Evolutionary Study and Structure Prediction of Peptidase M48 Ste24p of *Cyanothece Sp.* PCC 7822 Using Bioinformatics Tools

A

Thesis Submitted

To

Orissa University of Agriculture & Technology, Bhubaneswar

In partial fulfillment of the requirement of the degree of

MASTER OF SCIENCE IN BIOINFORMATICS

BY

MITALI MADHUMITA SAHOO

Adm. No-27BI/09

DEPARTMENT OF BIOINFORMATICS

CENTRE FOR POST GRADUATE STUDIES

ORISSA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY

BHUBANESWAR-751003

2011

Thesis Advisor — Mr. Surya Narayan Rath

DEDICATED TO

MY

BELOVED PARENTS

&

ADVISOR

Mr.Suryanarayan Rath

Orissa University of Agriculture & Technology
Department Of Bioinformatics
Centre for Post Graduate Studies
Bhubaneswar

Mr. Surya Narayan Rath
Assistant Professor

CERTIFICATE –I

This is to certify that thesis entitled "Evolutionary Study and Structure Prediction of Peptidase M48 Ste24p of Cyanothece Species PCC 7822 Using Bioinformatics Tools" submitted for award for the degree of Master of Science in the subject of Bioinformatics embodies a faithful bonafied research work carried out by Miss Mitali Madhumita Sahoo (Adm. No:27BI/09) under my guidance & supervision. No part of this thesis has been submitted by her for any other degree.
I further certify that any help or information received during the course of investigation have been duly acknowledged by her.

S.N.Rath 16/7/11

Place: Bhubaneswar

Date: 16/7/11

Mr. Surya Narayan Rath
Assistant Professor
Dept. of Bioinformatics

CERTIFICATE –II

This is to certify that the dissertation entitled "evolutionary study and structure prediction of peptidase M48 Ste24p of *cyanothece sp. PCC 7822* using bioinformatics tools "submitted by Mitali Madhumita Sahoo to the Orissa University Of Agriculture & Technology, Bhubaneswar in the partial fulfilment of the requirements for the award of the degree of Master of Science in Bioinformatics has been approved by the students advisory committee after an oral examination of the same in collaboration with external examiner.

ADVISORY COMMITTEE

1. Mr. Surya Narayan Rath — Chairman 16/7/11

   Asst. Professor

   Department of Bioinformatics

2. Mr. Sukanta Kumar Pradhan — Member 16/7/11

   Head of the Department

   Department of Bioinformatics

3. Mr. Rashmi R. Patro — Member 16/7/11

   Asst. Professor

   Department of CSA

External Examiner 16/7/2011

ACKNOWLEDGEMENT

I do feel great pleasure in expressing my deep sense of gratitude, sincerest respect and cordial thanks to **Mr. Surya Narayan Rath, Asst. Professor,** Dept of Bioinformatics, OUAT, Bhubaneswar, Orissa for his Guidance, Support, Advise and Co-operation. His affectionate personal attention has made my research a memorable experience.

. I am thankful to **Mr. Sukant Kumar Pradhan (HOD),** Dept of Bioinformatics, CPGS, OUAT, Bhubaneswar, Orissa for his support and encouragement to carry out this work successfully.

I would like to owe my sincere thanks to the Respected Dean,PGF-cum-DRI, **Dr. H.K. Senapati** for the encouragement.

I express my sincere and deep sense of gratitude to my esteemed guide **Mr. Mahesh Chandra Patra, Research Associate, Gopal Krushna Bhoi, Trainee, Dept. Of Bioinformatics, CPGS, OUAT**, for his constant supervision, scholarly guidance, timely advice and help throughout the period of my work for preparing this thesis.

I feel great pleasure in expressing deepest regards and cordial thanks to **Miss Sushma Rani Martha** and **Miss Sucharita Balbantaray** mam for aware me about career using their experience which help me in great way to focus on my career.

I would like to express my heartiest and cordial regards to my beloved Father, Mother, my dearest younger sister and brother, whose unbound emotional support and attachment always inspire me to face all challenges.

I feel honoured to be a part of this auspicious university for providing me a healthy atmosphere in these two years.

Thanks to all persons who had helped me directly or indirectly whose names could not find a separate place as duly acknowledged.

Above all, I owe a never-ending gratitude to that almighty for his grace and blessings throughout my life.

Mitali Madhumita Sahoo

Miss Mitali M. Sahoo

Name of the Student	: Mitali M. Sahoo
Admission No	: 27BI/09
Title Of Thesis	: Evolutionary analysis and structure prediction of peptidase M48 Ste24p of *cyanothece sp.* PCC 7822
Degree for which thesis is submitted	: Master of Science in Bioinformatics
Name of the Department	: Department of Bioinformatics
College & University	: Center for Post Graduate Studies, Orissa University Of Agriculture and Technology, Bhubaneswar-751003
Year of submission	: 2011
Name of the advisor	: Mr. Surya Narayan Rath Assistant Professor, Dept. Of Bioinformatics, CPGS, OUAT

ABSTRACT

This thesis elucidates the attempt of evolutionary study, secondary structure prediction and tertiary structure prediction of peptidase M48 Ste24p of *Cyanothece Sp.* of strain PCC 7822, which is a metal binding protein. This is rice plant protein,located in the root tip of the plant. This protein helps in nitrogen fixation, which helps to reduce pollution of environment. As the three dimensional structure of the protein is not available in PDB,MMDB,Modbase the protein has taken for the tertiary structure prediction and phylogenetic analysis. The phylogenetic analysis of protein is generated using MEGA, figtree, Bioedit.The secondary structure of the protein is predicted using SOPMA, GOR4, PSIPred. The tertiary structure is predicted using modeller 9.9. and the protein is visualized in DS visualizer 3.0. For the validation of protein PROCHECK and Verify 3D is used. Then energy minimisation is done by using YASARA. This work will enhance the further work for the productivity of the rice plant.

Mr. Surya Narayan Rath
ADVISOR

Mitali Madhumita Sahoo.
Mitali M. Sahoo
AUTHOR

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBRIVIATIONS

NCBI	National Center for Biotechnology Information
PDB	Protein Data Bank
BLAST	Basic Local Alignment
ExPACY	Expert Protein Analysis system
SOPMA	Self Optimised Prediction Method with Alignment
GOR4	Garnier Osgothrope Robson
MSA	Multiple Sequence Alignment
PROCHECK	Protein check

INTRODUCTION

CHAPTER-I

INTRODUCTION

Proteins regulate most of the activities in living organism. All proteins start out on a ribosome as a linear sequence of the amino acids. This linear sequence must fold during and after the synthesis so that the protein can take up its native confirmation. The native conformation of a protein is a stable three dimensional structure that strongly determines the protein's biological function. The ultimate goal of protein modelling is to predict a structure from its sequence with an accuracy that is comparable to the best results achieved experimentally. Computational modelling is the only way to obtain structural information if experimental techniques fail. Many proteins are simply too large for NMR analysis and cannot be crystallized for x-ray diffraction. Comparative modelling helps to generate insilico protein models in all the contexts where today only experimental structures provide a solid basis structure-based drug design, analysis of protein function, interactions, antigenic behaviour and rational design of proteins with increased stability or novel functions.

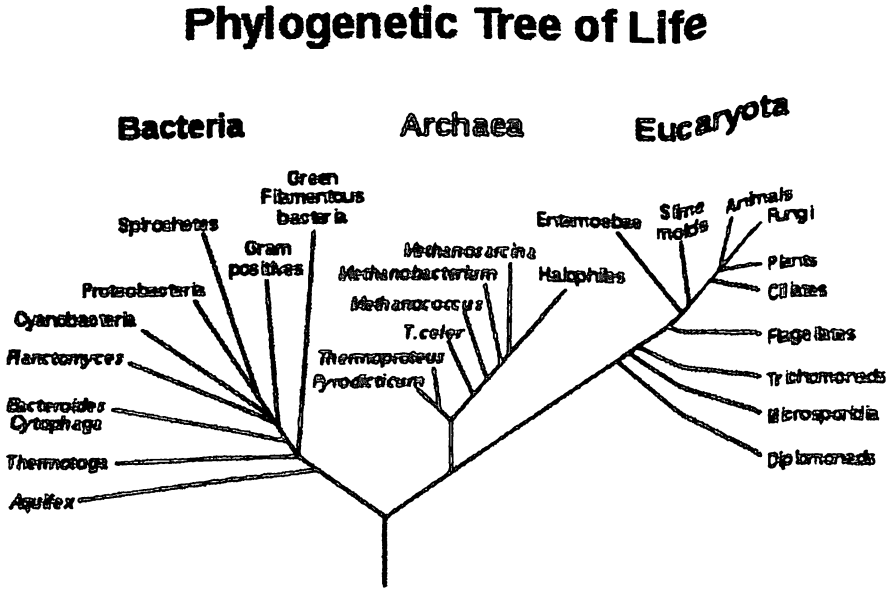


Fig 1: Display evolutionary tree of Bacteria, Archaea and Eucaryota

- The specific residue (nucleotide or amino acid) present in a given sequence is referred to as the character state.
- They are assumed to have been derived from a single common ancestor (this statement is actually redundant; by definition homologous sequences must be derived from a common ancestor).
- In most cases ancestral sequences are not known, and the ancestral states must be inferred.
- The ancestral sequences are assumed to have undergone *mutation.*
- Modeling mutation accurately is one of the challenges of phylogenetic analysis.
- They are assumed to be related by a dichotomously branching tree.

1. Accuracy of sequence

    - That the sequence itself is correct
    - That it was determined from the correct organism
    - Violations of this assumption are more common than one might suspect. Several kinds of laboratory errors can result in incorrect annotation of an otherwise legitimate sequence.

2. That homology has been correctly determined. This applies to both the sequences themselves and the alignment.

    - Paralogy can cause tremendous confusion.
    - The assumptions that went into making the multiple sequence alignment are among the assumptions of the phylogenetic analysis that is based on that alignment.

3. That sufficient similarity remains among the sequences that there is usable phylogenetic information present.

The information content of the sequences

    - Invariant sequences
    - Saturated sequences

Note that even if a gene phylogeny is correctly inferred, that phylogeny may not be helpful. For example, because of paralogy, hybridization, introgression, and horizontal gene transfer, gene phylogenies do not always correspond to the phylogeny of the genome as a whole.

Phylogenetic methods can be divided into three general categories

- Parsimony
- Minimum Distance
- Likelihood

Parsimony

- Part of a larger theoretical system refered to as "Cladistics".
- Emphasises shared derived character states.
- The idea is that monophyletic groups can be recognized because they share derived character states ("synapomorphies").
- Invariant, unique ("autapomorphic"), and ancestral character states are considered to be uninformative
- Search for the tree that requires the smallest number of character-state changes

Parsimony is easy to understand and can be a useful analytical method, but the method makes some assumptions that may not be immediately obvious. One of parsimony's most important assumptions is that it is relatively unusual for identical character-states to appear independently in different parts of the phylogenetic tree. In other words, it assumes that convergent evolution is a relatively rare phenomenon. Unfortunately this is not a valid assumption for biological sequence data.Although amino acid data have more character states than DNA and are therefore probably less. A model of sequence evolution can be used to relate the data to a hypothesis (typically a tree topology).

Maximum likelihood

- Search for the tree that maximizes the likelihood function
- The idea is to find the tree that is most likely given the data and the model

Bayesian analysis

- Typically uses a Monte Carlo algorithm
- Estimates probabilities for branch lengths and tree topologies [33]

Protein structure prediction

Protein structure prediction is the prediction of the three-dimensional structure of a protein from its amino acid sequence — that is, the prediction of its secondary, tertiary, and quaternary structure from its primary structure. Structure prediction is fundamentally different from the inverse problem of protein design. Protein structure prediction is one of the most important goals pursued by bioinformatics and theoretical chemistry; it is highly important in medicine (for example, in drug design) and biotechnology (for example, in the design of novel enzymes). Every two years, the performance of current methods is assessed in the CASP experiment (Critical Assessment of Techniques for Protein Structure Prediction).

The necessity for the 3D structure/confirmations of a protein is:

- Enhance understanding of how proteins function and hoe they interact with eachother.
- Explain antigenic property of protein.
- Provide an understanding about DNA-binding specificity, etc.[1]

Comparative modelling

Comparative modelling of protein refers to constructing an atomic resolution model of the "target" protein from its aminoacid sequence and an experimental 3d structure of a related homologous protein "template". Comparative modelling relies on the identification of the one or more known protein structures likely to resemble the structure of the query sequence, and on the production of an alignment that maps residues in the query sequence to residues in the template sequence. The sequence alignment and template structure are then used to produce a structural model of the target. Comparative modelling can produce high-quality structural models when the target and template are closely related [2].

Cyanobacteria

Cyanobacteria are often called "blue-green algae". This name is convenient for talking about organisms in the water that make their own food, but does not reflect any relationship between

the cyanobacteria and other organisms called algae. Cyanobacteria are relatives of the bacteria, not eukaryotes, and it is only the chloroplast in eukaryotic algae to which the cyanobacteria are related. Many plants, especially legumes, have formed symbiotic relations with nitrifying bacteria, providing specialized tissues in their roots or stems to house the bacteria, in return for organic nitrogen. This has been used to great advantage in the cultivation of rice, where the floating fern Azolla is actively distributed among the rice paddies. The fern houses colonies of the cyanobacterium Anabaena in its leaves, where it fixes nitrogen. The ferns then provide an inexpensive natural fertilizer and nitrogen source for the rice plants when they die at the end of +he season. Cyanobacteria also form symbiotic relationships with many fungi, forming complex symbiotic "organisms" known as lichens. Cyanobacteria are photosynthetic prokaryotes with important roles in diverse aqueous environments. Many cyanobacteria can also fix nitrogen, a process biochemically incompatible with oxygenic photosynthesis. We are interested in the strategies employed by unicellular nitrogen-fixing cyanobacteria to accommodate these processes, which the analysis of gene structure and regulation can help elucidate [3].Molecular hydrogen is one of the potential future energy sources as an alternative to the limited fossil fuel resources of today. Its advantages as fuel are numerous: it is eco-friendly, efficient, renewable, and during its production and utilization no CO_2 and at most only small amounts of NO_x are generated [9]. By the virtue of all these attributes the hydrogen gas can be used as an energy source. Hydrogen gas can be prepared in many conventional ways (including those of photoelectrochemical or thermochemical processes) for its large-scale utilization. In this review we aim to discuss about photobiological hydrogen production by cyanobacteria and the scientific and technical aspects of large-scale utilization of produced hydrogen for various applications. We also have described about salient features of cyanobacterial enzymatic system, different species and strains producing hydrogen, parameters controlling the hydrogen production and large-scale production utilizing photobioreactors. Cyanobacteria are thought to play a crucial role in the Precambrian phase by contributing oxygen to the atmosphere [10]. Under certain conditions the cyanobacterial species instead of reducing CO_2, consume biochemical energy to produce molecular hydrogen.

Cyanothece sp. PCC 7822

Cyanothece sp. PCC 7822 is a unicellular cyanobacterium isolated from rice fields in Cuttack, India. At 8-10??? m in length, *Cyanothece* 7822 is the largest known *Cyanothece* strain. This is an aerobic nitrogen fixing strain. In addition to chlorophyll and phycocyanin, these cells contain the light harvesting pigment phycoerythrin which imparts an unusual brown colour to the cells. These cells secrete large amounts of exopolysaccharides and tend to aggregate in culture [6].

Biological process- Proteolysis

The hydrolysis of proteins in to smaller polypeptides and/or amino acids by cleavage of their peptide bonds.

Molecular function- Binding

The selective, non-covalent, often stoichiometric, interaction of a molecule with one or more specific sites on another molecule.

Metalloendopeptidase activity

Catalysis of the hydrolysis of internal, alpha-peptide bonds in a polypeptide chain by a mechanism in which water acts as a nucleophile, one or two metal ions hold the water molecule in place, and charged amino acid side chains are ligands for the metal ions.

