IN SILICO MODELING, DRUG DESIGNING AND PHYLOGENETIC ANALYSIS FOR DIABETES MELLITUS TYPE2

BY

ANKUR MAHTE-041525
GOPAL NANDAN-041522
RAMENDRA SINGH SIKARWAR-041526

MAY-2008

Submitted in partial fulfillment of the Degree of Bachelor of Technology

DEPARTMENT OF BIOINFORMATICS AND BIOTECHNOLOGY
JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY-WAKNAGHAT

CERTIFICATE

This is to certify that the work entitled, "Insilico modeling, drug designing and phylogenetic analysis for diabetes mellitus type2" submitted by Ankur Mahte, Gopal Nandan and Ramendra Singh Sikarwar in partial fulfillment for the award of degree of Bachelor of Technology in bioinformatics 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. Anil Kant Thakur.

ACKNOWLEDGEMENT

We would like to extend our heartfelt thanks to our project guide **Dr.Anil Kant Thakur** for his continuous support in our project. He showed us different ways to approach a research problem and the need to be persistent to accomplish any goal. He was always there to meet and talk about our ideas to proofread and use to ask us good questions to help us think through our problem.

27/05/08

Gopal Nandan

Ankur Mahte

Ramendra

Ramendra Singh Sikarwar

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Serial Number	Abbreviated Form	Full Form
1	NIDDM	Non-Insulin Dependent Diabetes Mellitus.
2	SAVS	Structure Analysis And Validation Server.
3	MSFC	Molecular Structures File Convertor.
4	SDSC	San Diego Super Computer Systems.
5	MSA	Multiple Sequence Alignment.
6	RMSD	Root Mean Square Deviation.
7	BLAST	Basic Local Alignment Search Tool.
8	HIV	Human Immunodeficiency Virus.
9	NCBI	National Center For Biotechnology Information.
10	PDB	Protein Data Bank.
11	SPDBV	Swiss Protein Data Bank Viewer.

ABSTRACT

Dipeptidyl-peptidase 4 (DPP4) plays a major role in glucose metabolism. Inhibition of dipeptidyl peptidase-4 (DPP-4) offers a new potential therapeutic approach for type 2 diabetes. Using various drug designing tools and techniques, a best fit ligand was designed for the active site of the target protein (DPP4) which hinders its activity. Further phylogenetic analysis of the target protein was performed with different model organisms to study their phylogenetic relationship.

CHAPTER-1

INTRODUCTION:

Diabetes mellitus:

The term *diabetes* was coined by Aretaeus of Cappadocia. It is derived from the Greek that literally means "passing through," or "siphon," a reference to one of diabetes' major symptoms—excessive urine production. In 1675 Thomas Willis added *mellitus* from the Latin word for honey (*Mel* in the sense of "honey sweet") when he noted that the blood and urine of a diabetic has a sweet taste. This had been noticed long before in ancient times by the Greeks, Chinese, Egyptians, and Indians. In 1776 Matthew Dobson confirmed the sweet taste was because of an excess of a kind of sugar in the urine and blood of people with diabetes. The ancient Indians tested for diabetes by observing whether ants were attracted to a person's urine, and called the ailment "sweet urine disease" (Madhumehalai).

Diabetes mellitus is a group of metabolic diseases characterized by high blood sugar (glucose) levels, which result from defects in insulin secretion, or action, or both. Elevated levels of blood glucose (hyperglycemia) lead to spillage of glucose into the urine, hence the term sweet urine. While the term diabetes without a modifier usually refers to diabetes mellitus, there is another, rare condition named diabetes insipidus (unquenchable diabetes) in which the urine is not sweet; it can be caused by either kidney (nephrogenic DI) or pituitary gland (central DI) damage.

Insufficient production of insulin (either absolutely or relative to the body's needs), production of defective insulin (which is uncommon), or the inability of cells to use insulin properly and efficiently leads to diabetes. Glucose is a type of sugar found in certain foods such as honey and some, but not all, fruits. Glucose is used by the body to make energy. Normally, blood glucose levels are tightly controlled by insulin, a chemical signaling substance (hormone) that is produced by a gland near our stomach called the pancreas. Insulin lowers the blood glucose level because it stimulates the body to make use of glucose. When the amount of glucose in the blood increases, for example, after eating food, insulin is released from the pancreas to normalize the glucose level. However, in patients with diabetes mellitus, the elevated glucose levels cannot be normalized. This causes abnormally high

levels of blood glucose, which ultimately leads to the presence of glucose in the urine (glucosuria).

Type 2 or non-insulin dependent diabetes mellitus (NIDDM):

Type 2 diabetes mellitus - previously known as adult-onset diabetes, maturity-onset diabetes, or non-insulin dependent diabetes mellitus (NIDDM) - is due to a combination of defective insulin secretion and defective responsiveness to insulin (often termed insulin resistance or reduced insulin sensitivity), almost certainly involving the insulin receptor in cell membranes. In early stages, the predominant abnormality is reduced insulin sensitivity, characterized by elevated levels of insulin in the blood. In the early stages, hyperglycemia can be reversed by a variety of measures and medications that improve insulin sensitivity or reduce glucose production by the liver, but as the disease progresses the impairment of insulin secretion worsens and therapeutic replacement of insulin often becomes necessary. There are numerous theories as to the exact cause and mechanism for this resistance, but central obesity (fat concentrated around the waist in relation to abdominal organs, not it seems, subcutaneous fat) is known to predispose for insulin resistance, possibly due to its secretion of adipokines (a group of hormones) that impair glucose tolerance. Abdominal fat is especially active hormonally. Obesity is found in approximately 90% of developed world patients diagnosed with type 2 diabetes. Other factors may include ageing and family history, although in the last decade it has increasingly begun to affect children and adolescents. There is a rather stronger inheritance pattern for type 2 diabetes. Those with first-degree relatives with type 2 have a much higher risk of developing type 2. Concordance among monozygotic twins is close to 100%, and 25% of those with the disease have a family history of diabetes. Type 2 diabetes may go unnoticed for years in a patient before diagnosis, since the symptoms are typically milder (eg, lack of ketoacidotic episodes) and can be sporadic. However, severe complications can result from unnoticed type 2 diabetes, including renal failure, vascular disease (including coronary artery disease), vision damage, etc. Sometimes; people with Type II diabetes don't notice any symptoms or the symptoms are experienced gradually. These include blurry vision, cuts or sores those are slow to heal, itchy skin, yeast infections, increased thirst, dry mouth, need to urinate often and Leg pain. Type 2 diabetes is usually first treated by changes in physical activity, diet (generally decrease carbohydrate intake, especially glucose generating

carbohydrates), and through weight loss. These can restore insulin sensitivity, even when the weight loss is modest, for example, around 5 kg, most especially when it is in abdominal fat deposits. The next step, if necessary, is treatment with oral ant diabetic drugs. As insulin production is initially unimpaired, oral medication (often used in combination) can still be used that improves insulin production (e.g., sulfonylurea) and regulate inappropriate release of glucose by the liver (and attenuate insulin resistance to some extent (e.g., metformin), and substantially attenuate insulin resistance (e.g., thiazolidinediones). If these fail, insulin therapy will be necessary to maintain normal or near normal glucose levels. A disciplined regimen of blood glucose checks is recommended in most cases, when most of these medications are being taken.

Target protein:

Target protein was dipeptidyl-peptidase 4 (DPP4), which plays a major role in glucose metabolism. It is responsible for the degradation of incretins such as GLP-1. Inhibition of dipeptidyl peptidase-4 (DPP-4) offers a new potential therapeutic approach for type 2 diabetes.DPP4 degrades GLP-1 (glucagon-like peptide-1), an important hormone that is released in response to the intake of food and that stimulates pancreatic beta cells to increase the secretion of insulin and that has potential to improve beta cell function.

High DPP IV → Degrade Incretin → low insulin secretion

The aliases of the target protein are CD26; ADABP; ADCP2; DPPIV; TP103. Protein has molecular weight 88279 dalton. The sub cellular location is cell membrane. Its preferred name is dipeptidylpeptidase4 and other designation includes Dipeptidylpeptidase IV, T-cell activation antigen CD26, Dipeptidylpeptidase 4 (CD26, adenosine deaminase complexing protein 2), Dipeptidylpeptidase IV (CD26, adenosine deaminase complexing protein 2).

CD26, or dipeptidyl peptidase 4 (DPP-4) is a membrane-associated peptidase of 738 amino acids that is widely distributed in numerous tissues. DPPIV can be considered a moonlighting protein because it is a multifunctional protein that exerts different functions depending on cell type and intra- or extracellular conditions in which it is expressed. This

protein acts as a costimulatory protein, receptor and also acts as a proteolytic enzyme. Therefore, targeting of CD26/DPPIV and especially its proteolytic activity has much therapeutic potential. On the other hand, there are homologous proteins with overlapping proteolytic activity, which thus may prevent specific modulation of CD26/DPPIV. DPP-4 also binds the enzyme adenosine deaminase specifically and with high affinity. The significance of this interaction has yet to be established. It is a rather indiscriminate enzyme for which at least 62 substrates are known. Human DPP4 is ubiquitously expressed in epithelial and endothelial cells and serves multiple functions in cleaving the penultimate positioned prolyl bonds at the NH2 terminus of a variety of physiologically important peptides in the circulation. Recent studies showed a linkage between DPP4 and down-regulation of certain chemokines and mitogenic growth factors, and degradation of denatured collagens (gelatin), suggesting a role of DPP4 in the cell invasive phenotype. DPP-4 also exists as a soluble circulating form in plasma and significant DPP-4-like activity is detectable in plasma from humans and rodents. This protein also has its role in apoptosis, signal transduction and in adhesion. It appears to work as a suppressor in the development of cancer and tumours.

The CD26/DPPIV protein plays a major role in immune response. Abnormal expression is found in the case of autoimmune diseases, HIV-related diseases and cancer. DPP-4 (CD26) exerts its biological effects via two distinct mechanisms of action. First, as a membrane spanning protein, it binds adenosine deaminase and when activated, conveys intracellular signals independent of its enzymatic function via dimerization and activation of intracellular signaling pathways. The signaling properties of membrane-associated CD26 have been most extensively characterized in T cells. CD26 associates with several membrane proteins, including CD45, CXCR4.

The second principal biological activity of CD26 (DPP-4) is its enzymatic function. The enzymatic activity of CD26 is exhibited by the membrane-spanning form of the molecule, and by the circulating soluble form. DPP-4 prefers substrates with an amino-terminal proline or alanine at position 2, but may also cleave substrates with non-preferred amino acids at position 2.

This protein is also present in other organisms for example *Canis familiaris*(dog), *Felis catus*(domestic cat), *Pan troglodytes*, *Rattus norvegicus* (brown rat), *Musmusculus* (house mouse), *Sus scrofa domestica* (domestic pig), *Bos taurus* (cow), *Oxyuranus*

scutellatus (Australian snake), *Oxyuranus microlepidotus* (Small Scaled Snake and Fierce Snake), *Notechis scutatus* (Tiger snakes), *Gloydius blomhoffi brevicaudus* (*Japanese mamushi), Xenopus tropicalis* (western clawed frog), *Macaca mulatto* (rhesus monkey) and *Gallus gallus* (chicken).

Objective of the project:

The objective of present studies was to design the best fit ligand for the protein dipeptidyl peptidase 4 in human beings which act as a receptor for diabetes mellitus type 2 in human body and also to perform the phylogenetic analysis for the dipeptidyl peptidase 4.

CHAPTER-2

PROCEDURES:

Softwares and tools used:

Ncbi:

NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information, all for the better understanding of molecular processes affecting human health and disease.

Blast:

Basic Local Alignment Search Tool, or BLAST, is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences. A BLAST *search* enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold.

Modeller:

MODELLER is used for homology or comparative modeling of protein three-dimensional structures. The user provides an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms.

Swiss pdb viewer:

Swiss-Pdb Viewer is an application that provides a user friendly interface allowing to analyze several proteins at the same time. The proteins can be superimposed in order to deduce structural alignments and compare their active sites or any other relevant parts. Amino acid mutations, H-bonds, angles and distances between atoms are easy to obtain.

Procheck:

The aim of PROCHECK is to assess how normal, or conversely how unusual, the geometry of the residues in a given protein structure is, as compared with stereo chemical parameters derived from well-refined, high-resolution structures. The input to PROCHECK is a single file containing the coordinates of your protein structure.

Ligsite:

LIGSITE is a program for the automatic and time-efficient detection of pockets on the surface of proteins that may act as binding sites for small molecule ligands.

ChemSketch:

ChemSketch is an advanced chemical drawing tool and is the accepted interface into the industry's best NMR and molecular property predictions, nomenclature, and analytical data handling software.

Hex 4.5:

Hex is an interactive protein docking and molecular superposition program.

Ligbuilder:

LigBuilder is a powerful multiple-purposed program written for structure-based drug design procedure. Based on the three-dimensional structure of the target protein, it can automatically build ligand molecules within the binding pocket and subsequently screen them.

Autodock:

AutoDock is a suite of automated docking tools. It is designed to predict how small molecules, such as substrates or drug candidates, bind to a receptor of known 3D structure.

Protocol:

1. The role of various proteins involved in regulation of insulin level was investigated from the literature and DPP4 was selected as target protein which plays a major role in down regulation of insulin.

2. The sequence of target protein was retrieved from NCBI.

3. BLASTP was performed to get the maximum sequence similarity.

4. Protein structure was designed through homology modelling using MODELLER.

5. The Stability of the generated target protein was checked by using Ramachandran plot.

6. Procheck tool was used in structure analysis and validation server (SAVS) to check deform contacts in target protein.

7. The Modeled target protein was made stable by performing loop formation and energy minimization by using SPDB viewer.

8. The lead molecule was chosen and designed by using chemsketch.

9. An online tool LIGSITE was used to find the pockets of modeled protein.

10. Molecular structure file convertor (MSFC) was used for file conversion.

11. HEX 4.5 software was used for docking of the protein and lead molecule.

12. The lead molecule was grown with the help of LIGBUILDER.

13. Best fit ligand was chosen on the basis of OSIRIS property explorer.

14. Docking of best fit ligand with target protein was performed using AUTODOCK.

15. Phylogenetic analysis of target protein was performed using SDSC WORKBENCH.

CHAPTER-3

RESULTS AND DISCUSSIONS:

Sequence of the target protein from NCBI: Target protein (DPP4) sequence was taken from NCBI. Given below is the fasta format of target protein sequence.

gi|1352311|sp|P27487|DPP4_HUMAN Dipeptidyl peptidase 4 (Dipeptidyl peptidase IV) (DPP IV) (T-cell activation antigen CD26) (TP103) (Adenosine deaminase complexing protein 2) (ADABP) [Contains: Dipeptidyl peptidase 4 membrane form (Dipeptidyl peptidase IV membrane form); Dipeptidyl peptidase 4 soluble form (Dipeptidyl peptidase IV soluble form)]

MKTPWKVLLGLLGAAALVTIITVPVVLLNKGTDDATADSRKTYTLTDYLKNTYRLK
LYSLRWISDHEYYKQENNILVFNAEYGNSSVFLENSTFDEFGHSINDYSISPDGQFILLE
YNYVKQWRHSYTASYDIYDLNKRQLITEERIPNNTQWVTWSPVGHKLAYVWNNDIY
VKIEPNLPSYRITWTGKEDIIYNGITDWVYEEEVFAYSALWWSPNGTFLAYAQFNDTE
VPLIEYSFYSDESLQYPKTVRVPYPKAGAVNPTVKFFVVNTDSLSSVTNATSIQITAPA
SMLIGDHYLCDVTWATQERISLQWLRRIQNYSVMDICDYDESSGRWNCLVARQHIE
MSTTGWVGRFRPSEPHFTLDGNSFYKIISNEEGYRHICYFQIDKKDCTFITKGTWEVIG
IEALTSDYLYYISNEYKGMPGGRNLYKIQLSDYTKVTCLSCELNPERCQYYSVSFSKE
AKYYQLRCSGPGLPLYTLHSSVNDGLRVLEDNSALDKMLQNVQMPSKKLDFIILNET
KFWYQMILPPHFDKSKKYPLLLDVYAGPCSQKADTVRLNWATYLASTENIIVASFDG
RGSGYQGDKIMHAINRRLGTFEVEDQIEAARQFSKMGFVDNKRIAIWGSYGGYVTS
MVLGSGSGVFKCGIAVAPVSRWEYYDSVYTERYMGLPTPEDNLDHYRNSTVMSRAE
NFKEYLLIHGTADDNVHFQQSAQISKALVDVGVDFQAMWYTDEDHGIASSTAHQHI
YTHMSHFIKQCFSLP

Blastp result: Blastp was performed to get the maximum similarity by comparing the query sequence (i.e. target protein sequence) with a library or database of sequence. Results obtain by performing blastp is given below.

pdb|2BGR|A Chain A, Crystal Structure of Hiv-1 Tat Derived Nonapeptides
Tat (1-9) Bound To The Active Site of Dipeptidyl Peptidase IV (Cd26)
pdb|2BGR|B Chain B, Crystal Structure of Hiv-1 Tat Derived Nonapeptides

Tat (1-9) Bound To the Active Site of Dipeptidyl Peptidase IV (Cd26)

Length of the target sequence is 738.Results include Score: 1513 bits (3918), Identities: 730/738 (99%), Positives: 730/738 (99%), Gaps: 0/738 (0%) and Method: Composition-based stats.

Structure of target protein:

Modeler requires three files as an input, atom file, alignment file and python file. To make the atom file, all the atoms information was taken from the PDB file of target protein and by arranging the subject and query sequence alignment obtained from the blast result, alignment file was constructed. Python file includes only coding part, which instructs the modeller for the generation of the protein structure. Given below is the coding part:

```
from modeller.automodel import*
log.verbose()
env=environ()
env.io.atom_files_directory='./:../2abl.atm'
a=automodel(env,
alnfile='2abl.ali',
knowns='2abl',
sequence='query')
starting_model=1
ending_model=5
make()
```

The three files (i.e. atom, alignment and python file) used as an input by the modeler and generated protein structure of the target protein (shown in fig.1).

Ramachandran plot after modelling:

Ramachandran plot (Fig. 2) was used for checking the stability of modelled target protein. There are three regions in ramachandaran plot core region, allowed region and disallowed region. A protein to be stable maximum number of its amino acid residues should be in core and allowed region. Residue presents in disallowed region indicates given protein have some level of unstability and to work on such protein, it is necessary to make a protein stable i.e. all amino acid residue should be either in core or allowed region.

Structure analysis and validation (to find the deformities in target protein) was performed using procheck tool. Result obtained (Table 1) shows that 91.1% of amino acid residues are present in core region, 7.8% are present in allowed region, 0.4% amino acid residues are present in gener region, 0.7% are present in disallowed region and three bad contacts are also present. The bad contacts and the percentage of the amino acid residues present in the disallowed region indicate the unstability of the target protein.

Loop formation and energy minimization:

Loop formation and energy minimization was performed using spdb viewer to make the target protein stable. Loop formation was performed to move residue from disallowed region to allowed region and bad contacts were removed through energy minimization. Loop formation (Fig.3) was performed around the amino acid residue which is present in disallowed region to move them either in allowed or core region. During loop generation a window appear on the screen indicating three values clash score, pp value, force value for different loops that can be possible and among those possible loops we have to select that loop which have minimum ff (force field) value and in case if there is same ff value for two different loops, then loop is selected on the basis of clash score value whose clash score value is minimum.

Energy minimization:

Energy minimization (Fig. 4) was performed using spdb viewer to remove the bad contacts until and unless all the bad contacts were removed from the target protein.

