PREDICTION OF NOVEL ENZYME & CLASSIFICATION FROM SEQUENCE DERIVED FEATURES AND PSSM MATRIX USING ANN

By

MUKUL GUPTA - 031523
VIPLAV SHANKER MISHRA - 031502

MAY-2007

Submitted in partial fulfillment of the Degree of Bachelor of Technology

DEPARTMENT OF BIOTECHNOLOGY & BIOINFORMATICS
JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY-WAKNAGHAT, SOLAN, H.P, INDIA

CERTIFICATE

This is to certify that the work entitled, **"Prediction of novel enzyme & classification from sequence derived features and PSSM matrix using ANN"** submitted by **Mukul Gupta (031523)** and **Viplav Shanker Mishra (031502)** in partial fulfillment for the award of degree of Bachelor of Technology in Bio-Informatics of Jaypee University of Information Technology has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Dr. Pradeep Kumar Naik
(Project Coordinator)
Senior Lecturer
Dept. of Bioinformatics and Biotechnology
Jaypee University of Information Technology
Waknaghat, Solan, Himachal Pradesh, India

ACKNOWLEDGMENT

Many people have contributed to this project in a variety of ways over the past few months. To the individuals who have helped us, we again express our appreciation. We also acknowledge the many helpful comments received from our teachers of the concerned departments. We are indebted to all those who provided reviews & suggestions for improving the results and the topics covered in our project, and we extend our apologies to anyone we may have failed to mention.

Viplav Shanker Mishra
(031502)

Mukul Gupta
(031523)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
pI	Iso-electric Point
PSSM	Position Specific Scoring Matrix
PSI-BLAST	Position specific iterative BLAST
EC number	Enzyme Commission number
ORF	Open Reading Frame
PEPSTAT	Protein Statistics
SNNS	Stuttgart Neural Network Simulator
MCC	Mathew Correlation constant
ROC	receiver operating characteristic
PPV	Positive Predictive Values
CNS	Central Nervous System

ABSTRACT

The prediction of novel enzymes and their classification into six major classes remains one of the most important unsolved problems in biological world. Many previous algorithms have been developed to solve this intricacy. But, none of them has been able to provide a reliable method. The previous methods used a much complex method for prediction i.e from its 3-dimensional structure. It proved out to be a time consuming and a costly method with lot of limitation. A large number of data are constantly being generated thanks to several genome-sequencing projects throughout the world. However, the gap between the growth rate of biological sequences and the capability to characterize experimentally the roles and functions associated with these new sequences is constantly increasing. This results in an accumulation of raw data that can lead to an increase in our biological knowledge only if computational characterization tools are developed. Enzymes are a subclass of protein that are specialized in catalytic activity. They are large and complex molecules, present in all living beings, and play an essential role in biochemical reactions. They control several vital functions, including many metabolic processes that convert nutrients into energy and into other products necessary to cell functioning. We focus here on the annotation of novel protein as enzymes/non-enzymes and if it is an enzyme its classification into six measure classes.

In this study we have developed an automated tool (EnzymePred) that attempt to algorithmically predict enzymes from the primary sequence only. A generic approach to this problem consists of transferring the annotation from sequences of known enzymes to uncharacterised proteins. These approaches seemed initially to hold quite a bit of promise, that sequence derived features and position specific scoring matrix directly dictate the protein function. Classes of newly found enzyme sequences are usually determined either by biochemical analysis of eukaryotic and prokaryotic genomes or by microarray chips. However, with the explosion of protein sequences entering into databanks, it is highly desirable to explore the feasibility of selectively classifying newly found enzyme sequences into their respective enzyme classes by means of an automated method. This is indeed important because knowing which family or subfamily an enzyme belongs to may help deduce its catalytic mechanism and specificity, giving clues to the relevant biological function. Sequence similarity metrics are a useful approach to provide functional

annotation, but its use is sometimes limited, prompting the development and use of machine learning methods (MLMs). MLMs also have a certain degree of flexibility regarding data inputs, allowing them to expand progressively to meet the requirements of rapidly accumulating mountain of data generated from genomics research. Hence, in this study we have developed a two layer artificial neural network; the first layer is for binary prediction of enzymes/non-enzymes from the protein sequence only using sequence derived features and PSSM matrix. If it is predicted as enzyme than in the second layer it is classify into its specific class out of six major classes of enzymes. The tool is perfectly trained and validated and is predicting the result with more than 80% accuracy. It has been validated with five organisms also. Finally it is uploaded into the university web server; http://www.juit.ac.in/enzyme/tool.html for the public uses.

Chapter 1

Introduction

Enzymes are proteins that catalyze chemical reactions [Smith AD (Ed) *et. al.* (1997)].In enzymatic reactions, the molecules at the beginning of the process are called substrates, and the enzyme converts them into different molecules, the products. Enzymes are biological *catalysts* or assistants. Enzymes consist of various types of proteins that work to drive the chemical reaction required for a specific action or nutrient. Enzymes can either launch a reaction or speed it up.

Enzymes are mostly proteins, and range from just 62 amino acid residues in size for the monomer of 4-Oxalocrotonate tautomerase, to over 2,500 residues in the animal fatty acid synthase[Chen LH, Kenyon GL, Curtin F, Harayama S, Bembenek ME, Hajipour G, Whitman CP (1992)].The activities of enzymes are determined by their three-dimensional structure[Anfinsen C.B. (1973)]. Most enzymes are much larger than the substrates they act on, and only a very small portion of the enzyme (around 3–4 amino acids) is directly involved in catalysis[The Catalytic Site Atlas at The European Bioinformatics Institute]. The region that contains these catalytic residues, binds the substrate, and then carries out the reaction is known as the active site. Enzymes can also contain sites that bind cofactors, which are needed for catalysis. Some enzymes also have binding sites for small molecules, which are often direct or indirect products or substrates of the reaction catalyzed.

Like all proteins, enzymes are made as long, linear chains of amino acids that fold to produce a three-dimensional product. Each unique amino acid sequence produces a unique structure, which has unique properties. Individual protein chains may sometimes group together to form a protein complex. Most enzymes can be denatured—that is, unfolded and inactivated—by heating, which destroys the three-dimensional structure of the protein.

Generally, enzymes work on substrates in one of three ways: substrate orientation, physical stress, and changes in substrate reactivity. Substrate orientation occurs when an enzyme causes substrate molecules to align with each other and form a bond. When an enzyme uses physical stress on a substrate, it in effect grips the substrate and forces the molecule to break apart. An enzyme that causes changes in substrate reactivity alters the placement of the molecule's electrons, which influences the molecule's ability to bond with other molecules. Enzymes have active sites where they come into contact with particular substrates. The catalytic properties of enzymes are a cyclic process. Once a substrate has come into contact with the active site of an enzyme, it is modified by the enzyme to form the end product. Once the process is complete, the enzyme releases the product and is ready to begin the process with new substrates. Enzymes are never wasted and always recycled.

Function and structure

Enzymes are very efficient catalysts for biochemical reactions. They speed up reactions by providing an alternative reaction pathway of lower activation energy. Like all catalysts, enzymes take part in the reaction - that is how they provide an alternative reaction pathway. But they do not undergo permanent changes and so remain unchanged at the end of the reaction. They can only alter the rate of reaction, not the position of the equilibrium. Most chemical catalysts catalyse a wide range of reactions. They are not usually very selective. In contrast enzymes are usually highly selective, catalysing specific reactions only. This specificity is due to the shapes of the enzyme molecules. Many enzymes consist of a protein and a non-protein (called the **cofactor**) [http://www.chemsoc.org/networks/LearnNet/cfb/index.htm]. The proteins in enzymes are usually globular. The intra- and intermolecular bonds that hold proteins in their secondary and tertiary structures are disrupted by changes in temperature and pH. This affects shapes and so the catalytic activity of an enzyme is pH and temperature sensitive.

How enzymes work

For two molecules to react they must collide with one another. They must collide in the right direction (orientation) and with sufficient energy. Sufficient energy means that between them they have enough energy to overcome the energy barrier to reaction. This is called the **activation energy**. Enzymes have an **active site** [http://www.chemsoc.org/networks/LearnNet/cfb/index.htm]. This is part of the molecule that has just the right shape and functional groups to bind to one of the reacting molecules.

Lock and key hypothesis

This is the simplest model to represent how an enzyme works. The substrate simply fits into the active site to form a reaction intermediate [Fischer E. (1894). "Einfluss der Configuration auf die Wirkung der Enzyme". *Ber. Dt.* Chem. Ges. *27: 2985-2993.*].

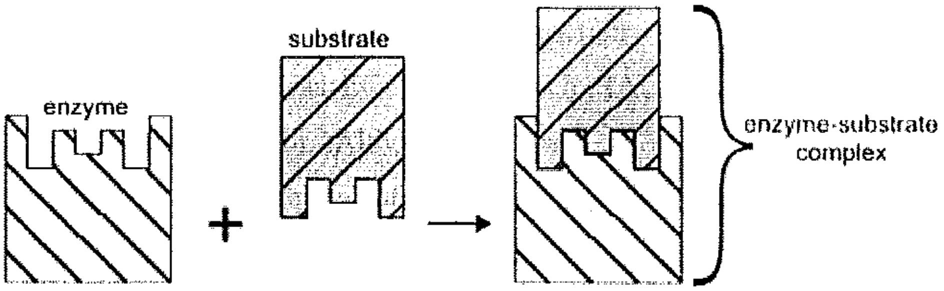


Fig 1.b. Lock and key hypothesis

Induced fit hypothesis

In this model the enzyme molecule changes shape as the substrate molecules gets close. The change in shape is 'induced' by the approaching substrate molecule [Koshland D. E. (1958)]. This more sophisticated model relies on the fact that molecules are flexible because single covalent bonds are free to rotate [Vasella A, Davies GJ, Bohm M. (2002)].

Different classes of enzymes

Basically all the enzymes available in the nature are classified into six major classes based on their mechanism of action. They are discussed below.

Class 1. Oxidoreductases.

To this class belong all enzymes catalysing oxidoreduction reactions. The substrate that is oxidized is regarded as hydrogen donor. The systematic name is based on donor:acceptor oxidoreductase. The common name will be dehydrogenase, wherever this is possible; as an alternative, reductase can be used. Oxidase is only used in cases where O2 is the acceptor. The second figure in the EC-Number of the oxidoreductases, it is 11, 13, 14 or 15, indicates the group in the hydrogen (or electron) donor that undergoes oxidation: 1 denotes a -CHOH- group, 2 a -CHO or -CO-COOH group or carbon monoxide, and so on, as listed in the key. The third figure, except in subclasses EC 1.11, EC 1.13, EC 1.14 and EC 1.15, indicates the type of acceptor involved: 1 denotes NAD(P)+, 2 a cytochrome, 3 molecular oxygen, 4 a disulfide, 5 a quinone or similar compound, 6 a nitrogenous group, 7 an iron-sulfur protein and 8 a flavin. In subclasses EC 1.13 and EC 1.14 a different classification scheme is used and subclasses are numbered from 11 onwards. It should be noted that in reactions with a nicotinamide coenzyme this is always regarded as acceptor, even if this direction of the reaction is not readily demonstrated. The only exception is the subclass EC 1.6, in which NAD(P)H is the donor; some other redox catalyst is the acceptor. Although not used as a criterion for classification, the two hydrogen atoms at carbon-4 of the dihydropyridine ring of nicotinamide nucleotides are not equivalent in that the hydrogen is transferred stereospecifically.

Class 2. Transferases.

Transferases are enzymes transferring a group, e.g. a methyl group or a glycosyl group, from one compound (generally regarded as donor) to another compound (generally regarded as acceptor). The systematic names are formed according to the scheme donor: acceptor group transferase. The common names are normally formed according to acceptor grouptransferase or donor grouptransferase. In many cases, the donor is a cofactor (coenzyme) charged with the group to be transferred. A special case is that of the transaminases. Some transferase reactions can be viewed in different ways. For example, the enzyme-catalysed reaction :X-Y + Z = X + Z-Y; may be regarded either as a transfer of the group Y from X to Z, or as a breaking of the X-Y bond by the introduction of Z. Where Z represents phosphate or arsenate, the process is often spoken of as 'phosphorolysis' or 'arsenolysis', respectively, and a number of enzyme names based on the pattern of phosphorylase have come into use. These names are not suitable for a systematic nomenclature, because there is no reason to single out these particular enzymes from the other transferases, and it is better to regard them simply as Y-transferases. In the above reaction, the group transferred is usually exchanged, at least formally, for hydrogen, so that the equation could more strictly be written as:

X-Y + Z-H = X-H + Z-Y. Another problem is posed in enzyme-catalysed transaminations, where the -NH2 group and -H are transferred to a compound containing a carbonyl group in exchange for the = O of that group, according to the general equation:

R1-CH(-NH2)-R2 + R3-CO-R4 R1-CO-R2 + R3-CH(-NH2)-R4.

The reaction can be considered formally as oxidative deamination of the donor (e.g. amino acid) linked with reductive amination of the acceptor (e.g. oxo acid), and the transaminating enzymes (pyridoxal-phosphate proteins) might be classified as oxidoreductases. However, the unique distinctive feature of the reaction is the transfer of the amino group (by a well-established mechanism involving covalent substrate-coenzyme intermediates), which justified allocation of these enzymes among the transferases as a special subclass (EC 2.6.1, transaminases). The second figure in the code number of transferases indicates the group transferred; a one-carbon group in EC 2.1, an aldehydic or ketonic group in EC 2.2, an acyl group in EC 2.3 and so on. The third figure gives further information on the group transferred; e.g. subclass EC 2.1 is subdivided into' methyltransferases (EC 2.1.1), hydroxymethyl- and formyltransferases (EC 2.1.2) and so on; only in subclass EC 2.7, does the third figure indicate the nature of the acceptor group.

Class 3. Hydrolases.

These enzymes catalyse the hydrolytic cleavage of C-O, C-N, C-C and some other bonds, including phosphoric anhydride bonds. Although the systematic name always includes hydrolase, the common name is, in many cases, formed by the name of the substrate with the suffix -ase. It is understood that the name of the substrate with this suffix means a hydrolytic enzyme. A number of hydrolases acting on ester, glycosyl, peptide, amide or other bonds are known to catalyse not only hydrolytic removal of a particular group from their substrates, but likewise the transfer of this group to suitable acceptor molecules. In principle, all hydrolytic enzymes might be classified as transferases, since hydrolysis itself can be regarded as transfer of a specific group to water as the acceptor. Yet, in most cases, the reaction with water as the acceptor was discovered earlier and is considered as the main physiological function of the enzyme. This is why such enzymes are classified as hydrolases rather than as transferases. Some hydrolases (especially some of the esterases and glycosidases) pose problems because they have a very wide specificity and it is not easy to decide if two preparations described by different authors (perhaps from different sources) have the same catalytic properties, or if they should be listed under separate entries. An example is vitamin A esterase (formerly EC 3.1.1.12, now believed to be identical with EC 3.1.1.1). To some extent the choice must be arbitrary; however, separate entries should be given only when the specificities are sufficiently different. Another problem is that proteinases have 'esterolytic' action; they usually hydrolyse ester bonds in appropriate substrates even more rapidly than natural peptide bonds. In this case, classification among the peptide hydrolases is based on historical priority and presumed physiological function. The second figure in the code number of the hydrolases indicates the nature of the bond hydrolysed; EC 3.1 are the esterases; EC 3.2 the glycosylases, and so on. The third figure normally specifies the nature of the substrate, e.g. in the esterases the carboxylic ester hydrolases (EC 3.1.1), thiolester hydrolases (EC 3.1.2), phosphoric monoester hydrolases (EC 3.1.3); in the glycosylases the O-glycosidases (EC 3.2.1), N-glycosylases (EC 3.2.2), etc. Exceptionally, in the case of the peptidyl-peptide hydrolases the third figure is based on the catalytic mechanism as shown by active centre studies or the effect of pH.

Class 4. Lyases.

Lyases are enzymes cleaving C-C, C-O, C-N, and other bonds by elimination, leaving double bonds or rings, or conversely adding groups to double bonds. The systematic name is formed according to the pattern substrate group-lyase. The hyphen is an important part of the name, and to avoid confusion should not be omitted, e.g. hydro-lyase not 'hydrolyase'. In the common names, expressions like decarboxylase, aldolase, dehydratase (in case of elimination of CO2, aldehyde, or water) are used. In cases where the reverse reaction is much more important, or the only one demonstrated, synthase (not synthetase) may be used in the name. Various subclasses of the lyases include pyridoxal-phosphate enzymes that catalyse the elimination of a b- or g-substituent from an a-amino acid followed by a replacement of this substituent by some other group. In the overall replacement reaction, no unsaturated end-product is formed; therefore, these enzymes might formally be classified as alkyl-transferases (EC 2.5.1...). However, there is ample evidence that the replacement is a two-step reaction involving the transient formation of enzyme-bound a,b(or b,g)-unsaturated amino acids. According to the rule that the first reaction is indicative for classification, these enzymes are correctly classified as lyases. Examples are tryptophan synthase (EC 4.2.1.20) and cystathionine b-synthase (EC 4.2.1.22). The second figure in the code number indicates the bond broken: EC 4.1 are carbon-carbon lyases, EC 4.2 carbon-oxygen lyases and so on. The third figure gives further information on the group eliminated (e.g. CO2 in EC 4.1.1, H2O in EC 4.2.1).

Class 5. Isomerases.

These enzymes catalyse geometric or structural changes within one molecule. According to the type of isomerism, they may be called racemases, epimerases, cis-trans-isomerases, isomerases, tautomerases, mutases or cycloisomerases. In some cases, the interconversion in the substrate is brought about by an intramolecular oxidoreduction (EC 5.3); since hydrogen donor and acceptor are the same molecule, and no oxidized product appears, they are not classified as oxidoreductases, even though they may contain firmly bound NAD(P)+. The subclasses are formed according to the type of isomerism, the sub-subclasses to the type of substrates.

Class 6. Ligases.

Ligases are enzymes catalysing the joining together of two molecules coupled with the hydrolysis of a diphosphate bond in ATP or a similar triphosphate. The systematic names are formed on the system X:Y ligase (ADP-forming). In earlier editions of the list the term synthetase has been used for the common names. Many authors have been confused by the use of the terms synthetase (used only for Group 6) and synthase (used throughout the list when it is desired to emphasis the synthetic nature of the reaction). Consequently NC-IUB decided in 1983 to abandon the use of synthetase for common names, and to replace them with names of the type X-Y ligase. In a few cases in Group 6, where the reaction is more complex or there is a common name for the product, a synthase name is used (e.g. EC 6.3.2.11 and EC 6.3.5.1). It is recommended that if the term synthetase is used by authors, it should continue to be restricted to the ligase group. The second figure in the code number indicates the bond formed: EC 6.1 for C-O bonds (enzymes acylating tRNA), EC 6.2 for C-S bonds (acyl-CoA derivatives), etc. Sub-subclasses are only in use in the C-N ligases.

Classification of Enzymes (EC Number)

Traditionally the enzymes are classified into six major classes based on their EC Number. The **Enzyme Commission number (EC number)** is a numerical classification scheme for enzymes, based on the chemical reactions they catalyze. As a system of **enzyme nomenclature**, every EC number is associated with a recommended name for the respective enzyme [ExPASy].Every enzyme code consists of the letters "EC" followed by four numbers separated by periods. Those numbers represent a progressively finer classification of the enzyme.