Peptidase M48 Ste24p of *Cyanothece Sp.* PCC 7822

Peptidase M48 Ste24p of *Cyanothece Sp.* PCC 7822 having accession no YP_003888914.1 (NCBI) and EOUH40 (SWISS-PROT) belongs to a group of metallopeptidase family M48(Ste 24p endopeptidase family, clad M- as described in MEROPS database; accession no MERO02635).They are integral membrane proteins are associated with the endoplasmic reticulum and golgi bodies, binds to zinc ion per subunit. Metalloproteases are the most diverse of the four main types of protease, with more than 50 families identified to date. In these enzymes, a divalent cation, usually zinc, activates the water molecule. The metal ion is held in place by amino acid ligands, usually three in number. The known metal ligands are His, Glu, Asp or Lys and at least one other residue is required for catalysis, which may play an electrophillic role. Of the known metalloproteases, around half contain an HEXXH motif, which has been shown in crystallographic studies to form part of the metal-binding site [5]. The HEXXH motif is relatively common, but can be more stringently defined for metalloproteases as ' HEXXH ', where 'a' is most often valine or threonine and forms part of the S1' subsite in thermolysin and

neprilysin, 'b' is an uncharged residue, and 'c' a hydrophobic residue. Proline is never found in this site, possibly because it would break the helical structure adopted by this motif in metalloproteases [5].In saccharomyces cerevisiae (Baker's Yeast) ste24p is required for the first NH2-terminal proteolytic processing event within the a-factor precursor, which takes place after COOH-terminal CAAX modification is complete. Ste24p contains multiple predicted membrane spans, a zinc metalloprotease motif (HEXXH) and a COOH-terminal ER retrieval signal (KKXX).The HEXXH protease motif is critical for Ste24p activity, since Ste24p activity, since Ste24p fails to function when conserved residues within this motif are mutated.The Ste24p homologues occur in a diverse group of organisms, including E.Coli, Schizosaccharomyces pombe(Fissio yeast), Haemophilus influenzae, Homo sapiens, which indicates that the gene is highly conserved throughout evolution. Ste24p and the proteins related to it define a subfamily of proteins that are likely to function as intracellular, membrane-associated zinc metalloproteases [6].The Ste24p endopeptidase is an integral membrane protein with 7 transmembrane domains. It resides in the membrane of ER, with the N-terminus in the lumen and the active site and the C-terminus in the cytoplasm.HtpX0- is a zinc-dependent endoprotease member of the membrane-localized proteolytic system in E. coli, which participates in the proteolytic quality control of membrane proteins in conjunction with FtsH, a membrane-bound and ATP-dependent protease. Biochemical characterisation revealed that HtpX undergoes self-degradation upon cell disruption or membrane solubilization. It can also degrade casein and cleaves solubilized membrane proteins, for example, SecY [7]. Expression of HtpX in the plasma membrane is under the control of CpxR, with the metalloproteinase active site of HtpX located on the cytosolic side of the membrane. This suggests a potential role for HtpX in the response to mis-folded proteins [8].

Objectives

- To construct evolutionary tree of different *Cyanothece* species using MEGA 5.0 and Bioedit.
- To predict the secondary structure of the sequence by using SOPMA, GOR4 and PSIPred.
- To predict the tertiary structure of the sequence was predicted by using modeller9.9.

RREVIEW OF LITERETURE

CHAPTER-II

REVIEW OF LITERETURE

Proteins are the most versatile macromolecules in living systems and also serve crucial function in essentially all biologically process. They function as catalysts, they transport and store oxygen and other molecules , provide mechanical support and immune protection , generate movement transmit nerve impulses, controls growth and differentiation and one of the most important function as stress resistance capacity.

Proteins perform these functions.

1. Proteins are linear polymers of monomer units called amino acids. The function of a protein is directly dependent on its three dimensional structure, that are determined by the sequence of amino acids in the protein polymer.

2. Proteins contain a wide range of functional groups. These functional groups include alcohols, thiols, thioethers, carboxylic acids, carboxamides, and a variety of basic groups. This array of functional groups accounts for the broad spectrum of protein function.

3. Proteins can interact with one another with other biological macromolecules to form complex assemblies. The protein with these assemblies can act synergistically to generate capabilities not afforded by the individual component proteins.

4. Some proteins are quite rigid, whereas others display limited flexibility. Rigid units can function as structural elements in the cytoskeleton or in connective tissues. Parts of proteins with limited flexibility may act as hinges, springs, and levers that are crucial to protein function (such as -: assembly of proteins with one another or with other to form complex units, and to the transmission of information within and between cells).

At neutral Ph., the amino acids are exists as dipolar ion (zwitter ion). In the dipolar form the amino group is protonated but the carboxyl group is deprotonated. The ionisation state of amino

acid varies with Ph. The dipolar form persists until the Ph. approaches 9.All the proteins in all species, bacterial, archaeal, eukaryotic are constructed from 20 amino acids. The remarkable range of functions mediated by proteins result from the diversity and versatility of these 20 building blocks. Among all amino acids glycine is unique in being achiral. The longer aliphatic side chains are hydrophobic-that is, they tend to cluster together rather than contact water. The three dimensional structure of water-soluble proteins are stabilized by this tendency of hydrophobic groups to come together, called the hydrophobic effect. The different sizes and shapes of these hydrocarbon side chains enable to pack together to form compact structures with few holes. A polypeptide chain consists of a regularly repeating part called main chain or backbone, and a variable part comprising the distinctive side chains. The backbone is rich in hydrogen-bonding potential. Each aminoacid unit in a polypeptide chain is called a residue. Each residue contains a carbonyl group (a good hydrogen bond acceptor) except proline (a good hydrogen donor) .Aminoacid sequence is the link between the genetic message in DNA and the three dimensional structure that performs a protein's biological function.

Geometry of protein backbone reveals several important features. Peptide chain is planar, for a pair of aminoacid linked by peptide bond, six atoms lie in the same plane (alpha group from first aminoacid, NH group and alpha carbon of second aminoacid). The peptide bond has double bond character which prevents rotation about this bond. The inability of rotation of bond constrains the conformation of the peptide backbone and accounts for the bond's polarity.

The peptide bond is uncharged, allowing polymers of aminoacid linked by peptide bonds to form tightly packed globular structures. Two configuration is possible for a planar peptide bond. In the trans configuration, 2 alpha carbon atoms are on opposite sides of peptide bond, where just opposite case in cis form. Almost all peptide bonds in proteins are trans. It is only due to the steric clashes between groups attached to the alpha carbon atoms hinder formation of the cis form. The bonds between alpha carbon atom and bonds between alpha carbon atom and carbonyl group are single bonds. These two single bonds of each amino acid rotate, which allows proteins to fold in many different ways. The rotations about these bonds canbe specified by dihedral angles. [Dihedral angles: A measure of the rotation about a bond, usually taken to lie between -180° and +180°. Dihedral angles sometimes also called as torsion angle].The angle of rotation about the bond between the alpha carbon atom and carbonyl carbon atoms is called Psi

(Ψ) angle. The angle of rotation about the bond between the alpha carbon atom and nitrogen atoms is called phi (Φ) angle. The clockwise rotation of either bond corresponds to a positive value. The clockwise rotation of either bond corresponds to a positive value. The phi and psi angles determine the Path of polypeptide chain. All combinations of phi and psi angles are possible. G.N. Ramachandran recognized that many combinations are forbidden because of steric collision between atoms. The allowed values can be visualized on a two-dimensional plot called a ramachandran diagram. Three quarters of the possible (phi, psi) combinations are excluded simply by local steric clashes. The combination should not be +90°,-90°. In the combination one torsion angle should be clockwise (+ve°) and another should be anticlockwise (-ve°). [Steric exclusions: The fact is that two atoms cannot be placed i n same position in sometime always.]. In ramachandran plot one-quarters s disfavoured region, where atoms are placed affected by steric clashes,(Ψ = -90° & Φ= +90°).The unfolded polymers existing as a random coil have many possible conformations and folded form has a unique conformation. The rigidity of the peptide unit and restricted set of the allowed phi and psi angles limits the number of structures understood the unfolded form to allow protein folding to occur.

Secondary structure

Polypeptide chains can fold in to regular structures such as the alpha helix, beta sheet, turns and loops. In 1951, linus Pauling and Robert Corey proposed two periodic structures called as alpha helix and beta pleated sheet. Subsequently beta turn and omega loop were identified. Corey and Pauling considered which conformations of peptides were sterically allowed and which NH and CO group has most exploited hydrogen bonding capacity.

Alpha Helix

It is a rod like structure. A tightly coiled backbone forms inner part of the rod and side chains extended outward in a helical array. The alpha helix is stabilised by hydrogen bonds between the NH and CO groups of the main chain.The CO group of each aminoacid forms a hydrogen bond with the NH group of the aminoacid present at the 4 residues ahead in the sequence. Only the CO and NH groups near the end of aminoacid of main-chain are not H-bonded. Each residue is related to next residue with a rise of 1.5 A° along the helix and rotation of 100 degrees, gives 3.6 aminoacid residues per turn of helix. Aminoacids are two apart in the sequence are situated on

opposite sides of the helix, so that makes contact with each other. The number of residues per turn (3.6) is 5.4 A°. In ramachandran plot, both left-handed and right handed helices are energetically more favourable, because of less steric clash between the side chains and backbone. However, essentially all alpha helices in the proteins are right-handed. These alpha helix coils found in myosin, tropomyosin in muscle, in fibrin in blood clots, in keratin in hair. The two alpha helices wind arround one another to form superhelix. Such structures are found in keratin in hair, quills, claws and horns.

Beta pleated sheet

Beta sheets are stabilized by H-bonding between polypeptide strands. After alpha helix structure Corey and Pauling discovered another periodic structural motif, named it as beta pleated sheet. The beta sheet is fully extended structure rather than coiled like alpha helix. The distance between adjacent amino acids along a beta strand is approximately 3.5A°, in contrast with a distance of 1.5A° along an alpha helix. The side chains of adjacent aminoacid points to opposite direction. A beta sheet is formed by linking two or more beta strands by hydrogen bonds. The adjacent beta sheet can run in opposite or same direction. In the antiparallel arrangement,the NH group and CO group of each aminoacid are respectively H-bonded to the patner of CO and NH in the adjacent chain. But in case of parallel arrangement, the H-bonding is quite complicated. The beta strand gathers 4 or 5 or many as 10 residues. The H-bonding occurs, between the NH of one aminoacid with CO of the adjacent strand's aminoacid. Beta sheets are relativelyflat and adopt a twisted shape. Beta sheets are important structural element in many proteins like fatty acid binding proteins, important for lipid metabolism, are built almost entirely from beta sheets.

Loops and Turns

In many reverse turns, the CO group of residue 'i' of a polypeptide is H-bonded to NH group of residue 'i+3'.This interaction stabilizes abrupt changes in direction of the polypeptide chain. Sometime other structure called loops or omega loops are responsible for chain reversals. Turns and loops are lie on the surface of proteins, so produces interaction between proteins and other molecules.

Tertiary Structure

Tertiary structure refers to the spatial arrangement of aminoacid residues that are far apart in the sequence and to the apart in the sequence and to pattern of disulfide bonds.Christian anfinsen in the 1950s perform work on the enzyme ribonuclease revealed the relation between amino acid sequence of a protein and its conformation.The contrasting distribution of polar and nonpolar residues reveals a key facet of protein architecture. In an aqueous environment, protein olding is driven by the strong tendency of hydrophobic residues to be excluded from water. A system is more thermodynamically stable when hydrophobic groups are clustered rather than extended into the aqueous surroundings. The polypeptide chain therefore folds so that its hydrophobic side chains are buried and its polar, charged chains are on the surface. Many alpha helices and beta strands are amphipathic; that is, the alpha helix or beta strand has a hydrophobic face, which points in to solution.The fate of the main chain accompanying the hydrophobic side chains is important, too.An unpaired peptide NH or CO group markedly prefers water to a nonpolar milieu. The secret of burying a segment of main chain in a hydrophobic environment is pairing all the NH and CO groups by the hydrogen bonding. This pairing is neatly accomplished in an alpha helix or beta sheet. Vander Waals interaction between tightly packed hydrocarbon side chains also contribute to the stability of proteins. Some proteins that spans biological membranes are "the exceptions that prove the rule" regarding the distribution of hydrophobic and hydrophilic amino acids throughout three-dimensional structures. For example: consider porins, proteins found in the outer membranes of many bacteria. The permeability barriers of membranes are built largely of alkane chains that are quite hydrophobic. Thus, porins are covered on the outside largely with hydrophobic residues that interact with the neighbouring alkane chains. In contrast, the center of the protein contains many charged and polar amino acids that surround a water-filled channel running through the middle of the protein. Thus, because porins function in hydrophobic environments, they are "inside out" relative proteins that function in aqueous solution. Some polypeptide chains fold into two or more compact regions that may be connected by a flexible segment of polypeptide chain, rather like pearls on a string. These compact globular units, called domains, range in size from about 30 to 400 aminoacid residues. From earlier studies, proteins can be denatured by heat and chemicals like urea or guanidium chloride. After denaturation the folded or native form of protein converted in to the unfolded form. But when the denaturants removed again the protein regains it's native conformation.

Protein folding and unfolding is thus largely an "all or none" process that results from a cooperative transition. The consequences of cooperative folding can be illustrated by considering the contents of a protein solution under conditions corresponding to the middle of transition between the folded and unfolded forms. Under these conditions, the protein is "half folded". Structures that are partly intact and partly disrupted are not thermodynamically stable and exist only transiently. Cooperative folding ensures that partly folded structures that might interfere with processes within cells do not accumulate.

Prediction of Three-Dimensional Structure

The amino acid sequence completely determines the three-dimensional structure of a protein. The local sequence appears to determine only between 60% and 70% of the secondary structure; long-range interactions are required to fix the full secondary structure and the tertiary structure. Investigators are exploring two fundamentally different approaches to predicting three-dimensional structure from amino acid sequence. The cooperative folding of proteins is a thermodynamic property; its occurrence reveals nothing about the kinetics and mechanism of protein folding. A protein make the transition from a diverse ensemble of unfolded structures into a unique conformation in the native form. One possibility a priori would be that all possible conformations are tried out to find the energetically most favourable one [34].

Cyanobacteria

Cyanobacteria are aquatic and photosynthetic, that is, they live in the water, and can manufacture their own food. Because they are bacteria, they are quite small and usually unicellular, though they often grow in colonies large enough to see. They have the distinction of being the oldest known fossils, more than 3.5 billion years old, in fact! the cyanobacteria are still one of the largest and most important groups of bacteria on earth. They are photosynthetic and aquatic, cyanobacteria are often called "blue-green algae". This name is convenient for talking about organisms in the water that make their own food, but does not reflect any relationship between the cyanobacteria and other organisms called algae. Cyanobacteria are relatives of the bacteria, not eukaryotes, and it is only the *chloroplast* in eukaryotic algae to which the cyanobacteria are related. The cyanobacteria have an extensive fossil record. The oldest known fossils, in fact, are cyanobacteria from Archaean rocks of western Australia, dated

3.5 billion years old. Cyanobacteria are among the easiest microfossils to recognize. Cyanobacteria are nitrogen fixing bacteria and also very anceint which absorbs nitrogen from environment and convert it to oxygen which is one of very important characterestics of cyanobacteria. So the cyanobacteria have also been tremendously important in shaping the course of evolution and ecological change throughout earth's history. The oxygen atmosphere that we depend on was generated by numerous cyanobacteria photosynthesizing during the Archaean and Proterozoic Era. Before that time, the atmosphere had a very different chemistry, unsuitable for life as we know it today.

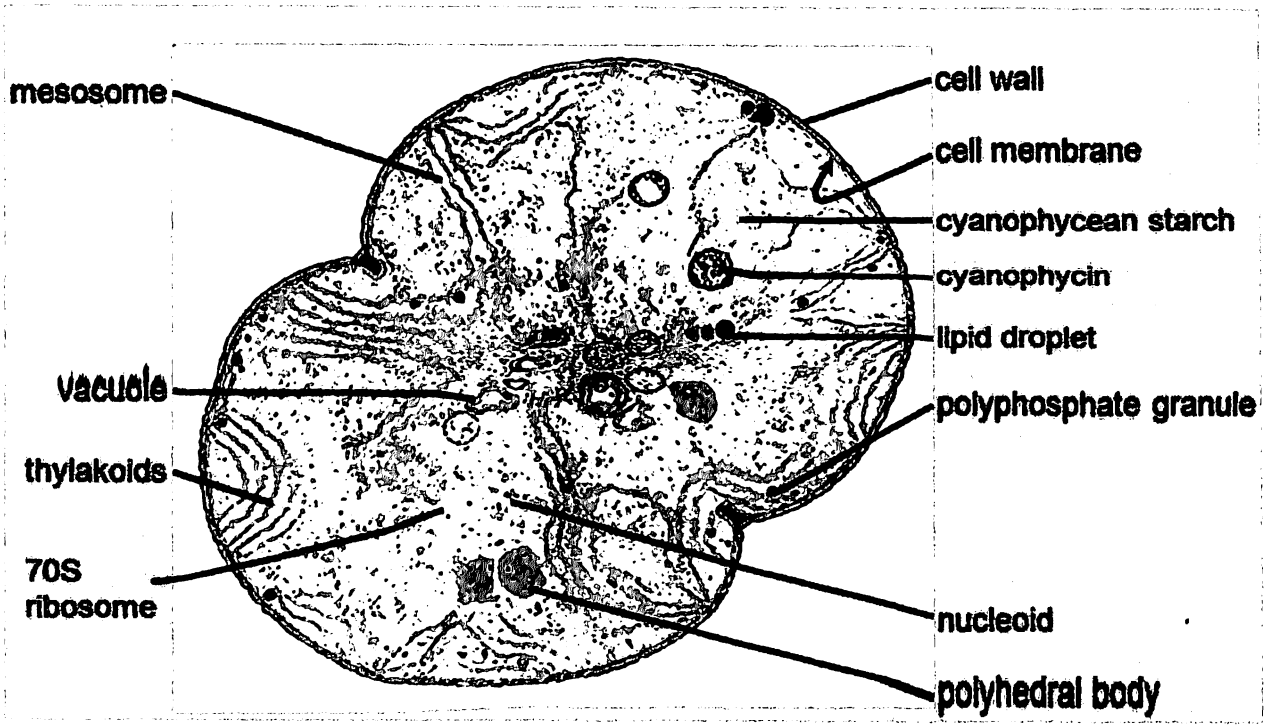


Fig.2: Shows cyanobacterial vegetative cell (photosynthesis).