Ramachandran plot after removal of deformities:

Results (Table 2) obtained after energy minimization and loop formation shows that 89.7% of amino acid residues are present in core region, 10.2% are present in allowed region, 0.2% amino acid residues are present in gener region and 0.0% are present in disallowed region and no bad contact is present. Absence of bad contacts and the percentage of the amino acid residues present in the disallowed region indicate that target protein is completely stable.

Lead molecule:

On the basis of the molecular structure of different ligands, which are already present for diabetes mellitus type2 and by taking following properties into account, lead molecule was chosen. The properties include, size should be small, No metal atom should be there, Maximum no. of possible growing sites and should not contain any kind of unsaturation.

Pocket identification of the target modelled protein:

PDB file of the target protein was taken for pocket identification using ligsite. There were three pockets identified by the ligsite indicated by three different colours; green, red and orange in SPDB viewer control panel window (Fig.7). The active site at the top of the window was taken as the main active site. An amino acid residue serine181 present very closely to the main active site (red colour) was located at the distance of 7.60 Å from the active site. It was chosen because it is nearest to the active site and lead molecule will be able to cover the active site, when it will grow after binding with it. This will ultimately block the active site of the target protein.

Binding of lead molecule to the closest amino acid residue:

The docking (Fig.8) of the lead molecule to the closest amino acid residue was performed using HEX 4.5 software. Lead molecule(red colour) was attached (Fig.9) to the amino acid residue(green colour) which was present closest to the active site(yellow colour).

Growing of the lead molecule:

LIGBUILDER was used to grow the lead molecule. In this three steps were performed pocketing, growing and processing. POCKET has two main functions: first, it analyze the binding pocket and prepare the information necessary for running GROW and second, it derive the key interaction sites within the binding pocket and suggest a pharmacophore model. The major function of GROW is to construct the ligand molecules for the target protein by applying the growing strategy. All the molecules originated from a "seed" structure were developed and evolved with a Genetic Algorithm procedure. All resultant molecules were collected in a file. Processing provides ability to analyze a ligbuilder LIG file, extracts the desired molecules, and converts them to viewable mol2 files. Input files for ligbuilder are HEX complex pdb file and lead molecule in mol2 file. Ten conformations (Fig.10a to10j) of ligand were obtained from ligbuilder.

Selection of best fit ligand (using OSIRIS property explorer):

Best fit ligand was chosen on the basis of OSIRIS property explorer. The molecular structures of ligands obtained from the ligbuilder were drawn on the OSIRIS property explorer (Fig.11a to 11j) to check their properties like toxicity risks, clogp value, solubility, molecular weight, drug likeness and drug score for each conformation. On the basis of these properties best fit ligand which was free from mutagenicity and any kind of side effects were chosen. The deciding parameter was maximum drug score value (Fig. 11h).

Best fit ligand:

Among all ten conformations of ligand, best fit ligand (Fig.12) was chosen on the basis of maximum drug score value having parameters clogp, solubility, molecular weight, drug likeness and drug score with values 2.44, -2.67, 297.0, -0.14 and 0.65 respectively.

Docking of best fit ligand with target protein:

Docking is a single autodock process, which carries out a number of independent docking runs, each of which begins with the same initial conditions. After all the docking runs have been completed in a given job, the cluster analysis or structure binning is performed. This is based on positional root mean square deviation of corresponding atoms, ranking the resulting families of docked conformations in order of increasing energy. The method for structure binning allows for symmetry rotations. The rmsd tolerance "rmstol" and reference structure "rmsref" filename should be specified while clustering the conformations. Typical value for rmstol ranges from 0.5 to 1.5Å.

Autodock was performed for docking best fit ligand (red colour) with target protein (green colour) (Fig.13). Autodock analysis tool compares all the docked conformations with one another, and if two conformations have rmsd that is less than the rmstol value, they are both stored in the same cluster (Table 4). This was repeated for all conformations, and the clusters were ranked in order of increasing energy.

Table 3 shows the final docked energy for each conformation and the rms difference between the lowest energy member of the cluster and every other member. The rms for the lowest member of the group by definition is zero. Each conformation has a set of remark records, one of which describes the rms difference between itself and the coordinates specified. This can be useful to compare docked conformation and experimentally determined position.

Phylogenetic analysis:

Phylogenetic analysis was performed to compare the DPP4 sequences from human being with that of *Rattus norvegicus* (brown rat), *Musmusculus* (house mouse), *Felis catus* (domestic cat), *Sus scrofa domestica* (domestic pig), *Bos Taurus* (cow), *Gallus gallus* (chicken), *Xenopus tropicalis* (western clawed frog) and *Macaca mulatta* (rhesus monkey). It was performed using SAN DIEGO SUPER COMPUTER SYSTEMS (SDSC) WORKBENCH. Pairwise sequence alignment, multiple sequence alignment, dendrogram, clustal distance matrix, boxshade and texshade were used for phylogenetic analysis.

Pairwise alignment:

Sequence alignment is a way of arranging the primary sequences of DNA, RNA, or protein and to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Aligned sequences of nucleotides or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that residues with identical or similar characters are aligned in successive columns. If two sequences in an alignment share a common ancestor, mismatches can be interpreted as point mutations and gaps as indels (that is, insertion or deletion mutations) introduced in one or both lineages in the time since they diverged from one another. In protein sequence alignment, the degree of similarity between amino acids occupying a particular position in the sequence can be interpreted as a rough measure of how conserved a particular region or sequence motif is among lineages.

Pairwise alignment (Table 5) of protein sequences of query sequence and the model organisms were performed. The highest alignment score (97) was between query sequence and that of *macaca mulatta* which shows close phylogenetic relationship between them. The sequence from *Xenopus tropicalis* is having lowest alignment score (58) with query sequence indicating distant phylogeny.

Multiple sequence alignment:

Multiple sequence alignment (MSA) is sequence alignment of three or more biological sequences, generally protein, DNA, or RNA. In general the input set of query sequences are assumed to have an evolutionary relationship by which they share a lineage and are descended from a common ancestor. Sequence homology can be inferred and phylogenetic analysis can be conducted to assess the sequences shared evolutionary origins from the resulting MSA. In this study multiple sequence alignment was performed between query sequence and model organisms. The results obtained shown in table 7 in which fully conserved residues, conservation of strong and weak groups have been represented with different colours.

Dendrogram:

A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by a clustering algorithm. A dendrogram is strictly defined as a binary tree with a distinguished root, which has all the data items at its leaves. Conventionally, all the leaves are shown at the same level of the drawing. This branching diagram shows the relative sequence similarity between different proteins or genes. Typically, horizontal lines indicate the degree of differences in sequences, but vertical lines are used to separate the branches. The sequence similarity between the query sequence and model organisms have been shown as a tree diagram in Fig.14. It indicates close phylogenetic relationship of query sequences with *macaca mulatta* whereas *Xenopus tropicalis* is most distantly located in the phylogenetic tree.

Clustal distance matrix:

This program calculates distances between sequences using a matrix. The clustal distance (Table 6) was calculated between query sequence and model organisms. Here query sequence is having minimum distance (0.30) with that of *macaca mulatta* as comparison to

rest of the model organisms. *Xenopus tropicalis* is having maximum clustal distance which is 0.413 showing distant phylogenetic relationship with query protein sequence.

Boxshade:

Boxshade is a program for pretty printing of multiple sequence alignment output. Various kinds of shadings are applied to identical and similar residues. This Result (Table 8) represents conserved, identical, similar and different amino acid residues in green, yellow, cyan and white colour respectively of query and model organisms protein sequence.

Texshade:

Texshade is TEX-based alignment shading software featuring standard identity and similarity shading. The Result (Table 9) shows non conserved (black colour), similar (pink colour), and conserved (sky blue colour) and matched (dark blue colour) amino acid residues in different colours of query sequence and model organisms.

CONCLUSION

The best fit ligand was designed for the target protein DPP4 which specifically binds to its active site and blocks its activity. As DPP4 is involved in down regulation of insulin so this could be the first step directed towards the designing and development of a therapeutic drug for the diabetes mellitus type 2. Phylogenetic analysis concludes that human DPP4 sequence has maximum sequence similarity with that of *macaca mulatta* and shares the closest evolutionary relationship whereas DPP4 sequences from *Xenopus tropicalis* showed lowest similarity with it among all model organisms tested.