For example, the tripeptide aminopeptidases have the code "EC 3.4.11.4", whose components indicate the following groups of enzymes:

- *EC 3* enzymes are hydrolases (enzymes that use water to break up some other molecule)
- *EC 3.4* are hydrolases that act on peptide bonds
- *EC 3.4.11* are those hydrolases that cleave off the amino-terminal amino acid from a polypeptide
- *EC 3.4.11.4* are those that cleave off the amino-terminal end from a tripeptide

Top Level EC numbers [Moss, G.P 2006-03-14]

Class	Reaction catalyzed	Typical reaction	Enzyme example(s) with trivial name
EC 1 ***Oxidoreductases***	To catalyze oxidation/reduction reactions; transfer of H and O atoms or electrons from one substance to another	AH + B → A + BH **(reduced)** A + O → AO **(oxidized)**	Dehydrogenase, oxidase
EC 2 ***Transferases***	Transfer of a functional group from one substance to another. The group may be methyl-, acyl-, amino- or phosphate group	AB + C → A + BC	Transaminase, kinase
EC 3 ***Hydrolases***	Formation of two products from a substrate by hydrolysis	AB + H_2O → AOH + BH	Lipase, amylase, peptidase
EC 4 ***Lyases***	Non-hydrolytic addition or removal of groups from substrates. C-C, C-N, C-O or C-S bonds may be cleaved	RCOCOOH → RCOH + CO_2	
EC 5 ***Isomerases***	Intramolecule rearrangement, i.e. isomerization changes within a single molecule	AB → BA	Isomerase, mutase
EC 6 ***Ligases***	Join together two molecules by synthesis of new C-O, C-S, C-N	X + Y+ ATP → XY + ADP + Pi	Synthetase

Need of Prediction and classification of Enzymes

Enzymes are substances that occur naturally in all living things, including the human body. If it's an animal or a plant, it has enzymes. Enzymes are critical for life. At present, researchers have identified more than 3,000 different enzymes in the human body. Every second of our lives these enzymes are constantly changing and renewing, sometimes at an unbelievable rate. Our body's ability to function, to repair when injured, and to ward off disease is directly related to the strength and numbers of our enzymes. That's why an enzyme deficiency can be so devastating. All life processes consist of a complex series of chemical reactions.

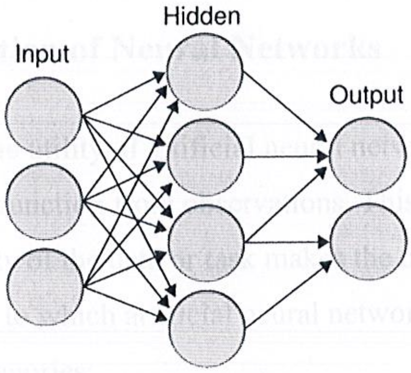
Using the protein engineering techniques, new enzymes are been created, ranging from food enzymes to the enzymes used for curing diseases. The large international genome sequence projects are gaining a great amount of public attention and huge sequence data bases are created it becomes more and more obvious that we are very limited in our ability to access functional data for the gene products - the proteins, in particular for enzymes. It seems quite improbable to experimentally determine function and structure of each candidate protein. So a revolutionary method is needed to solve this computation catastrophe. Primary sequence of these proteins are readily available, therefore a method using the sequence derived features will prove a much valuable and a cost effective process of determining and classifying these proteins into broader enzyme/non-enzyme and specifically into 6 major classes as defined by international enzyme commission.

Machine Learning in Classification

As a broad subfield of artificial intelligence, **machine learning** is concerned with the design and development of algorithms and techniques that allow computers to "learn". At a general level, there are two types of learning: inductive, and deductive. Inductive machine learning methods extract rules and patterns out of massive data sets. The major focus of Machine learning research is to extract information from data automatically by computational and statistical methods, hence, machine learning is closely related to data mining and statistics but also theoretical computer science [Christopher M. Bishop (2007)].

Machine learning has a wide spectrum of applications including natural language processing, syntactic pattern recognition, search engines, medical diagnosis, bioinformatics and cheminformatics, detecting credit card fraud, stock market analysis, classifying DNA sequences, speech and handwriting recognition, object recognition in computer vision, game playing and robot locomotion.

An **artificial neural network** (ANN) or commonly just **neural network** (NN) is an interconnected group of artificial neurons that uses a mathematical model or computational model for information processing based on a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network. In more practical terms neural networks are non-linear statistical data modeling tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data.



The original inspiration for the technique was from examination of the central nervous system and the neurons (and their axons, dendrites and synapses) which constitute one of its most significant information processing elements (see Neuroscience). In a neural network model, simple nodes (called variously "neurons", "neurodes", "PEs" ("processing elements") or "units") are connected together to form a network of nodes — hence the term "neural network." While a neural network does not have to be adaptive per se, its practical use comes with algorithms designed to alter the strength (weights) of the connections in the network to produce a desired signal flow. These networks are also similar to the biological neural networks in the sense that functions are performed collectively and in parallel by the units, rather than there being a clear delineation of subtasks to which various units are assigned (see also connectionism). Currently, the term Artificial Neural

Network (ANN) tends to refer mostly to neural network models employed in statistics, cognitive psychology and artificial intelligence. Neural network models designed with emulation of the central nervous system (CNS) in mind are a subject of theoretical neuroscience.

In modern software implementations of artificial neural networks the approach inspired by biology has more or less been abandoned for a more practical approach based on statistics and signal processing. In some of these systems neural networks, or parts of neural networks (such as artificial neurons) are used as components in larger systems that combine both adaptive and non-adaptive elements. While the more general approach of such adaptive systems is more suitable for real-world problem solving, it has far less to do with the traditional artificial intelligence connectionist models. What they do however have in common is the principle of non-linear, distributed, parallel and local processing and adaptation.

Application of Neural Networks

The utility of artificial neural network models lies in the fact that they can be used to infer a function from observations. This is particularly useful in applications where the complexity of the data or task makes the design of such a function by hand impractical. The tasks to which artificial neural networks are applied tend to fall within the following broad categories:

- Function approximation, or regression analysis, including time series prediction and modeling.
- Classification, including pattern and sequence recognition, novelty detection and sequential decision making.
- Data processing, including filtering, clustering, blind source separation and compression.

Application areas include system identification and control (vehicle control, process control), game-playing and decision making (backgammon, chess, racing), pattern recognition (radar systems, face identification, object recognition and more), sequence recognition (gesture, speech, handwritten text recognition), medical diagnosis, financial applications, data mining (or knowledge discovery in databases, "KDD"), visualization and e-mail spam filtering [*Neural Computing and Applications*, Springer-Verlag].

References

Smith AD (Ed) *et. al.* (1997) *Oxford Dictionary of Biochemistry and Molecular Biology* Oxford University Press ISBN 0-19-854768-4

Chen LH, Kenyon GL, Curtin F, Harayama S, Bembenek ME, Hajipour G, Whitman CP (1992). "4-Oxalocrotonate tautomerase, an enzyme composed of 62 amino acid residues per monomer". *J. Biol. Chem.* **267** (25): 17716-21. PMID 1339435.

Anfinsen C.B. (1973). "Principles that Govern the Folding of Protein Chains". *Science*: 223-230. PMID 4124164.

The Catalytic Site Atlas at The European Bioinformatics Institute Accessed 04 April 2007

Fischer E. (1894). "Einfluss der Configuration auf die Wirkung der Enzyme". *Ber. Dt.* Chem. Ges. *27: 2985-2993.*

Koshland D. E. (1958). "Application of a Theory of Enzyme Specificity to Protein Synthesis". *Proc. Natl. Acad. Sci.* **44** (2): 98-104. PMID 16590179.

Vasella A, Davies GJ, Bohm M. (2002). "Glycosidase mechanisms.". *Curr Opin Chem Biol.* **6** (5): 619-629. PMID 12413546.

http://www.chemsoc.org/networks/LearnNet/cfb/index.htm

ENZYME (Enzyme nomenclature database). ExPASy. Retrieved on 2006-03-14.

[Moss, G.P.. Recommendations of the Nomenclature Committee. International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes by the Reactions they Catalyse. Retrieved on 2006-03-14

Christopher M. Bishop (2007) *Pattern Recognition and Machine Learning*, Springer ISBN 0-387-31073-8.

Neural Computing and Applications, Springer-Verlag. (address: Sweetapple Ho, Catteshall Rd., Godalming, GU7 3DJ)

Objective:

With the explosion of protein sequences entering into databanks, it is highly desirable to explore the feasibility of selectively classifying newly found enzyme sequences into their respective enzyme classes by means of an automated method. This is indeed important because knowing which family or subfamily an enzyme belongs to, may help deduce its catalytic mechanism and specificity, giving clues to the relevant biological function. Sequence similarity metrics are a useful approach to provide functional annotation, but its use is sometimes limited, prompting the development and use of machine learning methods (MLMs). MLMs also have a certain degree of flexibility regarding data inputs, allowing them to expand progressively to meet the requirements of rapidly accumulating mountain of data generated from genomics research.

Hence, in this study an attempt has been taken to develop an automated tool using machine learning technique for annotation of protein sequence with following objectives.

1. To extract sequence derived features and selection of important features from protein sequence to be used for prediction and classification of enzymes.

2. To develop and optimized the artificial neural network for binary prediction of enzymes/non-enzymes using sequence derived features and PSSM matrix.

3. To develop a second layer of neural network for classifying the predicted enzymes into their corresponding classes; out of six major classes.

4. To validate the develop neural network model for predicting and classification of enzymes in some organisms.

5. To develop an automated tool for prediction and classification of enzymes from the protein sequence and to upload in the university web server for public uses.

Chapter 2

EnzymePred1: A tool for binary prediction of enzymes/non-enzymes from sequence derived features and PSSM matrix using ANN

(The developed tool is uploaded in the university web server and the tool is communicated for publication in *Journal of Computational Biology,* Wiley Publisher.)

ABSTRACT

The problem of predicting the enzymes and non-enzymes from the protein sequence information is still an open problem in bioinformatics. It is further becoming more important as the number of sequenced information grows exponentially over time. Sequence similarity metrics are a useful approach to provide functional annotation, but its use is sometimes limited, prompting the development and use of machine learning methods. We describe a novel approach for predicting the enzymes and non-enzymes from its amino-acid sequence using artificial neural network (ANN). The ANN used in this study is a feed-forward neural network with a standard back-propagation training' algorithm. Using 61 sequence derived features alone we have been able to achieve 79 % correct prediction of proteins into enzymes/non-enzymes (in the set of 660 proteins). For the complete set of 61 parameters using 5-fold cross-validated classification, ANN model reveal a superior model (accuracy = 78.79 ± 6.86 %, Q_{pred} = 74.734 ± 17.08 %, sensitivity = 84.48 ± 6.73 %, specificity = 77.13 ± 13.39 %). The second module of ANN is based on PSSM matrix. Using the same 5-fold cross-validation set, this ANN model predicts enzymes/non-enzymes with more accuracy (accuracy = 80.37 ± 6.59 %, Q_{pred} = 67.466 ± 12.41 %, sensitivity = 0.9070 ± 3.37 %, specificity = 74.66 ± 7.17 %). The elaborated ANN model based on the Artificial Neural Networks approach has been extensively validated and has confidently predicted enzymes/non-enzymes from the protein sequence to a number of annotated protein sequences from organisms.

Key words: Enzymes, non-enzymes, neural network, sequence derived features, PSSM..

1. INTRODUCTION

It is generally accepted that protein structure is determined by its amino acid sequence (Anfinsen et al., 1973) and that the knowledge of protein structures plays an important role in understanding their functions. To understand the rules relating amino acid sequence to three-dimensional protein structure is one of the major goals of contemporary molecular biology. A priori knowledge of protein as enzymes and non-enzymes has become quite useful from both an experimental and theoretical point of view.

One of the fundamental problems in bioinformatics is the prediction and classification of enzymes given only their primary sequence. Eisenberg et al. (2000) assessed some of the problems that researchers will face in the post-genome era. The number of proteins that are being made available to public and private databases is growing exponentially, and new methods must be found to understand and classify that information. The enormous task of function determination for every entry in GenBank has prompted the development of more sophisticated methods for protein automatic classification (Wu et al., 1995; DesJardins 1997; King et al., 2004; Pasquier et al., 2001; Cai et al., 2003). A computational method allowing for the automatic determination of protein function from its sequence alone is one of the prevailing problems in bioinformatics (Bork and Kroonin, 1998). Determination of three-dimensional structure is the traditional approach to functional classification of proteins that cannot be assigned a role based on homology to known proteins. This is a very time-consuming process, and the need for a faster method of classification is obvious (Baker et al., 2003). As structure determination is still another problem for itself, one of the most common tool for prediction of protein function has been usually to employ sequence alignment methods as PSI-BLAST (Altshul et al., 1997). Yet, Tian and Skolnick (2003) demonstrated that the E-value resulting from PSI-BLAST is weakly correlated to the enzyme function. Even for very low E-values (below e-100) the average conservation rate of the 4 digits of the EC classification is only 68 %. Furthermore these authors argue that to fully transfer the 4 digits of the EC number, to reach 90% accuracy, above 60 % sequence identity is required. This cautions the use of PSI-BLAST results for functional annotation. Other methods based on homology detection thorough motifs and conserved domains have been used with some success (Nagl, 2003), yet in many cases they have limited applicability. Pellegrini

(1999) provides a good overview of the most used techniques for determination of protein function.

Several machine-learning methodologies have provided good results without alignments. DesJardins et al. (1997) tested 3 machine-learning techniques (C4.5, instance based learning and discretized Naïve Bayes) to predict the first 2 levels of enzyme classification, using only sequence based data or information derived directly from it. Some of the results reached proved to be on par with sequence alignments. King et al. (2000) used inductive logic programming clustering and rule learning for functional classification of ORFs in *M. tuberculosis* and *E. coli* genomes. Dobson and Doig (2003)' addressed the topic of distinguishing enzyme structures from non-enzymes without using alignments. These authors used support vector machines tested over 36 protein features that included structural information and ligands. Using the same algorithm, yet focusing on enzyme classification, Cai et al. (2004) showed that it is possible to get good results even for distant enzymes and discriminating homologous enzymes of different functions. Several authors tested neural networks for the same end (Wu et al., 1995).

It is reported that structural classes of proteins correlate strongly with amino acid composition marked the onset of algorithm developments aimed at predicting the structural class of a protein from its amino acid composition alone (Nishikawa & Ooi, 1982; Nishikawa et al., 1983a; 1983b and Nakashima et al., 1986). There have been a number of algorithms proposed, such as the least Hamming distance (Chou et al., 1989), the least Euclidian distance (Nakashima et al., 1986), the discriminate analysis (Klein 1986), the vector decomposition.(Chou and Zhang, 1995), the component-coupled algorithm (Chou et al., 1998), and fuzzy structural vectors (Boberg et al., 1995). In general, due to the different datasets used in different studies, the evaluation for existing algorithms is still controversial. To improve structure prediction significantly, more information is required. In addition to amino acid composition, it might be expected that taking the sequence order along the primary structure of a protein into account would result in the improvement of predictive accuracy (Bu et al., 1999).

Hence, in this study we have develop two different neural networks which extract valuable information from protein sequence only for prediction into enzymes/non-enzymes. The first network used sequence derived features derived from PEPSTAT (EMBOSS suite) and the second network used PSSM profile obtained from PSI-BLAST, which would be useful for the systematic analysis of small or medium size protein sequences. Results are discussed, assessing the benefits of using this methodology in binary prediction of enzymes / non-enzymes. The preliminary results suggest that sequence derived feature can be used as a fast and effective classification methodology for proteins.

2. PREDICTION MODEL

2.1. Training data

To discriminate between the enzyme and non-enzymes, a data set of 660 proteins,' consisting of 330 non redundant enzymes and the same number of non redundant non-enzymes, were used for training and testing. The enzyme data set used in this study is obtained from the BRENDA database (http://www.brenda.uni-koeln.de) (Schomburg et al., 2004). This database is a comprehensive collection of enzyme and metabolic information, based on primary literature. It includes biochemical and molecular information on classification and nomenclature, they are more objective and reliable. The enzyme datasets used in this study consists of almost equal number of enzyme sequences for each of six major classes (56 class1, 56 class2, 56 class3, 56 class4, 56 class5 and 56 class6). The pairwise sequence identities in the datasets are less than 54 % for enzyme class and 45 % for non-enzyme class. Sequences of all the enzymes along with their name and function are included in the supplementary material.

2.2. Sequence derived parameters calculation and selection

To build a binary ANN model enabling effective prediction of enzymes/non-enzymes we initially calculated 61 parameters (Table 2.1) from the protein sequence alone using PEPSTAT (EMBOSS suite) ftp://emboss.open-bio.org/pub/EMBOSS (Rice et al., 2000) for all 660 protein sequences. The average values of these 61 parameters independently calculated for enzymes and non- enzymes have been plotted onto Figure 2.a. It showed clear distinction between enzymes and non-enzymes based on 61 parameters. The normalized values have been then used to generate ANN models for binary prediction.

2.3. Fivefold cross-validation

A prediction method is often developed by cross-validation or jack-knife method (Chou and Zhang, 1995). Because of the size of the dataset, the jack-knife method (individual testing of each enzyme in the data set) was not feasible. So a more limited cross-validation technique has been used, in which the dataset is randomly divided into five subsets, each containing equal number of enzyme sequences. Each set is a balanced set that consist of 50 percent of enzymes and 50 percent non-enzymes. The data set has been divided into training and testing set. The training set consists of five subsets. The network is validated for minimum error on testing set to calculate the performance measure for each fold of validation. This has been done five times to test for each subset. The final prediction results have been averaged over five testing sets.

2.4. ANN model for prediction of enzyme/non-enzyme using sequence derived features

Stuttgart Neural Network Simulator package (SNNS version 4.2) (Zell and Mamier, 1997) was used to implement the ANN model. In this study we have used the standard back-propagation ANN configuration consisting of 61 inputs and 1 output node in order to discriminate between enzymes/non-enzymes from the training sets (Figure 2b). For each sequence in the training and testing sets, we have transformed 61 network input parameters into the normalized values varying from 0 to 1. Similarly, the output parameters from the ANN were normalized to [0:1] range. The number of nodes in the hidden layer was varied from 0 to 6 in order to find the optimal network that allows most accurate separation of enzymes/non-enzymes in the training sets (Table 3.2). During the learning phase, a value of 1 was assigned for the enzyme sequence and 0 for non-enzyme. For each configuration of the ANN (with 0, 2, 4, and 6 hidden nodes respectively) 100 independent training runs were performed to evaluate the average predictive power of the network. The corresponding counts of the false/true positive and negative predictions were estimated using 0.1 and 0.9 cut-off values for non-enzymes and enzymes respectively. Thus, an enzyme from the testing set was considered correctly predicted by the ANN only when its output value ranged from 0.9 to 1.0. For each non-enzyme of the testing set the correct prediction was assumed if the corresponding ANN output lies between 0 and 0.1.

Thus, all network output values ranging from 0.2 to 0.9 have been ultimately considered as incorrect predictions (rather than undetermined or non-defined).

2.5. ANN model for prediction of enzyme/non-enzyme using PSSM matrix

In this module of the developed tool, with multiple alignment profile input, the position-specific scoring matrix generated by PSI-BLAST has been used as input to the neural network. The matrix has 20 x M real-number elements, where M is the length of the sliding window (M = 7). Each element represents the likelihood of that particular residue substitution at that position. Thus 20 real numbers rather than binary bits encode each residue. A standard back-propagation ANN configuration consisting of 140 inputs and 1 output node was used in order to discriminate between enzymes/non-enzymes from the training sets (Figure 2c). The number of nodes in the hidden layer was varied from 0 to 6 in order to find the optimal network that allows most accurate separation of enzymes/non-enzymes in the training sets (Table 2.2). The training and validation methods is similar as mentioned above. The corresponding counts of the false/true positive and negative predictions were estimated using 0.4 and 0.9 cut-off values for non-enzymes and enzymes respectively. Thus, an enzyme from the testing set was considered correctly predicted by the ANN only when its output value ranged from 0.9 to 1.0. For each non-enzyme of the testing set the correct prediction was assumed if the corresponding ANN output lies between 0.1 and 0.4.

2.6. Performance measures

The prediction results of ANN model developed in the study were evaluated using the following statistical measures.