Though cyanobacteria do not have a great diversity of form, and though they are microscopic, they are rich in chemical diversity. Cyanobacteria get their name from the bluish pigment phycocyanin, which they use to capture light for photosynthesis. They also contain chlorophyll a, the same photosynthetic pigment that plants use. In fact the chloroplast in plants is a symbiotic cyanobacterium, taken up by a green algal ancestor of the plants sometime in the Precambrian. However, not all "blue-green" bacteria are blue; some common forms are red or pink from the pigment phycoerythrin. These bacteria are often found growing

on greenhouse glass, or around sinks and drains. The Red Sea gets its name from occasional blooms of a reddish species of *Oscillatoria*, and African flamingos get their pink color from eating *Spirulina*. Whatever their color, cyanobacteria are photosynthetic, and so can manufacture their own food. This has caused them to be dubbed "blue-green algae", though they have no relationship to any of the various eukayotic algae. The term "algae" merely refers to any aquatic organisms capable of photosynthesis, and so applies to several groups [11].

Cyanobacteria are important in the nitrogen cycle.

Cyanobacteria are very important organisms for the health and growth of many plants. They are one of very few groups of organisms that can convert inert atmospheric nitrogen into an organic form, such as nitrate or ammonia. It is these "fixed" forms of nitrogen which plants need for their growth, and must obtain from the soil. Fertilizers work the way they do in part because they contain additional fixed nitrogen which plants can then absorb through their roots.

Nitrogen fixation

Nitrogen Fixation: Atmospheric dinitrogen (N_2) → Ammonia (NH_3)

Nitrogen-fixers – bacteria, free-living and symbiotic

$N_2 + 8H+ + 8e- + 16\ ATP = 2NH_3 + H_2 + 16ADP + 16\ Pi$

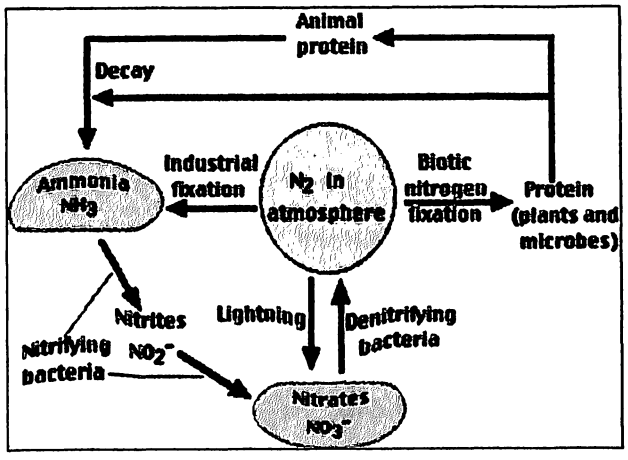


Fig.3: Shows the nitrogen fixation cycle

called stromatolites[13].Many cyanobacteria also form motile filaments, called hormogonia, that travel away from the main biomass to bud and form new colonies elsewhere. The cells in a hormogonium are often thinner than in the vegetative state, and the cells on either end of the motile chain may be tapered. In order to break away from the parent colony, a hormogonium often must tear apart a weaker cell in a filament, called a necridium.Each individual cell of a cyanobacterium typically has a thick, gelatinous cell wall. They lack flagella, but hormogonia and some species may move about by gliding along surfaces. Many of the multi-cellular filamentous forms of *Oscillatoria* are capable of a waving motion; the filament oscillates back and forth. In water columns some cyanobacteria float by forming gas vesicles, like in archaea. These vesicles are not organelles as such. They are not bounded by lipid membranes but by a protein sheath.Some of these organisms contribute significantly to global ecology and the oxygen cycle. The tiny marine cyanobacterium Prochlorococcus was discovered in 1986 and accounts for more than half of the photosynthesis of the open ocean. Many cyanobacteria even display the circadian rhythms that were once thought to exist only in eukaryotic cells (see bacterial circadian rhythms). Despite the fact that they lack flagella, many cyanobacteria are capable of moving about. *Oscillatoria* pictured above at right, gets its name from the "swaying" motion of its filaments when observed under the microscope. No one has yet determined how these filaments are able to do this [14].

Peptidase M48 Ste24p [cyanothece sp. PCC 7822]

LOCUS	YP_003888914 628 aa linear BCT 17-SEP-2010
DEFINITION	Peptidase M48 Ste24p [*Cyanothece sp.* PCC 7822].
ACCESSION	YP_003888914
VERSION	YP_003888914.1 GI: 307153530
DBLINK	Project: 52547
DBSOURCE	REFSEQ: accession NC_014501.1
SOURCE	*Cyanothece sp.* PCC 7822
ORGANISM	*Cyanothece sp.* PCC 7822 Bacteria; Cyanobacteria; Chroococcales; Cyanothece.
TITLE	Complete sequence of Chromosome of Cyanothece sp. PCC 7822
FEATURES	Location/Qualifiers
Source	1..628 /organism="*Cyanothece sp.* PCC 7822" /strain="PCC 7822"

/isolation_source="fresh water in rice field"
/db_xref="taxon:497965"
/country="India: Cuttack, Orissa"
/lat_lon="20.237556 N 84.270020 E"

Protein 1..628
/product="peptidase M48 Ste24p"
/calculated_mol_wt=70320

CDS 1..628
/locus_tag="Cyan7822_3702"
/coded_by="complement(NC_014501.1:4102418..4104304)"

ORIGIN

1 mklirktsli ilwsiigels pllvlaqtpk pidsvnlger swiaqkttet tstpsssqtp

61 enhqpnkpst padntttvne cwqpnetlpq pqpvaqiees sppqtvkens dssaasqeid

121 sqqsaeeeak qsqekekpvs fipkptseei alyqqlaqad ryyrcgqstv aeklyreakk

181 pfeaeikyqr elipqaiydp qylqpggavy wrlyqealke krlqskiiap lkllteqhpe

241 fipahleyak ilkeygqkeq aqevlenait lypneaplvk akveadmaak kwldasltar

301 rfalfnpdny laaeflqlan enlqrykndl rskltlgavg navlggigya ffgniggpis

361 aietvtlliq gekaigessa khfqkrltli edeevlsyvr eigkklaava grnefnyefy

421 vvmddnlnaf alpggkvfin agailktnse aelaglmahe lshavlshsf qimtggnlla

481 natqfvpylg ssagdlitln ysrdmeread ifgtrllaas gyaadgvrnl mvvlekeddp

541 sppawlsthp dtsdrvkyve kmlvqerlnr yayegverhw kirnrvgell dkyrqekeek

601 egkkskkppe ttpktqqiiq qqqenrpi//

Peptidase M48 ste24p

In the merops database peptidases are grouped in to clads and families.Clads are group of families of common ancestry based on common structural fold.Clads reprenting with 2 letters, the first representating the catalytic type of the families included the clad and the second representating some families cannot yet be assigned to the clads.Family is belonging to the clad A-, C-, M-, N-, S-, T-, or U-, according to the catalytic type.Peptidase families are grouped by their catalytic type, the first character represents the catalytic type.

A- Aspartic M- Metal T- Threonine C- Cystein N- Aspargine U-Unknown G- Glutamic Acid S- Serine.In case of S, T, C peptidases utilises the aminoacids as a nucleophile and form acyl intermidiate these peptidases can also readily act as transferases.In case of A, G, metallopeptidases, the nucleophile is an activated water molecule.In case of the aspargine endopeptidases, the nucleophile is aspargine and all are self-processing endopeptidases.In many instances the structural protein fold that characterises the clar or family may have lost its catalytic activity, yet retain its function in protein recognition and binding.In these enzymes, a divalent cation, usually zinc, activates the water molecule.The known metal ligands are His, Glu, Asp or Lys and atleast one other residue is required for catalysis, which may play an electrophilic role[15].

Methods of Phylogenetic Analysis

Phylogenetic analysis presents a unique problem in biology, because evolutionary history can never be known with certainty. The purpose of any phylogenetic analysis is to estimate the evolutionary relationships between a set of homologous taxa, which can be anything from morphological characteristics to molecular sequences (reviewed in Mount, 2001). The result is a tree composed of "nodes" and "branches", where the terminal nodes (or "leaves") correspond to the taxa being studied (henceforth assumed to be DNA sequences), the internal nodes represent ancestral sequences, and the branches represent the topological relationship between the nodes (reviewed in Saitou, 1996).While many different methods have been developed for inferring phylogeny, they can all be thought of as members of two broad classifications: (1) those methods that use an algorithm to directly build a tree through a series of defined steps; and (2) those methods that define a criterion to be maximized (or minimized), and then use an algorithm to evaluate potential trees based on this criterion (Swofford et al., 1996). The first class of programs works largely by converting the similarity between pairs of sequences into evolutionary distance, and then using a defined set of steps to build a tree.These methods are computationally very fast, but suffer from two major downfalls: (1) evolutionary information is lost when overall similarity between sequences is observed, rather than individual mutation events (Hendy and Penny, 1982); and (2) it is difficult to reliably assess the confidence of a tree produced by any given algorithm (Swofford et al., 1996). For these reasons, purely algorithmic methods will not be discussed further. In

contrast, the second group of methods proceeds in two steps. First, an optimality criterion is defined, which is simply a score used to assess the value of a particular tree. Second, an algorithm is used to compute the value of this function for various trees, while searching for the best tree (the one that maximizes the criterion) (Swofford et al., 1996). While these methods are appealing because they have the promise of finding the optimal tree according to the applied criterion, they can be computationally slow for even moderate numbers of taxa, to a point where the amount of time required for an exhaustive search is prohibitive. However, this limitation has led to the development of countless computational methods that attempt to reliably get as close to the optimal tree as possible, in a reasonable amount of computational time[32].

Maximum Parsimony

Maximum parsimony (or simply parsimony) analysis is intuitively appealing because it is based on finding the simplest solution to an observed set of data (reviewed in Swofford et al., 1996; Saitou, 1996). Essentially, parsimony attempts to build a tree that minimizes the number of evolutionary changes required to explain the observed data [31].

Maximum Likelihood

While parsimony methods seek phylogenetic solutions that minimize the amount of evolutionary change required to explain a data set, Maximum likelihood methods attempt to find solutions that have a maximum probability of being correct, given a particular evolutionary model (Swofford et al., 1996). This distinction may at first appear semantic, but it is extremely important when (as described above) the evolutionary time involved is long enough to produce a substantial number of multiply mutated positions within a sequence. Furthermore, unlike parsimony, maximum likelihood methods consider branch lengths when calculating the probability of a particular tree being correct [30].

Protein structure prediction

Protein structure prediction is one of the most important goals pursued by bioinformatics and theoritical chemistry. Protein structure prediction is of high importance in medicine (for example in drug design) and biotechnology. Modern prediction methods rely on prediction of

secondary structure. If the correct secondary structure is known, one can get the rules for packing elements of secondary structure against eachother in order to derive selective possible globular folds. Hence it is quite evident that secondary structure prediction is the key behind the tertiary structure prediction from amino acid sequence. The most basic functionality is providing structure visualisation. Analysis of protein structure canbe facilitated by softwares that aligns structures. In the absence of existing structures for a given protein sequence, there are methods to predict or to modelthe structure of such sequences based on known protein structures.And given models of known or predicted structures, one can use softwares to verify them for errors, predict protein conformational changes, or predict substrate binding sites[16].

Secondary structure

Secondary structure prediction is a set of techniques in bioinformatics that aim to predict the local secondary structures of proteins and RNA sequences based only on knowledge of their primary structure — amino acid or nucleotide sequence, respectively. For proteins, a prediction consists of assigning regions of the amino acid sequence as likely alpha helices, beta strands (often noted as "extended" conformations), or turns. Early methods of secondary structure prediction, introduced in the 1960s and early 1970s focused on identifying likely alpha helices and were based mainly on helix-coil transition models.Significantly more accurate predictions that included beta sheets were introduced in the 1970s and relied on statistical assessments based on probability parameters derived from known solved structures. These methods, applied to a single sequence, are typically at most about 60-65% accurate, and often underpredict beta sheets. The evolutionary conservation of secondary structures can be exploited by simultaneously assessing many homologous sequences in a multiple sequence alignment, by calculating the net secondary structure propensity of an aligned column of amino acids. Limitations are also imposed by secondary structure prediction's inability to account for tertiary structure; for example, a sequence predicted as a likely helix may still be able to adopt a beta-strand conformation if it is located within a beta-sheet region of the protein and its side chains pack well with their neighbors. Dramatic conformational changes related to the protein's function or environment can also alter local secondary structure.

GOR method

The GOR method, named for the three scientists who developed it — *G*arnier, *O*sguthorpe, and *R*obson — is an information theory-based method developed not long after Chou-Fasman. It uses a more powerful probabilistic technique of Bayesian inference. [17]The method is a specific optimized application of mathematics and algorithms developed in a series of papers by Robson and colleagues, [18][19]). The approach is both more sensitive and more accurate than that of Chou and Fasman because amino acid structural propensities are only strong for a small number of amino acids such as proline and glycine. The original GOR method was roughly 65% accurate and is dramatically more successful in predicting alpha helices than beta sheets, which it frequently mispredicted as loops or disorganized regions. Later GOR methods considered also pairs of amino acids, significantly improving performance. The SOPMA method is dramatically more successful in predicting beta sheets than alpha helices.

Tertiary structure

The practical role of protein structure prediction is now more important than ever. Massive amounts of protein sequence data are produced by modern large-scale DNA sequencing efforts such as the Human Genome Project. Despite community-wide efforts in structural genomics, the output of experimentally determined protein structures—typically by time-consuming and relatively expensive X-ray crystallography or NMR spectroscopy—is lagging far behind the output of protein sequences. The protein structure prediction remains an extremely difficult and unresolved undertaking. The two main problems are calculation of protein free energy and finding the global minimum of this energy. A protein structure prediction method must explore the space of possible protein structures which is astronomically large. These problems can be partially bypassed in "comparative" or homology modeling and fold recognition methods, in which the search space is pruned by the assumption that the protein in question adopts a structure that is close to the experimentally determined structure of another homologous protein. On the other hand, the de novo or ab initio protein structure prediction methods must explicitly resolve these problems.

Comparative protein modeling

Comparative protein modelling uses previously solved structures as starting points, or templates. This is effective because it appears that although the number of actual proteins is vast, there is a limited set of tertiary structural motifs to which most proteins belong. It has been suggested that there are only around 2,000 distinct protein folds in nature, though there are many millions of different proteins. steps for homology modelling:

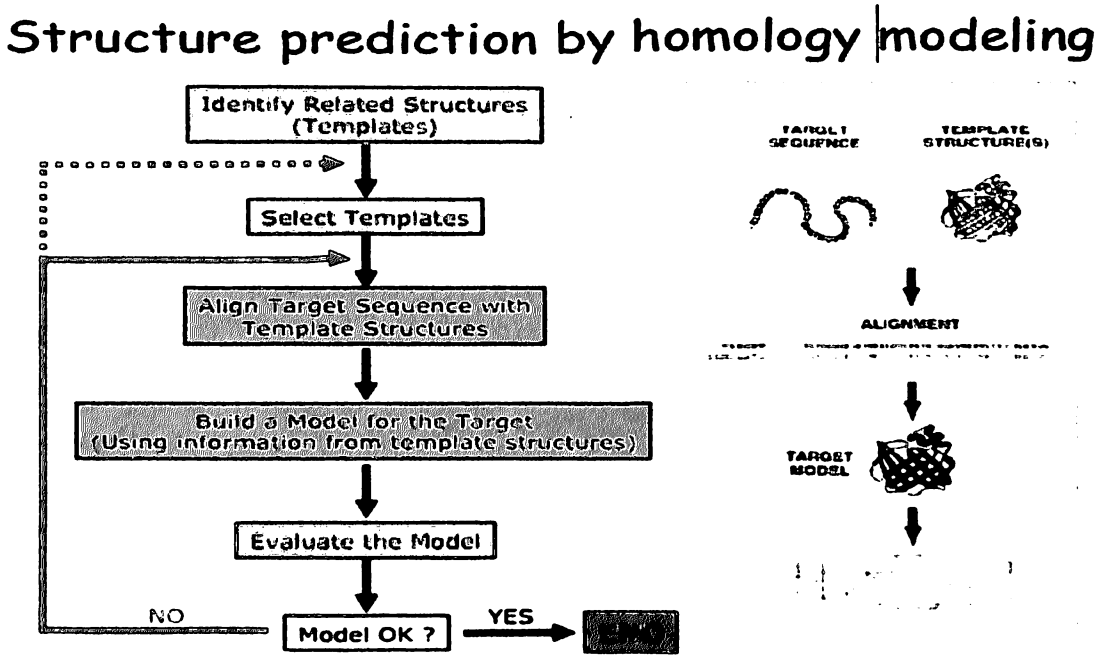


Fig.4: Shows the steps for structure prediction by homology modelling

These methods may also be split into two groups [20]:

Homology modeling

It is based on the reasonable assumption that two homologous proteins will share very similar structures. Because a protein's fold is more evolutionarily conserved than its amino acid sequence, a target sequence can be modeled with reasonable accuracy on a very distantly related template, provided that the relationship between target and template can be discerned through sequence alignment. It has been suggested that the primary bottleneck in comparative modelling arises from difficulties in alignment rather than from errors in structure prediction given a known-good alignment [20]. Unsurprisingly, homology modelling is most accurate when the

modeling, which worsen with lower sequence identity, derive from errors in the initial sequence alignment and from improper template selection. Like other methods of structure prediction, current practice in homology modeling is assessed in a biannual large-scale experiment known as the Critical Assessment of Techniques for Protein Structure Prediction, or CASP.