FIGURES

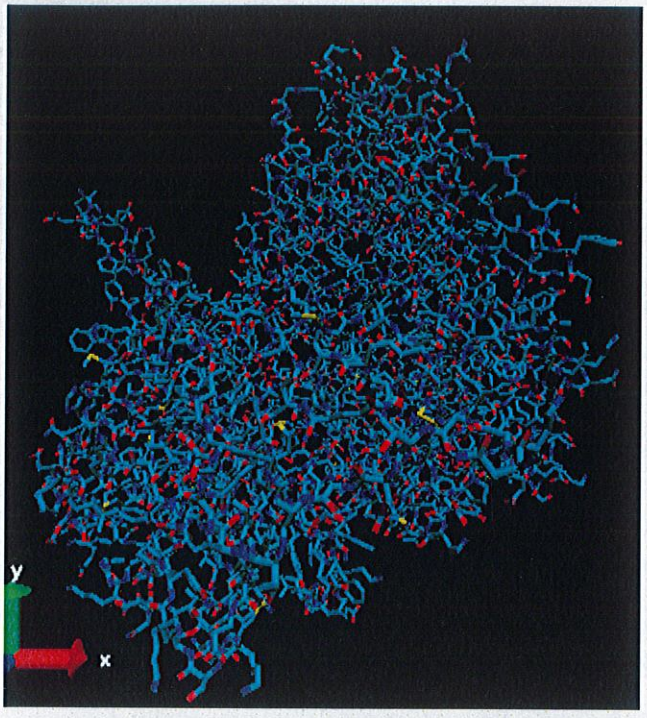


Fig.1 Structure of modeled target protein

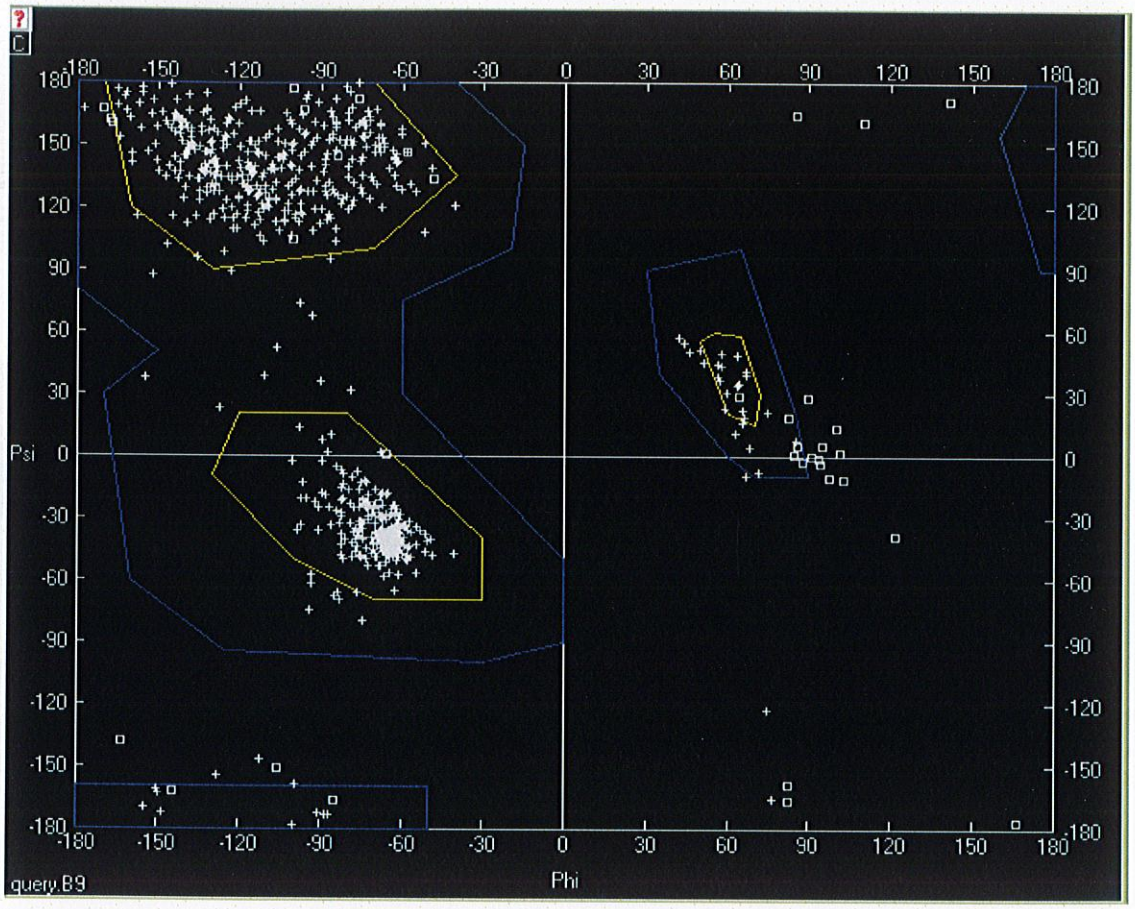


Fig. 2 Ramachandran plot after modeling

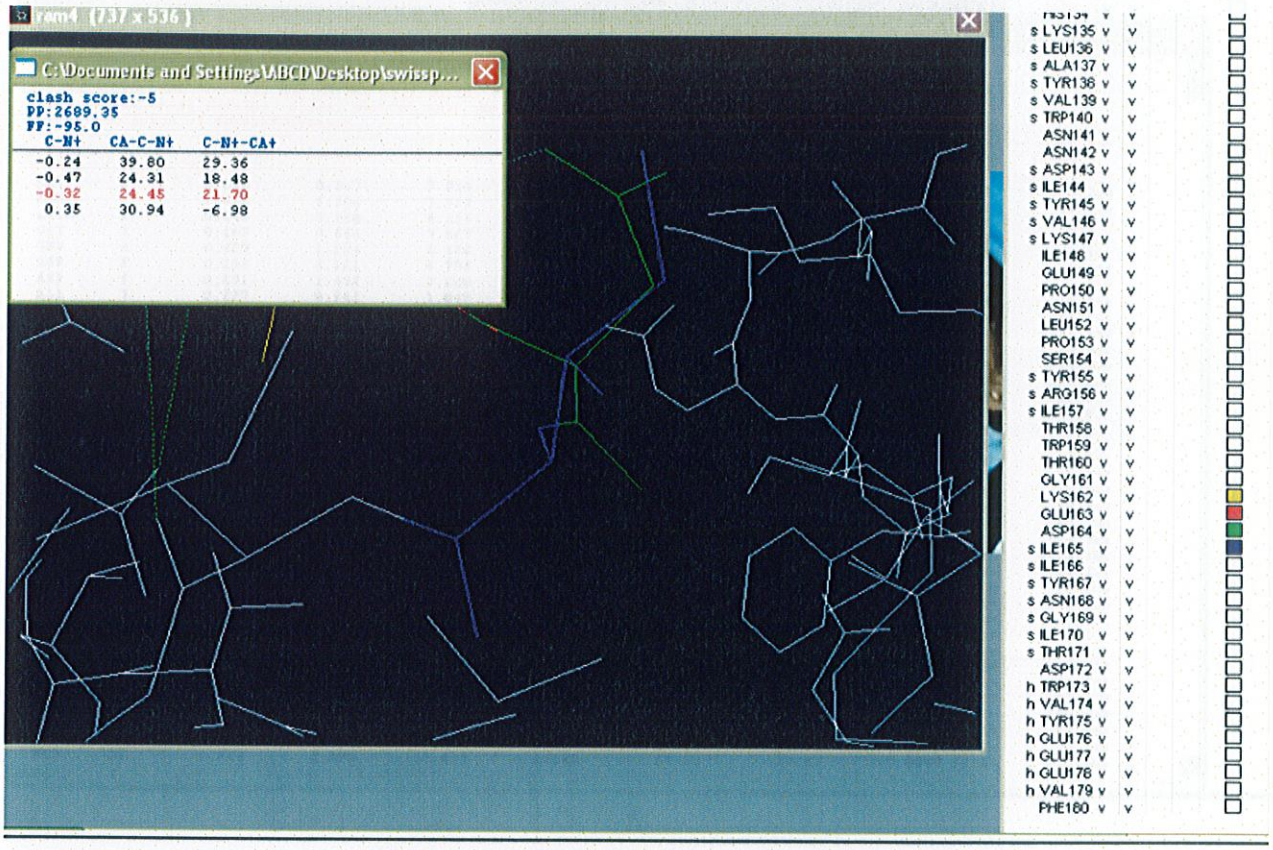


Fig. 3 Loop formation performed for attaining stability in protein

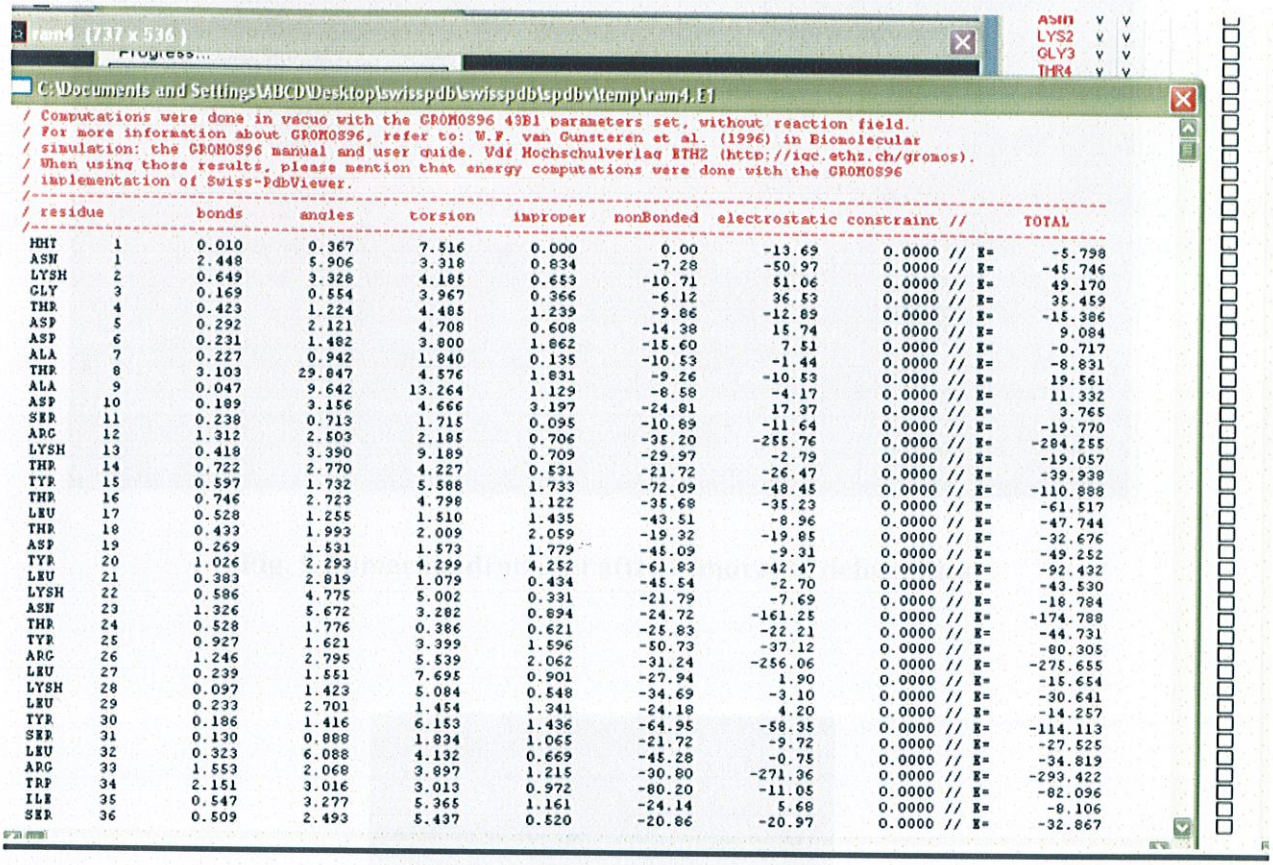


Fig. 4 Energy minimization

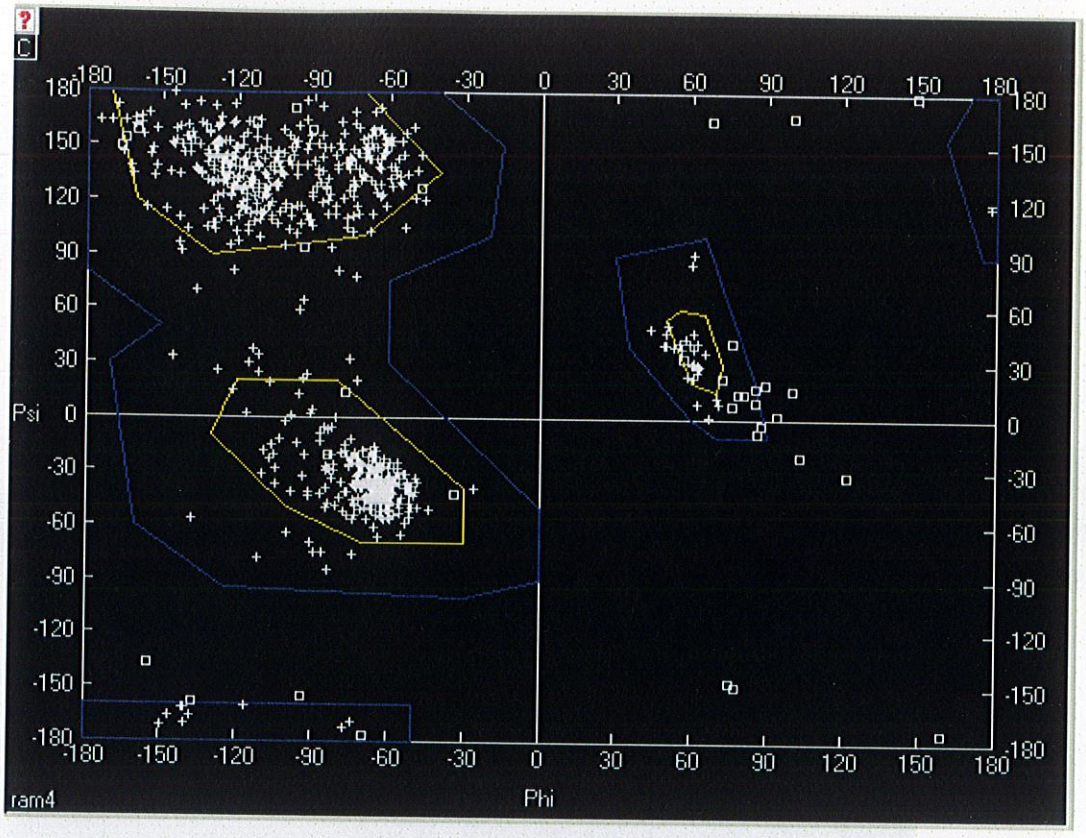


Fig. 5 Ramachandran plot after removal of deformities

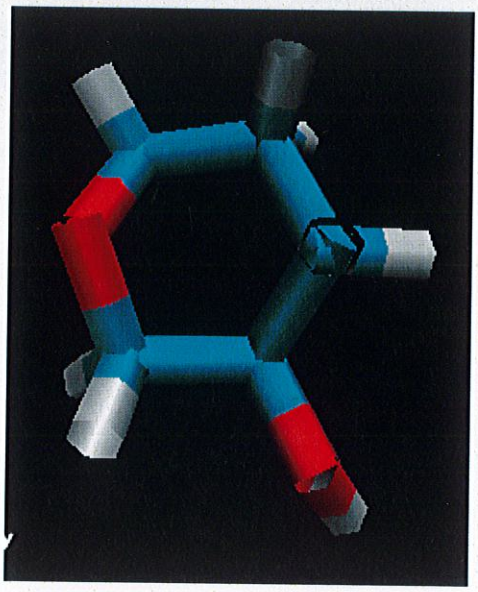


Fig. 6 Structure of lead molecule

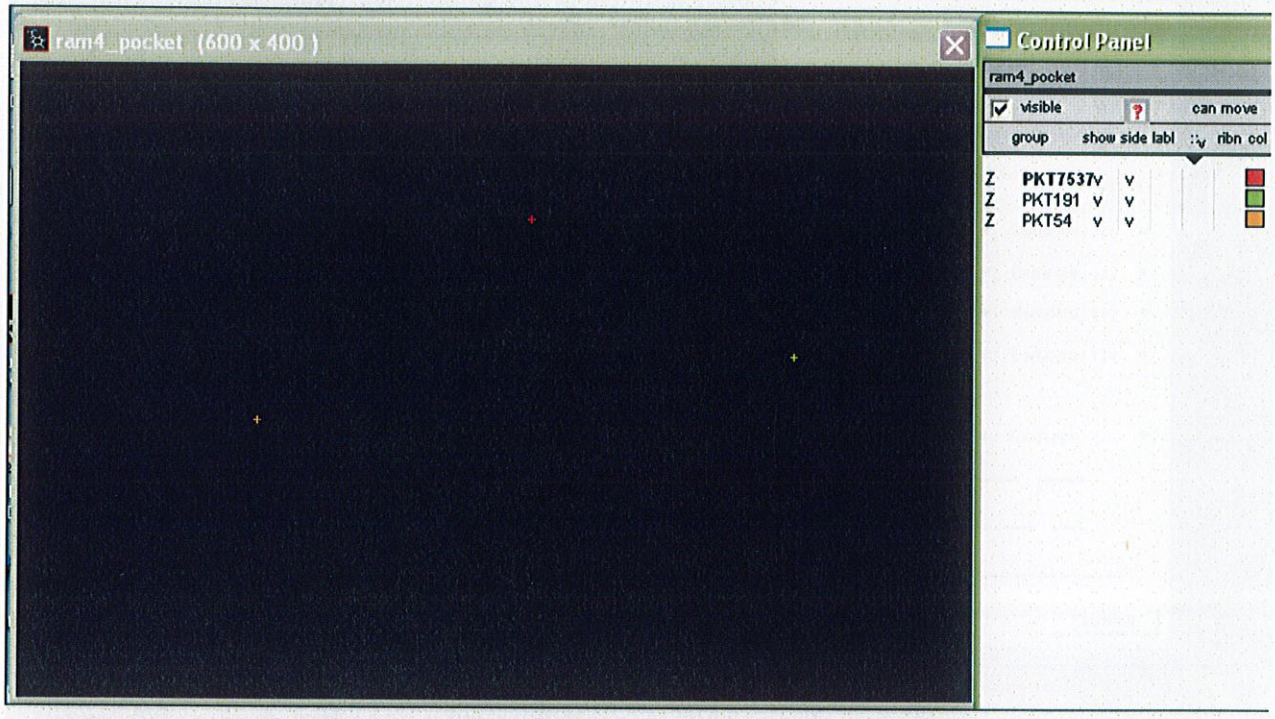


Fig. 7 Pocket identification of the target modelled protein

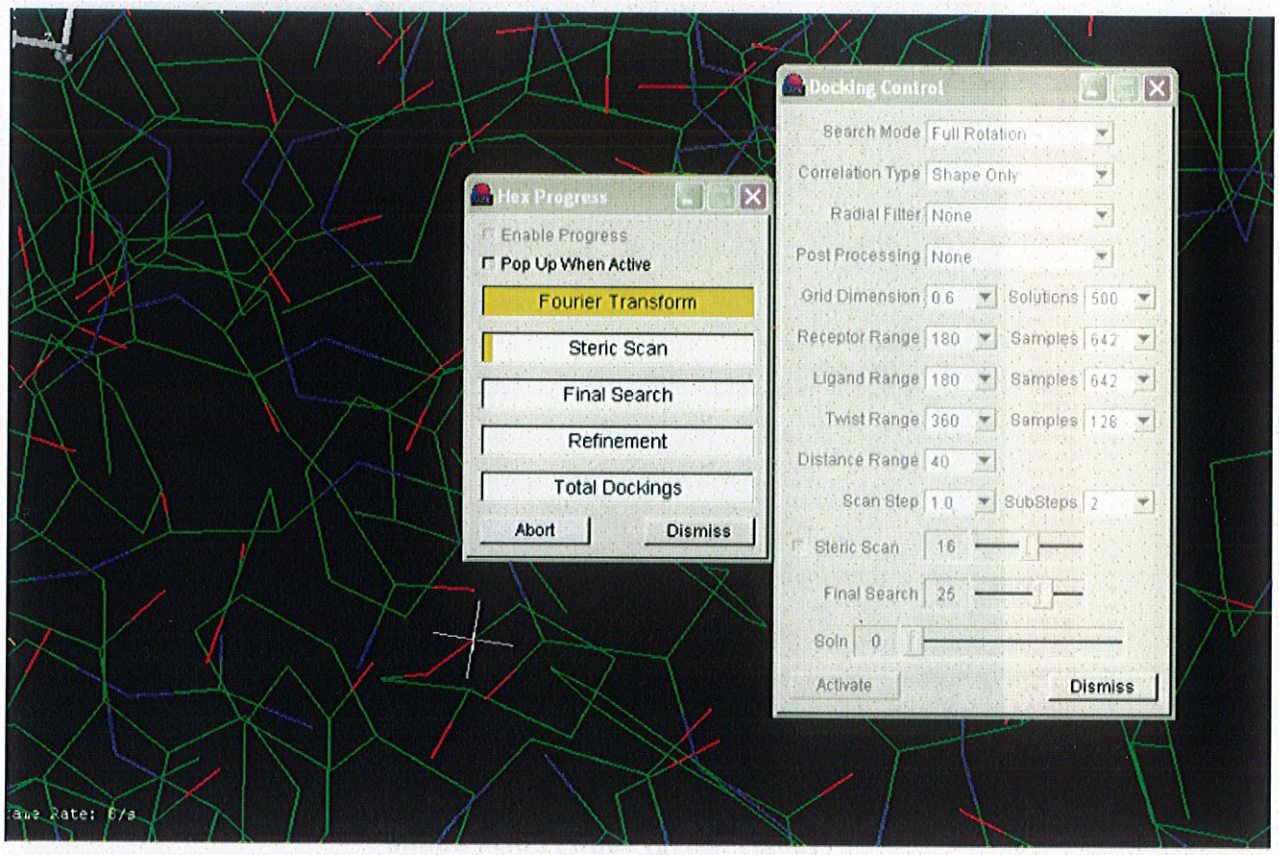


Fig. 8 Binding of lead molecule to the closest amino acid residue

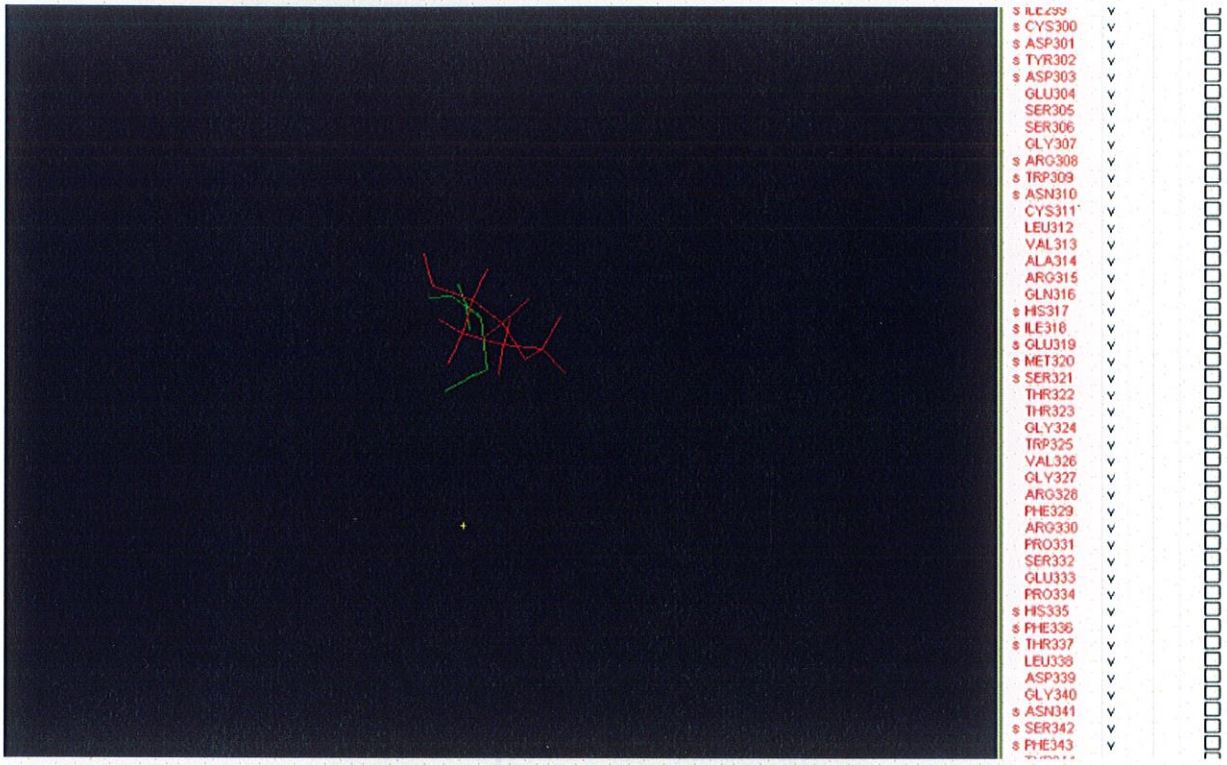


Fig. 9 lead molecule (red colour) get attached to the closest amino acid residue (green colour)

Ten conformations of ligands as results of ligbuilder:

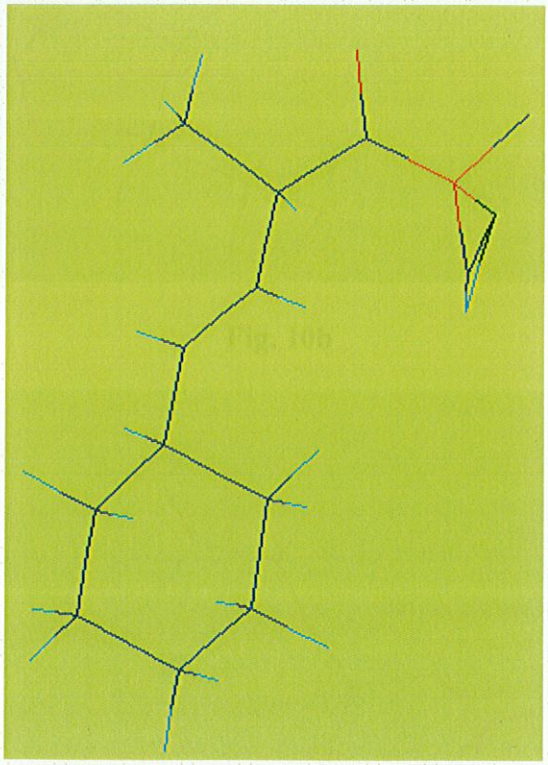


Fig. 10a

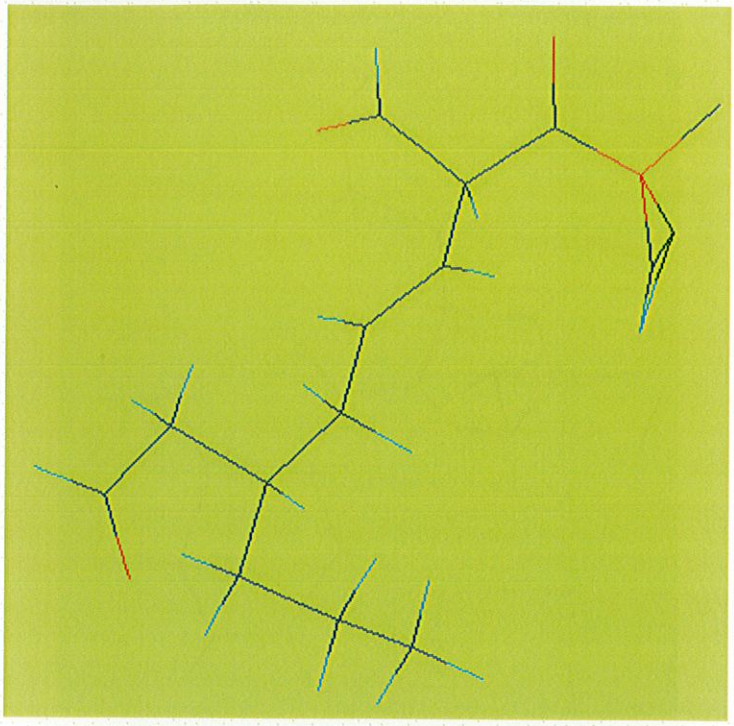


Fig. 10b

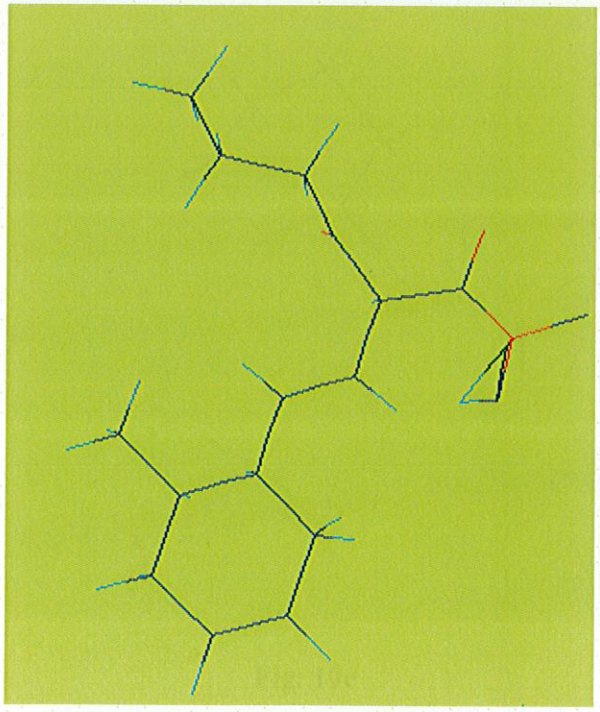


Fig. 10c

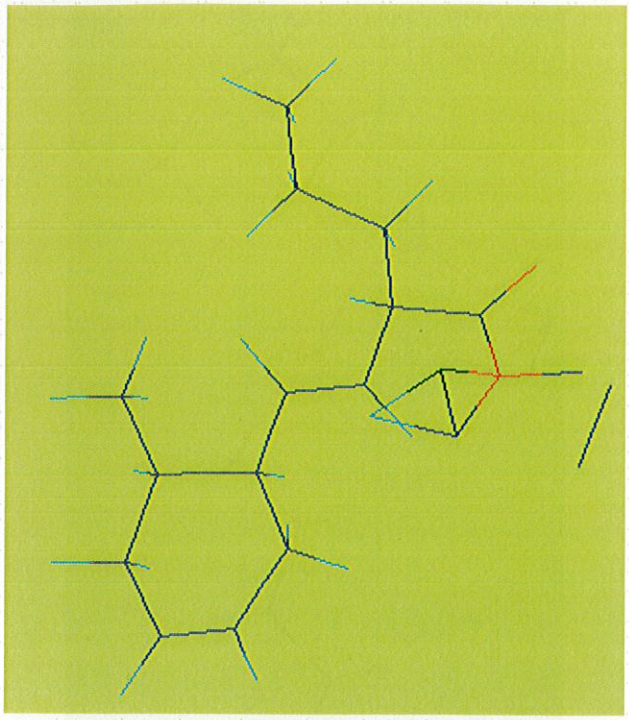


Fig. 10d

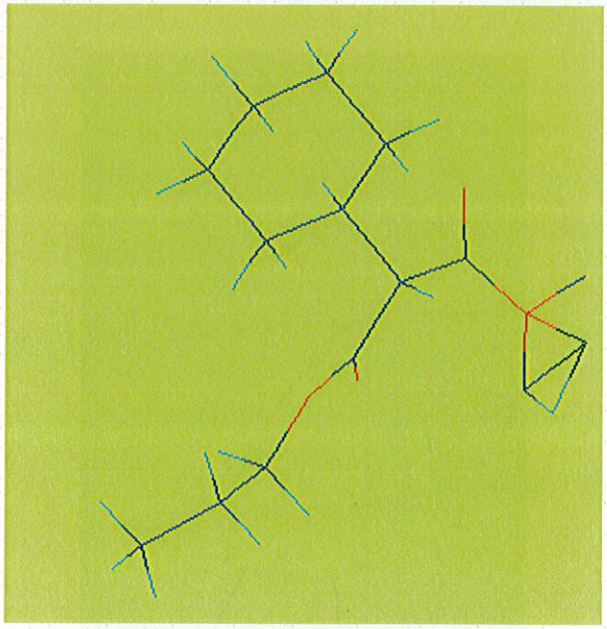


Fig. 10e

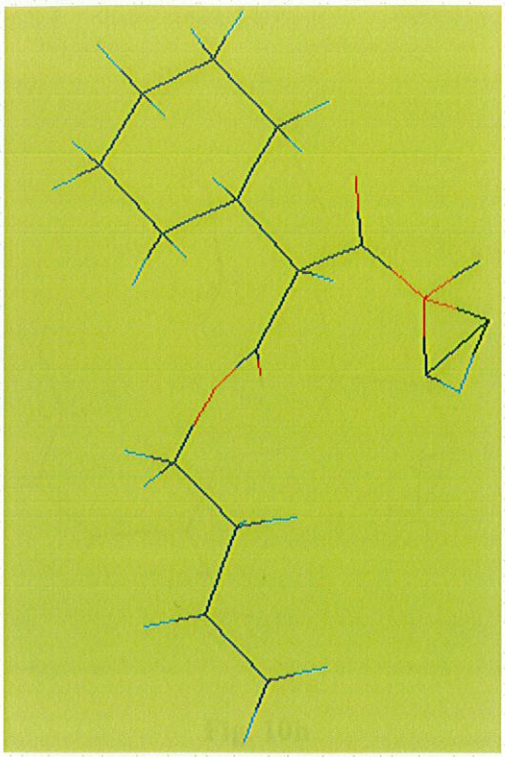


Fig. 10f

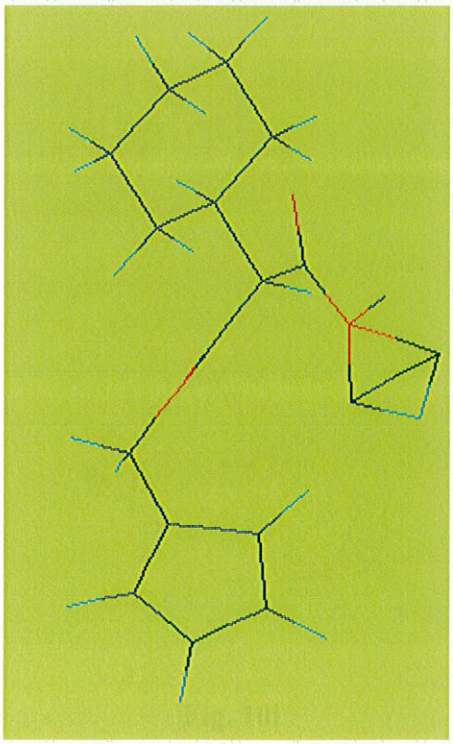


Fig. 10g

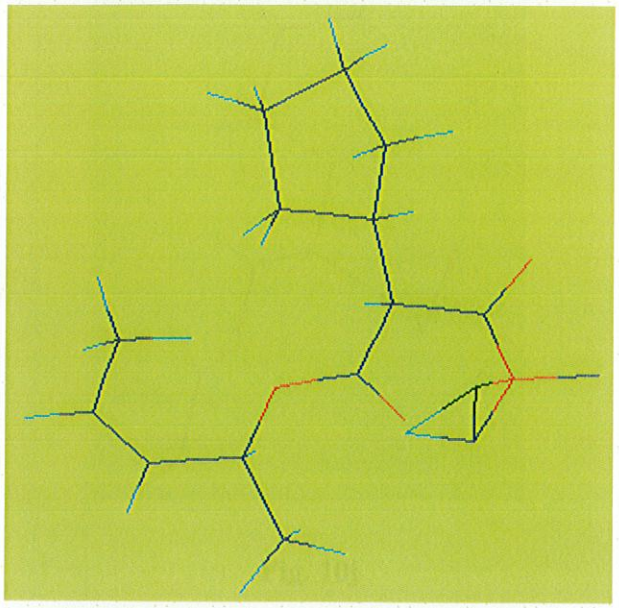


Fig. 10h

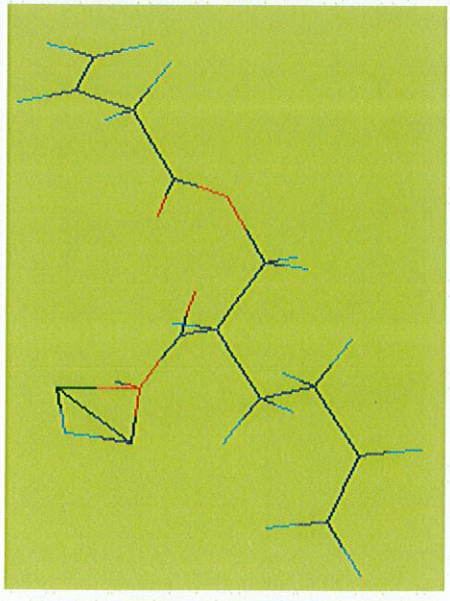


Fig. 10i

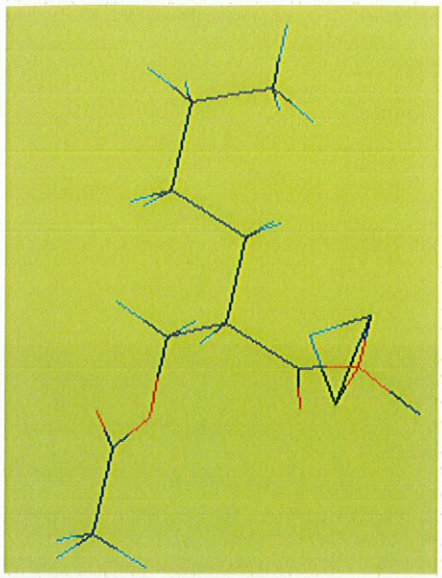


Fig. 10j

OSIRIS result (ligands molecular structure drawn on OSIRIS property explorer):

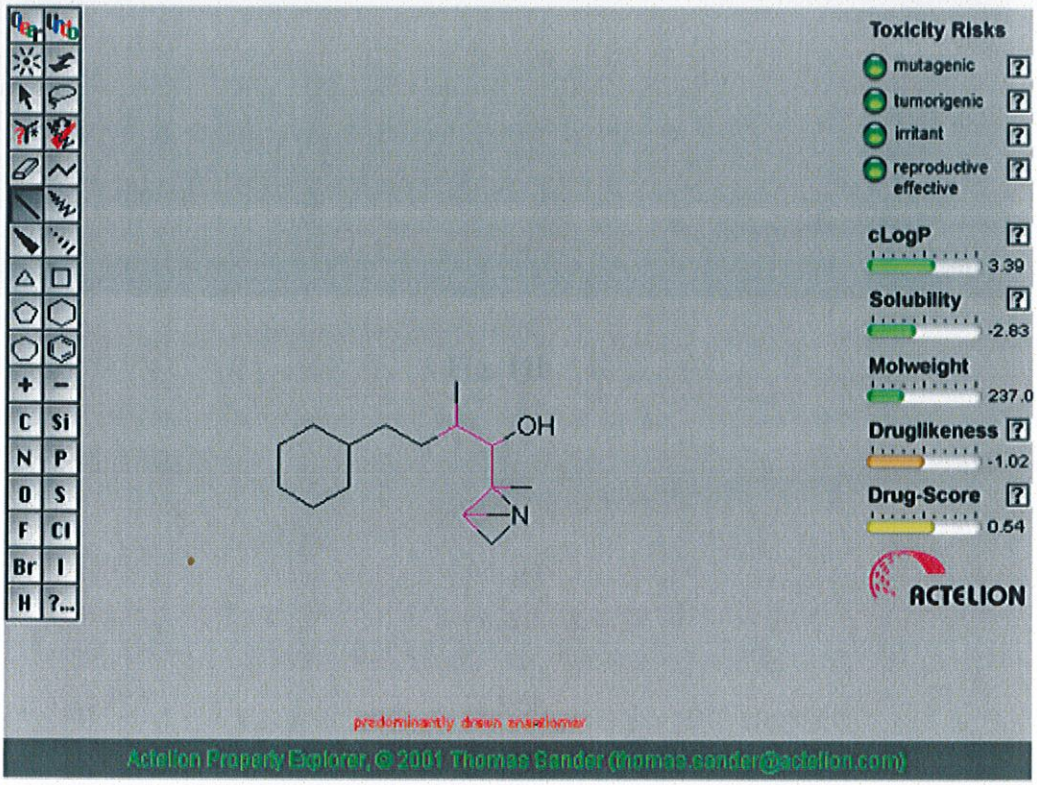


Fig. 11a

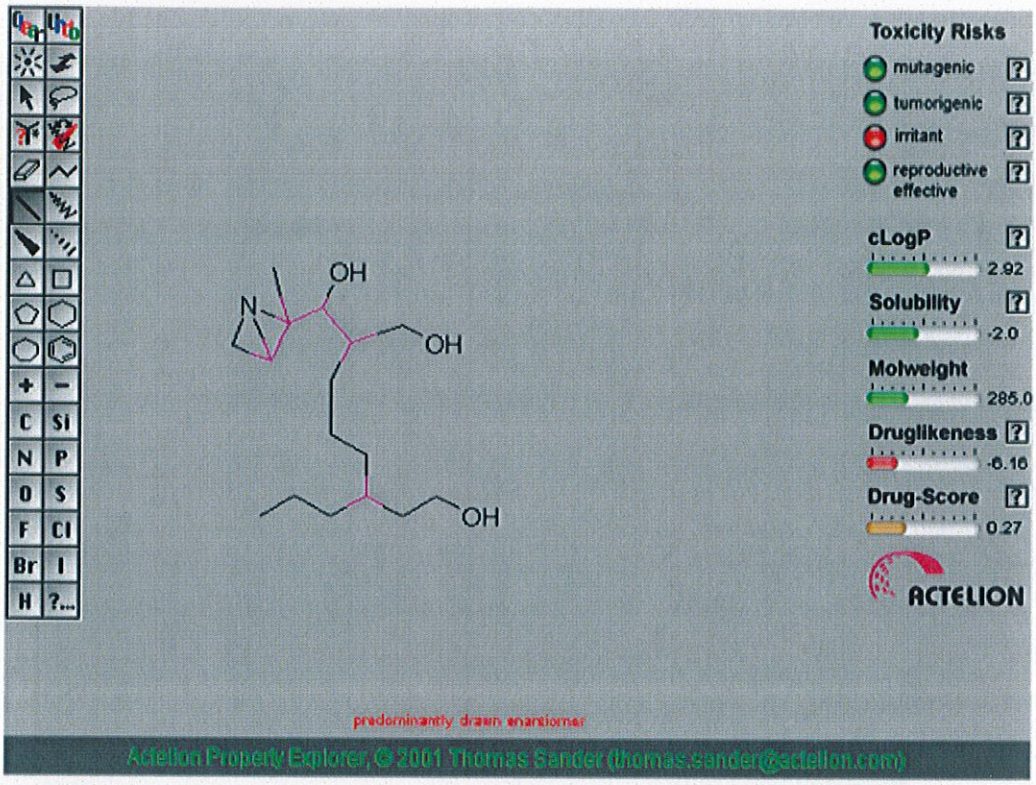

Toxicity Risks
mutagenic
tumorigenic
irritant
reproductive effective
cLogP
2.92
Solubility
-2.0
Molweight
285.0
Druglikeness
-0.16
Drug-Score
0.27
ACTELION
OH
N
OH
OH
predominantly drawn enantiomer
Actelion Property Explorer, © 2001 Thomas Sander (thomas.sander@actelion.com)

Fig. 11b

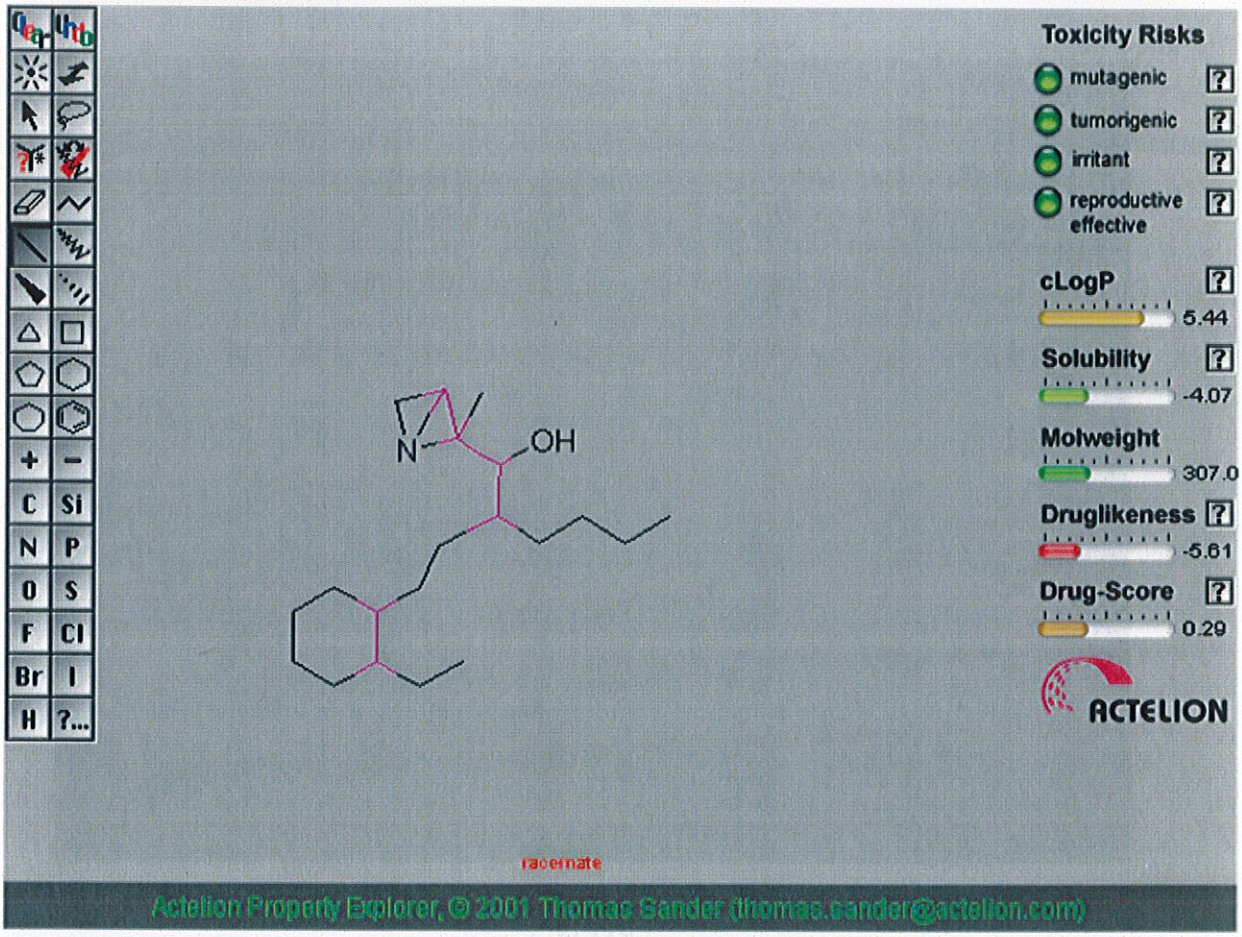

Toxicity Risks
mutagenic
tumorigenic
irritant
reproductive effective
cLogP
5.44
Solubility
-4.07
Molweight
307.0
Druglikeness
-5.61
Drug-Score
0.29
ACTELION
OH
N
C
Si
N
P
O
S
F
Cl
Br
I
H
?...
racemate
Actelion Property Explorer, © 2001 Thomas Sander (thomas.sander@actelion.com)

Fig. 11c

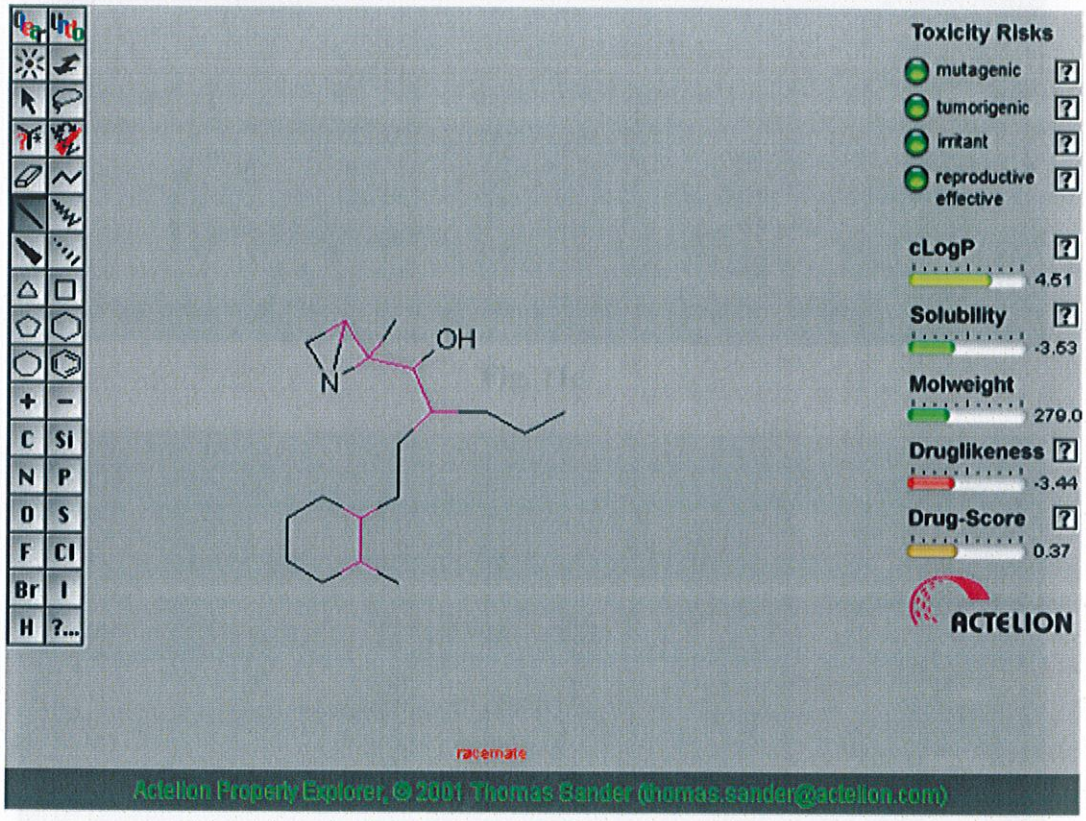

Toxicity Risks
mutagenic
tumorigenic
irritant
reproductive effective
cLogP
4.51
Solubility
-3.53
Molweight
279.0
Druglikeness
-3.44
Drug-Score
0.37
ACTELION
OH
N
C
Si
N
P
O
S
F
Cl
Br
I
H
?...
racemate
Actelion Property Explorer, © 2001 Thomas Sander (thomas.sander@actelion.com)