1. *Accuracy of the methods*: The accuracy of prediction for neural network models were calculated as follows:

$$Q_{ACC} = \frac{P+N}{T}, \text{ where T = (P+N+O+U)}$$

Where *P* and *N* refer to correctly predicted enzymes and non-enzymes, and *O* and *U* refer to over and under predictions, respectively.

2. The Matthews correlation coefficient (MCC) is defined as:

$$MCC = \frac{(P \times N) - (O \times U)}{\sqrt{(P+U) \times (P+O) \times (N+U) \times (N+O)}}$$

3. Sensitivity (Q_{sens}) and specificity (Q_{spec}) of the prediction methods are defined as:

$$Q_{sens} = \frac{P}{P+U}$$

$$Q_{spec} = \frac{N}{N+O}$$

4. Q_{Pred} (Probability of correct prediction) and Q_{obs} (Percentage over coverage) are defined as:

$$Q_{pred} = \frac{P}{P+O} \times 100$$

$$Q_{obs} = \frac{P}{P+U} \times 100$$

The receiver operating characteristic (ROC) curve is also used to evaluate the prediction accuracy of our system using both sequence derived features and PSSM matrix.

3. RESULTS

3.1. Predictability of enzymes with sequence derived features

The ANN model (61-4-1) is trained with the sequence derived features (61 parameters) calculated using PEPSTAT. When applying a fivefold cross-validation test using five data sets, we found that the network reached an overall accuracy of 78.79 ± 6.86 %. The prediction results are presented in Table 2.3. The net has achieved an MCC of 0.596 ± 0.135. The other performance measures are: Qpred = 67.466 ± 17.084 %, sensitivity = 90.70 ± 6.73 % and specificity = 74.66 ± 13.39 %. The value of the learning parameter was set to 0.1. Training was performed for 100 epochs for both the networks, after which the learning has been terminated when the error reached a stable value; differences between errors in subsequent steps become sufficiently small. Table 2.4 reveals the predictability of enzymes and non-enzymes of the network. Out of 67 in each cross validation set 35-67 enzymes were correctly predicted as enzymes. However, out of 66 in non-enzyme class; 52-64 were correctly predicted as non-enzymes. Prediction performance measures were averaged over five sets. Figure 3d features averaged frequencies of the output values for the five testing sets used in the study. As it can readily be seen from the graph, the vast majority of the predictions have been contained within (0.0 – 0.1) for non-enzymes and (0.9 – 1.0) for enzymes in case of sequence derived module. This illustrates that 0.1 and 0.9 cut-offs values provide very adequate separation of two bioactive classes using ANN. All network output values ranging from 0.1 to 0.9 have been ultimately considered as incorrect predictions (rather than undetermined or non-defined).

3.2. Predictability of enzymes with multiple alignment

To further enhance the prediction performance, the multiple sequence alignment is implemented for prediction. The network 7(20)-4-1 is trained on PSI-BLAST generated position-specific matrices (PSSM). The comparative results of network with sequence derived features are shown in Table 2.3. It is clear from the results that the performance is improved slightly when PSI-BLAST-generated scoring matrices are used as input, compared with single sequence. The prediction accuracy is improved from 78.79 % to 80.37%. There is improvement in MCC from 0.5959 with sequence derived features to 0.6299 with PSI-BLAST. However, most dramatic improvement is achieved in other parameters like Qpred, sensitivity and specificity as evident from Table 2.3. The predicted value range was 0.2-0.4 for non-enzyme and 0.9-1.0 for enzymes. All network output values ranging from 0.4 to 0.9 have been ultimately considered as incorrect predictions (rather than undetermined or non-defined) (Figure 3e). It is evident that using PSSM profile the prediction accuracy of enzymes/non-enzymes is better than the previous module. Out of 67 in each cross validation set 35-67 enzymes were correctly predicted as enzymes. However, out of 66 in non-enzyme class; 52-64 were correctly predicted as non-enzymes.

3.3. Evaluation of prediction accuracy

From a practical point of view the most important aspect of a prediction method is its ability to make correct predictions. As prediction methods are never perfect, one always faces the dilemma of choosing between making few false-positive predictions and having a high sensitivity, that is, correctly identifying as many positive examples as possible. This tradeoff can be visualized as what is known as the receiver output characteristic (ROC) curve, in which the sensitivity is plotted as a function of the 1-specificity by varying the score threshold used for making positive predictions. Figure 3.f shows the ROC curves for the two predictors included in our method. Performance of both networks has been evaluated by calculating the area under the ROC curve. The area under the curve is 0.90 for ANN-PSSM; revealing a better discrimination of network system than that of network which uses sequence derived features (area under curve is 0.78).

3.4. An application of the model

The reliability of developed model of binary prediction of enzymes and non-enzymes was tested on the complete annotated protein sequences of three organisms downloaded form GenBank (Benson et al. 2002). The predicted result is shown in Table 2.5. Prediction on new sequences is done by first running the PEPSTAT (EMBOSS) program to obtain the sequence-derived features, which are subsequently used as input for the prediction of enzymes/non-enzymes. Likewise, in the second module the new sequences are BLAST iteratively using PSIBLAST to obtain the PSSM matrix, which are subsequently used for the prediction.

3.5. Availability

The program is implemented on the Web server EnzymePred, available at http://www.juit.ac.in/enzyme/tool.html by using CGI/Perl script. The SNNS-generated network is converted into C program and is used as an interface. Users can enter primary amino acid sequence in fasta or free format. The protein sequence can be predicted as enzyme or non-enzyme.

4. DISCUSSIÒN

The two different ANN models developed in this study are based on sequence derived features and PSSM matrix method. Enzyme/non-enzyme prediction accuracy has also been assessed and it has been found that preduiction of enzyme/non-enzyme using PSSM matrix is more accurate for the same cross-valdated sets used for both the models. This is because it uses improved searching tool for multiple sequence alignment such as PSI-BLAST. PSI-BLAST searches the homologs against a larger database such as a nonredundant database. Typically, more divergent profiles yield better predictions. The developed ANN based on PSSM matrix uses multiple alignement sequence information in the form of PSI-BLAST-generated scoring matrices to improve the accuracy of prediction of enzymes/non-enzymes. From this study, it is clear that a combination of neural network and evolutionary information contained in multiple sequence alignment has improved the performance of prediction method. There could be two possible reasons for this: (1) use of large and recent data set for learning and (2) use of PSI-BLAST profiles, which finds more distantly related homologs than pair-wise search methods against a nonredundant database.

The results demonstrate that the developed ANN-based binary prediction of enzymes/non-enzymes is adequate and can be considered an effective tool for 'in silico' screening. The results also demonstrate that the sequence derived parameters as well as PSSM matrix readily accessible from the protein sequences only, can produce a variety of useful information to be used '*in silico*'; clearly demonstrates an adequacy and good predictive power of the developed ANN model. There is strong evidence, that the introduced sequence features do adequately reflect the structural properties of proteins. The structure of a protein is an important determinant for the detailed molecular function of proteins, and would consequently also be useful for prediction of enzymes and non-enzymes. Based on the analysis of limited sequence features from protein sequences, differences in the parameters between enzymes and non-enzymes have previously been shown to exist and used for prediction of enzymes/non-enzymes in archaeal (Jensen et al., 2002). This agrees well with our result that sequence derived features can be used for predicting enzymes. This observation is not surprising considering that the calculated parameters should cover a very broad range of proprieties of bound atoms and molecules related to their size, polarizability, electronegativity, compactness, mutual inductive and steric influence and distribution of electronic density, etc. As it can be seen that the average value for both the classes were clearly separated on the graph and, hence, the selected 61 parameters should allow building an effective ANN model for binary prediction (Fig. 2.a).

Considering that one of the most important implications for the "binary prediction" model is its potential use for identification of novel enzymes from electronic databases, we have calculated the parameters of the Positive Predictive Values (PPV) for the networks while varying the number of hidden nodes. Taking into account the PPV values for the networks with the varying number of the hidden nodes along with the corresponding values of sensitivity, specificity and general accuracy we have selected neural network with four hidden nodes as the most efficient among the studied (Table 2.3). The ANN with 61 input-, 4 hidden- and 1 output nodes has allowed the recognition of 79 % of enzymes and 79 % of non-enzymes, on average (Table 2.4). For the second module (PSSM matrix) the ANN with 140 input-, 4 hidden- and 1 output nodes has allowed the recognition of 80 % of enzymes and 80 % of non-enzymes, on average (Table 4). The output from this 61-4-1 and 140-4-1 network has also demonstrated very good separation on positive (enzymes) and negative (non-enzymes) predictions. Further, the reliability of developed ANN model for prediction of enzyme and non-enzyme were tested on the complete annotated protein sequences of three organisms downloaded form GenBank (Benson et al., 2002). The predicted results was shown in Table 2.5.

Presumably, accuracy of the approach operating by the sequence derived features can be improved even further by expanding the parameters or by applying more powerful classification techniques such as Support Vector Machines or Bayesian Neural Networks. Use of merely statistical techniques in conjunction with the sequence parameters would also be beneficial, as they will allow interpreting individual parameter contributions into "enzymes/non-enzymes-likeness".

The results of the present work demonstrate that both the sequence derived features and PSSM matrix with ANN appear to be a very fast protein classification mechanism providing good results, comparable to some of the current efforts in the literature. The developed ANN-based model for enzymes/non-enzymes prediction can be used as a powerful tool for filtering through the collections of genome sequences to discover novel enzymes.

References

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.

Anfinsen, C.B. 1973 Principles that govern the folding of protein chains. *Science* 181, 223–230.

Baker, E.N., Arcus, V.L., and Lott, J.S. 2003. Protein structure prediction and analysis as a tool for functional genomics. *Applied Bioinformatics* 2, (3 Suppl), s3-s10.

Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., Rapp, B., and Wheeler, D. 2002. GenBank. *Nucleic Acids Res.* 30, 17–20.

Boberg, J., Salakoski, T. & Vihinen, M. 1995. Accurate prediction of protein secondaryst ructural class with fuzzyst ructural vectors. *Protein Eng.* 8, 505–512.

Bork, P., and Koonin, E.V. 1998. Predicting functions from protein sequences - where are the bottlenecks? *Nat Genet.* 18, 313–318.

Bu, W.S., Feng, Z.P., Zhang, Z.D. and Zhang, C.T. 1999. Prediction of protein (domain) structural classes based on amino-acid index. *Eur. J. Biochem.* 266, 1043–1049.

Cai, C.Z., Wang, W.L., Sun, L.Z., and Chen, Y.Z. 2003. Protein function classification via support vector machine approach. *Math Biosci.* 185(2),111-122.

Chou, P.Y. 1989. Prediction of protein structural classes from amino acid composition. In Prediction of Protein Structures and the Principles of Protein Conformation (Fasman, G.D., ed.), pp. 549–586. Plenum Press, New York.

Chou, K.C. and Zhang, C.T. 1995. Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* 30, 275–349.

Chou, K.C., Liu, W.M., Maggiora, G.M. and Zhang, C.T. 1998. Prediction and classification of domain structural classes. *Proteins: Struct. Funct. Genet.* 31, 97–103.

desJardins, M., Karp, P.D, Krummenacker, M., Lee, T.J., and Ouzonis, C.A.1997. Prediction of enzyme classification from protein sequence without the use of sequence similarity, *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, Halkidiki, Greece.

Dobson, P.D., and Doig, A.J. 2003. Distinguishing Enzyme Structures from Non-Enzymes Without Alignments. *J. Mol. Biol.* 330, 771-783.

Eisenberg, D., Marcotte, C.A., Xenarios, I., and Yeates, T.O. 2000. Protein function in the post-genomic era. *Nature* 405, 823 – 826.

Jensen, L.J., Skovgaard, M., and Brunak, S. 2002. Prediction of novel archaeal enzymes from sequence-derived features. *Protein Sci.* 11, 2894-2898.

King, R.D, Paul, H., and Clare, A. 2004. Confirmation of data mining based predictions of protein function. *Bioinformatics* 20, 1110-1118.

King, R.D., Karwath, A, Clare, A., and Dehaspe, L 2000. Acurate prediction of protein functional class from sequence in the *M. tuberculosis* and *E. coli* genomes using data mining. *Yeast-Comparative and Functional Genomics* 17(4), 283 – 293.

Klein, P. 1986. Prediction of protein structural class bydi scriminant analysis. *Biochem. Biophys. Acta*. 874, 205–215.

Nagl, S. 2003. Function prediction from protein sequence. In Orengo, C.A., Jones, D.T. Thornton, J.M. (eds). *Bioinformatics - Genes, proteins and computers*. BIOS Scientific publishers. Oxford. 298 pp.

Nishikawa, K. and Ooi, T. 1982. Correlation of amino acid composition of a protein to its tructural and biological characters. *J. Biochem.* 91, 1821–1824.

Nishikawa, K., Kubota, Y. and Ooi, T. 1983. Classification of proteins into groups based on amino acid composition and other characters. I. Angular distribution. *J. Biochem.* 94, 981–995.

Nishikawa, K., Kubota, Y. and Ooi, T. 1983. Classification of the proteins into groups based on amino acid composition and other characters. II. Grouping into four types. *J. Biochem.* 94, 997–1007.

Nakashima, H., Nishikawa, K. and Ooi, T. 1986. The folding type of a protein is relevant to the amino acid composition. *J. Biochem.* 99, 152–162.

Pasquier, C., Promponas, V., and Hamodrakas, S.J. 2001. PRED-CLASS: Cascading Neural networks for generalized protein classification and genome wide applications. *Proteins* 44, 361-369.

Pellegrini, M., Marcotte, E., Thompson, M., Eisenberg, D., and Yeates, T. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci.* 38, 667–677.

Rice, P., Longden, I., and Bleasby, A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* 16, (6) pp276—277.

Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., and Schomburg, D. 2004. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.* Jan 1, 32(Database issue):D43, 1-3.

Tian, W., and Skolnick, J. 2003. How well is enzyme function conserved as a function of pairwise sequence identity?. *J. Mol. Biol.* 333, 863-882.

Wu, C., Berry, M., Shivakumar, S., and McLarty, J. 1995. Neural Networks for Full-Scale Protein Sequence Classification: Sequence Encoding with singular Value Decomposition. *J Mach. Learn.* 21 N(1-2), 177-193.

Zell, A., and Mamier, G. 1997. Stuggart Neural Network Simulator version 4.2. Universty of Stuttgart, Stuttgart, Germany.

Table 2.1. 61 'Pepstat(EMBOSS)' primary sequence descriptors used in the Study.

Sequence derived parameters	Enzyme		Non-Enzyme		Sequence derived parameters	Enzyme		Non-Enzyme	
	Max	Min	Max	Min		Max	Min	Max	Min
Molecular Weight	0.207588	0.00182	0.20947	0.00419	N_Mole %	0.7186	0.1200	0.9091	0.2300
Average Residue	0.11811	0.09159	0.1209	0.09186	N_DayhoffStat	0.1671	0.0987	0.2114	0.1078
Isoelectric Point	0.104656	0.0427	0.1288	0.03857	P_Mole %	0.9572	0.3450	3.6556	0.5680
Extinction Coefficient	0.29032	0.019	0.33257	0.027	P_DayhoffStat	0.1841	0.0089	0.703	0.02908
Extinction Coefficient (1 mg/ml)	0.275	0.024	0.376	0.036	Q_Mole %	0.585	0.0871	1.5106	0.1098
Improablity / Proability inclusion bodies	0.928	0.494	0.979	0.41	Q_DayhoffStat	0.15	0.0098	0.3873	0.0129
A_Mole %	0.18828	0.02881	0.21186	0.03	R_Mole %	1.0682	0.0088	2.1256	0.0187
A_DayhoffStat	0.2189	0.0335	0.2464	0.045	R_DayhoffStat	0.218	0.02389	0.434	0.0452
B_Mole %	0.1989	0.0017	0.0902	0.0011	S_Mole %	0.9035	0.1796	2.2034	0.0012
B_DayhoffStat	0.0292	0.001	0.0109	0.0009	S_DayhoffStat	0.1291	0.0257	0.3148	0.0389
C_Mole %	1	0.00659	2.0339	0.0089	T_Mole %	1.0497	0.3091	1.4352	0.1203
C_DayhoffStat	0.3448	0.02154	0.7013	0.0154	T_DayhoffStat	0.1721	0.0507	0.2353	0.0092
D_Mole %	0.8147	0.0154	1.206	0.0015	V_Mole %	0.15	0.04484	0.17647	0.0289
D_DayhoffStat	0.1481	0.0152	0.2193	0.0652	V_DayhoffStat	0.2273	0.0679	0.2674	0.0546
E_Mole %	1.018	0.0147	1.8615	0.0254	W_Mole %	0.4598	0.00245	0.4839	0.0254
E_DayhoffStat	0.1697	0.0215	0.3102	0.0145					
F_Mole %	0.9195	0.1277	1.0044	0.0596	W_DayhoffStat	0.3537	0.0021	0.3722	0.0215
F_DayhoffStat	0.2554	0.0355	0.279	0.0101	X_Mole %	0.4562	0.025	0.3262	0.0254
G_Mole %	0.25	0.00769	0.36923	0.00503					
G_DayhoffStat	0.2976	0.0092	0.4396	0.006	X_DayhoffStat	0.5263	0.0562	0.3215	0.025
H_Mole %	0.6513	0.00894	1.0271	0.021	Y_Mole %	0.6135	0.0159	2.4615	0.0521
H_DayhoffStat	0.3257	0.0456	0.5136	0.0598					
I_Mole %	1	0.2077	1.0377	0.0089	Y_DayhoffStat	0.1804	0.0154	0.724	0.00987
I_DayhoffStat	0.2222	0.0462	0.2306	0.0564	Z_Mole %	0.2222	0.0089	0.3262	0.0154
K_Mole %	1.018	0.0591	2.0455	0.00115					
K_DayhoffStat	0.1542	0.00213	0.3099	0.0002	Z_DayhoffStat	0.894	0.1256	0.265	0.03652
L_Mole %	0.19444	0.03139	0.19101	0.0321	Tiny Mole %	0.6	0.15569	0.6389	0.16239
L_DayhoffStat	0.2628	0.0424	0.2581	0.0021	Small Mole %	0.75	0.4012	0.77119	0.32479
M_Mole %	0.5169	0.0456	1.2346	0.0268	Aliphatic Mole %	0.31481	0.14808	0.32903	0.02542
M_DayhoffStat	0.3041	0.0154	0.7262	0.0158	Aromatic Mole %	0.24521	0.04918	0.29231	0.08541
Charged Mole %	0.33533	0.05	0.46986	0.01389	Non-polar Mole %	0.85	0.45521	0.86154	0.31818
Basic Mole %	0.17365	0.05	0.31624	0.00926	Polar Mole %	0.54479	0.15	0.68182	0.13846
Acidic Mole %	0.16168	0.00897	0.25	0.0154					

* The parameters are scaled down by appropriate scaling values.