Energy Minimization

In computational chemistry, energy minimization (also called energy optimization or geometry optimization) methods are used to compute the equilibrium configuration of molecules and solids. Stable states of molecular systems correspond to global and local minimum on their potential energy surface. Starting from a non-equilibriums molecular geometry, energy minimization employs the mathematical procedure of optimization to move atoms so as to reduce the net forces (the gradients of potential energy) on the atoms until they become negligible. Like molecular dynamics and Monte-Carlo approaches, periodic boundary conditions have been allowed in energy minimization methods, to make small systems. A well established algorithm of energy minimization can be an efficient tool for molecular structure optimization. Unlike molecular dynamics simulations, which are based on Newtonian dynamic laws and allow calculating atomic trajectory with kinetic energy, molecular energy minimization does not include the effect of temperature, and hence the trajectories of atoms during the calculation do not really make any physical sense, i.e. we can only obtain a final state of system that corresponds to a local minimum of potential energy. From a physical point of view, this final state of the system corresponds to the configuration of atoms when the temperature of the system is approximately zero.

MATERIALS & METHODS

CHAPTER-III

MATERIALS & METHODS

MATERIALS

The following databases and bioinformatics tools or servers were used for comparative proteomics analysis of peptidase M48 Ste24p [*cyanothece* sp. PCC 7822].

Databases used

1. NCBI

The Entrez Global Query Cross-Database Search System is a powerful federated search engine or web portal that allows users to search many discrete health sciences databases at the National Center for Biotechnology Information (NCBI) website. NCBI is a part of the National Library of Medicine (NLM) itself a department of the National Institutes of Health (NIH) of the United States government. Entrez also happens to be the French second person plural form of the verb "to enter", meaning literally "come in", Entrez Global Query is an integrated search and retrieval system that provides access to all databases simultaneously with a single query string and user interface. Entrez can efficiently retrieve related sequences, structures and references. The Entez system can provide views of gene and protein sequences and chromosome maps. Some textbooks also available online through the Entrez system [21].

2. UniprotKB

The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. Along with the basic information about a protein entry i.e. amino acid sequence, protein name or description, taxonomic data and citation information, this database also includes widely accepted biological ontologies, classifications and cross-references, and clear indications of the quality of annotation in the form of evidence attribution of experimental and computational data. The UniProt Knowledgebase consists of two sections: a section containing manually-annotated records with information extracted from literature and curator-evaluated computational

analysis, and a section with computationally analyzed records that await full manual annotation. For the sake of continuity and name recognition, the two sections are referred to as "UniProtKB/Swiss-Prot" (reviewed, manually annotated) and "UniProtKB/TrEMBL" (unreviewed, automatically annotated), respectively. More than 99% of the protein sequences provided by UniProtKB are derived from the translation of the coding sequences (CDS) which have been submitted to the public nucleic acid databases, the EMBL-Bank/GenBank/DDBJ databases. All these sequences, as well as the related data submitted by the authors, are automatically integrated into UniProtKB/TrEMBL. In order to have minimal redundancy and to improve sequence reliability, all protein sequences encoded by a same gene are merged into a single UniProtKB/Swiss-Prot entry. Differences found between various sequencing reports are analysed and fully described in the feature table. Once in UniProtKB/Swiss-Prot, a protein entry is removed from UniProtKB/TrEMBL.

3. RCSB PDB

The Protein Data Bank (PDB) is a repository for the 3-D structural data of large biological molecules, such as proteins and acids. The data, typically obtained by X-ray crystallography or NMR spectroscopy and submitted by biologists and biochemists from around the world, are freely accessible on the Internet via the websites of its member organisations (PDBe, PDBj, and RCSB). The PDB is overseen by an organization called the Worldwide Protein Data Bank, wwPDB. The PDB was established in 1971 at Brookhaven National Laboratory and originally contained 7 structures. In 1998, the Research Collaboratory for Structural Bioinformatics (RCSB) became responsible for the management of the PDB. In 2003, the wwPDB was formed to maintain a single PDB archive of macromolecular structural data that is freely and publicly available to the global community. It consists of organizations that act as deposition, data processing and distribution centers for PDB data [22].

Bioinformatics tools/servers used

1. ProtParam

It is a tool which allows the computation of various physical and chemical parameters for a given protein stored in swiss-prot or TrEMBL or for a user entered sequence. The computed parameters include the molecular weight, theoritical pl, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY)[23].

2. For phylogenetic analysis

2.1 MEGA

MEGA is a multi-threaded Windows application. It runs on all releases of Microsoft Windows operating system. MEGA is an integrated tool for conducting automatic and manual sequence alignment, inferring phylogenetic trees, mining web-based databases, estimating rates of molecular evolution, inferring ancestral sequences, and testing evolutionary hypotheses [24].

2.2 FIG TREE

Fig tree is a Graphical user interface(GUI) application for viewing phylogenies and producing the publication quality figures.Its features include cross-platform tree display , three different tree stages :polar, rectangular and radial, display of node height ,branch length, support values and other annotations .Node height range range bars, collapse of clades into triangles,colouring of branches and tip lable,colouring by annotation ,quick search for tip labels or partial tip labels[25].

2.3 Bioedit

BioEdit is one of the most common program used in molecular biology studies. It was developed initially as a biological sequence alignment editor written for Windows only. It contains many features for sequence alignments modes of easy hand alignment, Split window view, user defined color, and information based shading and auto integration with other programs such as ClustalW and Blast. However, in the last few years it was developed dramatically to integrate many other features and functions and useful molecular tools for molecular biologist such as several modes of hand alignment, plasmid drawing and annotation, restriction mapping and much more. It became one of the widely used programs in molecular biology with its multipurpose tools in molecular biology. The freeware license and its efficient up to date modules beside its quick ability to produce results make it one of the most popular programs for molecular biologist nowadays [26].

3. For secondary structure prediction

3.1. SOPMA

Recently a new method called the self-optimized prediction method with alignment (SOPMA) has been described to improve the success rate in the prediction of the secondary structure of proteins. The SOPMA is based on homologue method of Levin et al. (1986). The improvement takes place because SOPMA takes place into the account the information from an alignment of sequences belonging to the same family. It might be noted that the SOPMA algorithm may take several minutes to compute large sequences (>500 amino acids). This method can be accessed via NPSA (Network Protein Sequence Analysis) server [27].

3.2. GOR4

The GOR method, named for the three scienties who developed it – Garnier, Osguthorpe, and Robson – is an information theory-based method developed not long after Chou-Fasman that uses more powerful probabilistic techniques of Bayesian inference. The GOR method takes into account not only the probability of each amino acid having a particular secondary structure, but also the conditional probability of the amino acid assuming each structure given

that its neighbors assume the same structure. This method is both more sensitive and more accurate because amino acid structural propensities are only strong for a small number amino acids such as proline and glycine. The original GOR method is roughly 65% accurate and is dramatically more successful in predicting the alpha helices than beta sheets, which it frequently mispredicts as loops or disorganised regions. GOR4 is the fourth version of GOR secondary structure prediction methods based on the information theory. There is no defined decision constant. Gor4 uses all possible pair frequencies within the window of 17 aminoacid residues [28].

4. For tertiary structure prediction

4.1. MODELLER

MODELLER is used for homology or comparative modelling 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. MODELLER implements comparative protein structure modelling by satisfaction of spatial restraints and can perform many additional tasks, including de novo modelling of loops in protein structures, optimization of various models of protein structure with respect to a flexibly defined objective function, multiple alignment of protein sequences and/or structures, clustering, searching of sequence databases, comparison of protein structures, etc. MODELLER is available for download for most Unix/Linux systems, Windows, and Mac.

4.2. For structure visualisation

DISCOVERY STUDIO VISUALIZER 3.0

Molecular visualization is a key aspect of molecular modeling and data analysis. It provides an understanding about the implications of a molecule's structure on certain interactions and biochemical reactions, while divulging mechanistic insight about a biochemical pathway.

5. For validation

5.1. PROCHECK

These Operating Instructions describe how to run the PROCHECK suite of programs (*Laskowski et al., 1993*) for assessing the "stereo chemical quality" of a given protein structure. The instructions assume that the programs have already been installed on your computer system. If this is not the case, please refer to the separate Installation Guide which deals with the installation procedures for your particular system.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. Unusual regions highlighted by PROCHECK are not necessarily errors as such, but may be unusual features for which there is a reasonable explanation (e.g. distortions due to ligand-binding in the protein's active site). Nevertheless they are regions that should be checked carefully.

5.2. VERIFY3D

The quality of the models was assessed by using the structure verification program verify 3D(Luthy et al. 1992) which tests the compatibility of a protein structure with its amino acid sequence. Verify 3D constructs a profile for the three dimentional model in which each residue position is characterised by its environmental score. The verify 3D profile is graphically represented by the numerical scores as a function of the residue number in the model. For high-resolution, experimentally determined structures, the verify 3D scores are positive and consinently(>0.2).

6 For energy minimisation

6.1. YASARA

YASARA is a molecular-graphics, -modeling and -simulation program for Windows,Linux and Mac OS X developed since 1993, which finally makes it really easy to answer your questions. With an intuitive user interface, photorealistic graphics and support for affordable shutter glasses, auto stereoscopic displays and input devices, YASARA creates a new level of

interaction with the 'artificial reality', that allows you to focus on your goal and forget about the details of the program.YASARA is powered by PVL (Portable Vector Language), a new development framework that provides performance way above traditional software . PVL allows you to visualize even the largest proteins and enables true interactive real-time simulations with highly accurate force fields on standard PCs (see benchmarks). You can push and pull molecules around and work with dynamic models instead of static pictures. [29]

METHODS

STEP-I: Collection of protein

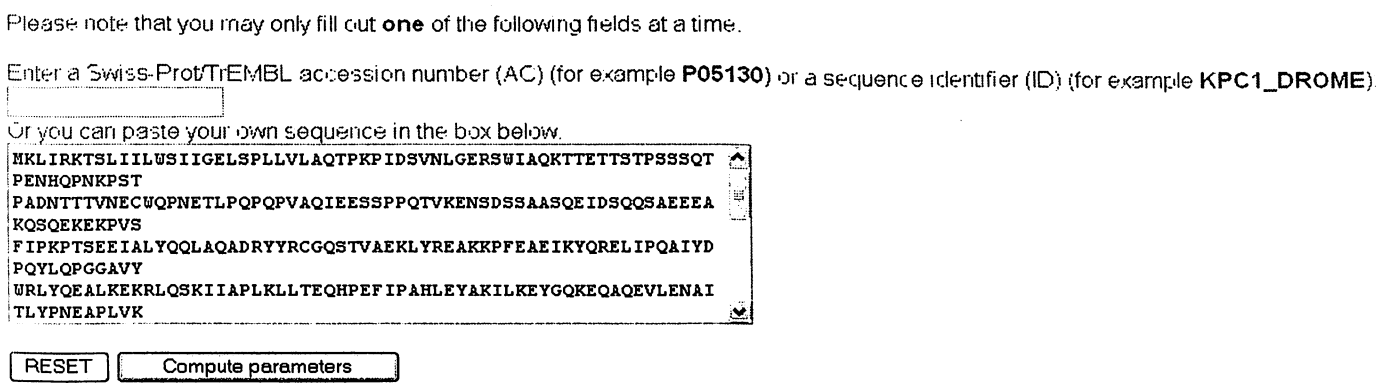
- The protein peptidase M48 Ste24p of cyanothece sp. PCC 7822 was retrieved from NCBI database having accession number YP_003888914.1. The sequence length of this protein is 628 aminoacid.
- The accession number of protein in UniprotKB/Swissprot is EOUH40.
- The protein was then converted in to fasta format and saved.

Steps of Protparam:

The home page of protparam was opened.

The raw sequence was pasted in the required box.

Fig.5: shows home page of protparam



Step II: Phylogenetic analysis of the protein

After the protein retrieved from the NCBI in fasta then processed for phylogenetic analysis for that purpose blastp of NCBI was used. The fasta sequence was submitted in the submission box. Then parameter set for similarity search was like this, database- nr database

Then homologous sequences were available in the result page. Then the sequences with lowest E value, identity between 25% and above, query coverage with 30% and above of different species other than cyanothece were selected. On the result page by clicking on the option selected species, the selected sequences were send to the ncbi. Then the sequences were saved in FASTA (text) format. Then all the selected sequences were saved in notedpad.

Steps for alignment and construction of tree in MEGA

- After collecting the sequences in the notepad and saved it in fasta format for example mega.fasta and save it.
- Now after installation of MEGA, open mega homepage then click on align.
- Then click on Edit/build alignment, and then click on retrieves the sequence from a file.
- Then the alignment was appeared, shown in the figure
- Then click on Alignment, then click on Align by clustalw, after selecting all sequences for alignment a clustalw parameter box was appeared.
- Then after setting all parameters as default.

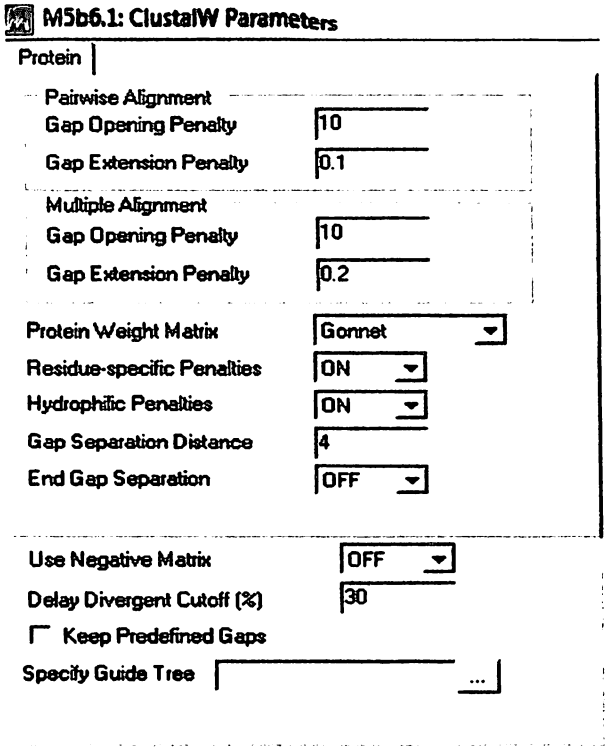
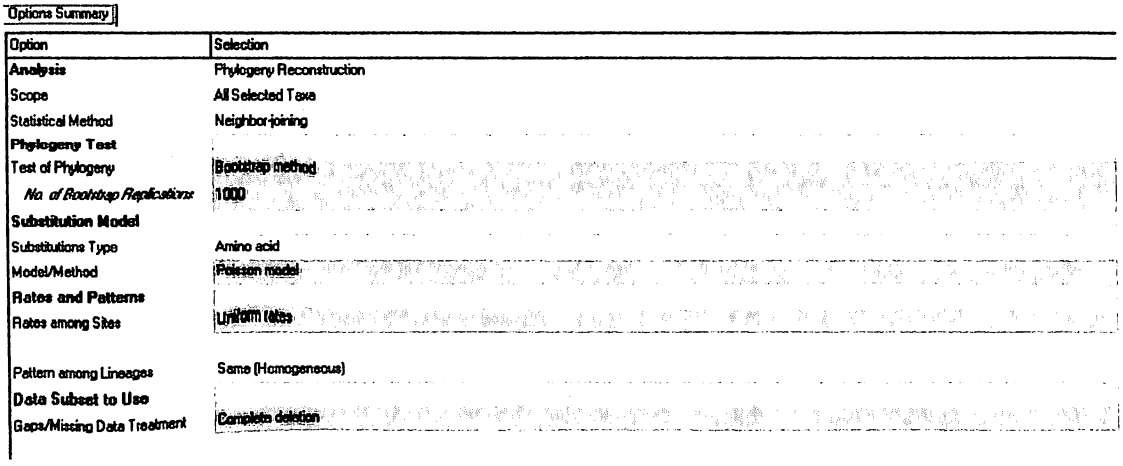


Fig.6: shows parameter of clustalw

- Then after clicking on ok button both pairwise and multiple alignment was done.
- Then that alignment was saved in mega format by clickjing on file, then on Data, then on export file as MEGA.
- Then from mega homepage, click on phylogeny Construct/Test neighbor-joining tree.
- Then set parameter as no. of bootstrap replications-1000.
- Then click on compute.



STEP-III: Steps for protparam:

From Google search go to homepage of protparam.

Then the raw sequence was submitted in the submission box.