Fig. 11d

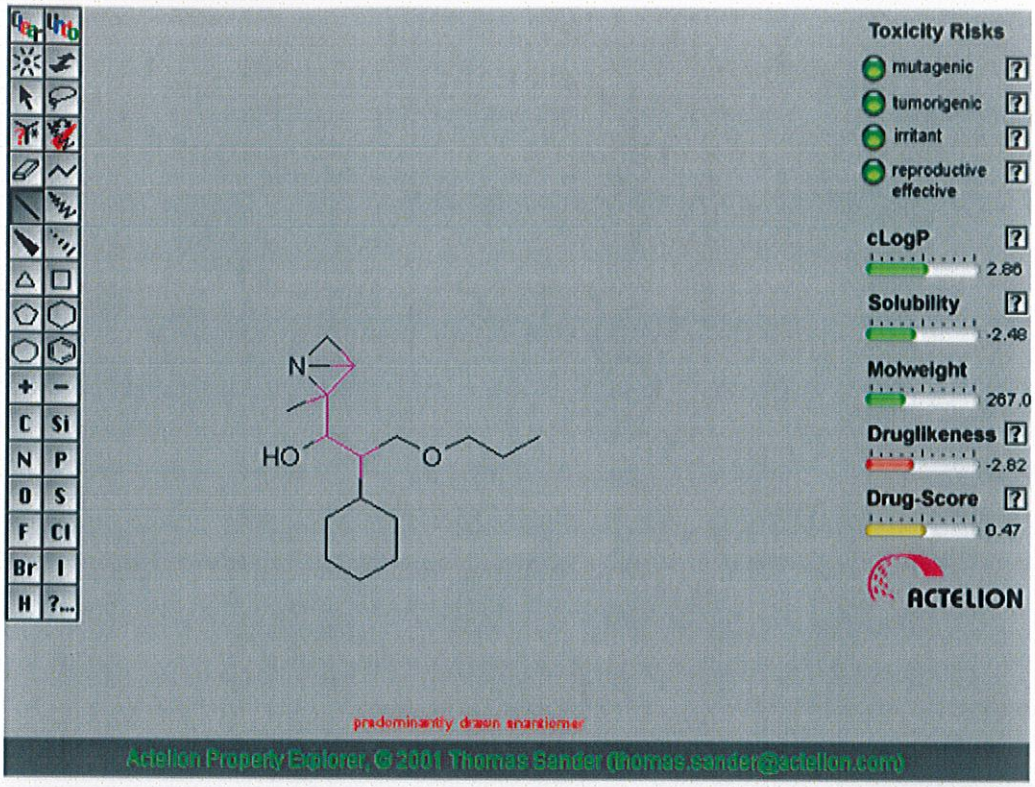


Fig. 11e

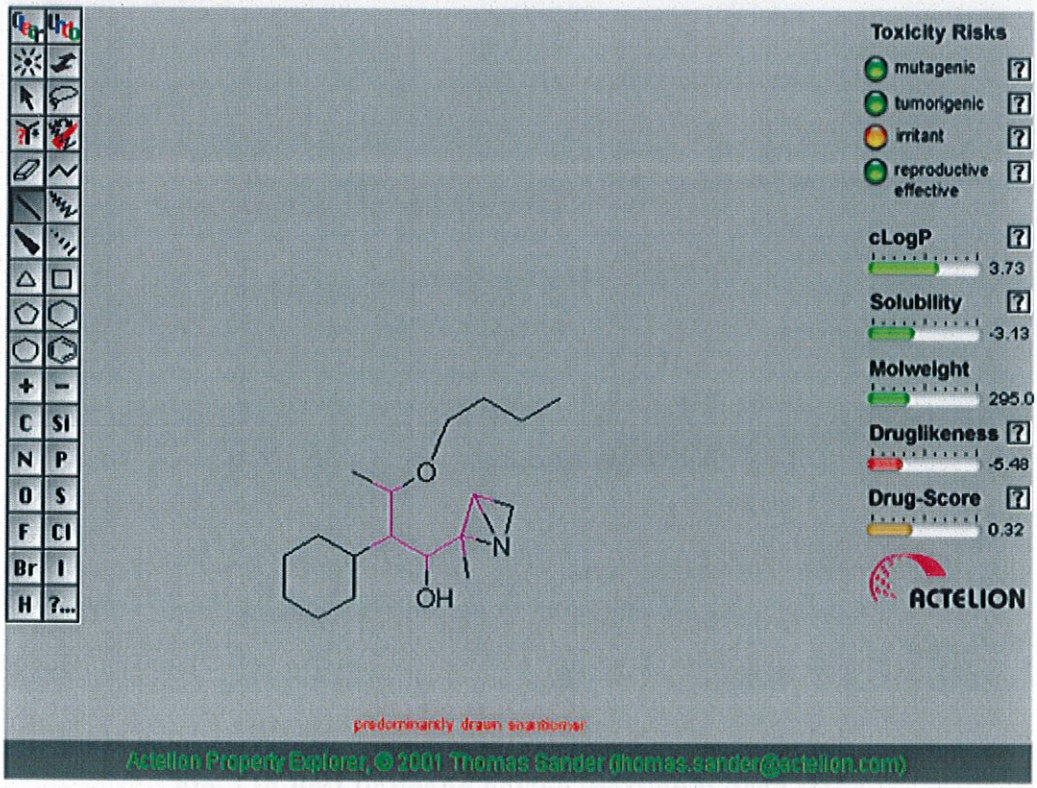


Fig. 11f

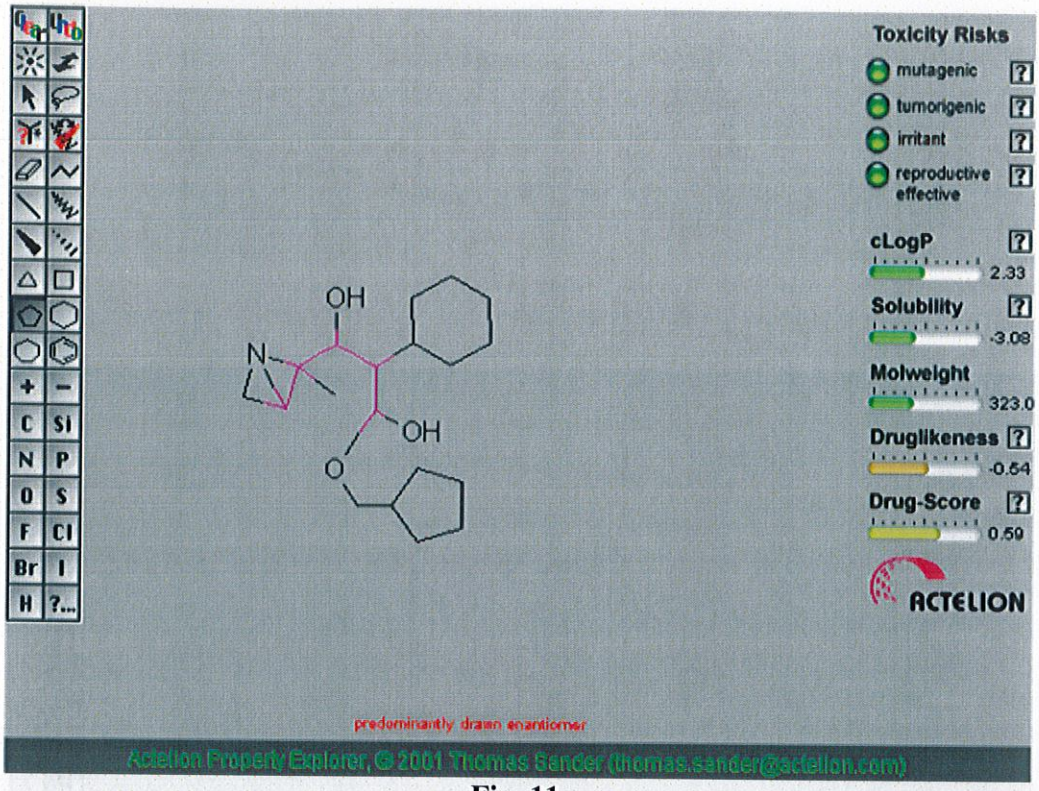


Fig. 11g

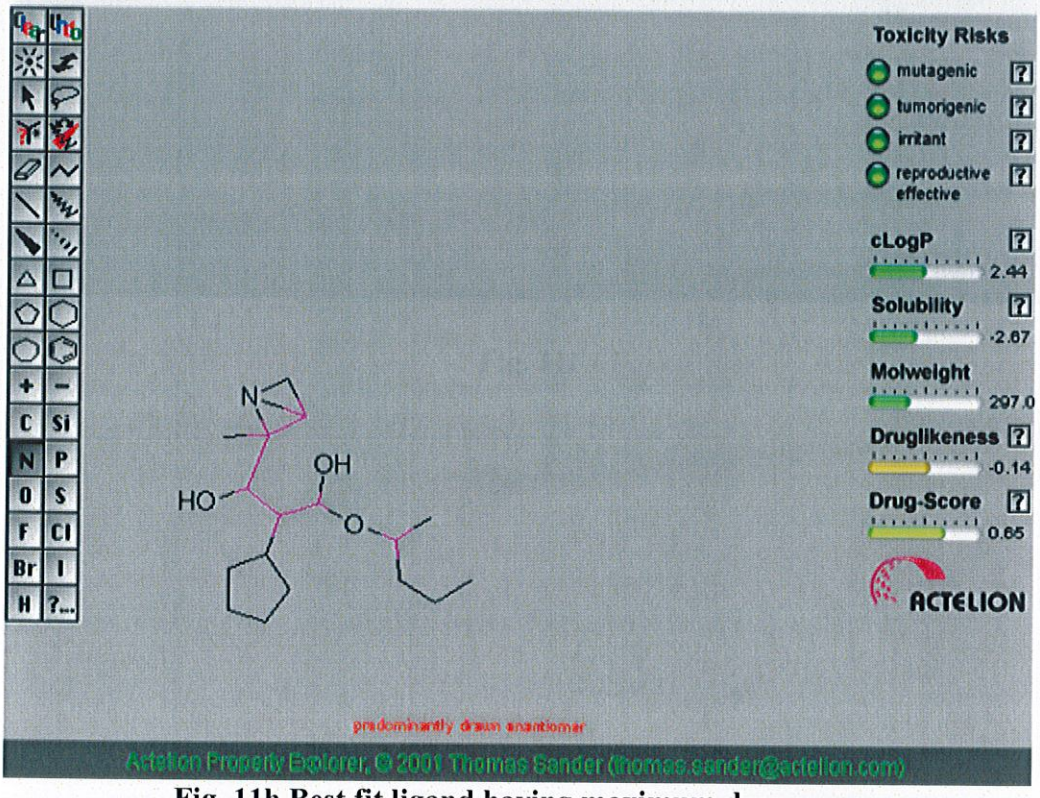


Fig. 11h Best fit ligand having maximum drug score

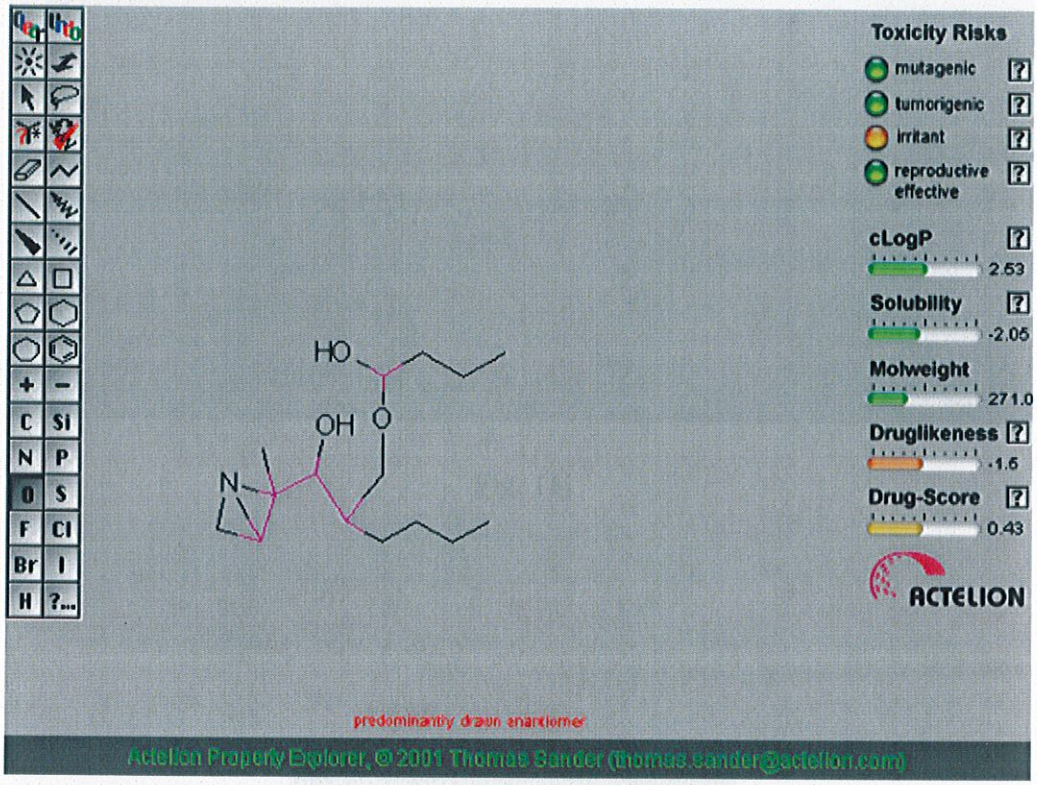

Toxicity Risks
mutagenic
tumorigenic
irritant
reproductive effective
cLogP
2.53
Solubility
-2.05
Molweight
271.0
Druglikeness
-1.6
Drug-Score
0.43
ACTELION
HO
OH
O
N
C
Si
N
P
O
S
F
Cl
Br
I
H
?...
predominantly drawn enantiomer
Actelion Property Explorer, © 2001 Thomas Sander (thomas.sander@actelion.com)

Fig. 11i

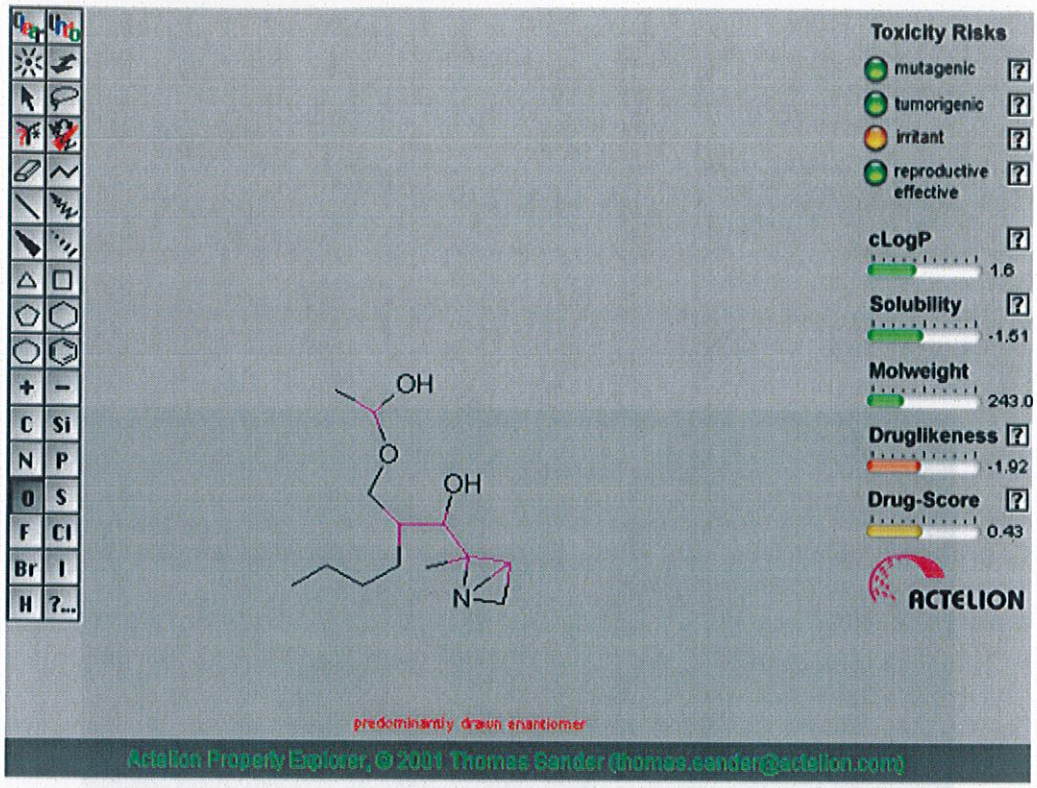


Fig. 11j

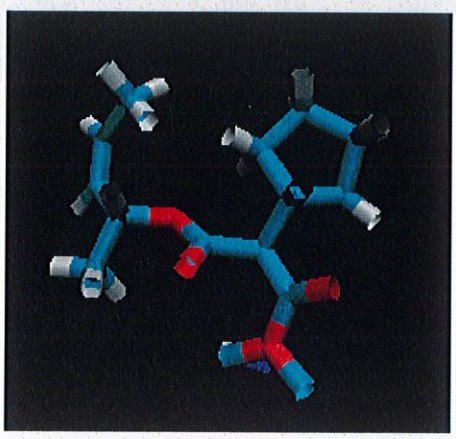


Fig.12 structure of best fit ligand

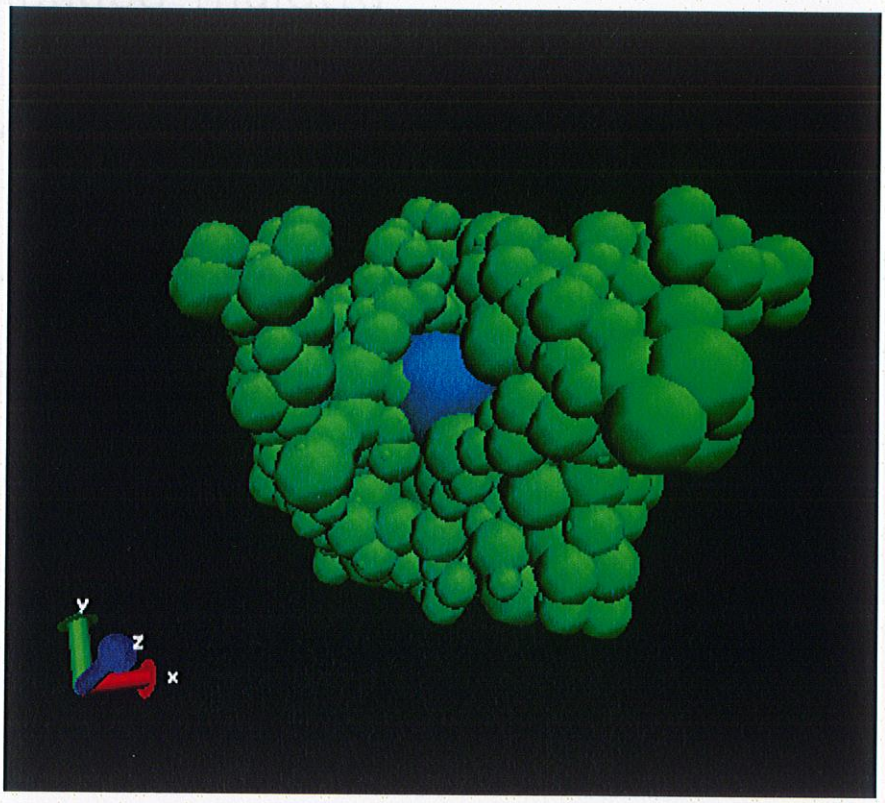


Fig.13 Docking of best fit ligand (blue colour) with target protein (green colour)

