772
- protein p11
 - heterologous associations, 514
- protein phosphatase 2A
 - evolution of proteins, 351
- protein S6
 - kinetics of folding, 702
- protein Z
 - peripheral membrane-bound proteins, 764
- protein-glutamine γ -glutamyltransferase
 - assembly of fibrin, 721
 - posttranslational modification, 122
- protein-tyrosine kinases
 - hydrogen bonds in crystallographic molecular models, 312
 - in rafts, 811
- protein-tyrosine kinase ZAP-70
 - domains, 390
- protein-tyrosine phosphatase
 - domains, 381
 - heterologous associations, 517
- proteoglycans
 - oligosaccharides on glycoproteins, 132–33
- proteolipid protein
 - membrane-bound proteins, 765
- prothrombin
 - diffusion coefficient, 578
 - frictional coefficient, 578
 - frictional ratio, 578
 - peripheral membrane-bound proteins, 764
 - sedimentation coefficient, 578
- protocatechuate 3,4-dioxygenase
 - assay, 16
 - axes of symmetry, 454
- protofibril
 - assembly of fibrin, 720
- protofilament
 - assembly of microtubules, 721

- protomer
 definition, 451
- proton exchange, 640–45
 α -amylase inhibitor HOE-467A, 644
 amido protons, 640
 basic trypsin inhibitor, 642–44
 buried hydrogen bonds, 641
 calmodulin, 645
 cooperative unfoldings, 645
 dihydrodipicolinate reductase, 645
 effect of pH, 644
 endopeptidolytic analysis, 642
 EX₁ limit, 643
 EX₂ limit, 643
 free energy of folding, 675
 hydrogen bonds, 641
 immunoglobulin, 645
 kinetic mechanism, 643
 kinetics of folding, 691–92, 698
 local vibrational modes, 645
 lysozyme, 645
 mass spectrometry, 642
 micrococcal nuclease, 645
 molten globule, 684
 myoglobin, 642
 [myosin-light-chain] kinase, 645
 neutron diffraction, 645
 nuclear magnetic resonance, 642
 protection factor, 644
 rate constants, 641
 ribonuclease, 645
 specific acid catalysis, 641
 specific base catalysis, 641
 trypsin, 645
 tryptophan, 640
- proto-oncogene protein c-fos
 heterologous associations, 519
- proto-oncogene protein-tyrosine kinase ABL1
 heterologous associations, 517
- PS
 nuclear magnetic resonance, 621
- pseudosymmetric, trimeric
 protomers
 quasi-equivalence, 492
- pseudosymmetry
 heterooligomers, 508
- purification of membrane-bound proteins
 nonionic detergents for, 768–71
- purification of peptides
 immunoabsorbents, 563
- purification of proteins, 20–32
 acetone powder, 23
 acetyl-CoA carboxylase, 49
N-acetylgalactosaminidemucin- β 1,3-galactosyltransferase, 29
N-acetylglucosamine kinase, 29
 [acyl-carrier-protein] S-malonyltransferase, 46
 adenylate cyclase, 29
 α_1 -adrenergic receptor, 29
 β -adrenergic receptor, 29, 816
 adsorption chromatography, 25
 affinity adsorption, 26
 affinity elution, 26
 α -ketoisocaproate oxygenase, 25
 ammonium sulfate precipitation, 23
 aryl-acylamidase, 21
 ATP diphosphatase, 773
 cathepsin D, 29
 choline *O*-acetyltransferase, 29
 coproporphyrinogen oxidase, 26
 cytochrome *b*₅, 773
N-deacetylheparin *N*-sulfotransferase, 29
 2-dehydro-3-deoxyphosphoheptonate aldolase, 29
 3-deoxy-7-phosphoheptulonate synthase, 25
 dihydrofolate reductase, 29
 dipeptidyl-peptidase IV, 773
 DNA polymerase, 30
 dolichyl-phosphate β - δ -mannosyltransferase, 773
 enrichment, 21
 formate-tetrahydrofolate ligase, 26
 5-formyltetrahydrofolate cyclo-ligase, 28–9
 fructose 1,6-bisphosphatase, 26
 glutamyl-tRNA reductase, 26, 31
 glyceraldehyde-3-phosphate dehydrogenase, 24, 48
 hexokinase, 29
 HLA histocompatibility antigen, 773
 HLA-linked B-cell antigen, 773
 homogenization, 20
 inhibitors of peptidases, 49
 ion exchange chromatography, 23
 isocitrate dehydrogenase, 26
 isocitrate dehydrogenase (NAD⁺), 29
 isoelectric precipitation, 23
 L-lactate dehydrogenase, 29
 malate dehydrogenase (oxaloacetate-decarboxylating) (NADP⁺), 29
 malate synthase, 30
 membrane alanine aminopeptidase, 773
 membrane-bound proteins, 768–73
 methylcrotonyl-CoA carboxylase, 25
 micrococcal nuclease, 26
 molecular exclusion chromatography, 24
 Na⁺/K⁺-exchanging ATPase, 768
 ornithine decarboxylase, 29
 phosphoglycerate dehydrogenase, 49
 phosphoglycerate kinase, 24
 phosphoglycerate mutase, 24
 plasminogen, 29
 poly(ethylene glycol) precipitation, 23
 precipitation of proteins, 22
 procollagen-proline dioxygenase, 29
 promoter-specific transcription factor Sp1, 28
 protein geranylgeranyltransferase, 29
 protein kinase N, 26
 specific activity, 21
 streptomycin sulfate precipitation, 23
 sucrose α -glucosidase/oligo-1,6-glucosidase, 773
 total activity, 21
 transketolase, 26
 trimethylamine oxide precipitation, 23
 UDP glucose 4-epimerase, 29
 unspecific monooxygenase, 773
 viral hemagglutinin, 773
 yield of activity, 21
- purine repressor
 nucleic acid, association of proteins with, 320
- purine-nucleoside phosphorylase
 space groups, 463
- pyridine
 electronic structure, 60
- pyroglutamate
 posttranslational modification, 117
- pyrrole
 electronic structure, 61
- pyruvate carboxylase
 assay, 17
- pyruvate decarboxylase
 molecular taxonomy, 396
- pyruvate dehydrogenase (acetyl-transferring)
 assembly of oligomers, 715
- pyruvate dehydrogenase complex
 assembly of oligomers, 715
 nonstoichiometric ratio of subunits, 512
 nuclear magnetic resonance, 638

pyruvate kinase
 aligning crystallographic molecular
 models, 375
 domains, 382–83
 molecular taxonomy, 394–95, 397
 sieving, 427
pyruvate oxidase
 domains, 385
 molecular rotational axes of
 pseudosymmetry, 477
N-2-pyruvylation
 posttranslational modification, 117