Then the programme was run.

STEP-IV: Secondary structure prediction

SOPMA:

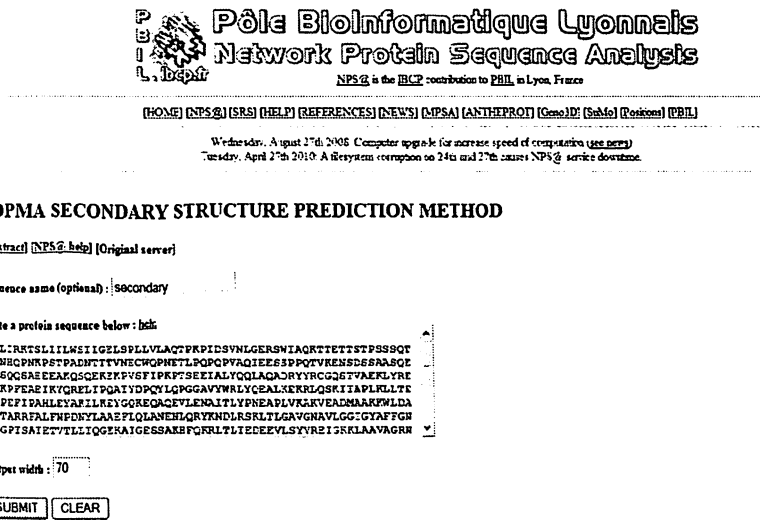


Fig.9: Shows homepage of SOPMA

Steps for SOPMA

- The fasta sequence of the peptidase M48 Ste24p[cyanothece sp. PCC 7822] was copied.
- EXPASY proteomic server was opened.
- From there secondary structure prediction tools and then SOPMA was selected and clicked. The homepage was obtained.
- There the fasta sequence was pasted in the specified position after submitting the request the result page was obtained .
- The result page had several para meters like window width, similarity threshold, number of states.
- It had information about alpha helix,310 helix,pie helix, beta bridge,extended strands ,beta term ,bend region ,random coil,ambigous states.

- Also it has some a graph & scale showing helix,sheets,turns,coils,taking those previously mentioned parameters .Also a text file was obtained showing several other protein information.

GOR4

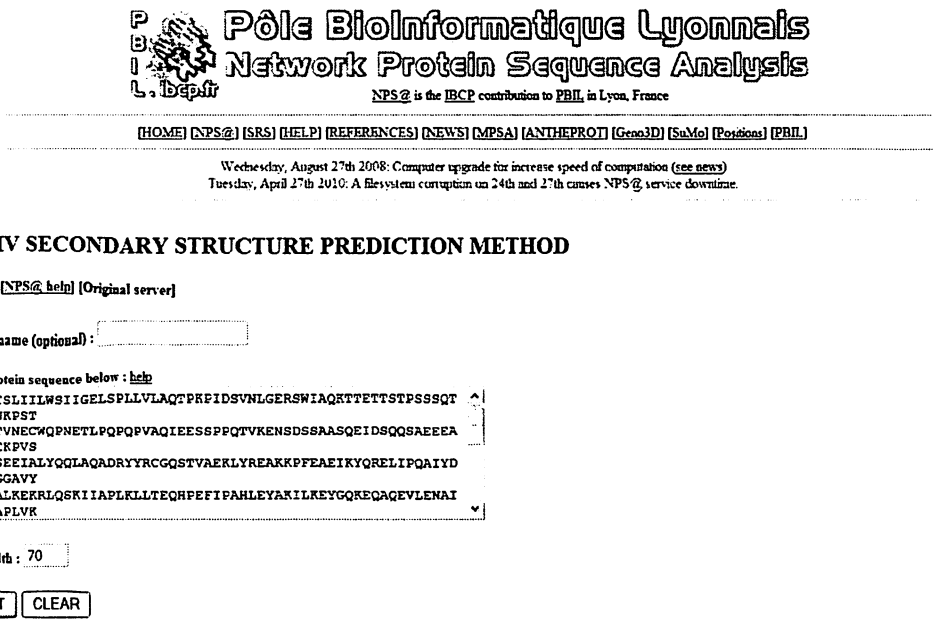


Fig.9: Shows homepage of GOR4

Steps for GOR4

- The fasta sequence of the peptidase M48 Ste24p[cyanothece sp. PCC 7822] was copied.
- The GOR4 home page was opened by mentioning its URL.The sequence was pasted in the submission box in home page of GOR4.
- After submitting the request by keeping the output width as 70 as default, the result page was obtained ,
- The result page had a color coded prediction, percentage of each secondary element, and the text file.

Steps for psipred

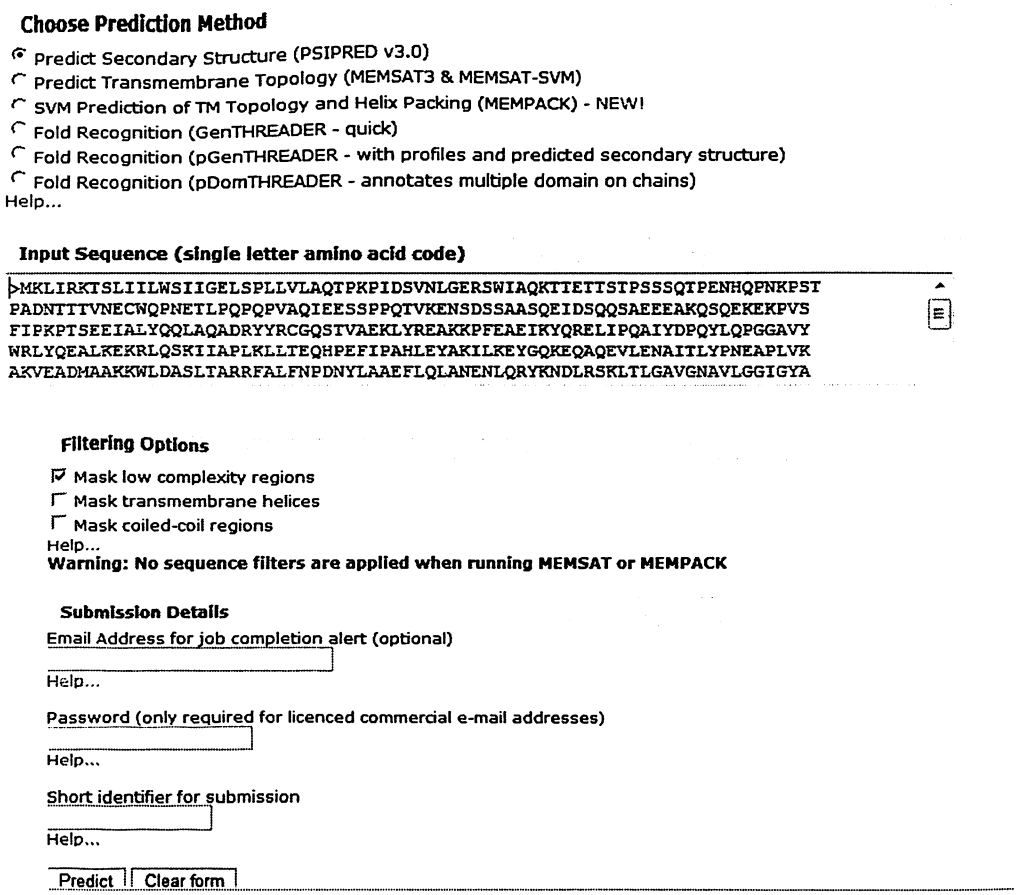


Fig. 10: Shows homepage of PSIPred

- As shown in the figure the steps for psipred is
- The homepage of the psipred was opened.
- Then by taking the default parameters the sequence was inserted inside the input box.
- Then the predict button was clicked.

For Alignment

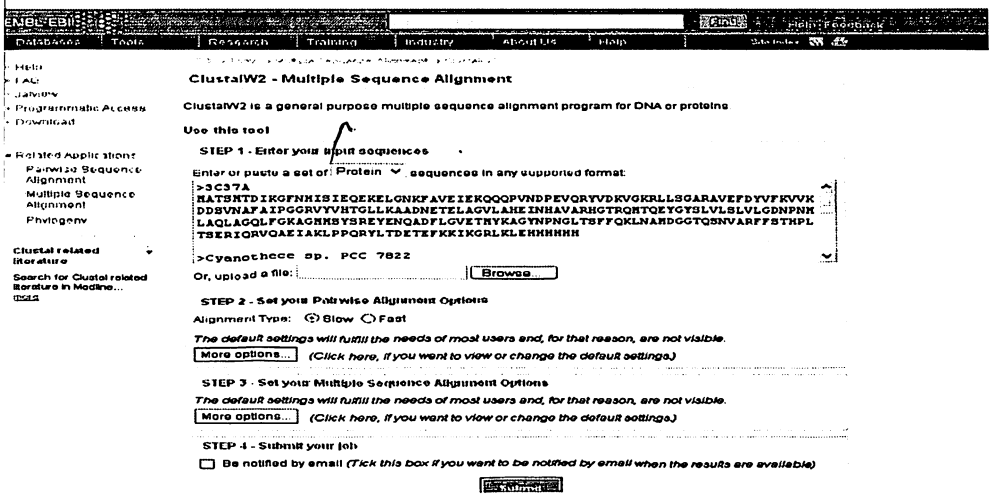


Fig.11: Shows homepage of clustalW alignment

Steps for ClustalW alignment

- The clustalW home page was opened.
- In the submission box the raw sequence and the template sequence was submitted in the fasta format.
- The by taking the default parameter alignment was done.

STEP 5: Tertiary structure prediction

COMPARATIVE MODELING USING MODELLER 9.9

PREPARING INPUT FILES

1. **Preparing alignment file.**
    - Generate the multiple sequence alignment file in PIR format.
    - The PIR file was modified as required and saved again as .ali file format.
    - The template structure was downloaded from PDB and hetero atoms and all chains were deleted except the required chain.
    - Then the script file was prepared by using python program.
    - Then in modeller command prompt (python file name).py command was run by giving appropriate path.

Target sequence in .pir format

```
>P1;target
sequence:target:::::::0.00: 0.00
TLLIQGEKAIGESSAKHFQKRLTLIEDEEVLSYVREIGKKLAAVAGRNEFNYEFYVVMDDNLNAF
ALPGGKVFINAGAILKTNSEAELAGLMAHELSHAVLSHSFQIMTGGNLLANATQFVPYLGSSAGD
LITLNYSRDMEREADIFGTRLLAASGYAADGVRNLMVVLEKEDDPSPPAWLSTHPDTSDRVKYVE
KMLVQERLNRYAYEGVERHWKIRNRVGELLDKYRQ*
```

Alignment script file (align2d.py)

```
from modeller import *

env = environ()
aln = alignment(env)
mdl = model(env, file='3C37', model_segment=('FIRST:A','LAST:A'))
aln.append_model(mdl, align_codes='3C37A', atom_files='3C37.pdb')
aln.append(file='target.ali', align_codes='target')
aln.align2d()
aln.write(file='alignment.ali', alignment_format='PIR')
aln.write(file='alignment.pap', alignment_format='PAP')
```

As gap is coming in the alignment again seven residues were deleted from starting position and eight residues at the ending position.

After that the new target was again taken for alignment, and finaly model was generated.

To generate model the script file wad generated using python programme.

```
from modeller import *
from modeller.automodel import *

env = environ()
a = automodel(env, alnfile='alignment.ali',
              knowns='3C37A', sequence='target',
              assess_methods=(assess.DOPE, assess.GA341))
a.starting_model = 1
a.ending_model = 100
a.make()

# Get a list of all successfully built models from a.outputs
ok_models = filter(lambda x: x['failure'] is None, a.outputs)

# Rank the models by DOPE score
key = 'DOPE score'
ok_models.sort(lambda a,b: cmp(a[key], b[key]))

# Get top model
m = ok_models[0]
print "Top model: %s (DOPE score %.3f)" % (m['name'], m[key])
```

- Hundred number of models were generated, out of which ten number of best models were selected according to their lowest dope score.
- The best model was superimposed with it's template to check whether loop is present or not.

Steps for superimposition

- The model was first opened in DS visualiser.
- Then the file was clicked, then the insert file was clicked and the template was selected.
- Then in upper panel the structure was opened their the superimpose option was clicked, their By Center of Geometry was clicked in the purpose of doing superimpose clearly.

Then five best models sorted out.

Then for each model validation was done by using procheck.

Steps for validation

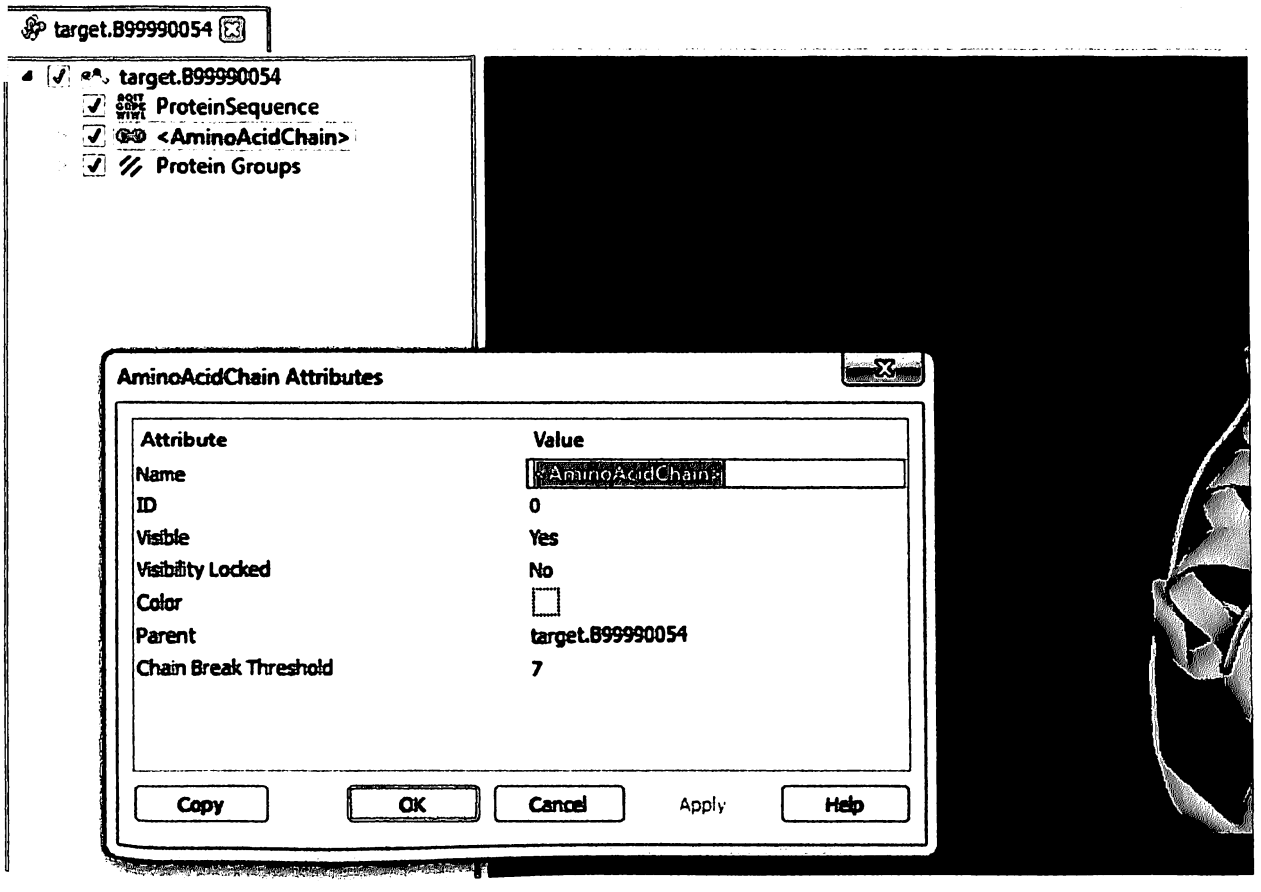
- Validation was done in the ADIT validation server, available in PDB.
- Here validation was done in procheck.
- The RSCB validation server allows the user to check the format of coordinate and structure factor files and also to create a variety of validation reports about a structure factor files and also to create a variety of validation reports about a structure.
- Before that we will save the pdb file in .gzip or .z format to speed up the file upload and validation process.
- First from the pull-down menu, the experimental method was selected as x-ray crystallography and after pressing the BEGIN button a new web page was appeared.
- Then the coordinate file was browsed, the coordinated file was specified as pdb.
- Then precheck was selected for operation and begin button was clicked.
- A brief was produced by procheck which identified changes need to be made in the data files.
- Then after few minutes result page was appeared, then after clicking on the click option that was like this click here to continue to validation, after few seconds the validation report page was appeared.
- Then the ramachandran plot was generated.

In this process the five top models were validated, basing upon the best validation report and lowest dope score the best model was selected.

As loops are located on the best model to avoid loop, loop modeling was done.