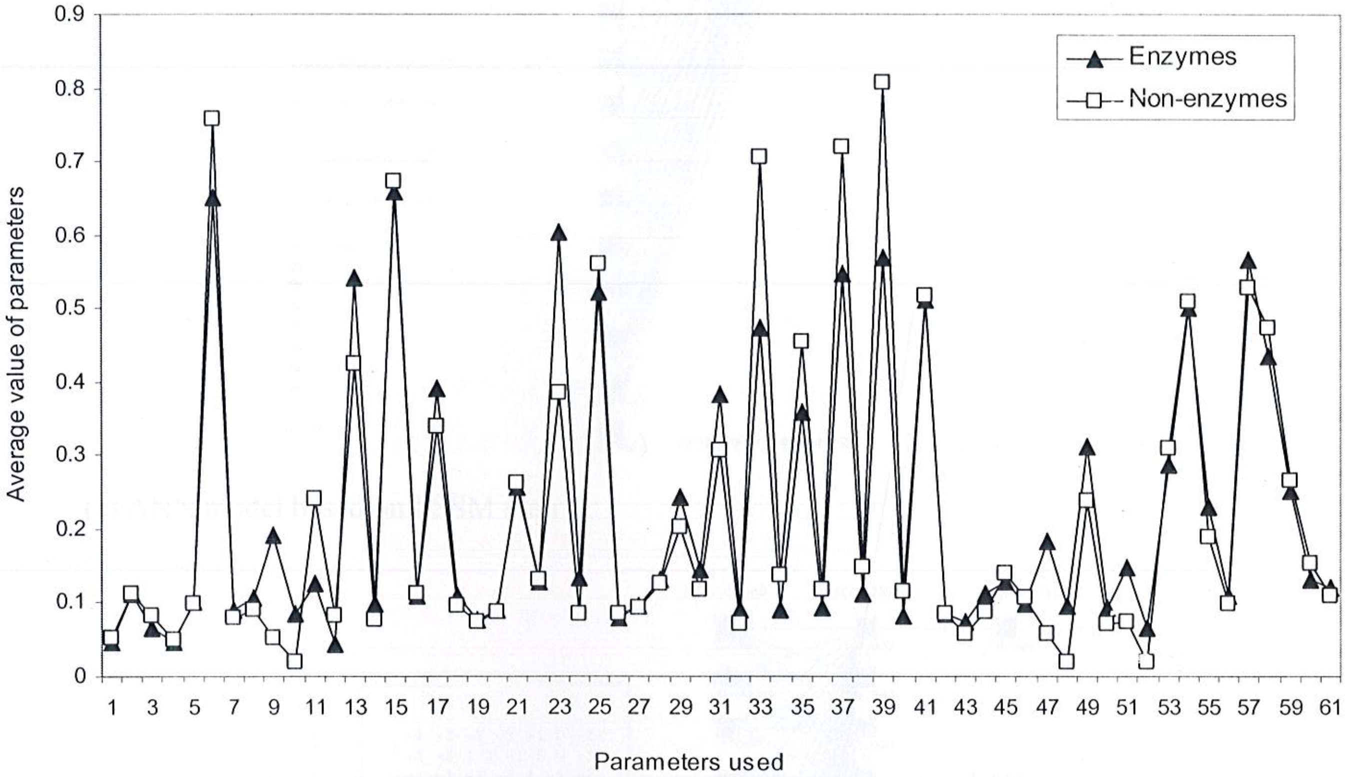
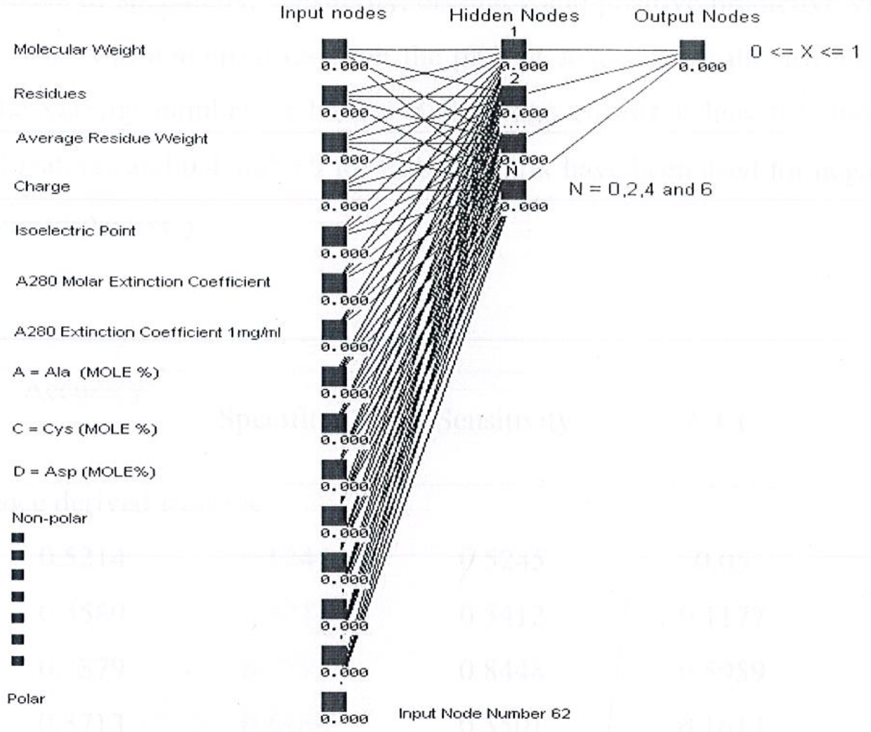


Figure 2.a. Averaged values of 61 sequence derived parameters calculated independently within studied sets of enzymes and non-enzymes.

(b) ANN model based on sequence derived features.



(c) ANN model based on PSSM matrix.

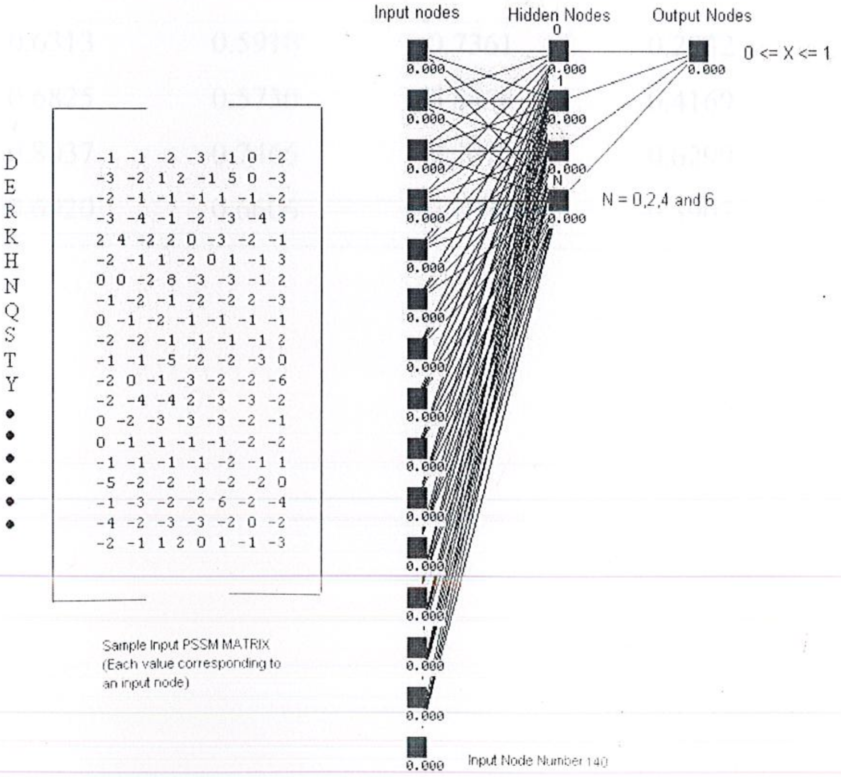


Figure 2 (b & c). Configuration of artificial neural network used to develop binary primary sequence descriptor model for enzyme / non-enzyme proteins.

Table 2.2. Parameters of specificity, sensitivity, accuracy and positive predictive values for prediction of enzymes and non-enzymes from the protein sequence by the artificial neural networks with the varying number of hidden nodes. The cut-off values 0.1 and 0.9 for sequence derived features and 0.4 and 0.9 for PSSM matrix have been used for negative and positive predictions respectively.

Hidden Nodes	Accuracy	Specificity	Sensitivity	MCC	Q(Pred)
(a) using sequence derived features					
0	0.5214	0.5244	0.5245	0.05	57.29
2	0.5589	0.6214	0.5412	0.1177	87.08
4	0.7879	0.7713	0.8448	0.5959	74.734
6	0.5713	0.6648	0.5501	0.1613	84.78
(b) using PSSM matrix					
0	0.6313	0.5910	0.7361	0.2932	40.95
2	0.6825	0.5730	0.8438	0.4169	57.30
4	0.8037	0.7466	0.9070	0.6299	67.466
6	0.6920	0.6606	0.7346	0.3907	61.46

Table 2.3. Results of enzymes / non-enzymes prediction methods, using five fold cross validation.

5-fold cross validation	Accuracy	Specificity	Sensitivity	MCC	Q(Pred)
(a) using sequence derived features					
C1	0.8947	1.00	0.8271	0.8072	100
C2	0.7969	0.7671	0.8333	0.5979	74.62
C3	0.7142	0.6794	0.7636	0.4364	76.68
C4	0.7443	0.6666	0.9495	0.5490	52.28
C5	0.7894	0.7934	0.8545	0.5891	70.14
Mean	0.7879 ± 0.0686	0.7713 ± 0.1339	0.8448 ± 0.0673	0.5959 ± 0.1345	74.734 ± 17.084
(b) using PSSM matrix					
C1	0.8230	0.7641	0.9158	0.6628	71.15
C2	0.8717	0.8148	0.9538	0.7560	78.13
C3	0.8521	0.8072	0.9123	0.7118	77.91
C4	0.7567	0.6988	0.8624	0.5368	61.09
C5	0.7153	0.6485	0.8911	0.4821	49.05
Mean	0.8037 ± 0.0659	0.7466 ± 0.0717	0.9070 ± 0.0337	0.6299 ± 0.1164	67.466 ± 12.411

Table 2.4. Output values from the neural network for the fivefold cross validation set's of enzymes/non-enzymes..

Testing 5 fold cross validation	Number of enzymes correctly predicted (out of 67)	Number of non-enzymes correctly predicted (out of 66)	Prediction range (enzymes)	Prediction range (non-enzymes)
(a) Using sequence features				
C1	67	52	0.9626-1.00	0.00-0.5340
C2	50	56	0.9579-1.00	0.00-0.6758
C3	42	53	0.9257-1.00	0.00-0.8786
C4	35	64	0.9692-1.00	0.00-0.8586
C5	47	58	0.9048-1.00	0.00-0.8236
(b) Using PSSM matrix as input				
C1	67	52	0.9237-0.9559	0.2180-0.2205
C2	50	56	0.9357-0.9443	0.3921-0.6006
C3	42	53	0.9061-0.9156	0.1626-0.7521
C4	35	64	0.9255-0.9272	0.3239-0.5133
C5	47	58	0.9123-0.9343	0.3005-0.4183

Table 2.5. Tha data set size and breakdown on organisms.

Organism	Annoted protein sequences	Assigned as enzymes (using sequence derived features)	Assigned as enzymes (using PSSM matrix)
Mycobacterium tuberculosis	4189	1659	1667
Mycobacterium leapre	1605	723	710
Methanococcus jannaschii	1770	805	798
Arabidopsis thaliana (chromosome 1)	6606	2885	2837
Total	14170	6072	6012

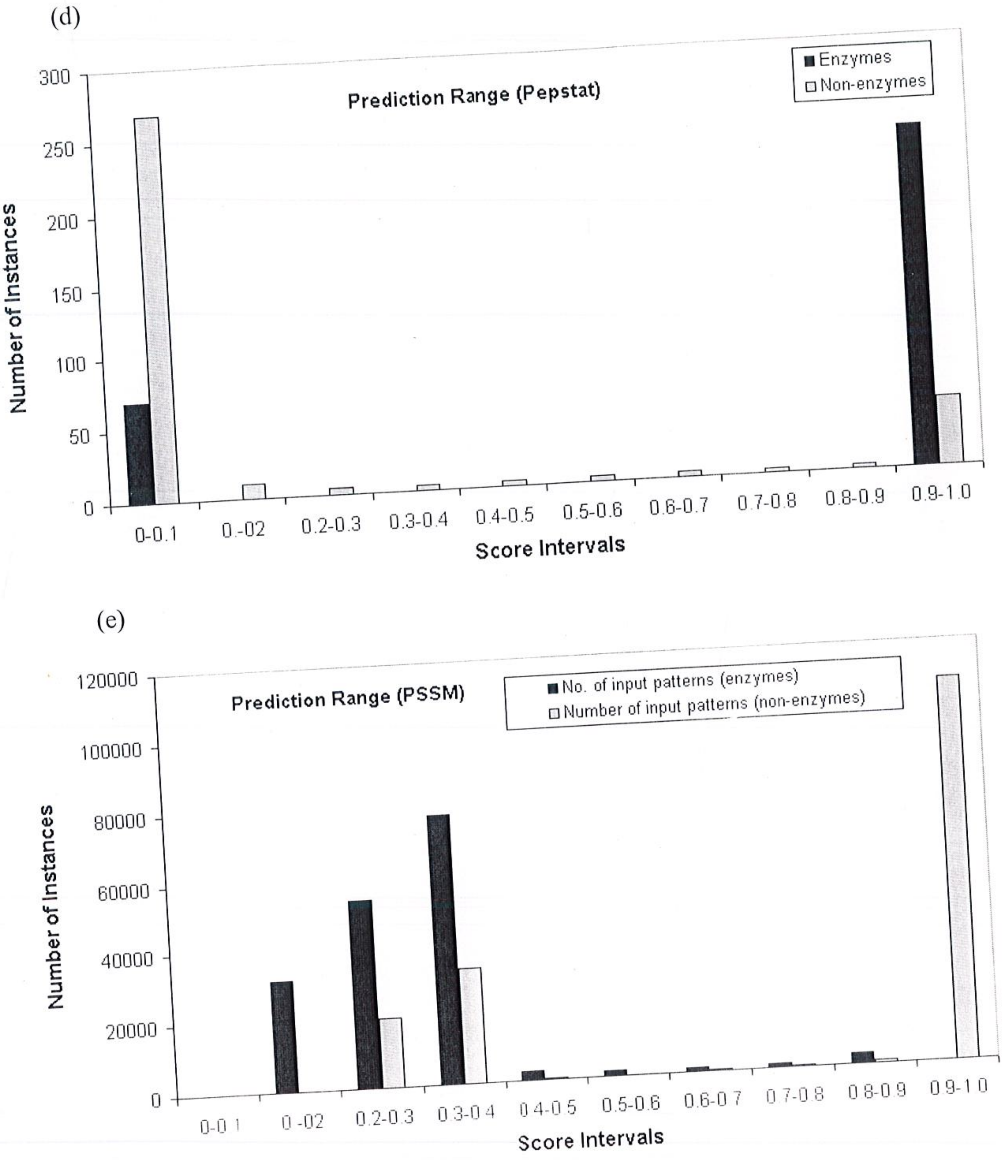


Figure 2 (d & e). Distribution of the output values from the ANN with four nodes in the hidden layer and trained on the set containing 90% of the studied protein sequences (a) using sequence derived features and (b) using PSSM matrix.

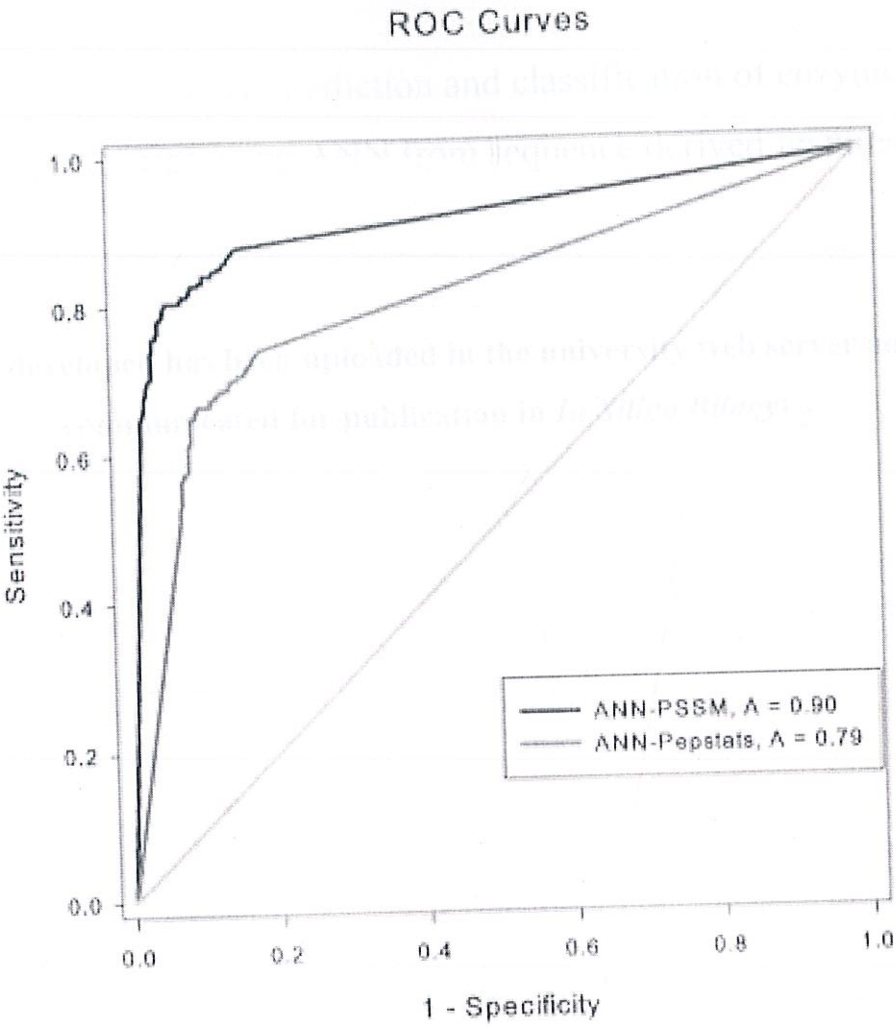


Figure 2.f. ROC curves for two different network systems.

Chapter 3

EnzymePred2: A Tool for prediction and classification of enzymes into six major classes using ANN from sequence derived features.

(The tool developed has been uploaded in the university web server and it is communicated for publication in *In Silico Bilogy*)

Abstract

Classes of newly found enzyme sequences are usually determined either by biochemical analysis of eukaryotic and prokaryotic genomes or by microarray chips. These experimental methods are both time-consuming and costly. With the explosion of protein sequences entering into databanks, it is highly desirable to explore the feasibility of selectively classifying newly found enzyme sequences into their respective enzyme classes by means of an automated method. This is indeed important because knowing which family or subfamily an enzyme belongs to may help deduce its catalytic mechanism and specificity, giving clues to the relevant biological function. In this study, we have developed a prediction method for detection and classification of enzymes from sequence alone (available at http://www.juit.ac.in/enzyme/tool2.html). The method does not make use of sequence similarity; rather, it relies on predicted protein features and simple physical/chemical properties. The tool has been validated in five different organisms and is proved to be very useful in prediction of novel enzymes with good accuracy.

Keywords: Function prediction; enzyme classification; sequence derived features.

Introduction

A large number of data are constantly being generated thanks to several genome-sequencing projects throughout the world. However, the gap between the growth rate of biological sequences and the capability to characterize experimentally the roles and functions associated with these new sequences is constantly increasing [1]. This results in an accumulation of raw data that can lead to an increase in our biological knowledge only if computational characterization tools are developed. Enzymes are a subclass of protein that are specialized in catalytic activity (Lehninger et al. 1998a). They are large and complex molecules, present in all living beings, and play an essential role in biochemical reactions. They control several vital functions, including many metabolic processes that convert nutrients into energy and into other products necessary to cell functioning. We focus here on the annotation of novel protein as enzymes/non-enzymes and if it is an enzyme its classification into six measure classes.

A generic approach to this problem consists of transferring the annotation from sequences of known enzymes to uncharacterized proteins [2]. The transfer mechanism might be subdivided in two steps: (i) to establish the list of known enzymes with significant sequence similarity to the uncharacterized sequence; (ii) to select the known sequence(s) from which the annotation is transferred [3]. The first step is usually performed with sequence alignment tools such as FASTA [4] or BLAST [5]. When sensitivity is critical, alternative tools such as PSI-BLAST [6] and hidden Markov models [7] can be used. Finding homologous proteins can also be accomplished using alignment independent sequence comparison tools, which have been developed to overcome the limitation arising from the assumption of contiguity between homologous segments [8,9]. Then, the challenge is the selection of true homologues from the list of similar sequences. Most of the above tools provide a score measuring the degree of similarity between the sequences compared. A simple criterion to single out a homologue is to choose the most similar sequence i.e. the highest scoring sequence. More elaborate methods have been designed to enhance the precision and reliability of the annotation process. These rely on the combination of the annotations of more than one homologue [10-13] .