Q

QAE cellulose
 chromatography, 9
quadrupole mass spectrometer
 mass spectrometry, 91
quantitative cross-linking, 443–45
 assembly of oligomers, 710–11, 713
 epidermal growth factor receptor,
 815
 integral membrane-bound
 proteins, 786
quantum yield
 fluorescence, 602
quasi-equivalence
 global rotational axis of symmetry,
 491
 icosahedral symmetry, 491–98
 interfaces, 492–94
 local 3-fold rotational axis of
 pseudosymmetry, 491
 protein coat of beanpod mottle
 virus, 492
 protein coat of black beetle
 nodavirus, 492
 protein coat of cowpea mosaic
 virus, 492
 protein coat of foot-and-mouth
 disease virus, 494
 protein coat of Mengo virus, 494
 protein coat of poliovirus, 494
 protein coat of primate calcivirus,
 492
 protein coat of rhinovirus, 494
 protein coat of tomato bushy stunt
 virus, 492
 protein coat of turnip yellow
 mosaic virus, 492
 pseudosymmetric, trimeric
 protomers, 492
quasi-equivalent interfaces
 protein coat of southern bean
 mosaic virus, 492–94
quaternary structure
 acetylcholine receptor, 407

chaperonin 60, 475
 closed structure, 454
 complementary faces, 455
 definition, 451
 3-deoxy-phosphogluconate
 aldolase, 407
 dihydrolipoyllysine residue
 (2-methylpropanoyl)
 transferase, 490
 dihydrolipoyllysine-residue
 acetyltransferase, 490
 dihydrolipoyllysine-residue
 succinyltransferase, 490
 DNA-directed RNA polymerase, 407
 evolution of, 455
 ferritin, 489–90
 glutamate–ammonia ligase, 475
 glycerate dehydrogenase, 483
 hemoglobin, 407
 histidine decarboxylase, 475
 HNRNP arginine methyltrans-
 ferase, 476
 IMP dehydrogenase, 475
 interface, 455
 L-lactate dehydrogenase, 407
 multifunctional endopeptidase, 475
 natural selection, 455
 oligomeric protein, 455
 open structure, 455
 polymeric protein, 455
 λ repressor, 476
 ribonucleoside-diphosphate
 reductase, 475
 serum amyloid P component, 475
 small heat shock proteins, 490
 transketolase, 482
quenching
 of fluorescence, 602
 serum album, 603

R

radial angle
 helical surface lattice, 500
radial molecular correlation function
 water, 193–94
radiationless decay, 602
radius of gyration
 cyclic AMP-dependent protein
 kinase, 581
 cytochrome *c*, 581
 myoglobin, 581
 scattering of electromagnetic
 radiation, 580
rafts
 caveolin, 811
 glycosphingolipids in, 811
 in biological membranes, 811

protein-tyrosine kinases in, 811
 saturated fatty acyl groups in, 811
 sphingomyelin in, 811
Ramachandran plot
 deoxyribonuclease, 255
 glycines, 255
 regions of lowest energy, 254
 secondary structure, 254
Raman effect
 absorption of light, 593
Raman infrared spectrum, 595–96
 resonance Raman infrared
 spectrum, 596
 ribonuclease, 596
 secondary structure, 597
random coil
 acid–base titration curve, 660
 amido proton exchange, 660
 circular dichroic spectrum, 660
 definition, 659
 intrinsic viscosity, 660
 ultraviolet spectrum, 660
random meander
 circular dichroism, 599
 crystallographic molecular mode,
 170
 nuclear magnetic resonance, 632
 secondary structure, 264
randomly coiled polypeptides
 sieving, 428
Raoult's law, 197
rapid mixing chamber
 kinetics of folding, 688
rate constant for folding
 kinetics of folding, 667
rate constants
 proline isomerization, 699, 701
 proton exchange, 641
rate of relaxation
 nuclear magnetic resonance, 614
rate sedimentation
 cell fractionation, 743
Rayleigh's ratio
 light scattering, 416, 579
reading frames, 98
RecA protein
 fluorescence resonance energy
 transfer, 609
 posttranslational modification,
 115
recent speciation
 aligning amino acid sequences,
 357
receptor for the Fc domain of
 immunoglobulinG
 glycosylphosphatidylinositol-
 linked proteins, 765

- receptor Tom 20
 - anchored membrane-bound proteins, 764
- receptors
 - assay, 15
- reconstitution
 - assay, 773
 - membrane-bound proteins, 771
- record of intolerance
 - aligning amino acid sequences, 348
- recovery of fluorescence following photobleaching
 - translational diffusion coefficient, 813
- recurring domain
 - domains, 382
- recurring structure
 - aligning crystallographic molecular models, 373
- red fluorescent protein
 - β barrel, 287
 - posttranslational modification, 126
- refinement
 - amino acid sequence, 178
 - coenzymes, 180
 - constraints, 174–75
 - deoxyribonuclease, 176, 178
 - global potential energy function, 177
 - image reconstruction, 793
 - local minimum, 176
 - molecular dynamics, 177
 - nuclear magnetic resonance, 632
 - omit maps of difference electron density, 178
 - potential energy, 175
 - simulated annealing, 177
 - variant surface glycoprotein, 179
 - water, 178
- refinement of crystallographic molecular models, 172–85
- reflecting faces
 - crystallography, 150
- refractive index
 - light scattering, 415
 - scattering of electromagnetic radiation, 579
- regular polyhedra
 - point groups, 486
- regulatory kinases
 - domains, 386
- regulatory protein Cro
 - association of proteins with nucleic acid, 315–16
- regulatory protein GAL4
 - nuclear magnetic resonance, 633
 - yeast two-hybrid assay, 518
- rehybridization
 - acids and bases, 63
- relative mobility
 - definition, 4
 - electrophoresis, 429
- relative permittivity
 - ionic interactions in crystallographic molecular models, 303
 - water, 190
- relaxation
 - nuclear magnetic resonance, 614
- renin
 - expressing DNA, 109
- replication protein A
 - association of proteins with nucleic acid, 322
- replication terminator
 - association of proteins with nucleic acid, 316
- replicative DNA helicase
 - point group, 469
- reporter group
 - ultraviolet absorption spectra, 601
- lac* repressor
 - interfaces, 480
- λ repressor
 - association of proteins with nucleic acid, 316
 - covalent modification, 547
 - kinetics of folding, 702–03
 - quaternary structure, 476
 - electrospray mass spectrometry, 417
- met* repressor
 - association of proteins with nucleic acid, 316
 - rotational axes of symmetry, 468
- trp* repressor
 - association of proteins with nucleic acid, 319, 321
- repressor protein CI
 - association of proteins with nucleic acid, 321
- resolution
 - chromatography, 4
 - crystallography, 158
- resonance
 - nuclear magnetic resonance, 613
- resonance Raman infrared spectrum, 596
 - cytochrome *d* ubiquinol oxidase, 596
 - halocyanin, 596
 - hemocyanin, 596
- resonance structures, 57
- restriction fragments
 - sequencing of DNA, 96
- restriction mapping
 - sequencing of DNA, 101
- restriction sites
 - expressing DNA, 109
 - sequencing of DNA, 96
- retardation coefficient
 - definition, 41
 - dodecyl sulfate gel electrophoresis, 422
 - electrophoresis, 426
- retinol-binding protein
 - β barrel, 287
 - domains, 384
- reverse-phase chromatography, 8
- reversibility
 - folding, 663
- reversible dissociation
 - interfaces, 471
- L-rhamnose
 - structure, 129
- rhinovirus
 - epitopes, 561
- rhodopsin
 - boundary layer of phospholipid, 786
 - covalent modification, 605
 - crystallization, 772
 - diffusion, 823
 - fluorescence resonance energy transfer, 605
 - membrane-spanning helices, 772
 - rotational diffusion coefficient, 813
 - size, 777
 - topography of membrane-spanning proteins, 807
 - translational diffusion coefficient, 813
- Rho-GDP dissociation inhibitor
 - fluorescence resonance energy transfer, 609
- rhombohedral lattice
 - crystallography, 151
- ribokinase
 - aligning crystallographic molecular models, 362
 - molecular taxonomy, 396
- ribonuclease
 - acid-base titration curve, 33
 - aligning amino acid sequences, 361
 - amino acid sequence, 85
 - covalent modification, 550
 - cross-linking, 548
 - cystine, 124
 - cystines, formation of, 708
 - electrophoresis, 45
 - equilibrium constant for folding, 687
 - fluorescence, 601
 - folding, 668, 678–679, 685