Steps for loop modelling using MODLOOP

- The modloop server was opened.
- The required fields were filled up, then in the submission box the position of the residues for the loop modelling was given along with its chain.
- 166:A:179:A:
- 166 is the starting residue's location from which loop is starting and ends with the 179 amino acid.
- Capital 'A' is the chain id of the sequence.
- Before going for loop modelling, in the pdb file of the target after the opening of the model in the DS visualiser in the hierarchy the amino acid heading was selected. Then by doing right click on it the properties was appeared, there the chain id was changed to A, instead of amino acid.
- Then the file was saved.



- Then after running the modloop the output file was appeared.
- Then the validation of the output file of modloop was validated.

Steps for removal of bumps

- For the removal of bumps the What if server home page was opened.
- First the process complete structure was done, in what if server the complete structure was clicked.
- In its homepage the model (best model) was uploaded and the programme was run.
- The resulted pdb fie was selected in the homepage of removal of bumps, and the programme was run.
- The resulted file was again validated in the procheck.

Steps for evaluation

In modeller evaluation was done by taking the script file for the target.

```
from modeller import *
from modeller.scripts import complete_pdb

log.verbose()     # request verbose output
env = environ()
env.libs.topology.read(file='$(LIB)/top_heav.lib') # read
topology
env.libs.parameters.read(file='$(LIB)/par.lib') # read
parameters

# read model file
mdl = complete_pdb(env, 'output.pdb')

# Assess with DOPE:
s = selection(mdl)    # all atom selection
s.assess_dope(output='ENERGY_PROFILE NO_REPORT',
file='target.profile',
               normalize_profile=True, smoothing_window=15)
```

Then command was given in modeller command prompt

mod9.9 evaluate_<filename for target evaluation>.py

The script file for the template

```
from modeller import *
from modeller.scripts import complete_pdb

log.verbose()     # request verbose output
env = environ()
env.libs.topology.read(file='$(LIB)/top_heav.lib') # read
topology
env.libs.parameters.read(file='$(LIB)/par.lib') # read
parameters

# directories for input atom files
env.io.atom_files_directory = './:../atom_files'

# read model file
mdl = complete_pdb(env, '3C37.pdb', model_segment=('FIRST:A',
'LAST:A'))

s = selection(mdl)
s.assess_dope(output='ENERGY_PROFILE NO_REPORT',
file='3C37A.profile',
              normalize_profile=True, smoothing_window=15)
```

Then command was given in modeller command prompt

mod9.9 evaluate_<filename for template evaluation>.py

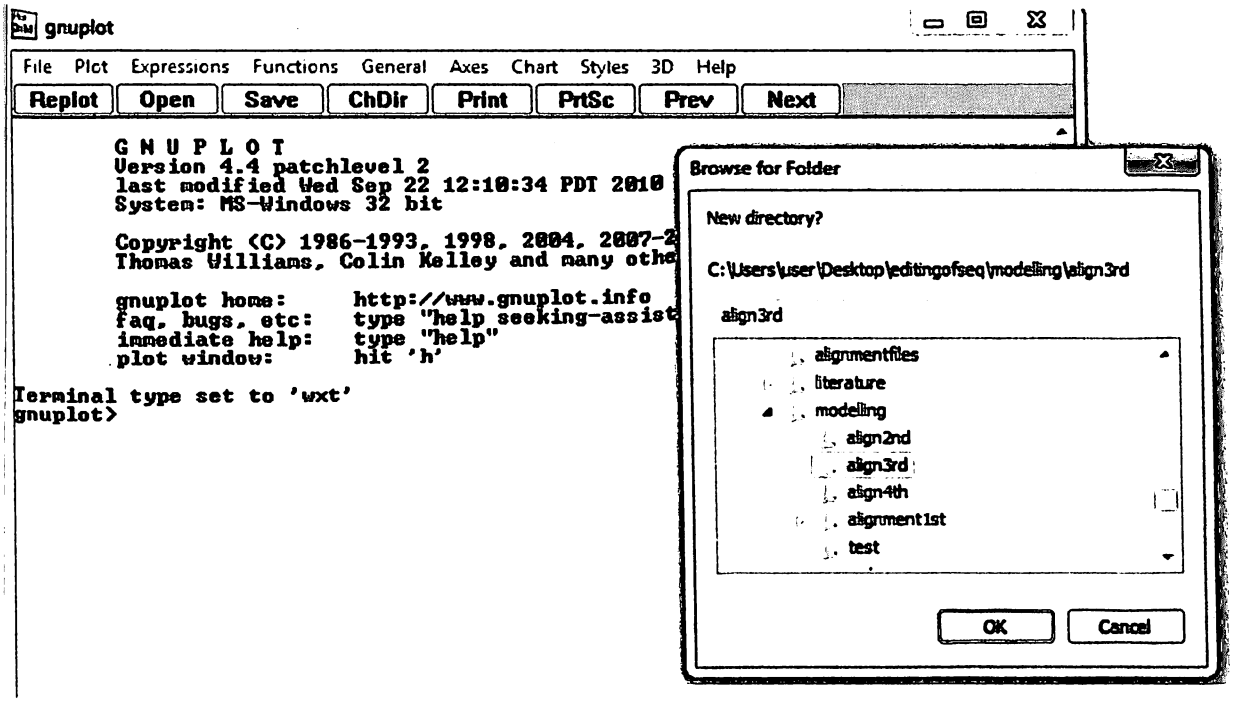


Fig.12: Shows gnu plot command prompt

- Then gnu plot was downloaded.
- In the gnu plot command prompt first the file was retrieved by clicking on ChDir button as given in the figure no. 12.
- Then in command prompt of gnu plot the commands: Plot "<filename>.profile" using 1:42 with lines was given, in file name the profile file name for target and template was given and was run.

STEP-5: Step for energy minimization

- For energy minimization the YASARA server was used.
- First the server homepage was opened, and then the model was uploaded.
- The Email id was given as for requirement.
- Then the programme was run.
- Then after that the YASARA was downloaded as the result was come in SCE format which can be open in YASARA server only.

STEP-6: step for identifying metal binding site

First in the target sequence residue which binds to metal Zn were identified, by identifying corresponding metalbinding residues in template.

The residues were identified in the alignment, of the modeller output file.

In template the metal binding residues are His 106, His 110, Glu 162, and His 208.

The patterns for these amino acids were identified in the sequence.

Then by taking patterns in to account, the metal binding residues were identified in the target sequence.

RESULTS AND DISCUSSION

Atomic composition:

Carbon	C	3141
Hydrogen	H	4973
Nitrogen	N	851
Oxygen	O	967
Sulfur	S	10

Formula: $C_{3141}H_{4973}N_{851}O_{967}S_{10}$
Total number of atoms: 9942

Aliphatic index: 84.89

Grand average of hydropathicity (GRAVY): -0.542

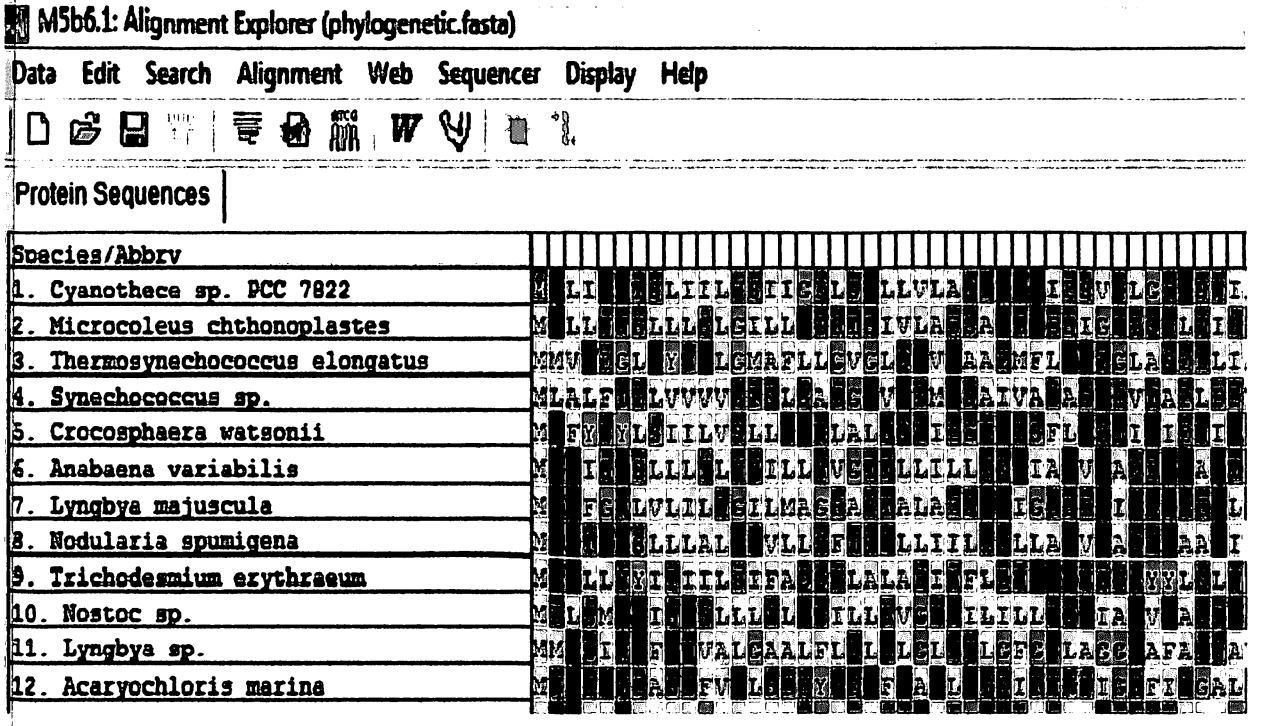


Fig.13: Shows Alignment page of MEGA

This figure represents the alignment produced by MEGA, by using clustalw algorithm.

The 64 number of different species with the cyanothece species of metallopeptidase family were aligned based on clustalw alignment.

Fig.14: Shows results of phylogenetic tree

The figure 14 represents the tree of 65 numbers of aligned species.

The tree was constructed on the basis of neighbour-joining method, where 1000 bootstrap was taken.

The tree was represented in the circular manner.

Fig.15: Shows the result of figtree.

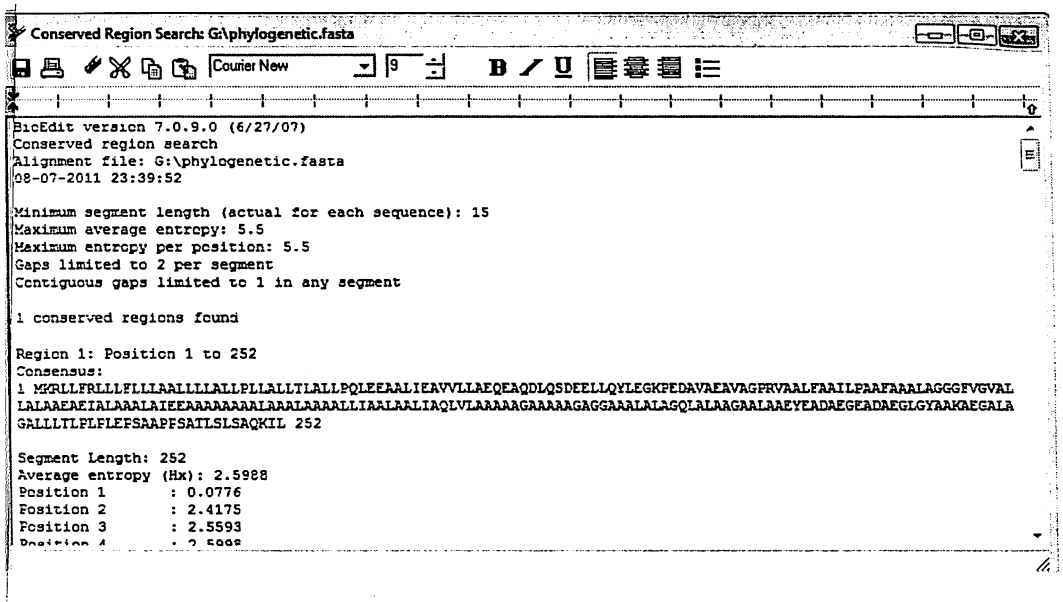
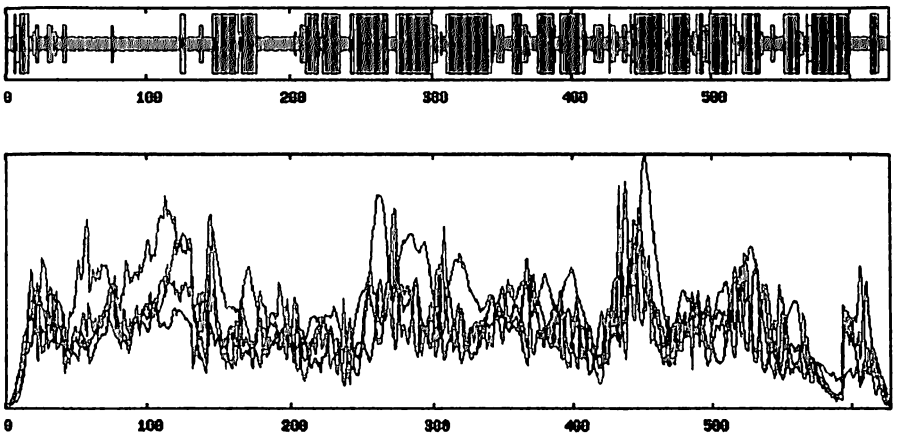


Fig.17: Shows the result for bioedit.

```
         10        20        30        40        50        60        70
          |         |         |         |         |         |         |
MKRLLFRRTSLIILWSIIGELSPLLVLAQTPKPIDSVNLGERSWIAQKTTETTSTPSSSQTPENHQPNKPST
hhhcccchheehhhhhhhccctеeeccccccceeeecttcccceeeccccccccccccccccccccccccc
PADNTTTVNECWQPNETLPQPQPVAQIEESSPPQTVKENSDSSAASQEIDSQQSAEEEAKQSQEKEKPVS
ccccccceeccccccccccccccccccccccccccccccccccccccccccccccccchhhhhcccccccccee
FIPKPTSEEIALYQQLAQADRYYRCGQSTVAEKLYREAKKPFEAEIKYQRELIPQAIYDPQYLQPGGAVY
ecccccchhhhhhhhhhhhhhhhhhtchhhhhhhhhhhhcccccccccccccccccccccccccccttccee
WRLYQEALKEKRLQSKIIAPLKLLTEQHPEFIPAHLEYAKILKEYGQKEQAQEVLENAITLYPNEAPLVK
ehhhhhhhhhhtchhhhhhhhhhhhhcccccchhhhhhhhhhhhhhhhhhhhhhhhhhhcccccchhhh
AKVEADMAAKKWLDASLTARRFALFNPDNYLAAEFLQLANENLQRYKNDLRSKLTLGAVGNAVLGGIGYA
hhhhhhhhhhhhhhhhhhhhheeeecttccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhetcceee
FFGNIGGPISAIETVTLLIQGEKAIGESSAKHFQKRLTLIEDEEVLSYVREIGKKLAAVAGRNEFNYEFY
eecccccchhhhhhheeeettcccccchhhhhhhhhhhhecccchhhhhhhhhhhhhhhhhhtcccccceeee
VVMDDNLNAFALPGGKVFINAGAILKTNSEAELAGLMAHELSHAVLSHSFQIMTGGNLLANATQFVPYLG
eeeccccheeecttcceeeettheehhhhhhhhhhhhhhhhhhhhhhhttchhhhhhhhhhhhhhhcctth
SSAGDLITLNYSRDMEREADIFGTRLLAASGYAADGVRNLMVVLEKEDDPSPPAWLSTHPDTSDRVKYVE
cchhheeeehhhhhhhhhhhhhhheeehttcchhhhhhhhhhhhhttcccccceeeccccccchhhhhhh
KMLVQERLNRYAYEGVERHWKIRNRVGELLDKYRQEKEEKEGKKSKKPPETTPKTQQIIQQQQENRPI
hhhhtttccheeehhhhhhhhhhhhhhhhhhhhhhhhhhhhhtcccccccccccccchhhhhhccccccc
Sequence length :   628
`0KSOPMA :
   Alpha helix        (Hh):   309 is  49.20%
   3₁₀ helix          (Gg):     0 is   0.00%
   Pi helix           (Ii):     0 is   0.00%
   Beta bridge        (Bb):     0 is   0.00%
   Extended strand    (Ee):    64 is  10.19%
   Beta turn          (Tt):    29 is   4.62%
   Bend region        (Ss):     0 is   0.00%
   Random coil        (Cc):   226 is  35.99%
   Ambigous states    (?) :     0 is   0.00%
   Other states           :     0 is   0.00%
```