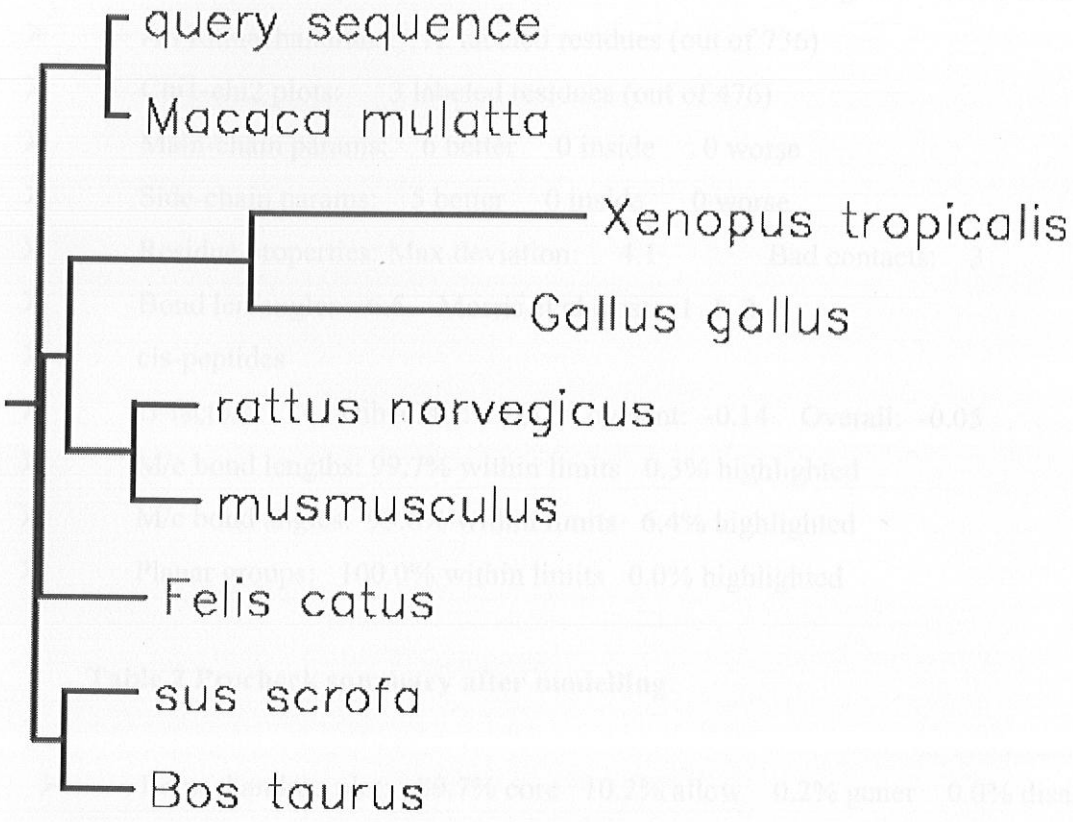


Fig.14 Dendrogram

TABULAR RESULTS

Table 1 Procheck summary before modelling:

- Ramachandran plot: 91.1% core 7.8% allow 0.4% gener 0.7% disall
- All Ramachandrans: 12 labeled residues (out of 736)
- Chi1-chi2 plots: 3 labeled residues (out of 476)
- Main-chain params: 6 better 0 inside 0 worse
- Side-chain params: 5 better 0 inside 0 worse
- Residue properties: Max.deviation: 4.1 Bad contacts: 3
- Bond len/angle: 6.5 Morris et al class: 1 1 2
- cis-peptides
- G-factors Dihedrals: 0.00 Covalent: -0.14 Overall: -0.05
- M/c bond lengths: 99.7% within limits 0.3% highlighted
- M/c bond angles: 93.6% within limits 6.4% highlighted
- Planar groups: 100.0% within limits 0.0% highlighted

Table 2 Procheck summary after modelling:

- Ramachandran plot: 89.7% core 10.2% allow 0.2% gener 0.0% disall
- All Ramachandrans: 13 labeled residues (out of 714)
- Chi1-chi2 plots: 2 labeled residues (out of 476)
- Main-chain params: 6 better 0 inside 0 worse
- Side-chain params: 5 better 0 inside 0 worse
- Residue properties: Max.deviation: 6.2 Bad contacts: 0
- Bond length/angle: 6.1 Morris et al class: 1 1 2
- 3 cis-peptides
- G-factors Dihedrals: -0.17 Covalent: 0.28 Overall: 0.02
- M/c bond lengths: 100.0% within limits 0.0% highlighted
- M/c bond angles: 97.7% within limits 2.3% highlighted
- Planar groups: 75.6% within limits 24.4% highlighted 13 off graph

Table 3 Clustering histogram:

Cluster rank	Lowest docked energy	Run	Mean docked energy	Number in cluster
1	-2.87	10	-2.87	1
2	-2.43	1	-2.43	1
3	-1.63	3	-1.63	1
4	-1.59	8	-1.59	1
5	-1.50	2	-1.50	1
6	-1.35	6	-1.35	1
7	-1.25	9	-1.25	1
8	-1.05	5	-1.05	1
9	-1.03	4	-1.03	1
10	-1.00	7	-100	1

Table 4 RMSD table:

Rank	Sub rank	Run	Docked energy	Cluster RMSD	Reference RMSD
1	1	10	-2.87	0.00	5.43
2	1	1	-2.43	0.00	5.60
3	1	3	-1.63	0.00	3.90
4	1	8	-1.59	0.00	4.75
5	1	2	-1.50	0.00	4.39
6	1	6	-1.35	0.00	4.71
7	1	9	-1.25	0.00	5.26
8	1	5	-1.05	0.00	4.55
9	1	4	-1.03	0.00	3.09
10	1	7	-1.00	0.00	5.60

Table 5 Pairwise alignment:

Sequence type explicitly set to Protein

Sequence format is Pearson:

Sequence 1: query_sequence 759 aa

Sequence 2: Xenopus_tropicalis 751 aa

Sequence 3: sus_scrofa 766 aa

Sequence 4: Macaca_mulatta 766 aa

Sequence 5: Gallus_gallus 759 aa

Sequence 6: Felis_catus 765 aa

Sequence 7: Bos_taurus 765 aa

Sequence 8: rattus_norvegicus 767 aa

Sequence 9: musmusculus 760 aa

Start of Pairwise alignments

Aligning...

Sequences (1:2) Aligned. Score: 58

Sequences (1:3) Aligned. Score: 88

Sequences (1:4) Aligned. Score: 97

Sequences (1:5) Aligned. Score: 63

Sequences (1:6) Aligned. Score: 88

Sequences (1:7) Aligned. Score: 88

Sequences (1:8) Aligned. Score: 84

Sequences (1:9) Aligned. Score: 84

Sequences (2:3) Aligned. Score: 59

Sequences (2:4) Aligned. Score: 59

Sequences (2:5) Aligned. Score: 61

Sequences (2:6) Aligned. Score: 59

Sequences (2:7) Aligned. Score: 59

Sequences (2:8) Aligned. Score: 59

Sequences (2:9) Aligned. Score: 58

Sequences (3:4) Aligned. Score: 87

Sequences (3:5) Aligned. Score: 64

Sequences (3:6) Aligned. Score: 86

Sequences (3:7) Aligned. Score: 90

Sequences (3:8) Aligned. Score: 82

Sequences (3:9) Aligned. Score: 82

Sequences (4:5) Aligned. Score: 64

Sequences (4:6) Aligned. Score: 87

Sequences (4:7) Aligned. Score: 88

Sequences (4:8) Aligned. Score: 84

Sequences (4:9) Aligned. Score: 84

Sequences (5:6) Aligned. Score: 63

Sequences (5:7) Aligned. Score: 63

Sequences (5:8) Aligned. Score: 63

Sequences (5:9) Aligned. Score: 63

Sequences (6:7) Aligned. Score: 86

Sequences (6:8) Aligned. Score: 83

Sequences (6:9) Aligned. Score: 82

Sequences (7:8) Aligned. Score: 83

Sequences (7:9) Aligned. Score: 83

Sequences (8:9) Aligned. Score: 91

Table 6 Clustal distance matrix:

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
query_sequence	(1)	0.00	0.030	0.123	0.120	0.124	0.153	0.154	0.360	0.413
Macaca_mulatta	(2)	0.030	0.000	0.124	0.118	0.127	0.148	0.151	0.362	0.411
sus_scrofa	(3)	0.123	0.124	0.000	0.092	0.135	0.174	0.175	0.362	0.408
Bos_taurus	(4)	0.120	0.118	0.092	0.000	0.135	0.168	0.166	0.371	0.407
Felis_catus	(5)	0.124	0.127	0.135	0.135	0.000	0.160	0.171	0.370	0.410
rattus_norvegicus	(6)	0.153	0.148	0.174	0.168	0.160	0.000	0.082	0.368	0.409
musmusculus	(7)	0.154	0.151	0.175	0.166	0.171	0.082	0.000	0.368	0.412
Gallus_gallus	(8)	0.360	0.362	0.362	0.371	0.370	0.368	0.368	0.000	0.384
Xenopus_tropicalis	(9)	0.413	0.411	0.408	0.407	0.410	0.409	0.412	0.384	0.000

Table 7 CLUSTAL W (1.81) multiple sequence alignment:

Consensus key

* - single, fully conserved residue

: - conservation of strong groups

. - Conservation of weak groups

- No consensus

```
query_sequence      MKTPWKVLLGLLGAAALVTIITVPVVLLNKGTDDATADSRKTYTLTDYLKNTYRLKLYSL
Macaca_mulatta      MKTAWKVLLGLLGAAALVTIITVPVVLLNKGTDDATADSRKTYILTDYLKNTYRLKLYSL
sus_scrofa          MKTPWKVLLGLLGIAALVTVITVPVVLLNKGTDDAAADSRRTYTLTDYLKSTFRVKFYTL
Bos_taurus          MKTPWKVLLGLLAIAALVTVITVPVVLLTK-GNDASTDSRRTYTLADYLKNTFRMKFYNL
Felis_catus         MKTPWKVLLGLLGLAALITIITVPVVLLNKG-NDAAADSRRTYTLTDYLKNTFRVKFYSL
rattus_norvegicus   MKTPWKVLLGLLGVAALVTIITVPVVLLNK--DEAAADSRRTYTLADYLKNTFRVKSYSL
musmusculus         MKTPWKVLLGLLGVAALVTIITVPIVLLSK--DEAAADSRRTYSLADYLKSTFRVKSYSL
Gallus_gallus       MKTLLKWLLGLVGVAVVITVIAVPLALLTG-ESIPESDSRSTYTLENYLNNDYVYKTHNL
Xenopus_tropicalis  MKTWLKWLLGILMGAVVVTVVAVPVALLATKGKE---DTRKTFTLEDYFGDEYRAKSFGL
                      *** * ***:: *.::*:::**:.**   .  *:* *: * :*: . : * . *

query_sequence      RWISDHEYYK-QEN-NILVFNAEYGNSSVFLENSTFDEFGHSINDYSISPDGQFILLEYN
Macaca_mulatta      RWISDHEYLYKQEN-NILVFNAEYGNSSVFLENSTFDEFGHSINDYSISPDGQFILLEYN
sus_scrofa          QWISDHEYLYKQEN-NILLFNAEYGNSSIFLENSTFDELGYSTNDYSVSPDRQFILFEYN
Bos_taurus          RWVSDHEYLYKQEN-NILLFNAEYGNSSIFLENSTFDEFGHSINDYSVSPDRQYILFEYN
Felis_catus         RWVSDHDYLYKQDN-NILLFNAEYGNSSIFLENSTFDEFEHSINDYSVSPDGQFILLEYN
rattus_norvegicus   RWVSDSEYLYKQEN-NILLFNAEHGNSSIFLENSTFEIFGDSISDYSVSPDRLFVLLEYN
musmusculus         WWVSDFEYLYKQEN-NILLLNAEHGNSSIFLENSTFESFG----YHSVSPDRLFVLLEYN
Gallus_gallus       QWISGNQYLHETSNGNILRFDAETGTSSVVLLNTTISIH--EATTAILSPDQRFALLQYK
Xenopus_tropicalis  KWVSENEFLFRTKD-NVLIYNVDNEKTTEMISNTTIYNSN--SSFYTLSDNRNYALLQYN
                      *:* ::  .: *:* :.: .:: .: *:*:       :* : : *::*:

query_sequence      YVKQWRHSYTASYDIYDLNKRQLITEERIPNNTQWVTWSPVGHKLAYVWNNDIYVKIEPN
Macaca_mulatta      YVKQWRHSYTASYDIYDLNKRQLITEERIPNNTQWVTWSPVGHKLAYVWNNDIYVKIEPN
sus_scrofa          YVKQWRHSYTASYDIYDLNKRQLITEERIPNNTQWITWSPVGHKLAYVWNNDIYVKNEPN
Bos_taurus          YVKQWRHSYTASYDIYDLNKRQLITEERIPNNTQWITWSSVGHKLAYVWNNDIYVKNEPN
Felis_catus         YVKQWRHSYTASYDIYDLNKRQLITEEKIPNNTQWITWSPEGHKLAYVWKNDVYVKNEPN
rattus_norvegicus   YVKQWRHSYTASYSIYDLNKRQLITEEKIPNNTQWITWSQEGHKLAYVWKNDIYVKIEPH
musmusculus         YVKQWRHSYTASYNIYDVNKRQLITEEKIPNNTQWITWSPEGHKLAYVWKNDIYVKVEPH
Gallus_gallus       YEKLWRHSYTASYHIYDFNTSSILDDALLPNDTQYISWSPVGHKLAYVWNNNIYIKASPT
Xenopus_tropicalis  YEKLWRHSYTASYHIYDTVKGEIVTANVLPNQIQYITWSPVGNKLAYVWENNIYIKETPG
                      * * ********* *** . .::   :**: *:::** *:******:*::*:*  *

query_sequence      LPSYRITWTGKEDIIYNGITDWVYEEEVF-AYSALWWSPNGTFLAYAQFNDTEVPLIEYS
Macaca_mulatta      LPGHRITSTGKEDMIYNGITDWVYEEEVFSAYSALWWSPNGTFLAYAQFNDTEVPLIEYS
sus_scrofa          LSSQRITWTGKENVIYNGVTDWVYEEEVFSAYSALWWSPNGTFLAYAQFNDTEVPLIEYS
Bos_taurus          SPSQRITWTGKKDVIYNGITDWVYEEEVFSAYSALWWSPNSTFLAYAQFNDTEVPLIEYS
Felis_catus         SSSHRITWTGEENAIYNGIADWVYEEEIFSAYSALWWSPKGTFLAYAQFNDTQVPLIEYS
rattus_norvegicus   LPSHRITSTGKENVIFNGINDWVYEEEIFGAYSALWWSPNGTFLAYAQFNDTGVPLIEYS
musmusculus         LPSHRITSTGEENVIYNGITDWVYEEEVFGAYSALWWSPNNTFLAYAQFNDTGVPLIEYS
Gallus_gallus       AAPVQITSNGEENKIFNGIPDWVYEEEMFGSHSALWWSPNGNFVAYAAFNDTEVPVIEYS
Xenopus_tropicalis  GSSIQITTNGEHNKILNGIPDWVYEEEMFSTNYALWWSPDAASLAYVEFNDTDVPAIEYS
                      . :** .*:.: * **: *******:*: ******.  :**. **** ** ****
```

```
query_sequence      FYSDESLQYPKTVRVPYPKAGAVNPTVKFFVVNTDSLSSVTNATSIQITAPASMLIGDHY
Macaca_mulatta      FYSDESLQYPKTVQVPYPKAGAVNPTVKFFVVNTDSLSSATNATSIQITAPASMLIGDHY
sus_scrofa          FYSDESLQYPKTVRIPYPKAGAENPTVKFFVVDTRTLSPNASVTSYQIVPPASVLIGDHY
Bos_taurus          FYSDESLQYPKTVKIPYPKAGAVNPTIKFFVVNISSLSPNINATSQQIVPPGSVLIGDHY
Felis_catus         FYSDESLQYPMTMRIPYPKAGAANPTVKLFVIKTDNLNPNTNATSVEITPPAAMLTGDYY
rattus_norvegicus   FYSDESLQYPKTVWIPYPKAGAVNPTVKFFIVNTDSLSSTTTTIPMQITAPASVTTGDHY
musmusculus         FYSDESLQYPKTVWIPYPKAGAVNPTVKFFIVNIDSLSSSSSAAPIQIPAPASVARGDHY
Gallus_gallus       FYSEDTLQYPKTIRIPYPKAGAKNPTVKFFIVDIQMLPDFN---STEISPPAEIKSGDHY
Xenopus_tropicalis  FYGEDSDQYPKTVVIPYPKAGAKNPTVRLFVVNTASLPAFN---SVEIPPPAELRTIDYY
                      **.::: *** *: :******* ***:::*::  *     .:* .*. :  *:*

query_sequence      LCDVTWATQERISLQWLRRIQNYSVMDICDYDESSGRWNCLVARQHIEMSTTGWVGRFRP
Macaca_mulatta      LCDVTWATQERISLQWLRRIQNYSVMDICDYDESSGRWNCLVARQHIETSTTGWVGRFRP
sus_scrofa          LCGVTWVTEERISLQWIRRAQNYSIIDICDYDESTGRWISSVARQHIEISTTGWVGRFRP
Bos_taurus          LCDVTWVTEERISLQWLRRIQNYSIMDICDYDRSTGRWISSVGRQHIEISTTGWVGRFRP
Felis_catus         LCDVTWANEERISLQWLRRIQNYSVMDIRDYNNSTGKWISSAAQEHIEMSTTGWVGRFRP
rattus_norvegicus   LCDVAWVSEDRISLQWLRRIQNYSVMAICDYDKTTLVWNCPTTQEHIETSATGWCGRFRP
musmusculus         LCDVVWATEERISLQWLRRIQNYSVMAICDYDKINLTWNCPSEQQHVEMSTTGWVGRFRP
Gallus_gallus       LSVVTWVTDERICLQWLRRIQNYSVLTICDFESATGNWTCHR-KNNFWKKVQLAGWQISA
Xenopus_tropicalis  INGMNWVTNGKMAVQWLRRIQNVSLLTVCDFAAPAWNCQPP----VYEQSTTGWVGYFQP
                      : : *.: ::.:**:** ** *:: : *:          ..    : .

query_sequence      SEPHFTLDGNSFYKIISNEEGYRHICYFQIDKKD---CTFITKGTWEVIGIEALTSDYLY
Macaca_mulatta      SEPHFTSDGNSFYKIISNEEGYRHICYFQINKKN---CTFITKGAWEVIGIEALTSDYLY
sus_scrofa          AEPHFTSDGNSFYKIISNEEGYKHICHFQTDKSN---CTFITKGAWEVIGIEALTSDYLY
Bos_taurus          AEPHFTSDGNSFYKIISNEEGYKHICHFQTDKRN---CTFITKGAWEVIGIEALTSDYLY
Felis_catus         AEPHFTSDGRNFYKIISNEDGYKHICRFQIDKKD---CTFITKGAWEVIGIEALTTDYLY
rattus_norvegicus   AEPHFTSDGSSFYKIVSDKDGYKHICQFQKDRKPEQVCTFITKGAWEVISIEALTSDYLY
musmusculus         AEPHFTSDGSSFYKIISDKDGYKHICHFPKDKKD---CTFITKGAWEVISIEALTSDYLY
Gallus_gallus       ICPLFAPDNTTYYKVFSNTEGYKHIHYINGTEAP----VPITEGKWEVISIAAVTKYFLY
Xenopus_tropicalis  SSPYFTQDGQKYYKVISNKEGYKHIHLFEGAKEP----VAITSGKWEVTSITTVASNLLY

query_sequence      YISNEYKGMPGGRNLYKIQLSDYTK-VTCLSCELNPERCQYYSVSFSKEAKYYQLRCSGP
Macaca_mulatta      YISNEYKGMPGGRNLYKIQLSDYTK-VTCLSCELNPERCQYYSVSFSKEAKYYQLRCSGP
sus_scrofa          YISNEHKGMPGGRNLYRIQLNDYTK-VTCLSCELNPERCQYYSASFSNKAKYYQLRCFGP
Bos_taurus          YISNEYKGMPGARNLYKIQLNDYTK-VTCLSCELNPDRCQYYSVSFSQEAKYYQLRCSGP
Felis_catus         YISNEYKGMPGGRNLYKIQLNDYTK-VACLSCELKPERCQYYSVSFSKEAKYYQLRCSGP
rattus_norvegicus   YISNEYKEMPGGRNLYKIQLTDHTN-KKCLSCDLNPERCQYYSVSLSKEAKYYQLGCRGP
musmusculus         YISNQYKEMPGGRNLYKIQLTDHTN-VKCLSCDLNPERCQYYAVSFSKEAKYYQLGCWGP
Gallus_gallus       YISNQNGEMPGGRNLYKMLLESSPKSTQCVSCDLNQERCQYYSASFSKDAQYYQLNCLGP
Xenopus_tropicalis  YISNE--GFPGRRQLYKIRLDGGSNSAQCVTCNIRQERCQQYSAYFSKDSKYYALNCNGP
                      ****:  :** *:**:: * . .:  *:.*:.. :*** *:. :*:.::** * * **

query_sequence      GLPLYTLHSSVNDG-LRVLEDNSALDKMLQNVQMPSKKLDFIILNETKFWYQMILPPHFD
Macaca_mulatta      GLPLYTLHSSVNDKGPRVLEDNSALDKMLQNVQMPSKTLDFIILNETKFWYQMILPPHFD
sus_scrofa          GLPLYTLHSSSSDKELRVLEDNSALDKMLQDVQMPSKKLDVINLHGTKFWYQMILPPHFD
Bos_taurus          GLPLYTLHNSNNDKELRVLENNSDLDQVLQDVQMPSKKLDFIHLHGTKFWYQMILPPHFD
Felis_catus         GLPLYTLHRSSNDEELRVLEDNSALDKMLQEVQMPSKKLDFIILNETKFWYQMILPPHFD
rattus_norvegicus   GLPLYTLHRSTDQKELRVLEDNSALDKMLQDVQMPSKKLDFIVLNETRFWYQMILPPHFD
musmusculus         GLPLYTLHRSTDHKELRVLEDNSALDRMLQDVQMPSKKLDFIVLNETRFWYQMILPPHFD
Gallus_gallus       GLPMSTLHRSSDDQVLRYLENNTELENSLKDIQMPSKKLGSITVGGYNLWYQMILPPHLD
Xenopus_tropicalis  GVPIYTLYNSSNDKEMHIMEDNEELKLLLDEVQMPTKENKSIIIDGFELWYQLTLPPNFD
                      *:*: **: * ..  : :*:*  *.  *.::***:*   * :  .:***: ***:.*
```