- free energy of folding, 676
- heterologous interfaces, 513
- hydration of a protein, 298–299
- hydrogen bonds in crystallographic molecular models, 310, 313
- infrared spectrum, 595
- kinetics of folding, 695
- molar mass, 418–419
- molecular charge, 35
- nuclear magnetic resonance, 635, 687
- proline isomerization, 699–701
- proton exchange, 645
- Raman spectrum, 596
- sieving, 424
- thermodynamics of folding, 665, 673, 682
- ribonuclease H
 - folding, 678
 - free energy of folding, 675–77
 - kinetics of folding, 688–90, 693, 697–98
 - local conformational changes, 678
- ribonuclease inhibitor
 - heterologous interfaces, 513
- ribonuclease T₁
 - crystallography, 184
 - effect of pH on folding, 663
 - fluorescence, 603
 - free energy of folding, 675, 677
 - hydropathy side chains, 275
 - kinetics of folding, 694
 - nuclear magnetic resonance, 636–37
 - proline isomerization, 700–01
 - stereochemistry of side chains, 267
 - water in crystallographic molecular models, 293
- ribonuclease U₂
 - water in crystallographic molecular models, 292
 - (put this flush with left margin)
 - ribonucleoproteins
 - association of proteins with nucleic acid, 323
- ribonucleoside-diphosphate reductase
 - electron nuclear double resonance, 650
 - electron paramagnetic resonance, 645, 649
 - molecular taxonomy, 397
 - packing of α helices, 285
 - packing of β structure, 286
 - quaternary structure, 475
- ribose-phosphate diphosphokinase assay, 18
- ribosomal protein S17
 - nuclear magnetic resonance, 629
- ribosome
 - assembly of oligomers, 715–17
 - cross-linking, 549
 - crystallography, 324
 - electron microscopy, 588
 - immunolectron microscopy, 567–69
 - neutron scattering, 583–84
 - nucleic acid, association of proteins with, 324
 - 30S subunit, 324
 - 50S subunit, 324
- ribulose-bisphosphate carboxylase
 - covalent modification, 536, 546
 - heterologous interfaces, 510
 - hydrogen bonds in crystallographic molecular models, 306–7
 - molecular rotational axes of symmetry, 465
 - sieving, 427
- ribulose-phosphate 3-epimerase
 - interfaces, 480
 - point group, 472–73, 475
- L-ribulose-phosphate 4-epimerase
 - point group, 469
- ring current
 - nuclear magnetic resonance, 617
- RNA recognition motif, 386
- RNA-directed DNA polymerase
 - cloning of DNA, 97
 - sedimentation equilibrium, 413
- root mean square deviation
 - aligning crystallographic molecular models, 362
- rose bengal
 - reagent for covalent modification, 536
- rotamer
 - definition, 272
 - stereochemistry of side chains, 272
- rotary shadowing
 - electron microscopy, 585
- rotational axes of pseudosymmetry
 - acetylcholine receptor, 787–88
 - aquaporin, 788
 - axes of symmetry, 452
 - cytidine deaminase, 484
 - double-helical DNA, 467
 - hemocyanin, 485
 - integral membrane-bound proteins, 789
 - photosynthetic reaction center, 777, 787
 - rotational axes of symmetry, 452
- bacteriorhodopsin, 787
- E2 DNA-binding domain, 468
- exo- α -sialidase, 787
- fibrin, 719
- gap junction connexon, 787
- glutathione synthase, 483
- integral membrane-bound proteins, 787
- methionine adenosyltransferase, 480
- nitric-oxide synthase, 480
- arc* repressor, 468
- met* repressor, 468
- viral hemagglutinin viral hemagglutinin, 787
- rotational conformation
 - stereochemistry of side chains, 267
- rotational correlation time
 - water, 195
- rotational diffusion coefficient
 - bacteriorhodopsin, 812
 - decay in the anisotropy, 812
 - in membranes, 812
 - rhodopsin, 813
- rotational relaxation time
 - molten globule, 684
- rough endoplasmic reticulum, 743
 - cell fractionation, 744
- running gel
 - electrophoresis, 44
- rusticyanin
 - electrospray mass spectrometry, 417
- ryanodine receptor
 - integral membrane-bound protein, 766
- S**
- salting in, 22
- salting out, 22
- SAND domain, 386
 - domains, 386
- saponins
 - detergents, 769
- saturated fatty acids
 - phospholipid, 745
- saturated fatty acyl groups
 - in rafts, 811
- saturation
 - chromatography, 2–3
 - nuclear magnetic resonance, 614
- SBC
 - nuclear magnetic resonance, 621
- scales of hydropathy, 244
 - from free energy of transfer, 274
- scanning calorimetry
 - folding, 664
 - thermodynamics of folding, 664

- scattering at small angles
 - hydration of a protein, 299
- scattering length
 - hydrogen, 583
 - neutron scattering, 583
- scattering of electromagnetic radiation, 579–85
 - forward scattering, 579
 - intramolecular interference, 579
 - optical constant, 579
 - radius of gyration, 580
 - Rayleigh ratio, 579
 - refractive index, 579
- scattering of X-radiation
 - kinetics of folding, 691
 - water, 193
- schistoside
 - glycosphingolipid, 748
- screen libraries
 - cloning of DNA, 99
 - immunoglobulins, 567
- screw axis of symmetry, 452
- searching data banks of amino acid sequences
 - aligning amino acid sequences, 353–54
- secondary structure, 251–67
 - α helix, 256–59
 - amide I band, 596
 - β -pleated sheet, 261
 - β structure, 260–61
 - β turn, 261–64
 - cis* peptide bonds, 251–52
 - crystallography, 165–67
 - deoxyribonuclease, 263
 - dihedral angles ϕ and ψ , 252–54
 - γ turn, 264
 - infrared spectrum, 596
 - peptide bond, 251
 - π helix, 260
 - polyproline helix, 259
 - Ramachandran plot, 254
 - Raman infrared spectrum, 597
 - random meander, 264
- sedimentation coefficient
 - cell fractionation, 743
 - protein coat of Hepatitis B virus, 611
 - sedimentation velocity, 576
- sedimentation equilibrium
 - and sedimentation velocity, 413
 - centrifugal potential, 411
 - chemical potential, 411
 - concentration of protein, 412
 - equilibrium between oligomers, 414
 - gradient of concentration, 411
 - molar mass, 411–14
 - partial specific volume, 412
 - protein coat of Hepatitis B virus, 610
 - RNA-directed DNA polymerase, 413
 - serum albumin, 412
 - virial coefficients, 412
- sedimentation velocity, 576–77
 - and sedimentation equilibrium, 413
 - aspartate carbamoyltransferase, 576
 - buoyant force, 576
 - buoyant mass, 576
 - desmin, 577
 - frictional coefficient, 576–78
 - frictional ratio, 577–78
 - sedimentation coefficient, 576
 - terminal velocity, 576
- selection rules
 - infrared spectroscopy, 595
- selective adsorption
 - chromatography, 3
- selenocysteine lyase
 - assay, 19
- self cleavage
 - posttranslational modification, 114
- self-charging energy
 - ion, 200
- self-diffusion coefficient
 - water, 194
- self-diffusion of water
 - hydration of a protein, 297
- self-rotation function
 - molecular rotational axes of symmetry, 465
- seminal ribonuclease
 - covalent modification, 545
- semipermeable membrane
 - osmotic pressure, 408
- separately unfolding domains, 388–89
- sequence of DNA
 - acetylcholine receptor, 106–7
 - Factor VIII, 106
 - frameshift, 106
 - genomic sequences, 108
 - initiation codon, 106
 - open reading frame, 106
 - termination codon, 106
- sequence of monosaccharides
 - oligosaccharides on glycoproteins, 133
- sequencing of DNA
 - acetylcholine receptor, 101–2
 - antisense strand, 98
 - blunt ends, 96
 - chemical method, 103, 105
 - cloning of DNA, 99
 - 2',3'-dideoxynucleotide, 104
 - electrophoresis, 101–3
 - 3'-end, 95
 - 5'-end, 95
 - end-labeled fluorescent fragments, 104
 - end-labeled fragments, 103
 - enzymatic method, 103–5
 - ladder, 102
 - nucleoside 5'-monophosphates, 95
 - polynucleotide 5'-hydroxyl-kinase, 96
 - polynucleotides, 96
 - restriction fragments, 96
 - restriction mapping, 101
 - restriction sites, 96
 - site-specific deoxyribonucleases, 95–96
 - sticky ends, 96
- sequencing of polypeptides, 85–95
 - carboxypeptidase A, 91
 - carboxypeptidase B, 91
 - denaturing proteins, 87
 - Edman degradation, 86
 - endopeptidases, 87–88
 - exopeptidases, 91
 - leucyl aminopeptidase, 91
 - mass spectrometry, 91–93
 - serine-type carboxypeptidase, 91
- sequencing oligosaccharides
 - β -N-acetylglucosamidase, 134
 - anion exchange chromatography, 134
 - β -elimination, 133
 - endo-1,4- β -galactosidase, 135
 - endo- α -sialidase, 135
 - endoglycosidases, 133
 - exo- α -2,3-sialidase, 134
 - exo- α -sialidase, 134
 - exoglycosidases, 134
 - α -L-fucosidase, 134
 - β -galactosidase, 134
 - mass spectrometry sequencing oligosaccharides, 136
 - methylation, 135
 - nuclear magnetic resonance spectroscopy, 136
 - periodic acid, 134, 136
 - Smith degradation, 134, 136
 - sodium borohydride, 135
- serine
 - acid dissociation constant, 75
 - electronic structure, 77
 - stereochemistry of side chains, 269
- serine peptidases, 49
- serine-type carboxypeptidase
 - sequencing of polypeptides, 91