However, annotating and assigning as enzymes/non-enzymes and their further classification into six major classes from their primary sequences requires highly automated computational methods linking experimental data. These methods must be able to discriminate the distinct catalytic function encapsulated in the protein's structure or in its primary sequences. To this end, the machine learning methods (MLMs) seem to be best suited for the task. Compared to the similarity-based methods such as BLAST or FASTA (Altschul *et al.*, 1990) and phylogeny-based method such as ClustalW, MLMs are widely applicable, and now frequently used in annotation of biological sequence analysis with relatively good accuracy. MLMs also have a certain degree of flexibility regarding data inputs, allowing them to expand progressively to meet the requirements of rapidly accumulating mountain of data generated from genomics research. The most often used methods of MLMs are support vector machine (SVM), neural network (NN), hidden markov model (HMM), decision tree (DT) and so on. Among these, NNs are particularly attractive due to its ability for pattern recognition (Raghava 2004), to handle large or small datasets, large input spaces (Narayanan et al. 2002), and its greater accuracy compared to simple BLAST or HMM methods (Bhasin and Raghava, 2004a; 2004b). Currently, there is no reliable systematic way for recognizing and classifying enzymes. Jensen et al. (2006) reported a method which classifies the enzymes into six major classes according to their sequence derived features, that is the co translational and posttranslational modifications, secondary structure, and simple physical/chemical properties. The limitation of this method is that it is confined only to archeal and they have developed six neural networks, each one for each of the six major classes of enzymes. Other methods make use of the structural features to classify enzymes into six major classes (Chou and Elrod 2003).

The natural encoding of the primary structure is a string of letters. However, this encoding is not appropriate for NNs, since it demands numerical (preferably, normalized) inputs. Therefore, proteins have to be encoded in a more suitable way. Proteins-including enzymes, in general, are composed by a variable number of amino acids, from tens to thousands. The encoding process proposed here allows differently sized enzymes to be processed by a predefined, fixed-size NN based on fixed number of sequence derived features from the protein sequence. The method does not make use of sequence similarity.

Strategically, we have develop a neural network, two-layer, fully automated computational method capable of recognizing enzymes first, and then classifying them into their subfamilies based on their protein sequences. A user-friendly program EnzymePred2 (http://www.juit.ac.in/enzyme/tool2.html) has been developed on the basis of this study to assist readers to distinguish enzymes and to annotate their subfamilies.

MATERIALS AND METHODS

Dataset for prediction of enzymes/non-enzymes

The sequence data on positive examples of enzymes used were obtained from the BRENDA database (Schomburg *et al.*, 2004). containing 360 protein sequences assigned to six classes according to their structural features. A non-redundant treatment was applied to eliminate the sequences which share a high degree of similarity (>90%) with others in order to avoid overtraining. The treatment was carried out using the program BLASTCLUST (http://www.ncbi.nlm.nih.gov/BLAST/), which used the BLAST algorithm to systematically cluster protein sequences on the basis of pair-wise matches. The default values were used for all BLAST parameters: matrix BLOSUM62, gap opening cost of 11, gap extension cost of 1, E-value threshold of 1e^{-6}. These sequences were used as positive examples for prediction as enzymes. The sequences data on negative examples were obtained from the SWISSPROT database (http://expasy.org/sprot/). Sequences related to enzymes were removed from the original dataset. A non-redundant treatment was applied (same as for positive datasets) such that no sequence had similarity higher than 25% to any others. Thus, 360 non-enzyme sequences were optimized as negative examples.

Dataset for classification of enzymes into six major classes

The above mentioned 360 sequences of enzymes were then grouped into six major classes Class 1 (Oxidoreductase) consist of 60 sequences, Class II (Transferase) consist of 60 sequences, Class III (Hydrolase) having 60 sequences, class IV (Lyase) with 60 sequences, class V (Transferase) consist of 60 sequences and class VI(Ligase) consist of 60 sequences. They were used for construction of neural networks training and validating the model for classification of novel enzymes into six classes.

Neural network architecture

The implementation of ANN was realized using the software package SNNS version 4.2 from Stuttgart University (Zell and Mamier 1997). We have used two feed-forward back-propagation neural networks with a single hidden layer. First layer of neural network is used for prediction of enzymes/non-enzymes from the protein sequence, whereas, the second layer is used for classifying the predicted enzyme into out of six major classes. The architecture of 1st neural network consisting of 61 inputs, 4 hidden nodes and 1 output node, whereas the 2nd neural network consisting of 61 inputs, 32 hidden nodes and six output nodes (each node is specified for each class of enzyme) (Figure 3.a). For each sequence in the training and testing sets, we have transformed 61 network input parameters into the normalized values varying from 0 to 1. Similarly, the output parameters from the ANN were in the range of 0 to 1. During the learning phase, a value of 1 was assigned for the enzyme sequence and 0 for non-enzyme. For configuration of the ANN, 100 independent training runs were performed to evaluate the average predictive power of the network. The corresponding counts of the false/true positive and negative predictions were estimated using 0.1 and 0.9 cut-off values for non-enzymes and enzymes respectively. Thus, an enzyme from the testing set was considered correctly predicted by the ANN only when its output value ranged from 0.9 to 1.0. For each non-enzyme of the testing set the correct prediction was assumed if the corresponding ANN output lies between 0 and 0.1. Thus, all network output values ranging from 0.2 to 0.9 have been ultimately considered as incorrect predictions (rather than undetermined or non-defined). For classifying the predicted enzymes into one

Sequence derived parameters calculation

To build a binary ANN model enabling effective prediction of enzymes/non-enzymes we initially calculated 61 parameters (Table 3.1 and 3.2) from the protein sequence alone using PEPSTAT (EMBOSS suite) ftp://emboss.open-bio.org/pub/EMBOSS (Rice et al., 2000) for all 660 protein sequences. The average values of these 61 parameters independently calculated for enzymes and non- enzymes have been plotted onto Figure 3.a. Similarly all the predicted 61 parameters for each class of enzyme independently is given in Table 3.1 and 3.2 & the average value of the parameters is given in Figure 3.a. It showed clear distinction between enzymes and non-enzymes based on 61 parameters. The normalized values have been then used to generate ANN models for binary prediction.

The input to second filtering network is the same input values used for the first layer and the predicted enzyme is classify into its particular class based on the maximum value obtained from the defined out put node for each class. For example to classify the predicted enzyme into class 1 the predicted output value is 1, 0, 0, 0, 0, 0 and so on.

Fivefold cross-validation

A prediction method is often developed by cross-validation or jack-knife method (Chou and Zhang, 1995). Because of the size of the dataset, the jack-knife method (individual testing of each enzyme in the data set) was not feasible. So a more limited cross-validation technique has been used, in which the dataset is randomly divided into five subsets, each containing equal number of enzyme sequences. Each set is a balanced set that consist of 50 percent of enzymes and 50 percent non-enzymes. The data set has been divided into training and testing set. The training set consists of five subsets. The network is validated for minimum error on testing set to calculate the performance measure for each fold of validation. This has been done five times to test for each subset. The final prediction results have been averaged over five testing sets.

Reliability index (RI)

RI is an assessment used to indicate the degree of confidence in the prediction to the user when using machine-learning techniques (Reinhardt and Hubbard, 1998; Emanuelsson *et al.*, 2000). RI is determined according to the difference (Δ) between the highest and the second highest value the SVMs gave in multi-class classification. The higher the RI is, the greater the probability that the prediction is accurate. In this study, RI is determined as follows:

$$RI = \begin{cases} 1 & \text{if } 0 \leq \Delta < 0.5 \\ INT(\Delta) \times 2 - 1 & \text{if } 0.5 \leq \Delta < 2.5 \\ 5 & \text{if } \Delta \geq 2.5 \end{cases}$$

RESULTS

Predictability of enzymes and enzyme classes

The ANN model develop in this study (61-4-1) is trained with the sequence derived features (61 parameters) calculated using PEPSTAT. When applying a fivefold cross-validation test using five data sets, we found that the network reached an overall accuracy of 78.79 ± 6.86 %. The prediction results are presented in Table 3.3. The net has achieved an MCC of 0.596 ± 0.135. The other performance measures are: Qpred = 67.466 ± 17.084 %, sensitivity = 90.70 ± 6.73 % and specificity = 74.66 ± 13.39 %. The value of the learning parameter was set to 0.1. Training was performed for 100 epochs for both the networks, after which the learning has been terminated when the error reached a stable value; differences between errors in subsequent steps become sufficiently small. Table 3.4 revealed the predictability of enzymes and non-enzymes of the network. Out of 67 enzymes in each cross validation set 35-67 enzymes were correctly predicted as enzymes. However, out of 66 in non-enzyme class; 52-64 were correctly predicted as non-enzymes. Prediction performance measures were averaged over five sets. Figure 3.a features averaged frequencies of the output values for the five testing sets used in the study. As it can readily be seen from the graph, the vast majority of the predictions has been contained within (0.0 – 0.1) for non-enzymes and (0.9 – 1.0) for enzymes in case of sequence derived module. This illustrates that 0.1 and 0.9 cut-offs values provide very adequate separation of two bioactive classes using ANN. All network output values ranging from 0.1 to 0.9 have been ultimately considered as incorrect predictions (rather than undetermined or non-defined).

An application of the model

The reliability of developed model of binary prediction of enzymes and non-enzymes and their classification into six major classes was tested on the complete annotated protein sequences of three organisms downloaded form GenBank (Benson et al. 2002). The predicted results are shown in Table 3.7. Prediction on new sequences is done by first running the PEPSTAT (EMBOSS) program to obtain the sequence-derived features, which are subsequently used as input for the prediction of enzymes/non-enzymes and the same values are again used for classification of the enzymes if it is predicted as enzyme from first layer.

Availability

The program is implemented on the Web server EnzymePred, available at http://www.juit.ac.in/enzyme/tool2.html by using CGI/Perl script. The SNNS-generated network is converted into C program and is used as an interface. Users can enter primary amino acid sequence in fasta or free format. The protein sequence can be predicted as enzyme or non-enzyme using first layer neural network and is further classify into its specific class using second network.

4. DISCUSSION

The tool EnzymePred2 has been developed in this study using two layered neural network based on sequence derived features. The results demonstrate that the developed ANN-based model for binary prediction of enzymes/non-enzymes and classification of enzymes into six major classes is adequate and can be considered an effective tool for 'in silico' screening. The results also demonstrate that the sequence derived parameters readily accessible from the protein sequences only, can produce a variety of useful information to be used '*in silico*'; clearly demonstrates an adequacy and good predictive power of the developed ANN model. There is strong evidence, that the introduced sequence features do adequately reflect the structural properties of proteins. The structure of a protein is an important determinant for the detailed molecular function of proteins, and would consequently also be useful for prediction of enzymes/non-enzymes and for their classification. Based on the analysis of limited sequence features from protein sequences, differences in the parameters between enzymes and non-enzymes have previously been shown to exist and used for prediction of enzymes/non-enzymes in archaeal (Jensen et al., 2002). This agrees well with our result that sequence derived features can be used for predicting enzymes. This observation is not surprising considering that the calculated parameters should cover a very broad range of proprieties of bound atoms and molecules related to their size, polarizability, electronegativity, compactness, mutual inductive and steric influence and distribution of electronic density, etc. As it can be seen that the average value for both the classes were clearly separated on the graph and, hence, the selected 61 parameters should allow building an effective ANN model for binary prediction (Fig.3.b). The average value of all the 61 sequence derived parameters used in the study for all the six classes of enzymes clearly separated and could be suitable for classification.

The ANN with 61 input-nodes, 4 hidden-nodes and 1 output nodes has allowed the recognition of 79 % of enzymes and 79 % of non-enzymes, on average (Table 3.3) and has also demonstrated very good separation on positive (enzymes) and negative (non-enzymes) predictions. The second layer of neural network with 61 input-,32 hidden and 6 output nodes able to correctly classify 67 % of the enzymes (Table 3.5) This result revealed a good prediction with accuracy of > 65 %. Further, the reliability of developed ANN model for prediction of enzyme and non-enzyme and their further classification were tested on the complete annotated protein sequences of three organisms downloaded form GenBank (Benson et al., 2002). The predicted results was shown in Table 3.7.

Presumably, accuracy of the approach operating by the sequence derived features can be improved even further by expanding the parameters or by applying more powerful classification techniques such as Support Vector Machines or Bayesian Neural Networks. Use of merely statistical techniques in conjunction with the sequence parameters would also be beneficial, as they will allow interpreting individual parameter contributions into "enzymes/non-enzymes-likeness".

The results of the present work demonstrate that the sequence derived features with ANN appear to be a very fast protein classification mechanism providing good results, comparable to some of the current efforts in the literature. The developed ANN-based model for enzymes/non-enzymes prediction and their classification into different classes can be used as a powerful tool for filtering through the collections of genome sequences to discover novel enzymes.

References

1. Janssen P, Audit B, Cases I, Darzentas N, Goldovsky L, Kunin V,Lopez-Bigas N, Peregrin-Alvarez JM, Pereira-Leal JB, Tsoka S,Ouzounis CA: Beyond 100 genomes. *Genome Biol* 2003, 4:402.

2. Andrade MA, Sander C: Bioinformatics: from genome data tobiological knowledge. *Curr Opin Biotechnol* 1997, 8:675-683.

3. Karp PD: What we do not know about sequence analysis and sequence databases. *Bioinformatics* 1998, 14:753-754.4. Pearson WR: Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol* 1990, 183:63-98.

4. Shah I, Hunter L: Predicting enzyme function from sequence: a systematic appraisal. *Proc Int Conf Intell Syst Mol Biol* 1997, 5:276-283.

5. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 1990, 215:403-410.

6. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997, 25:3389-3402.

7. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol* 1994, 235:1501-1531.

8. Vinga S, Almeida J: Alignment-free sequence comparison-a review. *Bioinformatics* 2003, 19:513-523.

9. Vries JK, Munshi R, Tobi D, Klein-Seetharaman J, Benos PV, Bahar I: A sequence alignment-independent method for protein classification. *Appl Bioinformatics* 2004, 3:137-148.

10. Abascal F, Valencia A: Automatic annotation of protein function based on family identification. *Proteins* 2003, 53:683-692.

11. Krebs WG, Bourne PE: Statistically rigorous automated protein annotation. *Bioinformatics* 2004, 20:1066-1073.

12. Leontovich AM, Brodsky LI, Drachev VA, Nikolaev VK: Adaptive algorithm of automated annotation. *Bioinformatics* 2002, 18:838-844.

13. Tatusov RL, Galperin MY, Natale DA, Koonin EV: The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 2000, 28:33-36.

14. Andrade MA, Brown NP, Leroy C, Hoersch S, de Daruvar A, Reich C, Franchini A, Tamames J, Valencia A, Ouzounis C, Sander C: Auto- mated genome sequence analysis and annotation. *Bioinformatics* 1999, 15:391-412.

15. Wilson CA, Kreychman J, Gerstein M: Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol* 2000, 297:233-249.

16. Kyrpides NC, Ouzounis CA: Whole-genome sequence annotation: 'Going wrong with confidence'. *Mol Microbiol* 1999, 32:886-887.

17. Bork P, Koonin EV: Predicting functions from protein sequences--where are the bottlenecks? *Nat Genet* 1998, 18:313-318.

18. Devos D, Valencia A: Intrinsic errors in genome annotation. *Trends Genet* 2001, 17:429-431.

19. Gerlt JA, Babbitt PC: Can sequence determine function? *Genome Biol* 2000, 1:REVIEWS0005.

20. Gilks WR, Audit B, De Angelis D, Tsoka S, Ouzounis CA: Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics* 2002, 18:1641-1649.

21. Cheng BY, Carbonell JG, Klein-Seetharaman J: Protein classification based on text document classification techniques. *Proteins* 2005, 58:955-970.

22. des Jardins M, Karp PD, Krummenacker M, Lee TJ, Ouzounis CA: Prediction of enzyme classification from protein sequence without the use of sequence similarity. *Proc Int Conf Intell Syst Mol Biol* 1997, 5:92-99.

23. Karchin R, Karplus K, Haussler D: Classifying G-protein coupled receptors with support vector machines. *Bioinformatics* 2002, 18:147-159.

24. Fillinger S, Boschi-Muller S, Azza S, Dervyn E, Branlant G, Aymerich S: Two glyceraldehyde-3-phosphate dehydrogenases with opposite physiological roles in a nonphotosynthetic bacterium. *J Biol Chem* 2000, 275:14031-14037.

25. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P: PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform* 2002, 3:265-274.

26. Wen Z, Morrison M: The NAD(P)H-dependent glutamate dehydrogenase activities of Prevotella ruminicola B(1)4 can be attributed to one enzyme (GdhA), and gdhA expression is regulated in response to the nitrogen source available for growth. *Appl Environ Microbiol* 1996, 62:3826-3833.

27. Itkor P, Tsukagoshi N, Udaka S: Nucleotide sequence of the rawstarch- digesting amylase gene from Bacillus sp. B1018 and its strong homology to the cyclodextrin glucanotransferase genes. *Biochem Biophys Res Commun* 1990, 166:630-636.

28. Shah I, Hunter L: Predicting enzyme function from sequence: a systematic appraisal. *Proc Int Conf Intell Syst Mol Biol* 1997, 5:276-283.

29. Devos D, Valencia A: Practical limits of function prediction. *Proteins* 2000, 41:98-107.

30. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel- Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: Gene ontology: tool for the unification of biology. *Nat Genet* 2000, 25:25-29.

31. Gattiker A, Michoud K, Rivoire C, Auchincloss AH, Coudert E, Lima T, Kersey P, Pagni M, Sigrist CJ, Lachaize C, Veuthey AL, Gasteiger E, Bairoch A: Automated annotation of microbial proteomes in SWISS-PROT. *Comput Biol Chem* 2003, 27:49-58.

32. Wieser D, Kretschmann E, Apweiler R: Filtering erroneous protein annotation. *Bioinformatics* 2004, 20 Suppl 1:I342-I347.

33. Murzin AG, Brenner SE, Hubbard T, Chothia C: SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995, 247:536-540.

34. Holm L, Sander C: Mapping the protein universe. *Science* 1996, 273:595-603.

35. Jaakkola T, Diekhans M, Haussler D: A discriminative framework for detecting remote protein homologies. *J Comput Biol* 2000, 7:95-114.

36. Bairoch A: The ENZYME database in 2000. *Nucleic Acids Res* 2000, 28:304-305.

37. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M: The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 2003, 31:365-370.

38. Promponas VJ, Enright AJ, Tsoka S, Kreil DP, Leroy C, Hamodrakas S, Sander C, Ouzounis CA: CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. *Bioinformatics* 2000, 16:915-922.