Parameters :
Window width : 17
Similarity threshold : 8
Number of states : 4

Fig.17: Shows the secondary structure of the protein sequence was predicted in the server SOPMA.

```
        10         20         30         40         50         60         70
         |          |          |          |          |          |          |
MKLIRKTSLIILWSIIGELSPLLVLAQTPKPIDSVNLGERSWIAQKTTETTSTPSSSQTPENHQPNKPST
      hhhhhhhccccchhhhccccccccceeeccceeeeeeccccccccccccccccccccccc
PADNTTTVNECWQPNETLPQPQPVAQIEESSPPQTVKENSDSSAASQEIDSQQSAEEEAKQSQEKEKPVS
    eeeeeecccccccccccchhhhcccccccceeecccccchhhhhhhhhhhhhhhhhhhhhhcccce
FIPKPTSEEIALYQQLAQADRYYRCGQSTVAEKLYREAKKPFEAEIKYQRELIPQAIYDPQYLQPGGAVY
e    chhhhhhhhhhhcceeecccchhhhhhhhhcchhhhhhhhhhhcccccccccccccccchhh
WRLYQEALKEKRLQSKIIAPLKLLTEQHPEFIPAHLEYAKILKEYGQKEQAQEVLENAITLYPNEAPLVK
hhhhhhhhhhhhhhhhhhhhhhhhcccccchhhhhhhhhhhhhchhhhhhhhhhhhhhcccccchhh
AKVEADMAAKKWLDASLTARRFALFNPDNYLAAEFLQLANENLQRYKNDLRSKLTLGAVGNAVLGGIGYA
hhhhhhhhhhhhhhhhhhhhhhhcccccchhhhhhhhhhhhhhhhhhhhhhhhhhhccccceeecccccee
FFGNIGGPISAIETVTLLIQGEKAIGESSAKHFQKRLTLIEDEEVLSYVREIGKKLAAVAGRNEFNYEFY
ee    cchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhcccccceee
VVMDDNLNAFALPGGKVFINAGAILKTNSEAELAGLMAHELSHAVLSHSFQIMTGGNLLANATQFVPYLG
eeee    ccccc eeeecchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhcccccccceeeccccc
SSAGDLITLNYSRDMEREADIFGTRLLAASGYAADGVRNLMVVLEKEDDPSPPAWLSTHPDTSDRVKYVE
    ccceeehhhhhhhhhhhhhhhhhhhhhhhcccccchhheeeecccccccceeeeccccccchhhhhh
KMLVQERLNRYAYEGVERHWKIRNRVGELLDKYRQEKEEKEGKKSKKPPETTPKTQQIIQQQQENRPI
hhhhhhhhhhhhhhcc hhhhhhhcccchhhhhhhhhhhhhhhhcccccccccccchhhhhhhhhcceec
Sequence length :   628
GOR4 :
   Alpha helix      (Hh) :   344 is  54.78%
   3₁₀ helix        (Gg) :     0 is   0.00%
   Pi helix         (Ii) :     0 is   0.00%
   Beta bridge      (Bb) :     0 is   0.00%
   Extended strand  (Ee) :    58 is   9.24%
   Beta turn        (Tt) :     0 is   0.00%
   Bend region      (  ) :     0 is   0.00%
   Random coil      (cc) :   226 is  35.99%
   Ambigous states  (?)  :     0 is   0.00%
   Other states          :     0 is   0.00%
```

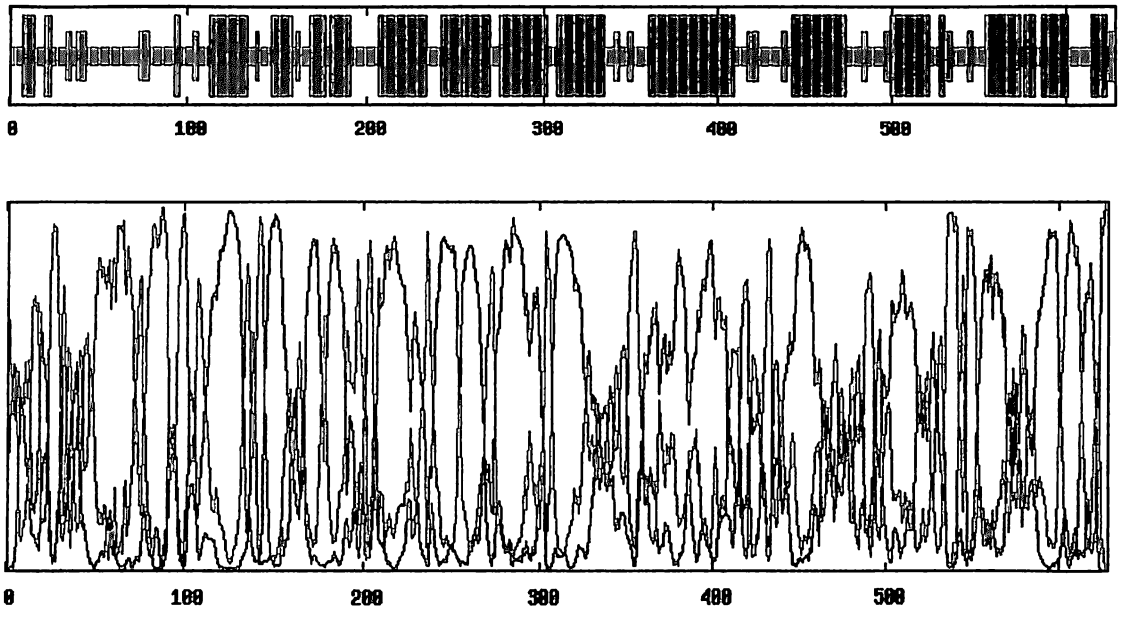


Fig.18: shows the secondary structure of the protein sequence was predicted in the server GOR4.

Tool used	Alpha helix Percentage	3_{10} helix	Pi helix	Beta turn
GOR IV	54.78%	0.00%	0.00%	0.00%
SOPMA	49.20%	0.00%	0.00%	4.62%

Table.1: shows the comparision between the sopma and the gor4 result basing upon the alpha helix and beta turn region.

PSIPRED HFORMAT (PSIPRED V3.0)

```
Conf: 942244788988878876245310024789887667676763434665323489999999
Pred: CCCHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
  AA: MKLIRKTSLIILWSIIGELSPLLVLAQTPKPIDSVNLGERSWIAQKTTETTSTPSSSQTP
              10        20        30        40        50        60

Conf: 888999999988886555668999999999988645545899987767888854432236
Pred: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
  AA: ENHQPNKPSTPADNTTTVNECWQPNETLPQPQPVAQIEESSPPQTVKENSDSSAASQEID
              70        80        90       100       110       120

Conf: 532035653220013688767999998999999999964477860561689999999469
Pred: CCCHHHHHHHHHHCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHCCCHHHHHHHHHHHCC
  AA: SQQSAEEEAKQSQEKEKPVSFIPKPTSEEIALYQQLAQADRYYRCGQSTVAEKLYREAKK
             130       140       150       160       170       180

Conf: 860245542102786777877668882156999899996311132101123455431698
Pred: CCCHHHHHHHCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHCCCCCCCCCHHHHHHHHCCC
  AA: PFEAEIKYQRELIPQAIYDPQYLQPGGAVYWRLYQEALKEKRLQSKIIAPLKLLTEQHPE
             190       200       210       220       230       240

Conf: 603457999999880994689999998551289981567662488999867999768665
Pred: CHHHHHHHHHHHHHCCCHHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHHHHHHHHHHHH
  AA: FIPAHLEYAKILKEYGQKEQAQEVLENAITLYPNEAPLVKAKVEADMAAKKWLDASLTAR
             250       260       270       280       290       300

Conf: 421269980537999882499799989877999999999999998898863147789976
Pred: HHHCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCC
  AA: RFALFNPDNYLAAEFLQLANENLQRYKNDLRSKLTLGAVGNAVLGGIGYAFFGNIGGPIS
             310       320       330       340       350       360

Conf: 311345699999999999999997359889999999999999999719999657999
Pred: CCCCCCCCHHHHHHHHHHHHHHHHHHHCCCCCCHHHHHHHHHHHHHHHHHHHCCCCCEEEE
  AA: AIETVTLLIQGEKAIGESSAKHFQKRLTLIEDEEVLSYVREIGKKLAAVAGRNEFNYEFY
             370       380       390       400       410       420

Conf: 982899881002899388726045309995799999999999865122999999727999
Pred: EEECCCCCCCCCCCCEEEECHHHHHCCCCHHHHHHHHHHHHHHHHCCCHHHHHHHHHHHH
  AA: VVMDDNLNAFALPGGKVFINAGAILKTNSEAELAGLMAHELSHAVLSHSFQIMTGGNLLA
```

```
            430        440        450        460        470        480
Conf: 99999889899999987714598134898999999999984999589899999999853999
Pred: HHHHHHHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHHHHHCCCCHHHHHHHHHHHHHCCCC
  AA: NATQFVPYLGSSAGDLITLNYSRDMEREADIFGTRLLAASGYAADGVRNLMVVLEKEDDP
            490        500        510        520        530        540
Conf: 9882002899965889999999786121014621157789999999976776567787 89
Pred: CCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHCCCCHHHHHHHHHHHHHHHHHHHHHHHHH
  AA: SPPAWLSTHPDTSDRVKYVEKMLVQERLNRYAYEGVERHWKIRNRVGELLDKYRQEKEEK
            550        560        570        580        590        600
Conf: 500289999999210788975020 4899
Pred: HCCCCCCCCCCCCCHHHHHHHHHHHCCCC
  AA: EGKKSKKPPETTPKTQQIIQQQQENRPI
            610        620
```

Fig.19: Shows the result for PSIPred

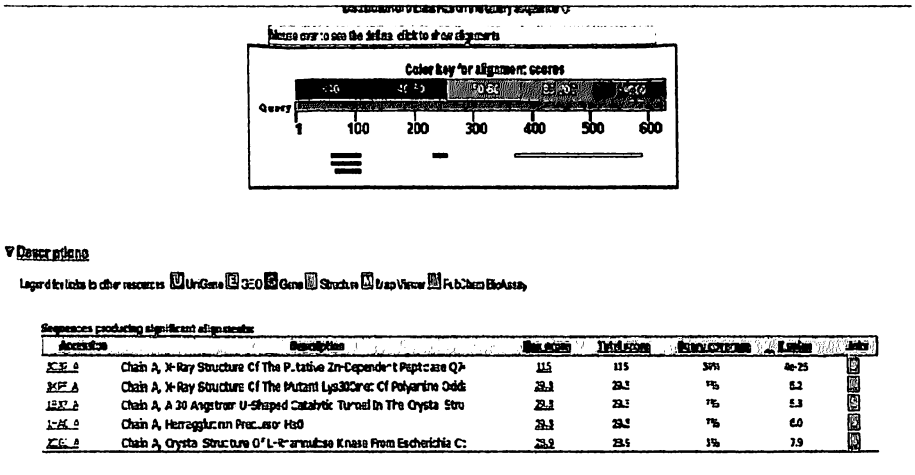


Fig .20: Shows the result for Blastp

Accession	Description	Max score	Total score	Query coverage	E value
3C37A	Chain A, X-Ray Structure Of The Putative Zn-Dependent Peptidase	115	115	34%	4e-25
3KPF A	Chain A, X-Ray Structure Of The Mutant Lys300met Of Polyamine	29.3	29.3	7%	5.2
1B37 A	Chain A, A 30 Angstrom U Shaped catalytic Tunnel In Crystal	29.3	29.3	7%	5.3
1HA0 A	Chain A, Hemagglutinin Precursor Ha0	29.3	29.3	7%	6.0
2CGJ A	Chain A, Crystal Structure Of L-Rhamnulose Kinase in Escherichia	28.9	28.9	3%	7.9

Table.2:

The table 2 represents the detail information about the homologous sequences predicted against pdb database, of peptidase M48 Ste24p of *cyanothece sp. PCC 7822.*

In this table the detail information about the homologous sequence is given, such as query coverage, e-value, total and maximum score.

The homologous sequence that has taken for structure prediction is said to be template.

Result for clutalW

```
Cyanothecesp.       MKLIRKTSLIILWSIIGELSPLLVLAQTPKPIDSVNLGERSWIAQKTTETTSTPSSSQTP 60
3C37A               ------------------------------------------------------------

Cyanothecesp.       ENHQPNKPSTPADNTTTVNECWQPNETLPQPQPVAQIEESSPPQTVKENSDSSAASQEID 120
3C37A               ------------------------------------------------------------
```

```
cyanothecesp.       SQQSAEEEAKQSQEKEKPVSFIPKPTSEEIALYQQLAQADRYYRCGQSTVAEKLYREAKK 180
3C37A               ------------------------------------------------------------

cyanothecesp.       PFEAEIKYQRELIPQAIYDPQYLQPGGAVYWRLYQEALKEKRLQSKIIAPLKLLTEQHPE 240
3C37A               ------------------------------------------------------------

cyanothecesp.       FIPAHLEYAKILKEYGQKEQAQEVLENAITLYPNEAPLVKAKVEADMAAKKWLDASLTAR 300
3C37A               ------------------------------------------------------------

cyanothecesp.       RFALFNPDNYLAAEFLQLANENLQRYKNDLRSKLTLGAVGNAVLGGIGYAFFGNIGGPIS 360
3C37A               --------------MATSMTDIKGFN------------------------------- 12
                                   :  :  ::: ::

cyanothecesp.       AIETVTLLIQGEKAIGESSAKHFQKRLTLIEDEEVLSYVREIGKKLAAVAGRNEFNYEFY 420
3C37A               -----MISIEQEKELGNKFAVEIEKQQQPVNDPEVQRYVDKVGKRLLSGARAVEFDYVFK 67
                          : *: ** :*:. * .::*:   ::* **  ** ::**:* : *   **:* *

cyanothecesp.       VVMDDNLNAFALPGGKVFINAGAILKTNSEAELAGLMAHELSHAVLSHSFQIMTGGN--- 477
3C37A               VVKDDSVNAFAIPGGRVYVHTGLLKAADNETELAGVLAHEINHAVARHGTRQMTQEYGYS 127
                    ** **.:****:***:*::::* :  ::.*:****::***:.***  *. : **

cyanothecesp.       ------LLANATQFVPYLGSSAGDLITLNYSRDMEREADIFGTRLLAASGYAADGVRNLM 531
3C37A               LVLSLVLGDNPNMLAQLAGQLFGKAGMMSYSREYENQADFLGVETMYKAGYNPNGLTSFF 187
                          *  *.. :.   *.  *.   :.***: *.:**::*.. :  :** .:*: .::

cyanothecesp.       VVLEKED---DPSPPAWLSTHPDTSDRVKYVEKMLVQERLNRYAYEGVERHWKIRNRVGE 588
3C37A               QKLNAMDGGTQSNVARFFSTHPLTSERIQRVQAEIAKLPPQRYLTD-ETEFKKIKGRLKL 246
                      *:  *   :.. . ::**** *:  :.:   :**  :   .. **:.*:
  :*::
cyanothecesp.       LLDKYRQEKEEKEGKKSKKPPETTPKTQQIIQQQQENRPI 628
3C37A               EHHHHHH---------------------------------- 253
                      .::::
```

Fig.21: Shows the alignment of the target and template sequence using clustaw.

Where '*' represents for conserved regions, '.' represents identity, ':' represents for semiconserved regions.

```
 _aln.pos          10        20        30        40        50        60
3C37A      KGFNMISIEQEKELGNKFAVEIEKQQQPVNDPEVQRYVDKVGKRLLSGARAVEFDYVFKVVKDDSVNA
target     ---TLL-IQGEKAIGESSAKHFQKRLTLIEDEEVLSYVREIGKKLAAVAGRNEFNYEFYVVMDDNLNA
 _consrvd         *  **  *   *    *       * **  **   ** *   *   ** * * ** **  **

 _aln.p    70        80        90        100       110       120       130
3C37A      FAIPGGRVYVHTGLLKAADNETELAGVLAHEINHAVARHGTRQMTQEYGYSLVLSLVLGDNMLAQLAG
target     FALPGGKVFINAGAILKTNSEAELAGLMAHELSHAVLSHSFQIMT---GGNLLANATQFVPYLGSSAG
 _consrvd ** *** *    *       * **** ***  ***  *    **   *  *          *   **

 _aln.pos  140       150       160       170       180       190       200
3C37A      QLFGKAGMMSYSREYENQADFLGVETMYKAGYNPNGLTSFFQKLN----------ATHPLTSERIQR
target     DLI----TLNYSRDMEREADIFGTRLLAASGYAADGVRNLMVVLEKEDDPSPPAWLSTHPDTSDRVKY
 _consrvd  *        ***  *  **  *       **   *       *             *** ** *

 _aln.pos    210       220       230       240
3C37A      VQAEIAKLPPQRYLTDETEFK-KIKGRLK--LE----
target     VEKMLVQERLNRYAYEGVERHWKIRNRVGELLDKYRQ
 _consrvd *          **     *   **  *    *
```

Fig.22: Represents the structural alignment done in the modeller by taking the target protein sequence and template sequence 3C37 A.