- serum albumin
 collisional quenching, 603
 diffusion coefficient, 578
 dodecyl sulfate gel electrophoresis, 422
 domains, 384
 Donnan effect, 419
 electrophoresis, 42, 45
 frictional coefficient, 578
 frictional ratio, 578
 hydration of a protein, 298–99
 light scattering, 416
 mean molar mass of an amino acid, 418
 molar mass, 412, 418
 osmotic pressure, 410
 preferential solvation, 661
 sedimentation coefficient, 578
 sedimentation equilibrium, 412
 sieving, 424, 427–28
 unfolding, 661
- serum amyloid P component
 point group, 469
 quaternary structure, 475
- sex-lethal protein
 domains, 390
- SH2 domain, 386
- SH3 domain, 386
- Shaker S4 K⁺ channel
 immunoadsorbent, 566
- shape
 of a protein, 573–92
- shape of a protein
 electron microscopy, 585–88
- SHC transforming protein
 heterologous associations, 517
- shear
 viscosity, 578
- shell of hydration
 water in crystallographic molecular models, 296
- shikimate kinase
 domains, 380
- sialic acids
 oligosaccharides of glycoproteins, 128
 structure, 129
- side chains of the amino acids, 74
- shape, 164
- sieving, 423–31
 α -amylase, 427–28
 β -amylase, 427
 apoferritin, 424, 427
 apparent surface area of a protein, 423
 aspartate kinase–homoserine dehydrogenase, 427
- catalase, 424
- chymotrypsinogen, 424
- complexes between dodecyl sulfate and polypeptides, 429
- cytochrome *c*, 424, 428
- definition, 423
- electrophoresis, 426–30
- extended polymers, 428
- frictional ratio, 425
- fructose-bisphosphate aldolase, 424, 427–28
- fumarate hydratase, 424
- β -galactosidase, 424
- glyceraldehyde-3-phosphate dehydrogenase, 424
- hemoglobin, 424, 428
- hexokinase, 427
- immunoglobulin G, 424, 428
- L-lactate dehydrogenase, 424, 427
- β -lactoglobulin, 428
- lactoperoxidase, 424
- lysozyme, 424
- malate dehydrogenase, 424, 428
- molecular exclusion
 chromatography, 423
- myoglobin, 424
- number of amino acids in a protein, 424
- ovalbumin, 424, 427–28
- ovotransferrin, 427
- pepsin, 424
- pyruvate kinase, 427
- randomly coiled polypeptides, 428
- ribonuclease, 424
- ribulose-bisphosphate carboxylase, 427
- serum albumin, 424, 427–28
- single-stranded nucleic acids, 428–29
- standard proteins, 426
- Stokes radius, 426
- transferrin, 424, 427–28
- urease, 424, 427
- xanthine oxidase, 427
- σ bonds, 59
- sigma factor rpoD
 covalent modification, 548
- σ lone pair of electrons
 electronic structure, 59
- σ - π stereochemical representation
 electronic structure, 56
- σ structure, 59
- signal sequences
 anchored membrane-bound proteins, 764
 posttranslational modification, 114
- Simha factor
 viscosity, 579
- simulated annealing
 refinement, 177
- single-stranded DNA
 nucleic acid, association of proteins with, 321–22
- single-stranded DNA binding protein
 association of proteins with nucleic acid, 322
- single-stranded nucleic acids
 sieving, 428–29
- site-directed mutation, 110–11, 544
 cassettes, 111
 detection of heterologous associations, 519
 effect on free energy of folding, 675
 ionic interactions in crystallographic molecular models, 303
 kinetics of folding, 703
 membrane-bound proteins, 773
 mutagenic oligonucleotide, 110
 topography of membrane-spanning proteins, 799
 tyrosyl-tRNA synthetase, 110
 unnatural amino acids, 111
- site-specific deoxyribonuclease
 association of proteins with nucleic acid, 318
- site-specific deoxyribonucleases
 sequencing DNA, 95–96
- skeletal representation
 crystallographic molecular mode, 167
- sliding fluctuations
 bilayer of phospholipid, 752
- small heat shock proteins
 quaternary structure, 490
- small nuclear ribonucleoprotein
 point group, 470
- Smith degradation
 sequencing oligosaccharides, 134, 136
- sn*-glycerol 3-phosphate
 phospholipid, 745
- snorkeling
 membrane-bound proteins, 780
- sodium
 metalloproteins, 328
- sodium borohydride
 sequencing oligosaccharides, 135
- sodium/proton antiporter NhaA
 image reconstruction, 793
- softness
 metal ion, 327
- solution scattering
 X-ray scattering, 582