Table 3.1 61 'Pepstat(EMBOSS)' primary sequence descriptors used in the Study for class 1,2 and 3

Parameter	Class 1		Class 2		Class 3	
	Max	Min	Max	Min	Max	Min
Molecular Weight	0.207589	0.001823	0.176616	0.01511	0.207012	0.006066
Average Residue	0.118111	0.091159	0.11967	0.100633	0.116209	0.103872
Isoelectric Point	0.104656	0.0427	0.109987	0.045698	0.105404	0.040801
Extinction Coefficient	0.29032	0.0162	0.18136	0.00128	0.22801	0.00256
Extinction Coefficient (1 mg/ml)	0.275	0.0165	0.225	0.006	0.24	0.014
Improablity / Proability inclusion bodies	0.928	0.494	0.881	0.497	0.838	0.5
A_Mole %	0.18828	0.02881	0.24116	0.02712	0.18182	0.01887
A_DayhoffStat	0.2189	0.0335	0.2804	0.0315	0.2114	0.0219
B_Mole %	0.3125	0.0268	0.1642	0.0343	0.9195	0.0335
B_DayhoffStat	0.165	0.0095	1.1321	0.2642	0.2554	0.0268
C_Mole %	1	0.0921	0.3448	0.01254	0.3774	0.0095
C_DayhoffStat	0.3448	0.0215	0.1189	0.00145	0.1301	0.0921
D_Mole %	0.8147	0.0621	0.7353	0.197	0.894	0.1887
D_DayhoffStat	0.1481	0.0254	0.1337	0.0358	0.1626	0.0343
E_Mole %	1.018	0.01254	1.3095	0.3167	1.3208	0.2642
E_DayhoffStat	0.1697	0.00145	0.2183	0.0528	0.2201	0.044
F_Mole %	0.9195	0.1277	0.7251	0.098	0.7419	0.1887
F_DayhoffStat	0.2554	0.0355	0.2014	0.0272	0.2061	0.0524
G_Mole %	0.25	0.00769	0.11189	0.02387	0.11861	0.02273
G_DayhoffStat	0.2976	0.0092	0.1332	0.0284	0.1412	0.0271
H_Mole %	0.6513	0.012	0.4545	0.0489	0.5994	0.0505
H_DayhoffStat	0.3257	0.0654	0.2273	0.0169	0.2997	0.0253
I_Mole %	1	0.2077	1.215	0.1173	1.2453	0.16
I_DayhoffStat	0.2222	0.0462	0.27	0.0261	0.2767	0.0356
K_Mole %	1.018	0.0315	1.0894	0.1095	1.1321	0.04
K_DayhoffStat	0.1542	0.0654	0.1651	0.0166	0.1715	0.0061
L_Mole %	0.19444	0.03139	0.16742	0.03623	0.152	0.0566

L_DayhoffStat	0.2628	0.0424	0.2262	0.049	0.2054	0.0765
M_Mole %	0.5169	0.0154	0.4483	0.0457	0.5415	0.0658
M_DayhoffStat	0.3041	0.0805	0.2637	0.0269	0.3185	0.0387
N_Mole %	0.7186	0.1031	0.6107	0.045	0.6792	0.0649
N_DayhoffStat	0.1671	0.0606	0.142	0.495	0.158	0.0151
P_Mole %	0.9572	0.1401	0.8403	0.1639	1.0239	0.2632
P_DayhoffStat	0.1841	0.0326	0.1616	0.0315	0.1969	0.0506
Q_Mole %	0.585	0.1662	0.8173	0.0683	0.6213	0.1227
Q_DayhoffStat	0.15	0.00175	0.2096	0.1359	0.1593	0.0315
R_Mole %	1.0682	0.0235	1.5249	0.1934	1.0452	0.1707
R_DayhoffStat	0.218	0.0478	0.3112	0.0395	0.2133	0.0348
S_Mole %	0.9035	0.1796	1.25	0.2797	1.1024	0.1342
S_DayhoffStat	0.1291	0.0257	0.1786	0.04	0.1575	0.0192
T_Mole %	1.0497	0.3091	0.8511	0.1173	0.8136	0.2208
T_DayhoffStat	0.1721	0.0507	0.1395	0.0192	0.1334	0.0362
V_Mole %	0.15	0.04484	0.12941	0.03812	0.11321	0.03019
V_DayhoffStat	0.2273	0.0679	0.1961	0.0578	0.1715	0.0457
W_Mole %	0.4598	0.1276	0.3636	0.35088	0.4027	0.34121
W_DayhoffStat	0.3537	0.8673	0.2797	0.18644	0.3098	0.16848
X_Mole %	0.5123	0.1422	0.9572	0.09712	0.85	0.09021
X_DayhoffStat	0.2654	0.10526	0.1841	0.07627	0.54479	0.07065
Y_Mole %	0.6135	0.1595	0.5966	0.0322	0.7143	0.0613
Y_DayhoffStat	0.1804	0.2528	0.1755	0.0095	0.2101	0.018
Z_Mole %	0.1548	0.1945	0.54479	0.0654	0.1291	0.07107
Z_DayhoffStat	1.2306	0.0215	0.33533	0.2077	1.0497	0.51247
Tiny Mole %	0.6	0.15569	0.40836	0.17288	0.37014	0.20134
Small Mole %	0.75	0.4012	0.51896	0.3871	0.6019	0.39245
Aliphatic Mole %	0.31481	0.14808	0.33163	0.1653	0.27925	0.16327
Aromatic Mole %	0.24521	0.04918	0.17488	0.03939	0.17532	0.06089
Non-polar Mole %	0.85	0.45521	0.6881	0.46001	0.66	0.52101
Polar Mole %	0.54479	0.15	0.53999	0.3119	0.47899	0.34
Charged Mole %	0.33533	0.05	0.34409	0.16995	0.33962	0.20601
Basic Mole %	0.17365	0.05	0.20235	0.0836	0.18868	0.08233
Acidic Mole %	0.16168	0	0.17262	0.05665	0.16429	0.04906

Table 3.2 61 ‘Pepstat(EMBOSS)’ primary sequence descriptors used in the Study for class 4,5 and 6.

Parameter	Class 4		Class 5		Class 6	
	Max	Min	Max	Min	Max	Min
Molecular Weight	0.067056	0.004217	0.044612	0.038203	0.075422	0.036616
Average Residue	0.117138	0.104423	0.115822	0.104421	0.116345	0.10953
Isoelectric Point	0.085963	0.045253	0.101241	0.04488	0.100789	0.046294
Extinction Coefficient	0.08506	0.00384	0.06657	0.02048	0.10526	0.01664
Extinction Coefficient (1 mg/ml)	0.172	0.031	0.172	0.048	0.145	0.045
Improablity / Proability inclusion bodies	0.871	0.497	0.848	0.503	0.89	0.495
A_Mole %	0.16102	0.02778	0.1601	0.02151	0.1165	0.02358
A_DayhoffStat	0.1872	0.0323	0.1862	0.025	0.1355	0.0274
B_Mole %	0.1332	0.098	0.9195	0.01254	0.258	0.0548
B_DayhoffStat	0.4545	0.0272	0.2554	0.00145	0.168	0.05412
C_Mole %	0.4	0.02387	0.2989	0.0489	0.1655	0.012
C_DayhoffStat	0.1379	0.0284	0.1031	0.0169	0.0571	0.0659
D_Mole %	0.9032	0.1754	0.8184	0.2792	0.8134	0.3814
D_DayhoffStat	0.1642	0.0319	0.1488	0.0508	0.1479	0.0693
E_Mole %	1.1321	0.4159	0.9384	0.2989	1.1224	0.2133
E_DayhoffStat	0.1887	0.0693	0.1564	0.0498	0.1871	0.0355
F_Mole %	0.5556	0.01254	0.5914	0.1662	0.716	0.2477
F_DayhoffStat	0.1543	0.00145	0.1643	0.0462	0.1989	0.0688
G_Mole %	0.0995	0.05263	0.11811	0.04986	0.09832	0.04502
G_DayhoffStat	0.1185	0.0627	0.1406	0.0594	0.117	0.0536
H_Mole %	0.6349	0.1596	0.6094	0.1344	0.3883	0.1185
H_DayhoffStat	0.3175	0.0798	0.3047	0.0672	0.1942	0.0592
I_Mole %	1.1475	0.1724	1.3172	0.3073	1.2322	0.2913
I_DayhoffStat	0.255	0.0383	0.2927	0.0683	0.2738	0.0647
K_Mole %	0.8333	0.08	1.5323	0.1359	1.1575	0.1699
K_DayhoffStat	0.1263	0.0121	0.2322	0.0206	0.1754	0.0257
L_Mole %	0.13889	0.03448	0.13966	0.05959	0.15291	0.07474

L_DayhoffStat	0.1877	0.0466	0.1887	0.0805	0.2066	0.101
M_Mole %	0.5556	0.2477	0.3652	0.1031	0.5419	0.1193
M_DayhoffStat	0.3268	0.0688	0.2148	0.0606	0.3187	0.0702
N_Mole %	0.8621	0.2581	0.8871	0.1401	0.8531	0.2184
N_DayhoffStat	0.2005	0.06	0.2063	0.0326	0.1984	0.0508
P_Mole %	0.8621	0.16	0.6793	0.1662	0.6616	0.1481
P_DayhoffStat	0.1658	0.0308	0.1306	0.032	0.1272	0.0285
Q_Mole %	0.8743	0.01254	0.8549	0.0587	0.6383	0.2211
Q_DayhoffStat	0.2242	0.00145	0.2192	0.015	0.1637	0.0567
R_Mole %	0.7742	0.2135	0.8683	0.1344	0.9223	0.2545
R_DayhoffStat	0.158	0.0436	0.1772	0.0274	0.1882	0.0519
S_Mole %	0.9346	0.2477	0.8929	0.3571	0.963	0.3061
S_DayhoffStat	0.1335	0.0688	0.1276	0.051	0.1376	0.0437
T_Mole %	0.8743	0.1754	0.8673	0.2241	0.6715	0.2786
T_DayhoffStat	0.1433	0.0288	0.1422	0.0367	0.1101	0.0457
V_Mole %	0.152	0.04082	0.10526	0.04032	0.09108	0.00948
V_DayhoffStat	0.2303	0.0618	0.1595	0.0611	0.138	0.0144
W_Mole %	0.2551	0.0702	0.2528	0.0326	0.2188	0.38889
W_DayhoffStat	0.1962	0.2184	0.1945	0.1662	0.1683	0.17241
X_Mole %	0.51896	0.0508	0.8531	0.00175	0.2648	0.07627
X_DayhoffStat	0.33163	0.1481	0.1984	0.0235	0.6325	0.52459
Y_Mole %	1.0345	0.2542	0.544	0.2073	0.5687	0.35088
Y_DayhoffStat	0.3043	0.0748	0.16	0.061	0.1673	0.0357
Z_Mole %	0.51896	0.3061	0.51896	0.006	0.8954	0.0251
Z_DayhoffStat	0.33163	0.0437	0.33163	0.497	0.2478	0.0258
Tiny Mole %	0.36441	0.13889	0.3866	0.17473	0.32374	0.20379
Small Mole %	0.59658	0.38889	0.59021	0.36828	0.52143	0.39286
Aliphatic Mole %	0.30556	0.17241	0.27566	0.20716	0.26699	0.18281
Aromatic Mole %	0.16667	0.07627	0.1662	0.07107	0.14692	0.08978
Non-polar Mole %	0.64912	0.52459	0.65879	0.51247	0.56796	0.48104
Polar Mole %	0.47541	0.35088	0.48753	0.34121	0.51896	0.43204
Charged Mole %	0.27778	0.18644	0.32258	0.16848	0.33163	0.23223
Basic Mole %	0.14286	0.09712	0.18817	0.09021	0.17961	0.11893
Acidic Mole %	0.15094	0.07627	0.14691	0.07065	0.16099	0.08294

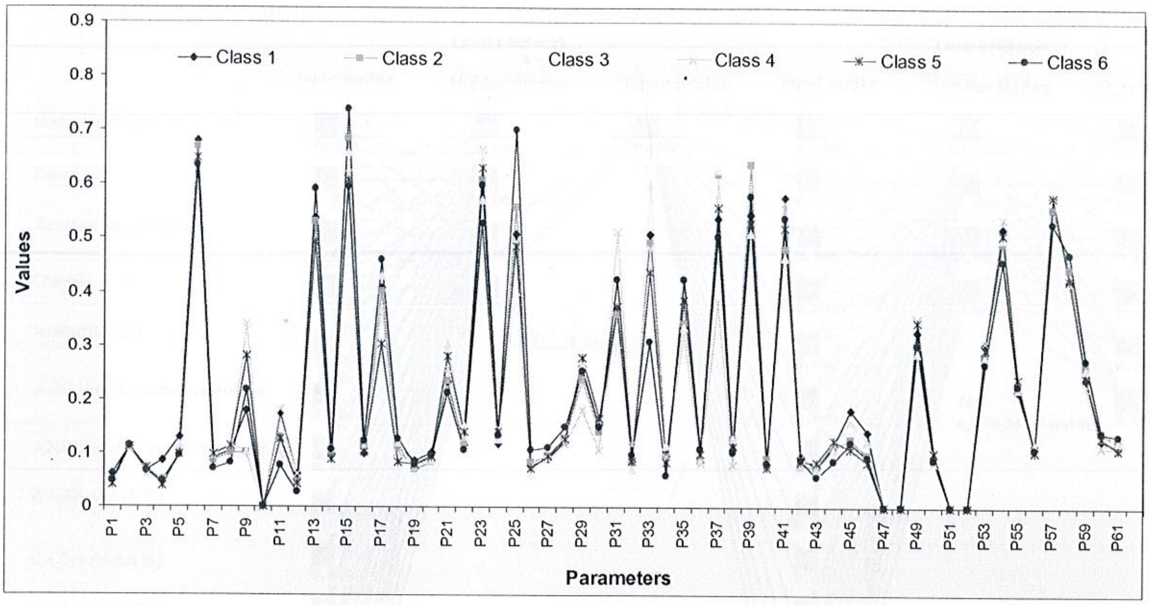


Figure 3.a. Averaged values of 61 sequence derived parameters calculated independently within studied sets of enzymes and non-enzymes.

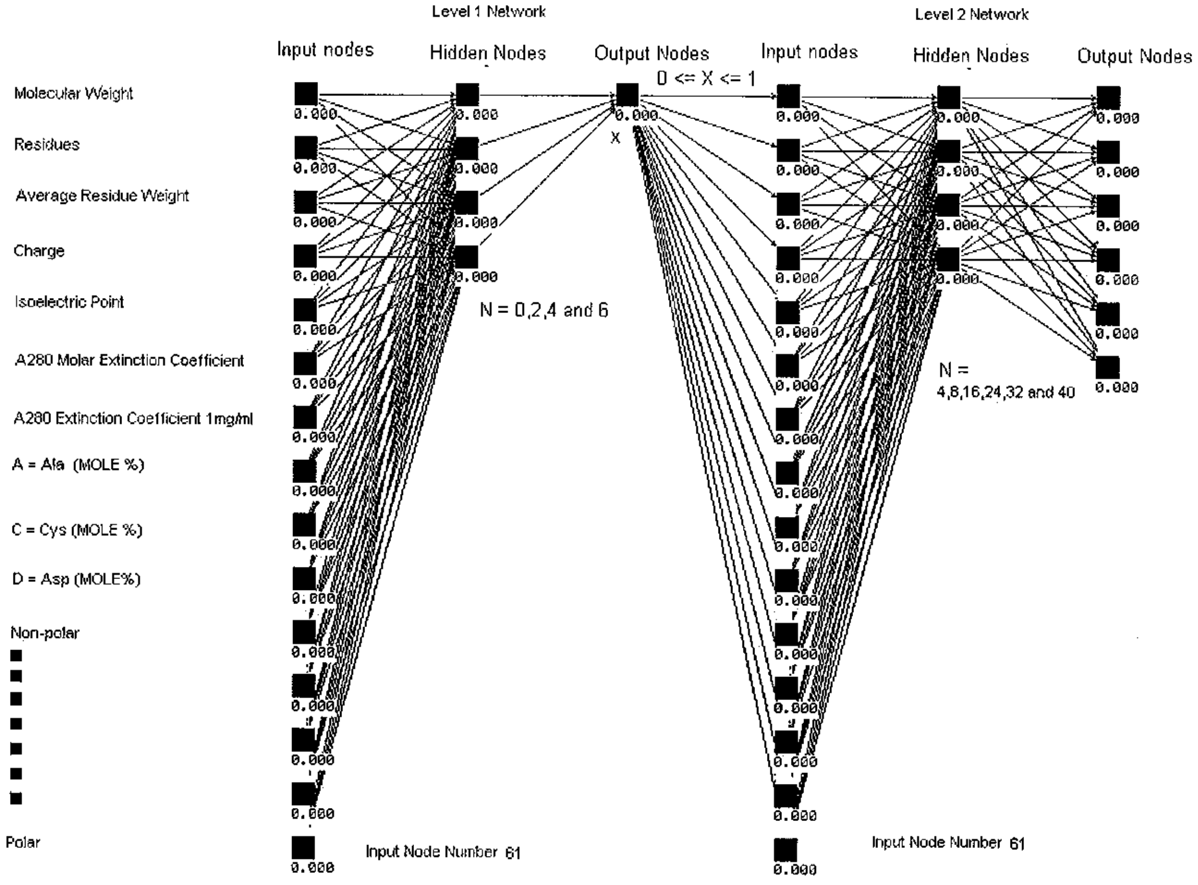


Figure 3.b. Configuration of artificial neural network used to develop binary primary sequence descriptor model for enzyme / non-enzyme proteins.

Table 3.3. Results of enzymes / non-enzymes prediction methods, using five fold cross validation.

5-fold cross validation	Accuracy	Specificity	Sensitivity	MCC	Q(Pred)
(a) using sequence derived features					
C1	0.8947	1.00	0.8271	0.8072	100
C2	0.7969	0.7671	0.8333	0.5979	74.62
C3	0.7142	0.6794	0.7636	0.4364	76.68
C4	0.7443	0.6666	0.9495	0.5490	52.28
C5	0.7894	0.7934	0.8545	0.5891	70.14
Mean	0.7879 ± 0.0686	0.7713 ± 0.1339	0.8448 ± 0.0673	0.5959 ± 0.1345	74.734 ± 17.0

Table 3.4. Values for prediction of enzymes and non-enzymes from the protein sequence by the artificial neural networks.

Testing 5 fold cross validation	Number of enzymes correctly predicted (out of 67)	Number of non-enzymes correctly predicted (out of 66)	Prediction range (enzymes)	Prediction range (non-enzymes)
(a) Using sequence features				
C1	67	52	0.9626-1.00	0.00-0.5340
C2	50	56	0.9579-1.00	0.00-0.6758
C3	42	53	0.9257-1.00	0.00-0.8786
C4	35	64	0.9692-1.00	0.00-0.8586
C5	47	58	0.9048-1.00	0.00-0.8236

Table 3.5. Prediction values from the neural network for the fivefold cross validation set's of enzyme classes.

5-Fold cross validation	Total cases taken	Correctly predicted enzymes
C1	72	48
C2	72	46
C3	72	53
C4	72	43
C5	72	44
Total	360	234

Table 3.6. Values for prediction of enzyme classes from the protein sequence by the artificial neural networks with the varying number of hidden nodes.

Number of hidden node	Number enzymes taken	Correctly predicted enzymes
4	360	134
8	360	148
16	360	189
24	360	212
32	360	234
40	360	219

Table 3.7. The results of enzymepred2 for the following organisms.

Organism	Annoted protein sequences	Assigned as enzymes (using sequence derived features)	Assigned as class 1 enzyme	Assigned as class 2 enzyme	Assigned as class 3 enzyme	Assigned as class 4 enzyme	Assigned as class 5 enzyme	Assigned as class 6 enzyme
Mycobacterium tuberculosis	4189	1659	444	186	492	155	326	56
Mycobacterium leapre	1605	723	184	91	185	109	130	24
Methanococcus jannaschii	1770	805	80	209	97	251	50	118
Arabidopsis thaliana (chromosome 1)	6606	2885	515	969	263	316	291	531
Total	14170	6072	1223	1455	1037	831	797	729

Chapter 4

Conclusion

From a practical point of view the most important aspect of a prediction method is its ability to make correct predictions. Till date most of the available methods use the 3-d structure of the protein to predict and classify enzymes. This is a very tedious job and requires much costlier endeavors. The sequence of a protein is an important determinant for the detailed molecular function of proteins, and would consequently also be useful for prediction of enzymes and enzyme classes. Additionally much encouraging results have been predicted using the PSSM technique. Therefore, a much accurate and reliable method is that which predicts the enzymes and enzyme classes based on both strategies.