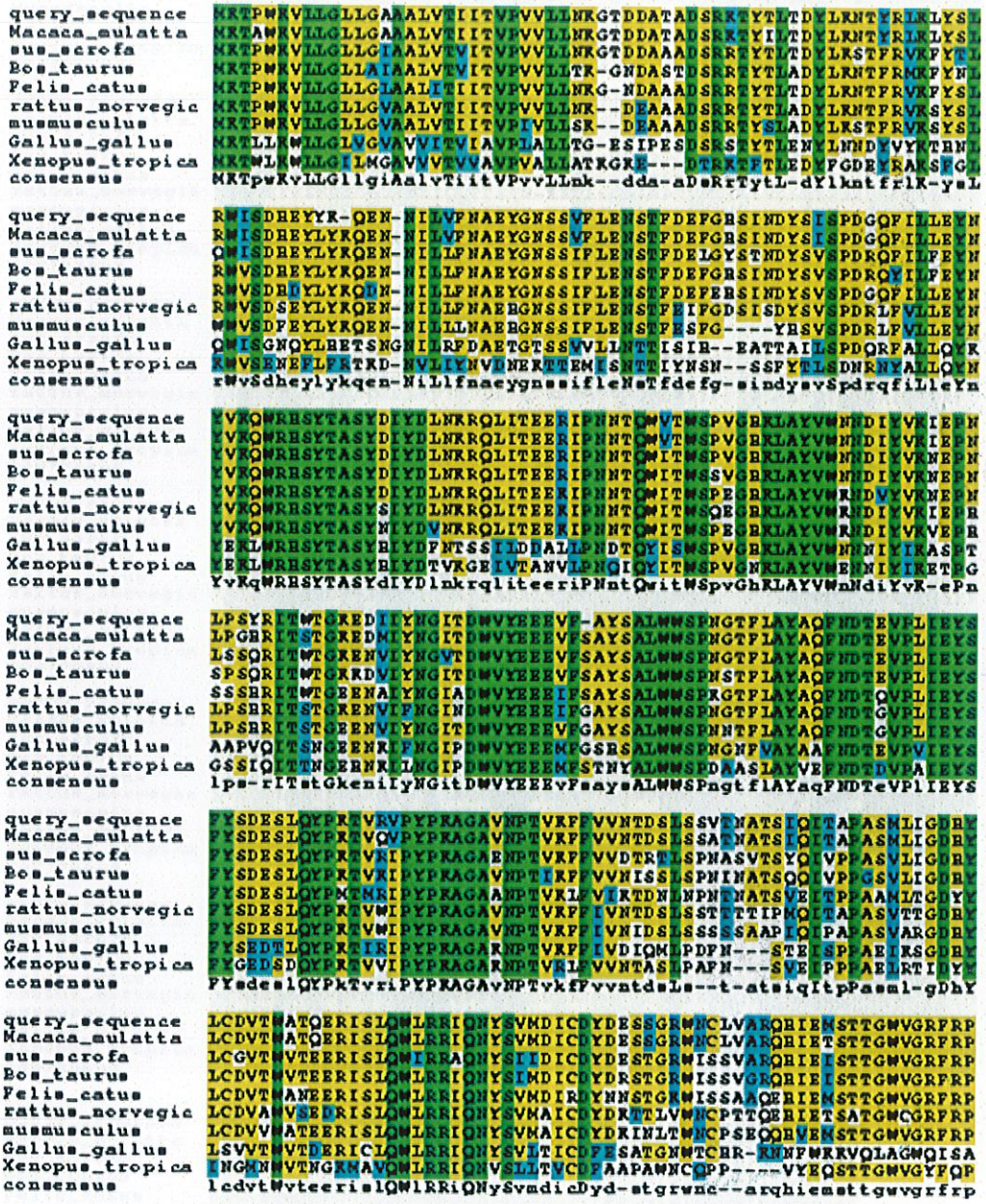
```
query_sequence      KSKKYPLLLDVYAGPCSQKADTVR-LNWATYLASTENIIVASFDGRGSGYQGDKIMHAIN
Macaca_mulatta      KSKKYPLLLDVYAGPCSQKADAVFRLNWATYLASTENIIVASFDGRGSGYQGDKIMHAIN
sus_scrofa          KSKKYPLLIEVYAGPCSQKVDTVFRLSWATYLASTENIIVASFDGRGSGYQGDKIMHAIN
Bos_taurus          KSKKYPLLLEVYAGPCSQKADAIFRLNWATYLASTENIIVASFDGRGSGYQGDKIMHAIN
Felis_catus         TSKKYPLLIDVYAGPCSQKADAIFRLNWATYLASTENIIVASFDGRGSGYQGDKIMHAVN
rattus_norvegicus   KSKKYPLLIDVYAGPCSQKADAAFRLNWATYLASTENIIVASFDGRGSGYQGDKIMHAIN
musmusculus         KSKKYPLLLDVYAGPCSQKADASFRLNWATYLASTENIIVASFDGRGSGYQGDKIMHAIN
Gallus_gallus       SSKKYPLLLEVYAGPCSQKVDHVFRINWATYLASTEQIIVASFDGRGSGYQGDEIMHAIN
Xenopus_tropicalis  KSKKYPLLIDVYGGPCSQKVDQYFRLNWATYLASTEKIIVASFDGRGSGYQGDKIMHQLY
                      .*******::**.******.*   :.*********:****************.***:

query_sequence      RRLGTFEVEDQIEAARQFSKMGFVDNKRIAIWGS-YGGYVTSMVLGSGSGVFKCGIAVAP
Macaca_mulatta      RRLGTFEVEDQIEAARQFSKMGFVDNKRIAIWGWSYGGYVTSMVLGSGSGVFKCGIAVAP
sus_scrofa          RRLGTFEVEDQIEATRQFSKMGFVDDKRIAIWGWSYGGYVTSMVLGAGSGVFKCGIAVAP
Bos_taurus          RRLGTFEVEDQIEATRQFSKMGFVDDKRIAIWGWSYGGYVTSMVLGAGSGVFKCGIAVAP
Felis_catus         RRLGTFEVEDQIEAARQFSKMGFVDDKRIAIWGWSYGGYVTSMVLGAGSGVFKCGIAVAP
rattus_norvegicus   KRLGTLEVEDQIEAARQFLKMGFVDSKRVAIWGWSYGGYVTSMVLGSGSGVFKCGIAVAP
musmusculus         RRLGTLEVEDQIEAARQFVKMGFVDSKRVAIWGWSYGGYVTSMVLGSGSGVFKCGIAVAP
Gallus_gallus       RRLGTYEVEDQISAARTFSEMSFVDKDRIAIWGWSYGGYVTSMVLGSGSGVFKCGIAVAP
Xenopus_tropicalis  HKLGTLEVQDQIAAARYFITLGFVDPKRIAIWGWSYGGYVTSMVLGSGSGLFKCGIAVAP
                      ::*** **:*** *:* * :.*** .*:**** ***********:***:*********

query_sequence      VSRWEYYDSVYTERYMGLPTPEDNLDHYRNSTVMSRAENFK--EYLLIHGTADDNVHFQQ
Macaca_mulatta      VSRWEYYDSVYTERYMGLPTPEDNLDHYRNSTVMSRAENFKQVEYLLIHGTADDNVHFQQ
sus_scrofa          VSKWEYYDSVYTERYMGLPTPEDNLDYYRNSTVMSRAENFKQVEYLLIHGTADDNVHFQQ
Bos_taurus          VSKWEYYDSVYTERYMGLPTPEDNLDSYRNSTVMSRAENFKQVEYLLIHGTADDNVHFQQ
Felis_catus         VSRWEYYDSVYTERYMGLPTPQDNLDYYKNSTVMSRAENFKQVEYLLIHGTADDNVHFQQ
rattus_norvegicus   VSRWEYYDSVYTERYMGLPTPEDNLDHYRNSTVMSRAENFKQVEYLLIHGTADDNVHFQQ
musmusculus         VSRWEYYDSVYTERYMGLPIPEDNLDHYRNSTVMSRAEHFKQVEYLLIHGTADDNVHFQQ
Gallus_gallus       VSRWQYYDSIYTERYMGLPTESDNLRNYNSSTVMARAEKFKEVEYLLIHGTADDNVHFQQ
Xenopus_tropicalis  VSSWEYYDSIYTERYMGLPTPEDNLENYLSSTVMARAQKFKDVEYLLIHGTADDNVHFQQ
                      ** *:****:********* .*** * .****:**::** *****************

query_sequence      SAQISKALVDVGVDFQAMWYTDEDHGIASSTAHQHIYTHMSHFIKQCFSLP
Macaca_mulatta      SAQISKALVDAGVDFQAMWYTDEDHGIASSTAHQHTYTHMSHFIKQCFSLP
sus_scrofa          SAQLSKALVDAGVDFQTMWYTDEDHGIASNMAHQHIYTHMSHFLKQCFSLP
Bos_taurus          SAQISKALVDAGVDFQSMWYTDEDHGIASSTAHQHIYTHMSHFLKQCFSLL
Felis_catus         SAQISKALVDAGVDFQAMWYTDEDHGIASGPAHQHIYTHMSHFIKQCFSLP
rattus_norvegicus   SAQISKALVDAGVDFQAMWYTDEDHGIASSTAHQHIYSHMSHFLQQCFSLR
musmusculus         SAQISKALVDAGVDFQAMWYTDEDHGIASSTAHQHIYSHMSHFLQQCFSLH
Gallus_gallus       AAQISKALVDAEVDFQAMWYTDKDHGISG-QAHKHIYTHMSHFIKQCFSLP
Xenopus_tropicalis  AAHISKALVDAQVDFDTMWYTDKDHGIG-GTANRHIYTHMSHYLKQCFNIQ
                    :*::******. ***::*****:****.  *::* *:****:::***.:
```

Table 8 Boxshade:



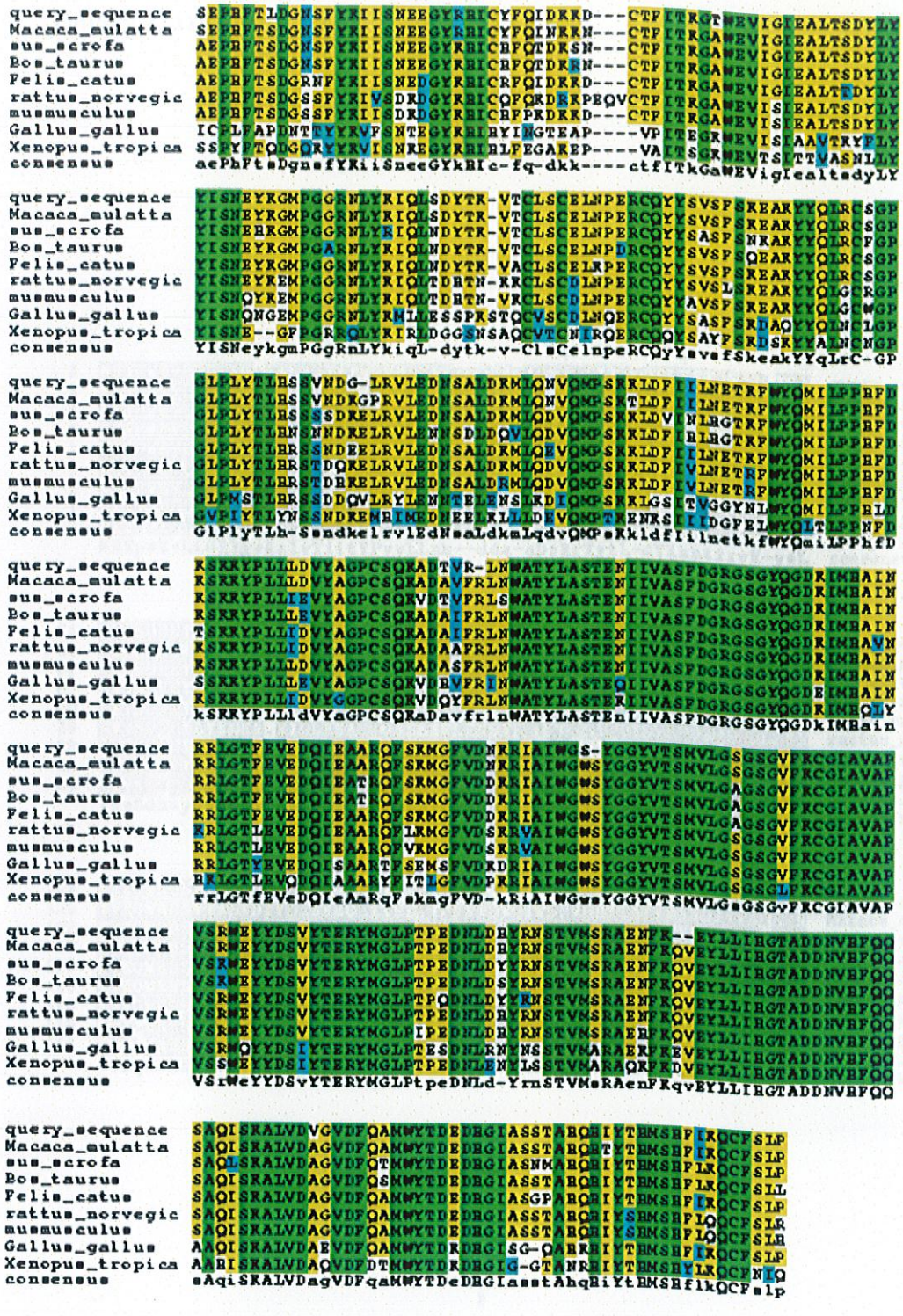
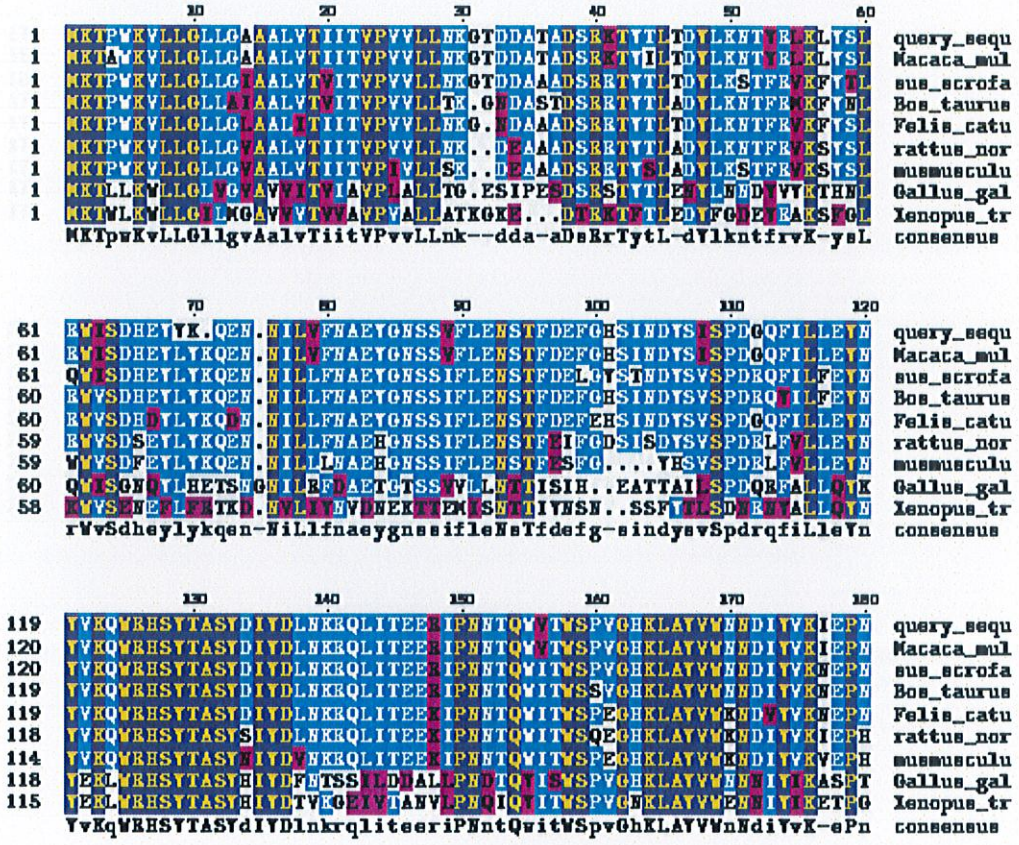
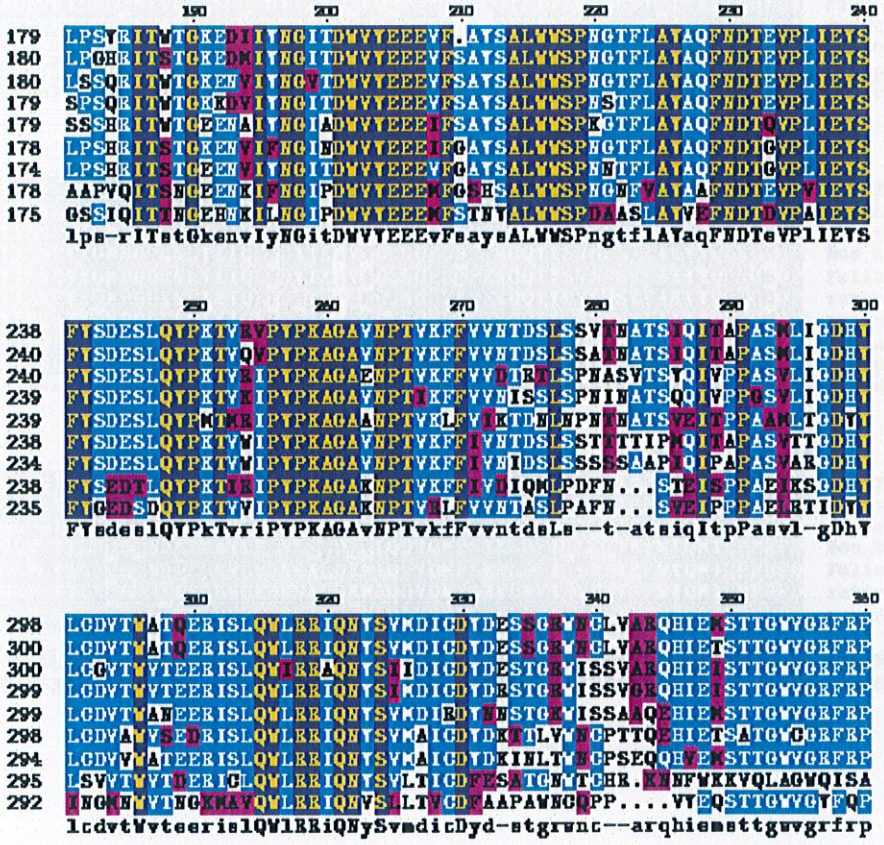
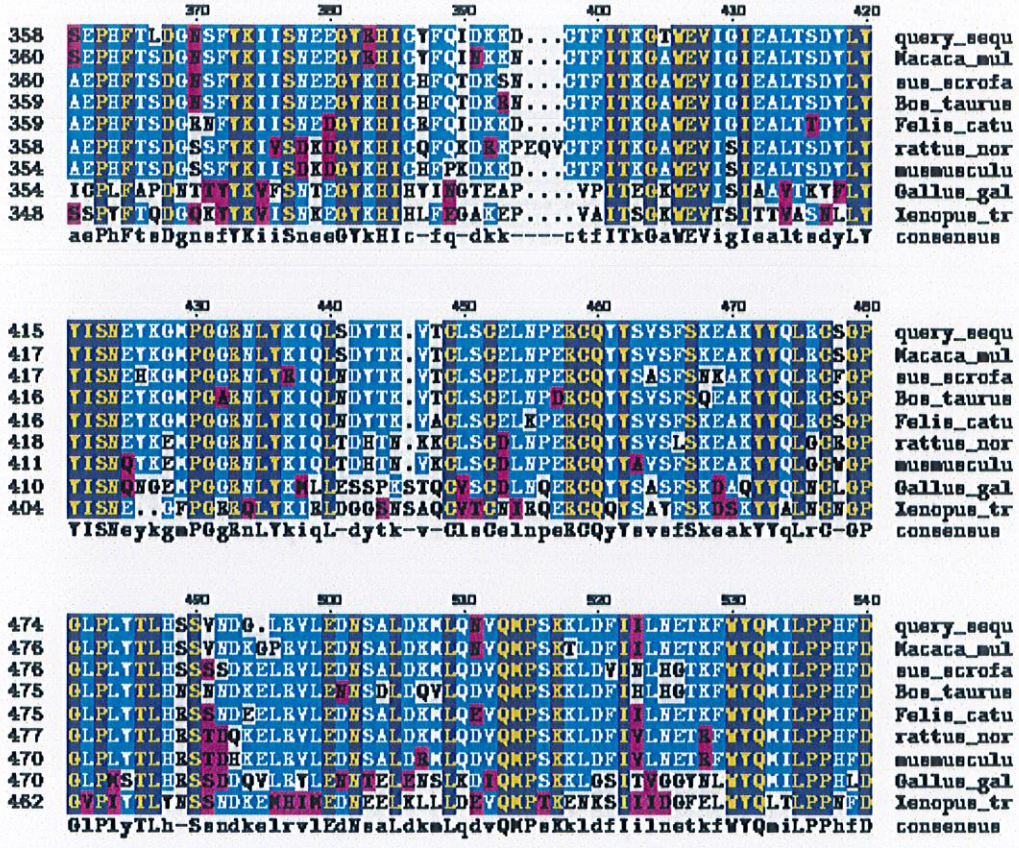
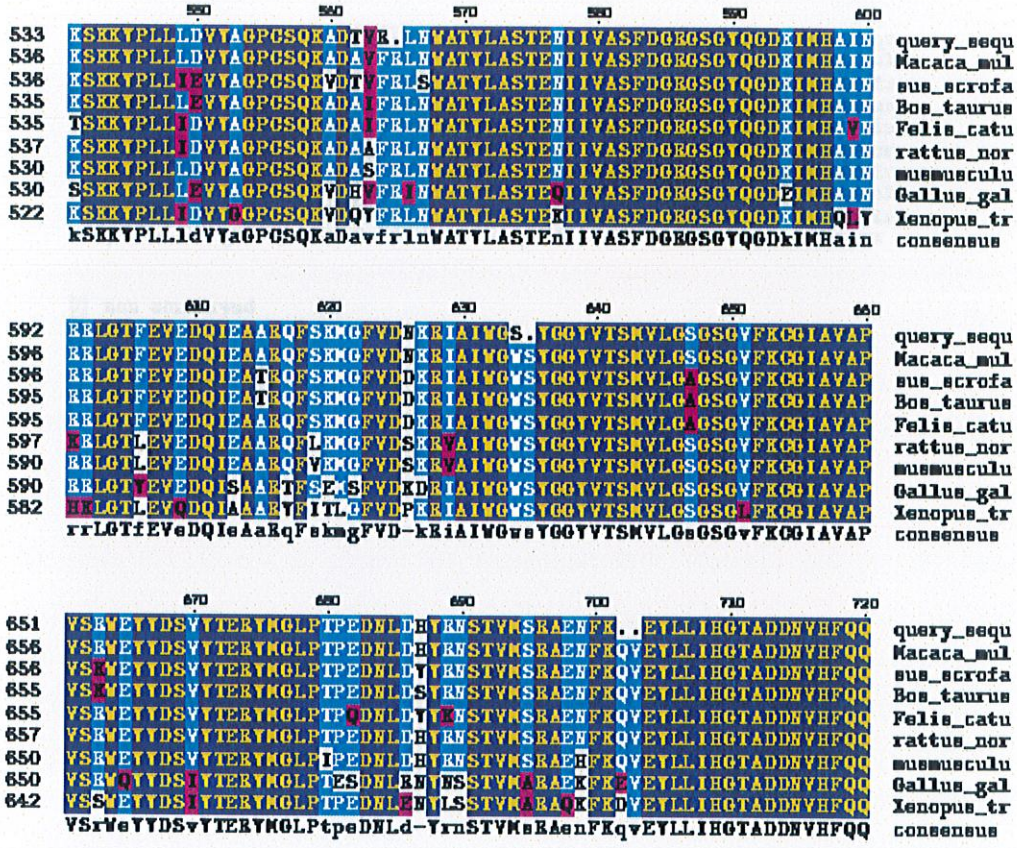

query_sequence
Macaca_mulatta
sus_scrofa
Bos_taurus
Felis_catus
rattus_norvegicus
musmusculus
Gallus_gallus
Xenopus_tropicalis
consensus

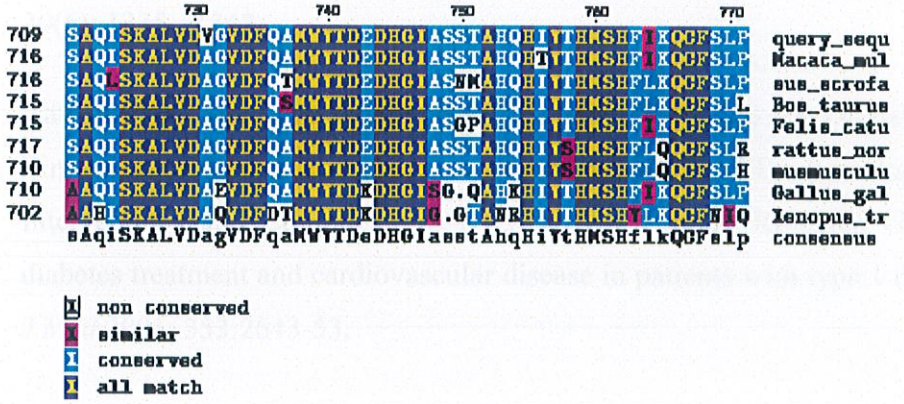
Table 9 Texshade:




query_sequ
Macaca_mul
sus_scrofa
Bos_taurus
Felis_catu
rattus_nor
musmusculu
Gallus_gal
Xenopus_tr
consensus




query_sequ
Macaca_mul
sus_scrofa
Bos_taurus
Felis_catu
rattus_nor
musmusculu
Gallus_gal
Xenopus_tr
consensus


730
740
750
760
770
709 SAQISKALVDVGVDFQAMVTTDEDHGIASSTAHQHIYTHMSHFIKQGFSLP query_sequ
716 SAQISKALVDAGVDFQAMVTTDEDHGIASSTAHQHTYTHMSHFIKQGFSLP Macaca_mul
716 SAQLSKALVDAGVDFQTMVTTDEDHGIASNMAHQHIYTHMSHFLKQGFSLP sus_scrofa
715 SAQISKALVDAGVDFQSMVTTDEDHGIASSTAHQHIYTHMSHFLKQGFSLL Bos_taurus
715 SAQISKALVDAGVDFQAMVTTDEDHGIASGPAHQHIYTHMSHFIKQGFSLP Felis_catu
717 SAQISKALVDAGVDFQAMVTTDEDHGIASSTAHQHIYSHMSHFLQQGFSLR rattus_nor
710 SAQISKALVDAGVDFQAMVTTDEDHGIASSTAHQHIYSHMSHFLQQGFSLH musmusculu
710 AAQISKALVDAEVDFQAMVTTDKDHGISG.QAHKHIYTHMSHFIKQGFSLP Gallus_gal
702 AAHISKALVDAQVDFDTMVTTDKDHGIG.GTANRHIYTHMSHYLKQGFHIQ Xenopus_tr
sAqiSKALVDagVDFqaMVTTDeDHGIasstAhqHiYtHMSHflkQGFslp consensus
non conserved
similar
conserved
all match

BIBLIOGRAPHY

D. J. Drucker., Dipeptidyl Peptidase-4 Inhibition and the Treatment of Type 2 Diabetes: Preclinical biology and mechanisms of action. Diabetes Care, June 1, 2007; 30(6): 1335 - 1343.

Nathan DM, Cleary PA, Backlund JY, Genuth SM, Lachin JM, Orchard TJ, Raskin P, Zinman B; Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications (DCCT/EDIC) Study Research Group. Intensive diabetes treatment and cardiovascular disease in patients with type 1 diabetes. *N Engl J Med* 2005; 353:2643-53.

Von Mehring J, Minkowski O. (1890). "Diabetes mellitus nach pankreasexstirpation." *Arch Exp Pathol Pharmakol* **26**: 371-387.

Banting FG, Best CH, Collip JB, Campbell WR, Fletcher AA (1922). "Pancreatic extracts in the treatment of diabetes mellitus". *Canad Med Assoc J* **12**: 141–146.

Him worth (1936). "*Diabetes mellitus: its differentiation into insulin-sensitive and insulin-insensitive types".* Lancet ***i****: 127–130.*

Vinik AI, Fish wick DT, Pittenger G. Advances in diabetes for the millennium: toward a cure for diabetes. MedGenMed 2004; 6:12.

Mentlein **R** 1999 Dipeptidyl-peptidase IV (CD26): role in the inactivation of regulatory peptides. Regul Pept 85:9–24.

Pacini G, Thomaseth K, Ahrén B 2001 Contribution to glucose intolerance of insulin-independent vs. insulin-dependent mechanisms in mice. Am J Physiol 281:E693–E703.

Pro B, Dang N (2004). "CD26/dipeptidyl peptidase IV and its role in cancer". *Histol Histopathol* **19** (4): 1345–51.

McIntosh CH, Demuth HU, Pospisilik JA, Pederson R. Dipeptidyl peptidase IV inhibitors: how do they work as new antidiabetic agents? Regul Pept. 2005 Jun 15; 128(2):159-65. PMID: 15780435.

Masur K et al. (2006). "DPPIV inhibitors extend GLP-2 mediated tumour promoting effects on intestinal cancer cells". *Regul Pept.* **137** (3): 147-55.

Wesley UV et al. (2005). "Dipeptidyl peptidase inhibits malignant phenotype of prostate cancer cells by blocking basic fibroblast growth factor signaling pathway". *Cancer Res.* **65** (4): 1325-3.

Deadman, J. (2000) Proteinase inhibitors and activators strategic targets for therapeutic intervention. J. Pept. Sci. **6**, 421 .

Southan, C. (2001) A genomic perspective on human proteases as drug targets. Drug Discov. Today **6**, 681–688.

Mentlein, R. (1999) Dipeptidyl-peptidase IV (CD26) – role in the inactivation of regulatory peptides. Regul. Pept. **85**, 9–24.

Yaron, A. and Naider, F. (1993) Proline-dependent structural and biological properties of peptides and proteins. Crit. Rev. Biochem. Mol. Biol. **28**, 31–81.

Rawlings, N. D., Polgar, L. and Barrett, A. J. (1991) A new family of serine-type peptidases related to prolyl oligopeptidase. Biochem. J. **279**, 907–908.

Polgar, L. (2002) The prolyl oligopeptidase family. Cell. Mol. Life Sci. **59**, 349.

Hansotia T, Baggio LL, Delmeire D, Hinke SA, Yamada Y, Tsukiyama K, Seino Y, Holst JJ, Schuit F, Drucker DJ: Double incretin receptor knockout (DIRKO) mice reveal an essential role for the enteroinsular axis in transducing the glucoregulatory actions of DPP-IV inhibitors. *Diabetes* 53:1326–1335, 2004.

Ahren B, Hughes TE: Inhibition of dipeptidyl peptidase-4 augments insulin secretion in response to exogenously administered glucagon-like peptide-1, glucose-dependent insulinotropic polypeptide, pituitary adenylate cyclase-activating polypeptide, and gastrin-releasing peptide in mice. *Endocrinology* 146:2055–2059, 2005.

Sudre B, Broqua P, White RB, Ashworth D, Evans DM, Haigh R, Junien JL, Aubert ML: Chronic inhibition of circulating dipeptidyl peptidase IV by FE 999011 delays the occurrence of diabetes in male Zucker diabetic fatty rats. *Diabetes* 51:1461–1469.

Zhou YP, Grill VE: Long-term exposure of rat pancreatic islets to fatty acids inhibits glucose-induced insulin secretion and biosynthesis through a glucose fatty acid cycle. *J Clin Invest* 93:870–876, 1994.

Unger RH: Lipotoxicity in the pathogenesis of obesity-dependent NIDDM: genetic and clinical implications. *Diabetes* 44:863–870., Lupi R, Dotta F, Marselli L, Del Guerra S, Masini M, Santangelo C, Patane G, Boggi U, Piro S, Anello M, Bergamini E, Mosca F, Di Mario U, Del Prato S, Marchetti .

Pospisilik JA, Martin J, Doty T, Ehses JA, Pamir N, Lynn FC, Piteau S, Demuth HU, McIntosh CH, Pederson RA: Dipeptidyl peptidase IV inhibitor treatment stimulates ß-cell survival and islet neogenesis in streptozotocin-induced diabetic rats. *Diabetes* 52:741–750, 2003.

Pospisilik JA, Stafford SG, Demuth HU, Brownsey R, Parkhouse W, Finegood DT, McIntosh CH, Pederson RA: Long-term treatment with the dipeptidyl peptidase IV inhibitor P32/98 causes sustained improvements in glucose tolerance, insulin sensitivity, hyperinsulinemia, and ß-cell glucose responsiveness in VDF (fa/fa) Zucker rats. *Diabetes* 51:943–950, 2002.

Montana E, Bonner-Weir S, Weir GC: Beta cell mass and growth after syngeneic islet cell transplantation in normal and streptozocin diabetic C57BL/6 mice. *J Clin Invest* 91:780–787, 1993.

Pick A, Clark J, Kubstrup C, Levisetti M, Pugh W, Bonner-Weir S, Polonsky KS: Role of apoptosis in failure of ß-cell mass compensation for insulin resistance and ß-cell defects in the male Zucker diabetic fatty rat. *Diabetes* 47:358–364, 1998.

Lacy PE, Kostianovsky M: Method for the isolation of intact islets of Langerhans from the rat pancreas. *Diabetes* 16:35–39, 1967.

Luo J, Quan J, Tsai J, Hobensack CK, Sullivan C, Hector R, Reaven GM: Nongenetic mouse models of non-insulin-dependent diabetes mellitus. *Metabolism* 47:663–668, 1999.

Zimmerman BR: Sulfonylureas (Review). *Endocrinol Metab Clin North Am* 26:511–522, 1997.

www.googlepatentdearch.com
www.google.com
www.altavista.com
www.ncbi.nlm.nih.org