- solution scattering curves, complete
 X-ray scattering, 582
 solution scattering curves, theoretical
 X-ray scattering, 582
 solvation
 definition, 189
 hydrogen bond, 208
 solvent flattening
 crystallography, 161
 somatotropin
 heterologous associations, 519
 somatotropin receptor
 heterologous associations, 519
 space group
 definition, 456
 space group *C*2, 458
 space group *P*2, 457
 space group *P*2₁2₁2₁, 460
 space group *P*3₂21, 462
 space groups, 456–65
 alcohol dehydrogenase, 463
 chloramphenicol *O*-
 acetyltransferase, 463
 crystallographic asymmetric unit,
 457
 crystallographic axis of symmetry,
 461
 cystathionine, 464
 deoxyribonuclease, 458
 designation of, 459
 dihydrolipoyllysine-residue
 acetyltransferase, 463
 dihydrolipoyllysine-residue
 succinyltransferase, 463
 exact rotational axis of symmetry,
 461
 ferredoxin, 464
 general control protein GCN4, 464
 glutathione peroxidase, 463
 glyceraldehyde-3-phosphate
 dehydrogenase
 (phosphorylating), 463
 (*S*)-2-hydroxy-acid oxidase, 463
 L-lactate dehydrogenase, 463
 lectin, 460
 malate dehydrogenase, 461
 mitochondrial H⁺-transporting
 two-sector ATPase, 461
 molecular axis of symmetry, 461
 packing in, 457
 phosphorylase *b*, 463
 porin, 461, 463
 protein HPr, 465
 purine-nucleoside phosphorylase,
 463
 sets of axes of symmetry, 457
 telokin, 462
 triose-phosphate isomerase, 463
 unit cell, 459
 space-filling representation
 crystallographic molecular mode,
 170
 speciation of organisms
 aligning amino acid sequences,
 354
 speciation of proteins
 molecular taxonomy, 393
 species of domains
 molecular taxonomy, 393
 specific acid catalysis
 proton exchange, 641
 specific activity
 definition, 21
 specific base catalysis
 proton exchange, 641
 specific viscosity
 definition, 578
 specificity
 iodoacetamide, 532
 spectrin
 cytoskeleton, 820
 domains, 384–85, 390
 peripheral membrane-bound
 proteins, 764
 spermadhesin
 X-ray scattering, 583
 spheroplasts, 744
 sphingomyelin
 in rafts, 811
 phospholipid, 748
 sphingosine, 748
 spider dragline silk
 amino acid sequence, 106
 spin diffusion
 nuclear magnetic resonance, 617
 spin quantum number
 electron paramagnetic resonance,
 646
 nuclear magnetic resonance,
 613–14
 spin states
 nuclear magnetic resonance, 613
 spin-labeled phospholipids
 integral membrane bound
 proteins, 795
 spin-labeled probes
 biological membranes, 808
 spin–spin coupling
 electron paramagnetic resonance,
 647
 nuclear magnetic resonance, 615
 splicing of messenger RNA
 evolution of proteins, 350
 spongiform encephalopathy, 508
 SSEARCH, 354
 evaluation of, 368
 stability of a protein
 hydrogen bonds in crystallographic
 molecular models, 311
 stable intermediates
 molten globules, 683
 stable moving boundaries
 electrophoresis, 42
 stacking
 dodecyl sulfate gel electrophoresis,
 422
 electrophoresis, 43
 stacking gel
 electrophoresis, 43
 staggered conformation
 stereochemistry of side chains, 267
 stain for enzymatic activity
 electrophoresis, 46
 standard enthalpy of formation
 ion pair, 200
 standard free energy of solvation, 198
 standard free energy of transfer
 concentration, units of, 198
 definition, 198
 standard proteins
 sieving, 426
 standard states, 196–99
 entropy of mixing, 196
 START domain, 386
 start site
 evolution of proteins, 350
 stationary phase
 definition, 2
 statistical significance
 aligning amino acid sequences, 353
 stearic acid, 745
 stereochemistry of side chains, 267–72
 alternative conformations, 267
 α -lytic endopeptidase, 268
 aromatic amino acids, 269
 asparagine, 270
 aspartate, 270
 carboxypeptidase C, 271
 cystine, 271
 dihedral angle χ_1 , 267
 dihedral angle χ_2 , 269
 glutamate, 270
 glutamine, 270
 glutathione reductase, 267
 hemoglobin, 267
 isoleucine, 268
 methionines, 270
 methyl group, 270
 ovomucoid inhibitor, 268
 penicillopepsin, 268
 proline, 270

- ribonuclease T₁, 267
- rotamer, 272
- rotational conformation, 267
- serine, 269
- staggered conformation, 267
- streptogrisin A, 268
- streptogrisin B, 268
- threonine, 268
- valine, 267
- steric crowding
 - assembly of oligomers, 715
- steric effects
 - hydrogen bonds in crystallographic molecular models, 309
- steric exclusion
 - definition, 510
 - heterooligomers, 510
 - immunoglobulin ϵ receptor, 510
 - transthyretin, 510–11
- steric repulsion
 - bilayers of phospholipid, 751
- steroid Δ -isomerase
 - assembly of oligomers, 712
 - fluorescence resonance energy transfer, 607
- sterol carrier protein
 - membrane-bound proteins, 766
- sticky ends
 - sequencing of DNA, 96
- stoichiometric ratio of subunits
 - cross-linking, 445
- stoichiometry of the subunits
 - definition, 407
- Stokes' radius, 38
 - definition, 37
 - sieving, 426
- stop site
 - evolution of proteins, 350
- stopped-flow apparatus
 - dead time, 688
 - kinetics of folding, 688
- streptococcal protein G
 - hydrogen bonds in crystallographic molecular models, 312
- streptogrisin A
 - stereochemistry of side chains, 268
- streptogrisin B
 - stereochemistry of side chains, 268
- streptomycin sulfate precipitation
 - purification, 23
- stretching frequency
 - hydrogen bond, 209
 - water, 195
- string of spherical beads
 - frictional coefficient, 576
- strongest possible hydrogen bond, 212
- structural alignment, 366
- structural domain
 - domains, 387
- structural swapping, 480–81
- structure factor
 - crystallography, 155
 - definition, 155
- subtilisin
 - aligning amino acid sequence, 360
 - circular dichroism, 600
 - folding, 679
 - ionic interactions, 300
 - molecular taxonomy, 395
 - nuclear magnetic resonance, 635
- subunit
 - definition, 407
- succinate dehydrogenase
 - crystallization, 772
 - membrane-spanning helices, 772
 - membrane-spanning segments, 777
- succinate-CoA ligase
 - molecular taxonomy, 395
- succinate-CoA ligase (ADP-forming)
 - cross-linking, 445
- succinate-propionate CoA-transferase
 - aligning amino acid sequences, 360
- succinic anhydride
 - reagent for covalent modification, 535
- succinyldiaminopimelate
 - transaminase assay, 20
- sucrose α -glucosidase/oligo-1,6-glucosidase
 - embedded anchor, 774
 - purification, 773
- sucrose porin
 - crystallization, 772
- sulfate
 - electronic structure, 82
 - preferential solvation, 22
- sulfate adenyltransferase
 - point group, 474
- sulfate-binding protein
 - hydrogen bonds in crystallographic molecular models, 309
- sulfenic acid
 - cysteine, 81–82
 - metalloproteins, 330
- sulfenyl halides
 - reagents for covalent modification, 538
- sulfhydryl peptidases, 49
- sulfinic acid
 - cysteine, 81–82
- metalloproteins, 330
- sulfite oxidase
 - domains, 377
- sulfite reductase
 - molecular rotational axes of pseudosymmetry, 476
- N*-(2-sulfoethyl)cyclohexylamine
 - buffer, 68
- N*-(2-sulfoethyl)morpholine
 - buffer, 68
- sulfonate
 - cysteine, 81–82
- sulfone
 - methionine, 81–82
- N*-(3-sulfopropyl)-2-amino-1,3-dihydroxy-2-hydroxymethylpropane
 - buffer, 68
- N*-(3-sulfopropyl)morpholine
 - buffer, 68
- sulfoxide
 - methionine, 81–82
- sulfur
 - hydrogen bond, 207
 - hydropathy, 241
- superfamily
 - molecular taxonomy, 396
- superoxide dismutase
 - aligning crystallographic molecular models, 363
 - convergent evolution, 373
 - hydrogen bonds in crystallographic molecular models, 306
 - interfaces, 480
 - packing of side chains, 281
 - point group, 475
 - X-ray scattering, 583
- superposed α carbons
 - molecular rotational axes of symmetry, 465
- superposition
 - aligning crystallographic molecular models, 362
 - definition, 362
- surface area
 - hydrophobic effect, 238
- surface area at zero pressure
 - monolayer of lipids, 761
- surface potential
 - bilayer of phospholipid, 754
- surface pressure
 - monolayer of lipids, 761
- surface tension
 - water, 190
- Swiss-Prot Sequence Database
 - aligning amino acid sequences, 354