This thesis contains detailed work on enzyme prediction and classification. We achieved an accuracy of ~ 79 % for the prediction of enzyme and non-enzyme based on non-redundant dataset of over 660 proteins. The neural network architecture used for the prediction was optimized for maximum accuracy. This was achieved by gradually testing networks with variable hidden nodes and retaining the one with highest true predictions. This is at par with best prediction tools available till date, but to the contrary, uses a much simpler and efficient prediction method based on sequence features and PSSM. This application not only gives optimum result with the dataset used, but also predicts enzymes from complex genomes to a very high satisfactory level. A much elaborate analysis has been done, which is evident from the extracted data, figures and tables compiled.

In addition to this tool, we further elaborated on classification of enzymes to their major classes. The first level of network imitates the binary model, and the second level uses the predicted result of the former to provide a much detailed and useful classification. We achieved an accuracy of ~ 67 % for classifying the protein sequences predicted as enzymes from the first network to their major classes. A similar methodology was used to optimize the network. The neural network architecture used for the classification was optimized for maximum accuracy. This was achieved by gradually testing networks with variable hidden nodes and retaining the one with highest true predictions.

LIST OF PUBLICATIONS

Research Papers Published

1. Dipankar SenGupta, Deeptak Verma, **Viplav S. Mishra** and Dr. P. K. Naik (2006). *Clustering of HIV-1 Subtype: Study of Molecular Diversity using Phylogenetic Analysis.* Bioinformatics Trend, Jan-Mar. 2006, Vol. I (1).

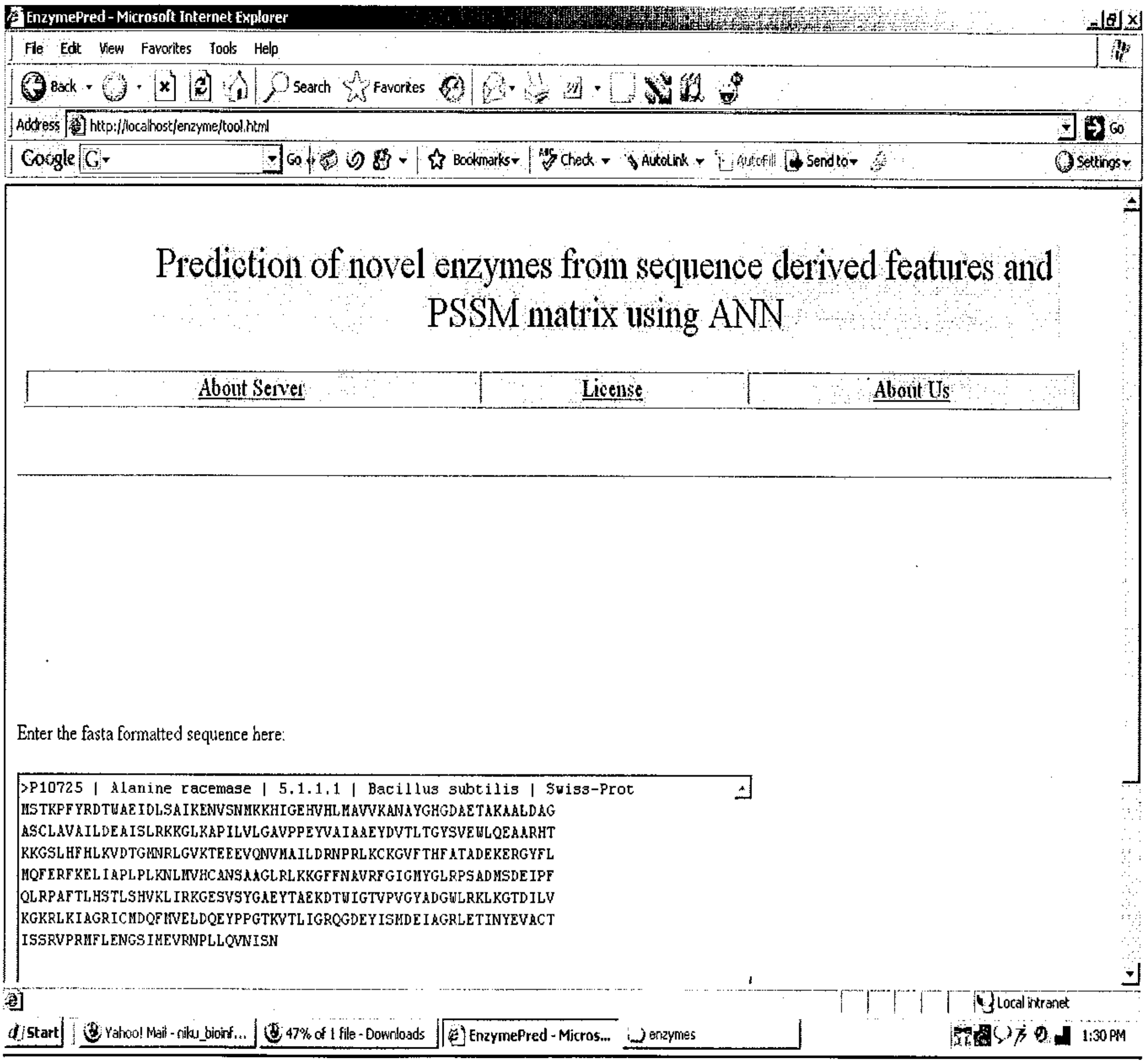
Research Papers Communicated

1. Viplav S. Mishra, Mukul Gupta, Kunal Jaiswal and Pradeep K. Naik (2007). Prediction and binary classification of Enzymes and Non-Enzymes from protein sequence using Artificial Neural Network. Communicated to *Journal of Computational Biology.*

2. Viplav S. Mishra, Mukul Gupta, Kunal Jaiswal and Pradeep K. Naik (2007). EnzymePred2: A Tool for prediction and classification of enzymes into six major classes using ANN from sequence derived features. *Communicated to In Silico Biology.*

APPENDIX

Snapshot of Enzymepred1 – tool for prediction of novel enzymes using sequence derived features and PSSM matrix.


EnzymePred - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://localhost/enzyme/tool.html
Go
Google G Go Bookmarks Check AutoLink AutoFill Send to Settings
Local intranet
Start
Yahoo! Mail - niku_bioinf...
47% of 1 file - Downloads
EnzymePred - Micros...
enzymes
1:30 PM

Prediction of novel enzymes from sequence derived features and PSSM matrix using ANN

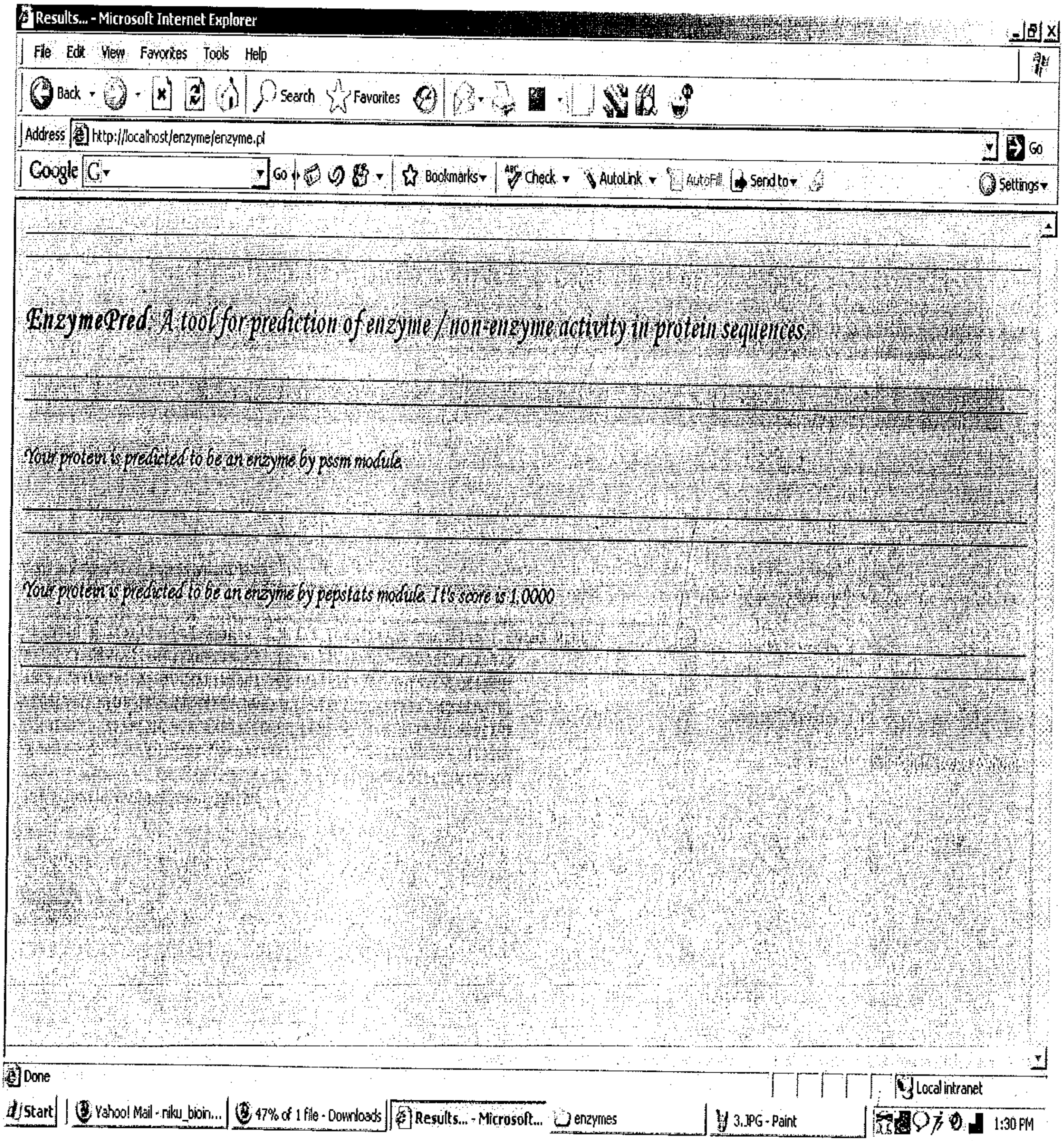
About Server | License | About Us

Enter the fasta formatted sequence here:

```
>P10725 | Alanine racemase | 5.1.1.1 | Bacillus subtilis | Swiss-Prot
MSTKPFYRDTWAEIDLSAIKENVSNMKKHIGEHVHLMAVVKANAYGHGDAETAKAALDAG
ASCLAVAILDEAISLRKKGLKAPILVLGAVPPEYVAIAAEYDVTLTGYSVEWLQEAARHT
KKGSLHFHLKVDTGMNRLGVKTEEEVQNVMAILDRNPRLKCKGVFTHFATADEKERGYFL
MQFERFKELIAPLPLKNLMVHCANSAAGLRLKKGFFNAVRFGIGMYGLRPSADMSDEIPF
QLRPAFTLHSTLSHVKLIRKGESVSYGAEYTAEKDTWIGTVPVGYADGWLRKLKGTDILV
KGKRLKIAGRICMDQFMVELDQEYPPGTKVTLIGRQGDEYISMDEIAGRLETINYEVACT
ISSRVPRMFLENGSIMEVRNPLLQVNISN
```

Snapshot of Output of

Enzymepred1

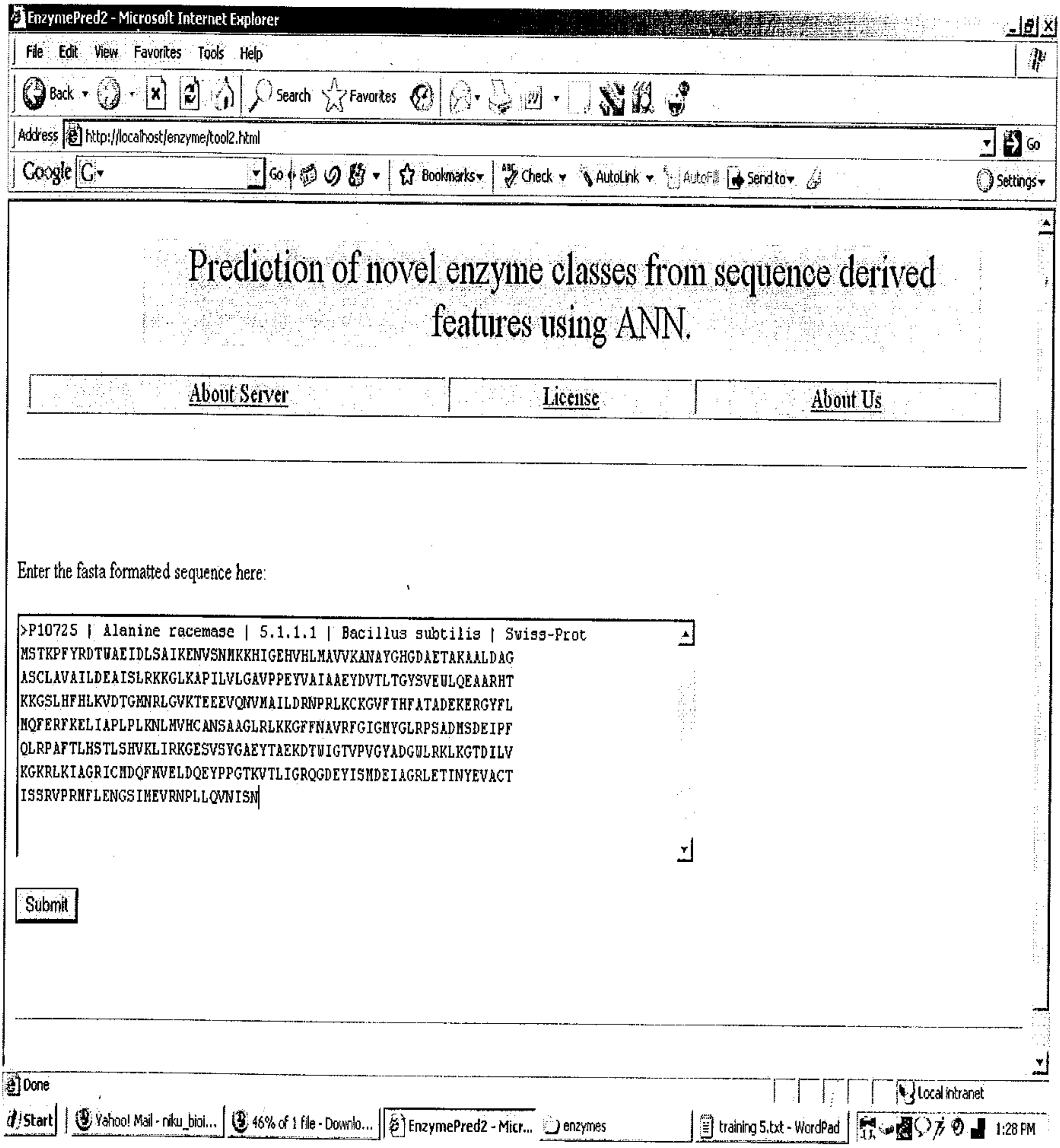

Results... - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://localhost/enzyme/enzyme.pl
Go
Google Go Bookmarks Check AutoLink AutoFill Send to Settings
Done
Local intranet
Start Yahoo! Mail - niku_bioin... 47% of 1 file - Downloads Results... - Microsoft... enzymes 3.JPG - Paint 1:30 PM

EnzymePred: A tool for prediction of enzyme / non-enzyme activity in protein sequences.

Your protein is predicted to be an enzyme by pssm module.

Your protein is predicted to be an enzyme by pepstats module. It's score is 1.0000

Snapshot of Enzymepred2 – A tool for prediction of novel enzymes and classification using sequence derived features.


EnzymePred2 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address http://localhost/enzyme/tool2.html
Go
Google
Bookmarks
Check
AutoLink
AutoFill
Send to
Settings
Done
Local intranet
Start
Yahoo! Mail - niku_bioi...
46% of 1 file - Downlo...
EnzymePred2 - Micr...
enzymes
training 5.txt - WordPad
1:28 PM

Prediction of novel enzyme classes from sequence derived features using ANN.

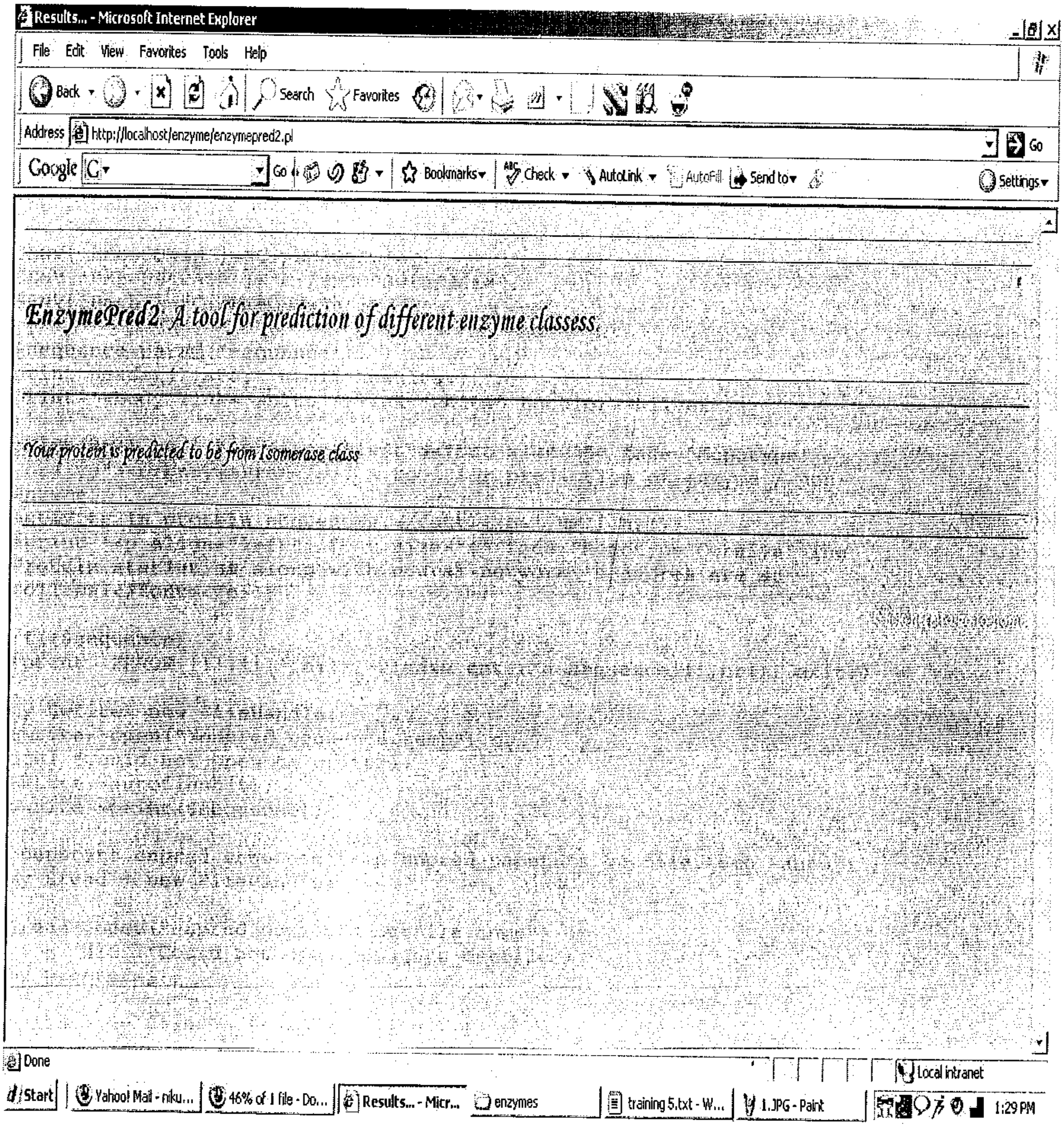
About Server	License	About Us

Enter the fasta formatted sequence here:

```
>P10725 | Alanine racemase | 5.1.1.1 | Bacillus subtilis | Swiss-Prot
MSTKPFYRDTWAEIDLSAIKENVSNMKKHIGEHVHLMAVVKANAYGHGDAETAKAALDAG
ASCLAVAILDEAISLRKKGLKAPILVLGAVPPEYVAIAAEYDVTLTGYSVEWLQEAARHT
KKGSLHFHLKVDTGMNRLGVKTEEEVQNVMAILDRNPRLKCKGVFTHFATADEKERGYFL
MQFERFKELIAPLPLKNLMVHCANSAAGLRLKKGFFNAVRFGIGMYGLRPSADMSDEIPF
QLRPAFTLHSTLSHVKLIRKGESVSYGAEYTAEKDTWIGTVPVGYADGWLRKLKGTDILV
KGKRLKIAGRICMDQFMVELDQEYPPGTKVTLIGRQGDEYISMDEIAGRLETINYEVACT
ISSRVPRMFLENGSIMEVRNPLLQVNISN
```

Submit

Snapshot of Output of Enzymepred2



Results... - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://localhost/enzyme/enzymepred2.pl
Go
Google G Go Bookmarks Check AutoLink AutoFill Send to Settings
Done
Local intranet
Start Yahoo! Mail - niku... 46% of 1 file - Do... Results... - Micr... enzymes training 5.txt - W... 1.JPG - Paint 1:29 PM

EnzymePred2: A tool for prediction of different enzyme classess.