- symmetric hydrogen bond, 212
 symmetry operations
 axes of symmetry, 451
 synapsin Ia
 molecular taxonomy, 397
 synaptobrevin-II
 heterologous interfaces, 513
 synaptotagmin
 heterologous associations, 516
 synchrotron
 crystallography, 156
syn lone pair, 69
 syntaxin-1A
 heterologous interfaces, 513
 synthetic hydrophobic peptide
 membrane-spanning α helices, 774
 synthetic peptide
 antigen, 562–64
 synthetic peptides
 coiled coil of α helices, 284
 systemic amyloidosis, 508
- T**
- $T = 3$ icosahedral symmetry, 496
 $T = 4$ icosahedral symmetry, 496
 protein coat of Nudaurelia
 ω Capensis virus, 496
 protein coat of Sindbis virus, 496
 $T = 7$ icosahedral symmetry, 497
 protein coat of bacteriophage P22,
 497
 protein coat of polyoma virus, 497
 protein coat of simian virus 40, 497
 $T = 13$ icosahedral symmetry, 498
 protein coat of Bluetongue virus,
 498
 protein coat of reovirus, 498
 $T = 16$ icosahedral symmetry, 498
 protein coat of herpes simplex
 virus, 498
 $T = 25$ icosahedral symmetry, 498
 protein coat of adenovirus, 498
 tandem mass spectrometer
 mass spectrometry, 93
 tartronate-semialdehyde synthase
 aligning amino acid sequence, 360
 TATA-box-binding protein
 nucleic acid, association of proteins
 with, 320
 tautomer
 definition, 69
 tautomeric equilibrium constants, 71
 tautomeric interactions
 ionic interactions in
 crystallographic molecular
 models, 300
 tautomers, 69–74
- β -xylanase, 73
 conformational isomers, 69
 equilibrium constants, 71
 ethyl acetoacetate, 70
 heterocycle, 73
 histidine, 78
 keto-enol, 70
 macroscopic acid dissociation
 constant, 71
 thioredoxin, 72–73
 titration curve, 73
 uridine, 69
 taxonomic system of the proteins, 393
 T-cell receptor
 heterologous associations, 513
 telokin
 space groups, 462
 telomere end-binding protein
 association of proteins with
 nucleic acid, 322
 temperature
 effect on folding, 662
 temperature jump
 kinetics of folding, 695
 temperature of maximum stability
 thermodynamics of folding, 672
 template
 polymerase, 97
 terminal velocity
 electrophoresis, 37
 sedimentation velocity, 576
 termination codon
 sequence of DNA, 106
 tertiary structure
 nucleic acid structure, 323
 tetracycline repressor
 covalent modification, 544
 tetragonal lattice
 crystallography, 151
 tetrahedral point group 23, 486–87
 tetrahedral symmetry
 bromoperoxidase, 487
 3-dehydroquinone dehydratase, 487
 hexamers of dimers, 487
 ornithine carbamoyltransferase,
 487
 phaseolin, 487
 tetramers of trimers, 487
 3,4,5,6-tetrahydrophthalic anhydride
 reagent for covalent modification,
 536
 2,3,4,5-tetrahydropyridine-2,6-
 dicarboxylate *N*-
 succinyltransferase
 β helix, 263
 tetraloop
 nucleic acid structure, 322
- tetramers of trimers
 tetrahedral symmetry, 487
 tetranitromethane, 547
 reagent for covalent modification,
 538
 theoretical plate
 chromatography, 4–5
 thermitase
 metalloproteins, 329
 thermodynamics of folding, 659–88
 calorimeter, 671
 chymotrypsinogen, 673
 compressibility of folding, 673
 condensation model, 683
 configurational entropy, 681–82
 cystine, 681
 enthalpy of folding, 671
 equilibrium constant for folding,
 664
 excluded volume, 681
 free energy of folding, 673
 heat capacity change of folding, 671
 immunoglobulin G, 682
 α -lactalbumin, 682
 β -lactoglobulin, 671
 lysozyme, 682
 myoglobin, 673
 oligomer, 670
 ribonuclease, 665, 673, 682
 scanning calorimetry, 664
 temperature of maximum stability,
 672
 two-state assumption, 665
 volume change of folding, 673
 thermolysin
 cleavage of polypeptide, 88
 domains, 389
 hydrogen bonds in crystallographic
 molecular models, 311
 thiamin pyridinylase
 aligning crystallographic molecular
 models, 363
 molecular taxonomy, 396
 thiazolines
 posttranslational modification, 114
 thick filament of myosin
 bare zone, 731–32
 helical surface lattice, 731–32
 length, 732
 structure, 731
 thickness of the double layer, 39
 thin filament
 actin, 506
 assembly of actin, 729
 thin-layer chromatography, 4
 thiocyanate
 mild denaturant, 713

- thiol:disulfide interchange protein
 cystines, formation of, 708
 reduction potential, 709
- thiol:disulfide interchange protein
 dsbA
 circular permutation, 680
- thioredoxin
 cystine, 125
 free energy of folding, 675, 677
 molecular taxonomy, 395
 sequencing by mass spectrometry, 92
 tautomers, 72
 water in crystallographic molecular models, 293
- thioredoxin-disulfide reductase
 domains, 388
 molecular taxonomy, 393
- thiosulfate sulfurtransferase
 domains, 383
 frictional coefficient, 588
 molecular rotational axes of
 pseudosymmetry, 476–77
 molecular taxonomy, 395
- 2-S-[¹⁴C]thiuroniummethanesulfonate
 impermeant reagent for covalent modification, 799
- three-dimensional spectroscopy
 nuclear magnetic resonance, 617–38
- threonine
 acid dissociation constant, 75
 electronic structure, 77
 nuclear magnetic resonance, 636
 stereochemistry of side chains, 268
 water in crystallographic molecular models, 296
- α -thrombin
 heterologous associations, 513
 metalloproteins, 328
- thrombomodulin
 cystine, 125
 heterologous associations, 513
- thrombospondin I
 domains, 386
- thymidine, 95
- thymidylate synthase
 domains, 380
- thyroglobulin
 oligosaccharides of glycoproteins, 138
- time of mixing
 nuclear magnetic resonance, 626
- time-of-flight mass spectrometer
 mass spectrometry, 91
- titin
 domains, 385
- heterologous associations, 513
 length, 85
- titration curve
 acetic acid, 67
 acids and bases, 66
 histidine, 79
 ionic interactions in crystallographic molecular models, 300
 tautomers, 73
- tobacco mosaic virus
 helical surface lattice, 499–500
 X-ray diffraction, 502
- TOCSY
 nuclear magnetic resonance, 621
- tonofilaments
 intermediate filaments, 506
- topography of membrane-spanning proteins, 798–803
- acetylcholine receptor, 802
- ADP, ATP carrier, 799
- alkaline phosphatase, 800
- band 3 anion transport protein, 800, 801–2
- Ca²⁺-transporting ATPase, 802–3
- chloramphenicol O-acetyltransferase, 800
- cytochrome-c oxidase, 802
- definition, 798
- endopeptidases, 800
- glutamine γ -glutamyltransferase, 799
- glycosylation, 800
- H⁺/K⁺-exchanging ATPase, 802–3
- H⁺-exchanging ATPase, 802–3
- immunoglobulins G, 800
- impermeant reagents, 798
- intact cells, 798
- K⁺-transporting ATPase, 802–3
- β -lactamase, 800
- lactoperoxidase, 799
- lactose permease, 800
- MalG protein, 800
- mitochondria, 798
- MotA protein, 799
- Na⁺/K⁺-exchanging ATPase, 799, 802
- oriented, sealed vesicles, 798
- rhodopsin, 807
- site-directed mutation, 799
- topoisomerase I
 association of proteins with nucleic acid, 316, 321
- total activity, 21
- transcription factor IIIA
 nucleic acid, association of proteins with, 324
- transcription factor AP-1
 association of proteins with nucleic acid, 316–17
- fluorescence resonance energy transfer, 609
- heterologous associations, 519
- transcription factor AREA
 nucleic acid, association of proteins with, 319
- transcription factor Rob
 nucleic acid, association of proteins with, 317
- transcription initiation factor TFIID
 heterologous associations, 517
- transcriptional activator NTRC1
 point group, 470
- transfer between water and the gas
 hydropathy, 241
- transfer of saturation
 nuclear magnetic resonance, 616
- transfer RNA
 nucleic acid structure, 322–23
- transferrin
 hydrogen bonds in crystallographic molecular models, 312
 sieving, 424, 427–28
- transforming growth factor β 1
 nuclear magnetic resonance, 627
- transitional endoplasmic reticulum
 ATPase
 point group, 469
- transketolase
 purification, 26
 quaternary structure, 482
- translational diffusion coefficient, 811
- acetylcholine receptor, 813
- bacteriorhodopsin, 813
- cytochrome-c reductase, 814
- endoplasmic reticulum Ca²⁺-transporting ATPase, 813
- in membranes, 812
- phospholipids, 814
- recovery of fluorescence following photobleaching, 813
- rhodopsin, 813
- ubiquinol-cytochrome-c reductase, 814
- transthyretin
 steric exclusion, 510–11
- trc* promoter
 expressing DNA, 109
- triacylglycerol lipase
 assay, 16
 water in crystallographic molecular models, 294
- trifluoroacetic anhydride
 reagent for covalent modification, 535

- 3-(trifluoromethyl)-3-(*m*-[¹²⁵I]iodophenyl)diazirine
hydrophobic reagent for covalent modification, 797
- trimethylamine oxide precipitation
purification, 23
- trimethylamine-*N*-oxide reductase (cytochrome *c*)
multiple isomorphous replacement, 160
- 2,4,6-trinitrobenzenesulfonate, 546
reagent for covalent modification, 536
- triose-phosphate isomerase
domains, 382
molecular rotational axes of symmetry, 465
molecular taxonomy, 394
space groups, 463
- tripeptidyl-peptidase II
aligning amino acid sequence, 360
- triple helix
collagen, 504
- tris(2-carboxyethyl)phosphine
cystine, 125
- 1-tritiospiro[adamantane-4,3'-diazirine]
hydrophobic reagent for covalent modification, 797
- Triton X-100
detergent, 770
- Tritons
detergents, 769
- tRNA-intron endonuclease
helical polymer, 455
- tropomyosin
viscosity, 589
- troponin
fluorescence resonance energy transfer, 609
- troponin C
aligning amino acid sequences, 360
heterologous associations, 514
hydrogen bonds in
crystallographic molecular models, 312
X-ray scattering, 583
- troponin I
heterologous associations, 514
- TROSY
nuclear magnetic resonance, 621
- trypsin
cleavage of polypeptide, 88
electrophoresis, 41, 45
hydrogen bonds in
crystallographic molecular models, 309
molecular taxonomy, 396
proton exchange, 645
- trypsinogen
aligning amino acid sequences, 360
- tryptase
aligning crystallographic molecular models, 362–63
- tryptic digests
peptide map, 432
- tryptophan
acid dissociation constant, 75
covalent modification, 538–39
electronic structure, 76
fluorescence, 601
free energy of transfer, 276
hydrogen bonds in crystallographic molecular models, 308
membrane-spanning α helices, 774
nuclear magnetic resonance, 636
proton exchange, 640
ultraviolet absorption spectra, 601
- tryptophan synthase
assembly of oligomers, 714
folding, 685
- tryptophanase
molecular charge, 34
- tryptophans
circular dichroism, 598
- tryptophan-tRNA ligase
assay, 14
- tube of electron density
crystallography, 162
- tubulin
assembly of microtubules, 722
evolution of proteins, 351
image reconstruction, 502
protomer of microtubule, 722
structure, 722
X-ray scattering, 583
- tumor necrosis factor
molecular taxonomy, 393
- tumor necrosis factor receptor-associated factor 2
heterologous associations, 514
- tungsten
metalloproteins, 330
- two-dimensional crystalline array
image reconstruction, 790
- two-dimensional gel electrophoresis, 567–68
- two-dimensional solution
biological membranes, 811
- two-dimensional spectroscopy
nuclear magnetic resonance, 617–38
- two-state assumption
thermodynamics of folding, 665
- tyrosine
acid dissociation constant, 75
circular dichroism, 598
covalent modification, 536–38
electronic structure, 77
free energy of transfer, 276
membrane-spanning α helices, 774
peptide map, 433
ultraviolet absorption spectra, 601
water in crystallographic molecular models, 296
- tyrosine phenol-lyase
molecular taxonomy, 396
- tyrosyl-tRNA synthetase
site-directed mutation, 110
- U**
- U₁ small nuclear ribonucleoprotein
nucleic acid, association of proteins with, 324
- ubiquinol-cytochrome-*c* reductase
bound phospholipid, 784
crystallization, 772, 775
crystallographic molecular model, 781
electrospray mass spectrometry, 417
membrane-spanning helices, 772, 777
short subunits, 777
size, 777
translational diffusion coefficient, 814
- ubiquitin
evolution of proteins, 351
expressing DNA, 109
kinetics of folding, 702
- UDP glucose 4-epimerase
purification, 29
- UDP-glucose 6-dehydrogenase
domains, 384
- UDP-*N*-acetylglucosamine 1-carboxyvinyltransferase
covalent modification, 546
- ultrasound
molten globule, 684
- ultraviolet absorption spectra
absorption of light, 601
cystine, 601
nitrotyrosine, 601
phenylalanine, 601
random coil, 660
reporter group, 601
tryptophan, 601
tyrosine, 601
- ultraviolet light
absorption of, 594