Your protein is predicted to be from Isomerase class

Code in CGI PERL for the tool to classify the given enzyme into its major class from sequence derived features.

ENZYMEPRED1

```
#!c:/perl/bin/perl.exe
#!c:/perl/lib

use CGI qw(:standard);
use FileHandle;

$ENV{'EMBOSSWIN'}="C:/EMBOSSwin";
$ENV{'EMBOSS_DATA'}="C:/EMBOSSwin/data";
$ENV{'Path'}="C:/EMBOSSwin";

$sequence=param('sequence');
print header(),start_html("Results...");
print '<body link="white" vlink="white" bgcolor="silver">';

print hr(),hr(),'<p align="left"><font size="5" face="Monotype
Corsiva"><b>EnzymePred</b>: A tool for prediction of enzyme / non-
enzyme
activity in protein sequences.</font></p>',hr(),hr();
#print '<p align="left"><font size="4" face="Monotype Corsiva">The
Protein statistics along with neural network\'s scores are as
follows:</font></p>';

if(!$sequence)
{print 'ERROR!!!!!!!!',br(),'Please enter a sequence!!',br(); exit;}

my $write= new FileHandle;
$write->open(">enpred_sequence.temp") or
die( "Could not open to write");
$write->autoflush(1);
$write->print($sequence);

`pepstats enpred_sequence.temp enpred_pepstats_outfile.temp -auto 1`;
my $read = new FileHandle;

$read->open("enpred_pepstats_outfile.temp") or
      die ("Could not open pepstats_outfile");
my @vector=();

while ( my $line = $read->getline() )
{
my @array=();
@array=split(' ',$line);
chomp(@array);

if($line=~ /Molecular.weight/)
{push(@vector,$array[3]/1000000);}    #scaling the features and making
the classification vector

if($line=~ /Average/)
```

```
{push(@vector,$array[4]/1000);}

if($line=~ /Isoelectric/)
{push(@vector,$array[3]/100);}

if($line=~ /A280.Molar/)
{push(@vector,$array[5]/1000000);}

if($line=~ /A280.Extinction/)
{push(@vector,$array[5]/10);}

if(($line=~ /Improbability/) || ($line=~ /Probability/))
{push(@vector,$array[7]);}

if( ($line=~ /A.=.Ala/)|| ($line=~ /G.=.Gly/) || ($line=~ /L.=.Leu/) ||
($line=~ /V.=.Val/)  )
{push(@vector,$array[4]/100);
push(@vector,$array[5]/10);}

if( ($line=~ /B.=.Asx/) || ($line=~ /X.=.Xaa/) || ($line=~ /Z.=.Glx/) )
{push(@vector,$array[4]);
push(@vector,$array[5]);}

if( ($line=~ /C.=.Cys/) || ($line=~ /Y.=.Tyr/) || ($line=~ /D.=.Asp/)||
($line=~ /E.=.Glu/) || ($line=~ /F.=.Phe/) || ($line=~ /H.=.His/) ||
($line=~ /I.=.Ile/) || ($line=~ /K.=.Lys/) || ($line=~ /M.=.Met/) ||
($line=~ /N.=.Asn/) || ($line=~ /P.=.Pro/) || ($line=~ /Q.=.Gln/) ||
($line=~ /R.=.Arg/) || ($line=~ /S.=.Ser/) || ($line=~ /T.=.Thr/) ||
($line=~ /W.=.Trp/) )
 {push(@vector,$array[4]/10);
push(@vector,$array[5]/10);}

if($line=~ /Tiny/)
{push(@vector,$array[3]/100);}

if($line=~ /Small/)
{push(@vector,$array[3]/100);}

if($line=~ /Aliphatic/)
{push(@vector,$array[3]/100);}

if($line=~ /Aromatic/)
{push(@vector,$array[3]/100);}

if($line=~ /Non-polar/)
{push(@vector,$array[3]/100);}

if($line=~ /Polar/)
{push(@vector,$array[3]/100);}

if($line=~ /Charged/)
{push(@vector,$array[3]/100);}

if($line=~ /Basic/)
{push(@vector,$array[3]/100);}

if($line=~ /Acidic/)
```

```
{push(@vector,$array[3]/100);}

}#while

my $write= new FileHandle;

$write->open(">enpred_infile_pepstats.temp") or
                die( "Could not open to write");

$write->autoflush(1);
$write->print("#Input_pattern_1:\n@vector");

`enzyme`;

`blastpgp -j 3 -d enpred -i enpred_sequence.temp -Q enpred_pssm.temp`;

{
my $zero='0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0';
my @array=();
my @pssmarray=();

foreach(1..3)
{push(@array,$zero);}

my $read = new FileHandle;
$read->open("enpred_pssm.temp") or
      die ("Could not open enpred_pssm.temp");
while ( my $line = $read->getline() )
{
@pssmarray=();
@pssmarray=split(' ',$line);
my $twenty = '';

      if($pssmarray[0]=~ /\d/)
      {
      $twenty = '';
      foreach(2..21)
      {$twenty=$twenty." ".$pssmarray[$_];}
      }
if($twenty)
{push(@array,$twenty);}
}#while

foreach(1..3)
{push(@array,$zero);}

my $write= new FileHandle;
$write->open(">enpred_infile_pssm_ann.temp") or
die( "Could not open to write");
$write->autoflush(1);
chomp(@array);

my $l=scalar(@array);
$l=$l-6;
my $counter=0;
my $lcounter=0;
my $index=1;
```

```
foreach(1..$l)
{
$lcounter=$counter+7;
$write->print("\n#Input_pattern_$index:\n");
      foreach($counter..$lcounter)
      {$write->print("$array[$_] ");}
$counter++;
$index++;
}

my $cmd = 'enpred_ann_pssm.exe';
my $out = `$cmd`;
my $read = new FileHandle;
$read->open("enpred_outfile_pssm_ann.temp") or
      die ("Could not open enpred_pssm.temp");
$result_pssm=0;
$counter_pssm=0;
$nines_pssm=0;
$less_nine_pssm=0;
while ( my $line = $read->getline() )
{
chomp($line);
      if($line>=0.9)
      {
      $nine_pssm++;
      }
      if($line<0.9)
      {
      $less_nine_pssm++;
      }
}

if($nine_pssm >= $less_nine_pssm)
{
print '<p align="left"><font size="4" face="Monotype Corsiva">Your
protein is predicted to be an enzyme by pssm module.</font></p>';
print hr(),hr(),br();
}
if($nine_pssm < $less_nine_pssm)
{
print '<p align="left"><font size="4" face="Monotype Corsiva">Your
protein is predicted to be an non-enzyme by pssm module.</font></p>';
print hr(),hr(),br();
}




}



my $read = new FileHandle;

$read->open("enpred_result.temp") or
      die ("Could not open pepstats_outfile");
```

```
my $result=$read->getline();

if($result >=0.9)
{
print '<p align="left"><font size="4" face="Monotype Corsiva">Your
protein is predicted to be an enzyme by pepstats module. It\'s score is
'.$result.'</font></p>';
print hr(),hr(),br();
}
if(($result < 0.9))
{
print '<p align="left"><font size="4" face="Monotype Corsiva">Your
protein is predicted to be an non-enzyme by pepstats module. It\'s
score is '.$result.'</font></p>';
print hr(),hr(),br();
}



print br(),'<p align="right"><a style="TEXT-DECORATION: none" href =
"http://www.juit.ac.in/">Click here to go to home</a></p>';

print end_html();
```

Code in CGI PERL for the tool to classify the given enzyme into its major class from sequence derived features.

ENZYMEPRED2

```
#!c:/perl/bin/perl.exe
#!c:/perl/lib

use CGI qw(:standard);
use FileHandle;

$ENV{'EMBOSSWIN'}="C:/EMBOSSwin";
$ENV{'EMBOSS_DATA'}="C:/EMBOSSwin/data";
$ENV{'Path'}="C:/EMBOSSwin";

$sequence=param('sequence');
print header(),start_html("Results...");
print '<body link="white" vlink="white" bgcolor="silver">';

print hr(),hr(),'<p align="left"><font size="5" face="Monotype
Corsiva"><b>EnzymePred2</b>: A tool for prediction of different enzyme
classess.</font></p>',hr(),hr();

if(!$sequence)
{print 'ERROR!!!!!!!!',br(),'Please enter a sequence!!',br(); exit;}
my $write= new FileHandle;
$write->open(">enpred_sequence.temp") or
die( "Could not open to write");
$write->autoflush(1);
$write->print($sequence);

`pepstats enpred_sequence.temp enpred_pepstats_outfile.temp -auto 1`;
my $read = new FileHandle;

$read->open("enpred_pepstats_outfile.temp") or
      die ("Could not open pepstats_outfile");
my @vector=();

while ( my $line = $read->getline() )
{
my @array=();
@array=split(' ',$line);
chomp(@array);

if($line=~ /Molecular.weight/)
{push(@vector,$array[3]/1000000);}     #scaling the features and making
the classification vector

if($line=~ /Average/)
{push(@vector,$array[4]/1000);}

if($line=~ /Isoelectric/)
{push(@vector,$array[3]/100);}
if($line=~ /A280.Molar/)
{push(@vector,$array[5]/1000000);}

if($line=~ /A280.Extinction/)
{push(@vector,$array[5]/10);}

if(($line=~ /Improbability/) || ($line=~ /Probability/))
{push(@vector,$array[7]);}

if( ($line=~ /A.=.Ala/)|| ($line=~ /G.=.Gly/) || ($line=~ /L.=.Leu/) ||
($line=~ /V.=.Val/)  )
{push(@vector,$array[4]/100);
push(@vector,$array[5]/10);}

if( ($line=~ /B.=.Asx/) || ($line=~ /X.=.Xaa/) || ($line=~ /Z.=.Glx/) )
{push(@vector,$array[4]);
push(@vector,$array[5]);}

if( ($line=~ /C.=.Cys/) || ($line=~ /Y.=.Tyr/) || ($line=~ /D.=.Asp/)||
($line=~ /E.=.Glu/) || ($line=~ /F.=.Phe/) || ($line=~ /H.=.His/) ||
($line=~ /I.=.Ile/) || ($line=~ /K.=.Lys/) || ($line=~ /M.=.Met/) ||
($line=~ /N.=.Asn/) || ($line=~ /P.=.Pro/) || ($line=~ /Q.=.Gln/) ||
($line=~ /R.=.Arg/) || ($line=~ /S.=.Ser/) || ($line=~ /T.=.Thr/) ||
($line=~ /W.=.Trp/) )
 {push(@vector,$array[4]/10);
push(@vector,$array[5]/10);}

if($line=~ /Tiny/)
{push(@vector,$array[3]/100);}

if($line=~ /Small/)
{push(@vector,$array[3]/100);}

if($line=~ /Aliphatic/)
{push(@vector,$array[3]/100);}

if($line=~ /Aromatic/)
{push(@vector,$array[3]/100);}

if($line=~ /Non-polar/)
{push(@vector,$array[3]/100);}

if($line=~ /Polar/)
{push(@vector,$array[3]/100);}

if($line=~ /Charged/)
{push(@vector,$array[3]/100);}

if($line=~ /Basic/)
{push(@vector,$array[3]/100);}

if($line=~ /Acidic/)
{push(@vector,$array[3]/100);}

}#while
```

```
my $write= new FileHandle;

$write->open(">enpred_infile_pepstats.temp") or
            die( "Could not open to write");

$write->autoflush(1);
$write->print("#Input_pattern_1:\n@vector");

`blastpgp -j 3 -d enpred -i enpred_sequence.temp -Q enpred_pssm.temp`;

{
my $zero='0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0';
my @array=();
my @pssmarray=();

foreach(1..3)
{push(@array,$zero);}

my $read = new FileHandle;
$read->open("enpred_pssm.temp") or
      die ("Could not open enpred_pssm.temp");
while ( my $line = $read->getline() )
{
@pssmarray=();
@pssmarray=split(' ',$line);
my $twenty = '';

      if($pssmarray[0]=~ /\d/)
      {
      $twenty = '';
      foreach(2..21)
      {$twenty=$twenty." ".$pssmarray[$_];}
      }
if($twenty)
{push(@array,$twenty);}
}#while

foreach(1..3)
{push(@array,$zero);}

my $write= new FileHandle;
$write->open(">enpred_infile_pssm_ann.temp") or
die( "Could not open to write");
$write->autoflush(1);
chomp(@array);

my $l=scalar(@array);
$l=$l-6;
my $counter=0;
my $lcounter=0;
my $index=1;

foreach(1..$l)
{
$lcounter=$counter+7;
$write->print("\n#Input_pattern_$index:\n");
      foreach($counter..$lcounter)
      {$write->print("$array[$_] ");}
$counter++;
$index++;
}

my $cmd = 'enpred_ann_pssm.exe';
my $out = `$cmd`;
my $read = new FileHandle;
$read->open("enpred_outfile_pssm_ann.temp") or
      die ("Could not open enpred_pssm.temp");
$result_pssm=0;
$counter_pssm=0;
$nines_pssm=0;
$less_nine_pssm=0;
while ( my $line = $read->getline() )
{
chomp($line);
      if($line>=0.9)
      {
      $nine_pssm++;
      }
      if($line<0.9)
      {
      $less_nine_pssm++;
      }
}

#if($nine_pssm >= $less_nine_pssm)
#{
#print '<p align="left"><font size="4" face="Monotype Corsiva">Your
protein is predicted to be an enzyme by pssm module.</font></p>';
#print hr(),hr(),br();
#}
#if($nine_pssm < $less_nine_pssm)
#{
#print '<p align="left"><font size="4" face="Monotype Corsiva">Your
protein is predicted to be an non-enzyme by pssm module.</font></p>';
#print hr(),hr(),br();
#}

}

`enzyme.exe`;

my $read = new FileHandle;

$read->open("enpred_result.temp") or
      die ("Could not open pepstats_outfile");

my $result=$read->getline();



my $write= new FileHandle;
$write->open(">enpred2_sequence.temp") or
die( "Could not open to write");
$write->autoflush(1);
```

```
$write->print($sequence);

`pepstats enpred2_sequence.temp enpred2_pepstats_outfile.temp -auto 1`;
my $read = new FileHandle;

$read->open("enpred2_pepstats_outfile.temp") or
      die ("Could not open pepstats_outfile");
my @vector=();

while ( my $line = $read->getline() )
{
my @array=();
@array=split(' ',$line);
chomp(@array);

if($line=~ /Molecular.weight/)
{push(@vector,$array[3]/1000000);}    #scaling the features and making
the classification vector

if($line=~ /Average/)
{push(@vector,$array[4]/1000);}

if($line=~ /Isoelectric/)
{push(@vector,$array[3]/100);}

if($line=~ /A280.Molar/)
{push(@vector,$array[5]/1000000);}

if($line=~ /A280.Extinction/)
{push(@vector,$array[5]/10);}

if(($line=~ /Improbability/) || ($line=~ /Probability/))
{push(@vector,$array[7]);}

if( ($line=~ /A.=.Ala/)|| ($line=~ /G.=.Gly/) || ($line=~ /L.=.Leu/) ||
($line=~ /V.=.Val/)  )
{push(@vector,$array[4]/100);
push(@vector,$array[5]/10);}

if( ($line=~ /B.=.Asx/) || ($line=~ /X.=.Xaa/) || ($line=~ /Z.=.Glx/) )
{push(@vector,$array[4]);
push(@vector,$array[5]);}

if( ($line=~ /C.=.Cys/) || ($line=~ /Y.=.Tyr/) || ($line=~ /D.=.Asp/)||
($line=~ /E.=.Glu/) || ($line=~ /F.=.Phe/) || ($line=~ /H.=.His/) ||
($line=~ /I.=.Ile/) || ($line=~ /K.=.Lys/) || ($line=~ /M.=.Met/) ||
($line=~ /N.=.Asn/) || ($line=~ /P.=.Pro/) || ($line=~ /Q.=.Gln/) ||
($line=~ /R.=.Arg/) || ($line=~ /S.=.Ser/) || ($line=~ /T.=.Thr/) ||
($line=~ /W.=.Trp/) )
 {push(@vector,$array[4]/10);
push(@vector,$array[5]/10);}

if($line=~ /Tiny/)
{push(@vector,$array[3]/100);}

if($line=~ /Small/)
{push(@vector,$array[3]/100);}
```

```
if($line=~ /Aliphatic/)
{push(@vector,$array[3]/100);}

if($line=~ /Aromatic/)
{push(@vector,$array[3]/100);}

if($line=~ /Non-polar/)
{push(@vector,$array[3]/100);}

if($line=~ /Polar/)
{push(@vector,$array[3]/100);}

if($line=~ /Charged/)
{push(@vector,$array[3]/100);}

if($line=~ /Basic/)
{push(@vector,$array[3]/100);}

if($line=~ /Acidic/)
{push(@vector,$array[3]/100);}

}#while

my $write= new FileHandle;

$write->open(">enpred2i.txt") or
            die( "Could not open to write");

$write->autoflush(1);
$write->print("@vector");

if( ($nine_pssm >= $less_nine_pssm) || ($result >=0.9))
{
`enpred2.exe`;

my $read = new FileHandle;

$read->open("enpred2o.txt") or
      die ("Could not open pepstats_outfile");

my $result=$read->getline();

if($result ==1)
{
print '<p align="left"><font size="4" face="Monotype Corsiva">Your
protein is predicted to be from Oxidoreductase class</font></p>';
print hr(),hr(),br();
}
if($result ==2)
{
print '<p align="left"><font size="4" face="Monotype Corsiva">Your
protein is predicted to be from Transferase class</font></p>';
print hr(),hr(),br();
}
if($result ==3)
{
```

```
print '<p align="left"><font size="4" face="Monotype Corsiva">Your
protein is predicted to be from Hydrolase class</font></p>';
print hr(),hr(),br();
}
if($result ==4)
{
print '<p align="left"><font size="4" face="Monotype Corsiva">Your
protein is predicted to be from Lyase class</font></p>';
print hr(),hr(),br();
}
if($result ==5)
{
print '<p align="left"><font size="4" face="Monotype Corsiva">Your
protein is predicted to be from Isomerase class</font></p>';
print hr(),hr(),br();
}
if($result ==6)
{
print '<p align="left"><font size="4" face="Monotype Corsiva">Your
protein is predicted to be from Ligase class</font></p>';
print hr(),hr(),br();
}


}
else
{
print '<p align="left"><font size="4" face="Monotype Corsiva">Your
protein is predicted to be an non-enzyme.</font></p>';
}



print br(),'<p align="right"><a style="TEXT-DECORATION: none" href =
"http://www.juit.ac.in/">Click here to go to home</a></p>';

print end_html();
```