- unfolded polypeptides
 - dodecyl sulfate gel electrophoresis, 421
- unfolded state
 - definition, 659
- unfrozen water
 - hydration of a protein, 297
- unit cell
 - crystallography, 150
 - definition, 150
 - space groups, 459
- unnatural amino acids
 - site-directed mutation, 111
- unrefined map of electron density
 - crystallography, 161
- unsaturated fatty acids
 - phospholipid, 745
- unspecific monooxygenase
 - expression, 776
 - membrane-spanning α helices, 796
 - purification, 773
- uracil
 - electronic structure, 65
- urate oxidase
 - interfaces, 480
- urea
 - denaturant, 660
 - free energies of transfer, 660
 - preferential solvation, 661
- urease
 - frictional ratio, 426
 - metalloproteins, 330
 - sieving, 424
- uridine
 - tautomers, 69
- UTP-hexose-1-phosphate
 - uridylyltransferase
 - metalloproteins, 330–31
- V**
 - vacuolar H^+ -transporting two-sector ATPase
 - aligning amino acid sequence, 360
 - posttranslational modification, 115
 - valence electrons
 - electronic structure, 56
 - valine
 - electronic structure, 76
 - stereochemistry of side chains, 267
 - van der Waals forces
 - hydrophobic effect, 235–37
 - van der Waals radius
 - definition, 277
 - values for, 277
 - vanadate
 - reagent for covalent modification, 544
 - vanadium
 - metalloproteins, 330
 - vapor
 - water, 190
 - variant surface glycoprotein
 - glycosylphosphatidylinositol-linked proteins, 765
 - interfaces, 480
 - refinement, 179
 - vector equations
 - multiple isomorphous replacement, 159
 - vectorial insertion
 - band 3 anion transport protein, 767
 - membrane-bound proteins, 767
 - of glycolipids into membranes, 767
 - of glycoproteins into membranes, 767
 - very low density lipoprotein
 - lipoproteins, 804
 - vesicles of phospholipid
 - large, unilamellar, 749
 - multibilayer, 749–50
 - small, unilamellar, 749
 - vibrational energy levels
 - absorption of light, 592–95
 - villin
 - assembly of actin, 730
 - vimentin
 - aligning amino acid sequences, 360
 - vimentin filaments
 - distribution in cell, 507
 - intermediate filaments, 506
 - vinculin
 - assembly of actin, 730
 - frictional ratio, 577
 - 2-vinylpyridine
 - reagent for covalent modification, 537
 - viral hemagglutinin
 - embedded anchor, 774
 - purification, 773
 - rotational axes of symmetry, 787
 - viral protein coats
 - electron microscopy, 588
 - virial coefficients
 - light scattering, 416
 - lysozyme, 419
 - osmotic pressure, 409
 - sedimentation equilibrium, 412
 - viruses
 - immune complexes, 561
 - viscosity, 578–79
 - effect on kinetics of folding, 697, 703
 - laminar flow, 578
 - of membrane, 813
 - shear, 578
 - Simha factor, 579
 - tropomyosin, 589
 - water, 193, 194
- visible light
 - absorption of, 594
- vitellogenin
 - amino acid sequence, 106
 - lipoproteins, 804
- vitronectin
 - heterologous associations, 516
- void volume
 - definition, 4
- voltage-gated chloride channel
 - immunoabsorbent, 566
- volume
 - of hydrodynamic particle, 573
- volume change of folding
 - thermodynamics of folding, 673
- volume fraction
 - units of concentration, 196
- volume of a molecule of protein
 - packing of side chains, 278
- W**
 - water, 190–96
 - α helix, 256
 - boiling point, 190
 - coating molecule of protein, 573
 - compressibility, isothermal, 192
 - configurational heat capacity, 192
 - cubic expansion coefficient, 193
 - dielectric relaxation, 195
 - dimers of water, 190
 - effect on hydrogen bond, 216–20
 - enthalpy of fusion, 190
 - enthalpy of vaporization, 190
 - heat capacity, 191
 - hydrogen bond, 190
 - hydrogen-bonded nearest neighbors, 193
 - ice Ih, 190
 - infrared absorption, 195
 - interstitial molecules of water, 194
 - liquid water, 190
 - melting point, 190
 - molar volume, 192
 - nuclear magnetic resonance, 632
 - nucleic acid, association of proteins with, 317
 - radial molecular correlation function, 193–94
 - refinement, 178
 - relative permittivity, 190
 - rotational correlation time, 195
 - scattering of X-radiation, 193
 - self-diffusion coefficient, 194

stretching frequency, 195
 surface tension, 190
 vapor, 190
 viscosity, 193–94
 water constant
 acids and bases, 66
 water in crystallographic molecular
 models, 190–96
 alcohol dehydrogenase, 294
 arginines, 296
 asparagine, 296
 aspartic acid, 296
 Bence-Jones protein, 294
 buried in the interior, 293
 chitinase B, 294
 chloramphenicol *O*-
 acetyltransferase, 294
 cholesterol oxidase, 295
 conservation, 293
 cytochrome *b*₅₆₂, 293
 cytochrome *f*, 293, 294
 deoxyribonuclease, 292
 dihydrofolate reductase, 299
 disordered side chains, 296
 disordered water, 290
 fatty-acid-binding protein, 294
 ferredoxin–NADP⁺ reductase, 293
 glutamic acid, 296
 glutamine, 296
 hydrogen bonds, 295
 hydrogen bonds in
 crystallographic molecular
 models, 308
 immune complex, 560
 interleukin 1 β , 294
 in the interior of integral
 membrane-bound proteins,
 781
 β -lactamase, 293
 location for a molecule of water,
 292
 lysine, 296
 lysozyme, 296, 299
 α -lytic endopeptidase, 294
 networks of water, 295
 nitrate reductase, 293
 peaks of positive electron density,
 291

penicillopepsin, 293, 295
 phthalate-dioxygenase reductase,
 293
 ribonuclease U₂, 292
 ribonuclease T₁, 293
 shell of hydration, 296
 thioredoxin, 293
 threonine, 296
 triacylglycerol lipase, 294
 tyrosine, 296
 weighting schemes
 aligning amino acid sequences, 351
 wells of potential energy
 hydrogen bond, 208
 WU-BLAST2
 evaluation of, 368
 WW domains
 folding, 683
 kinetics of folding, 703

X

xanthine oxidase
 metalloproteins, 330
 sieving, 427
 X-ray diffraction
 bilayer of phospholipid, 750–51
 oriented helical polymeric
 proteins, 502
 tobacco mosaic virus, 502
 X-ray scattering, 581–83
 aspartate carbamoyltransferase,
 584
 catalase, 583
 chymotrypsinogen, 583
 cyclic AMP-dependent protein
 kinase, 582, 584
 distance distribution function, 581
 fibronectin, 583
 Guinier plot, 581
 hydration, 583
 immunoglobulin M, 584
 myoglobin, 583
 nitrite reductase, 583
 ovalbumin, 583
 small-angle, 581–83
 solution scattering, 582
 solution scattering curve,
 theoretical, 582

solution scattering curves,
 complete, 582
 spermadhesin, 583
 superoxide dismutase, 583
 troponin C, 583
 tubulin, 583
 β -xylanase
 tautomers, 73
 D-xylose
 structure, 129
 xylose isomerase
 ionic interactions in
 crystallographic molecular
 models, 302
m-xylylene diisocyanate
 cross-linking, 440

Y

yeast
 expression of DNA, 109
 yeast two-hybrid assay
 detection of heterologous
 associations, 518–19
 large screenings, 519
 yield of activity, 21

Z

zero net proton charge, point of,
 33
 zero-point energy
 hydrogen bond, 208
 Zimm plot
 light scattering, 416
 zinc
 metalloproteins, 331–32
 zinc finger
 metalloproteins, 326, 331
 nucleic acid, association of
 proteins with, 324–25
 zinc finger protein GLI1
 nucleic acid, association of
 proteins with, 324
 zinc-binding protein TroA
 metalloproteins, 331–32
 zonal chromatography
 definition, 4