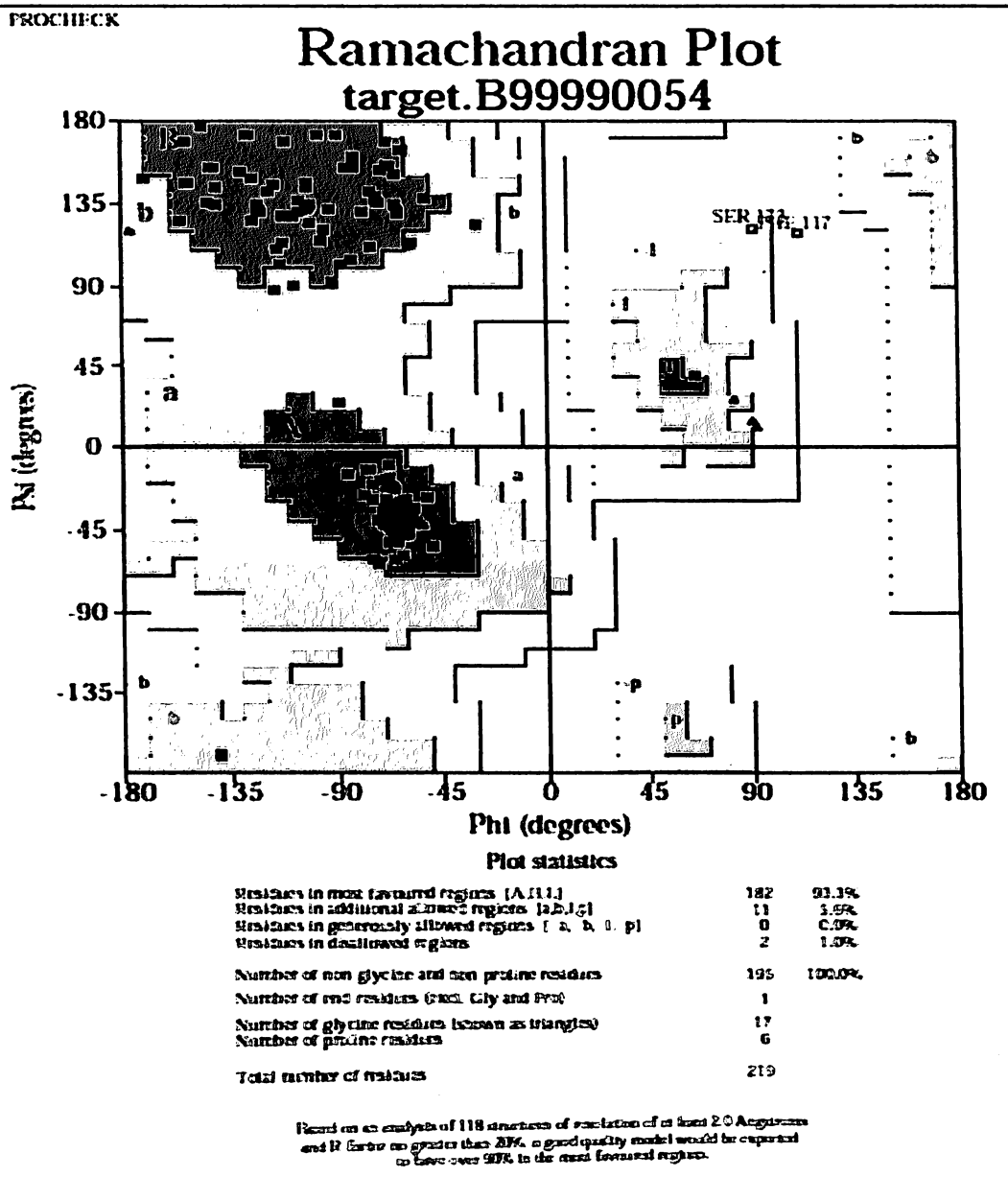


Fig.23: Shows ramachandran plot represents the validation report for the top model generated by the modeller.

Models	Number of residues in Favoured region	Number of residues in Additional Allowed region	Number of residues in disallowed region	Dope Score
target.B99990018	93.8%	4.6%	1.0%	-23774.36523
target.B99990037	93.3%	4.1%	2.1%	-23782.65430
target.B99990054	93.3%	5.6%	1.0%	-23936.23438
target.B99990077	93.8%	4.6%	1.0%	-23904.82031
target.B99990089	96.9%	2.6%	0.5%	-23796.74219

Table.3: Represents validation report for top five models, generated in homology modelling, using modeller 9.9.

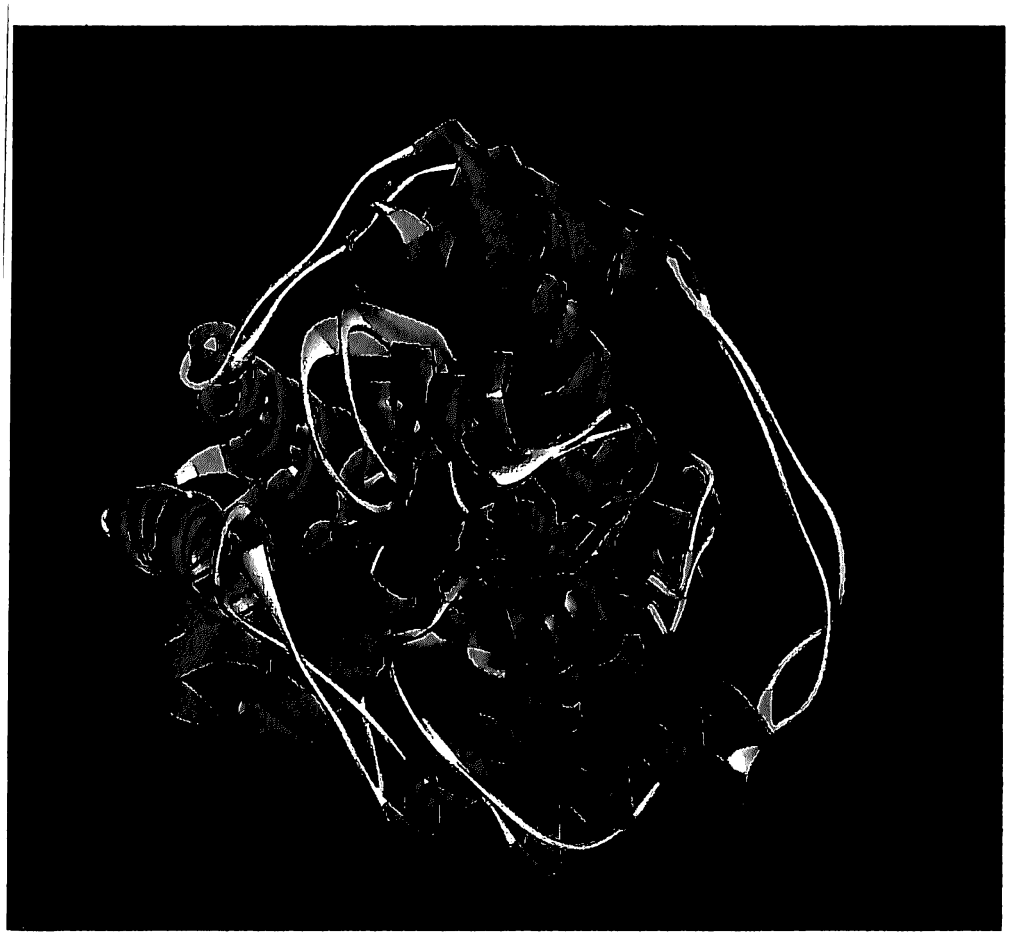


Fig.24: Shows the superimposed structure of the top model generated in the modeller and the template sequence.

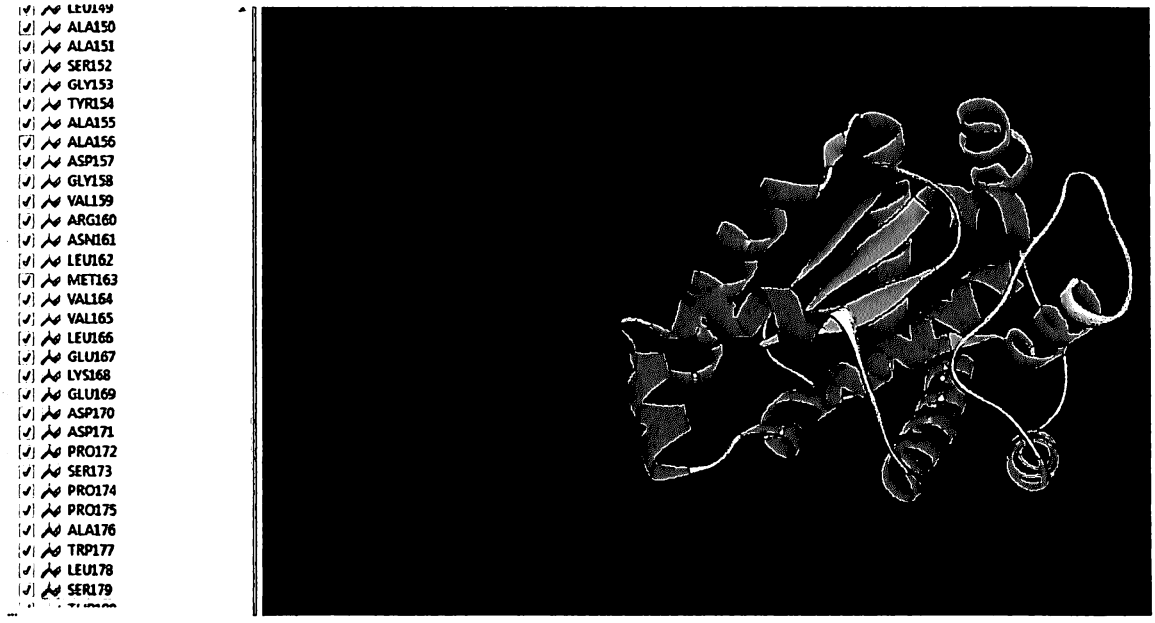


Fig.25: Figure of output of the modloop.

Pink- target, blue-output

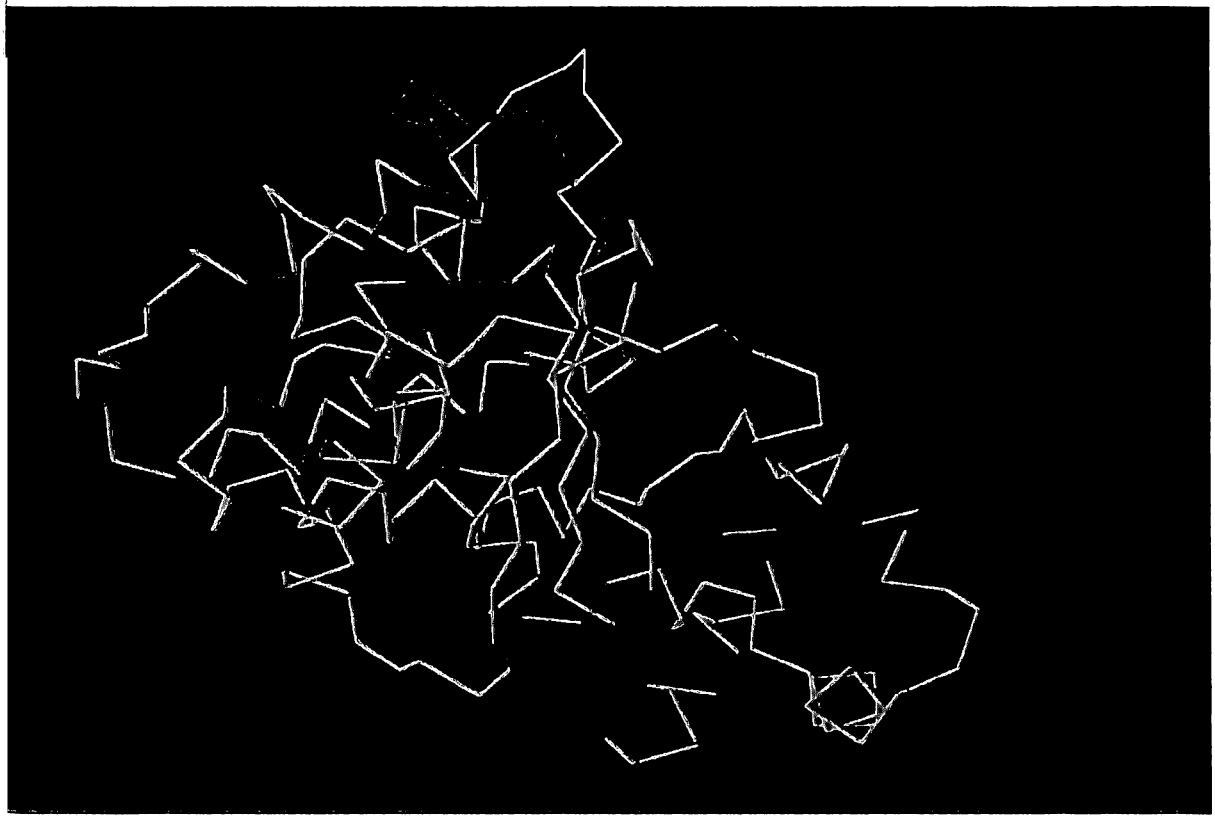


Fig.26: Shows the superimposed structure of modeller output and modloop output.

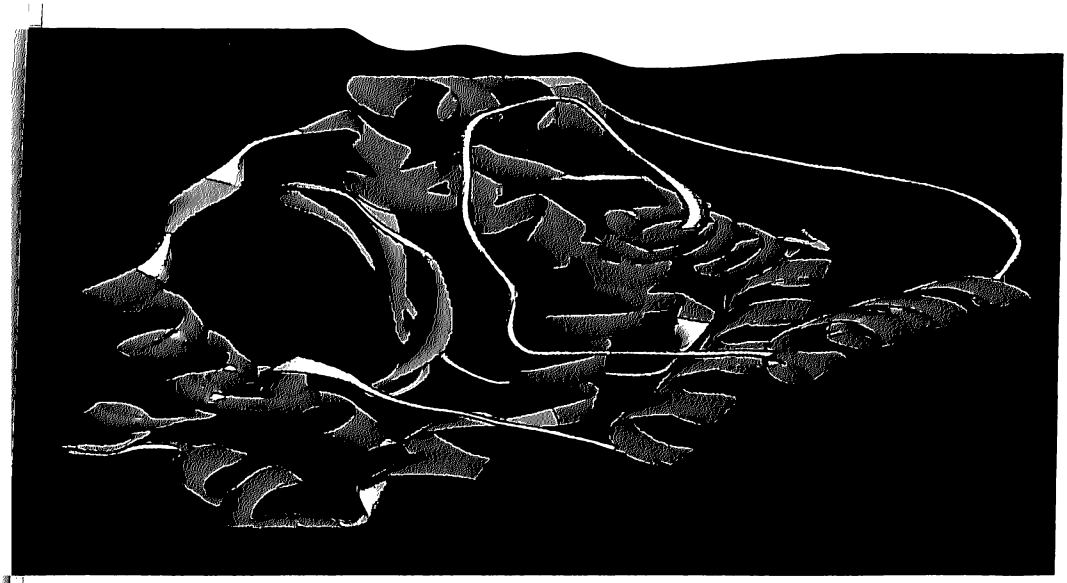


Fig.27: shows output of the complete structure.

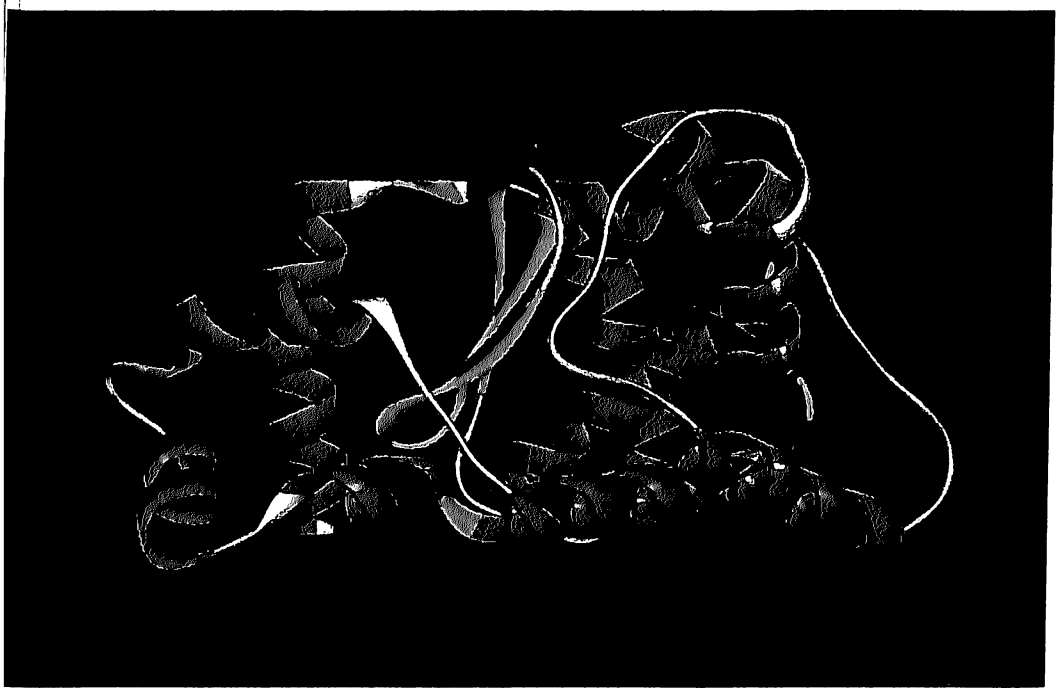


Fig.28: Shows the output of the model after removal of bumps.

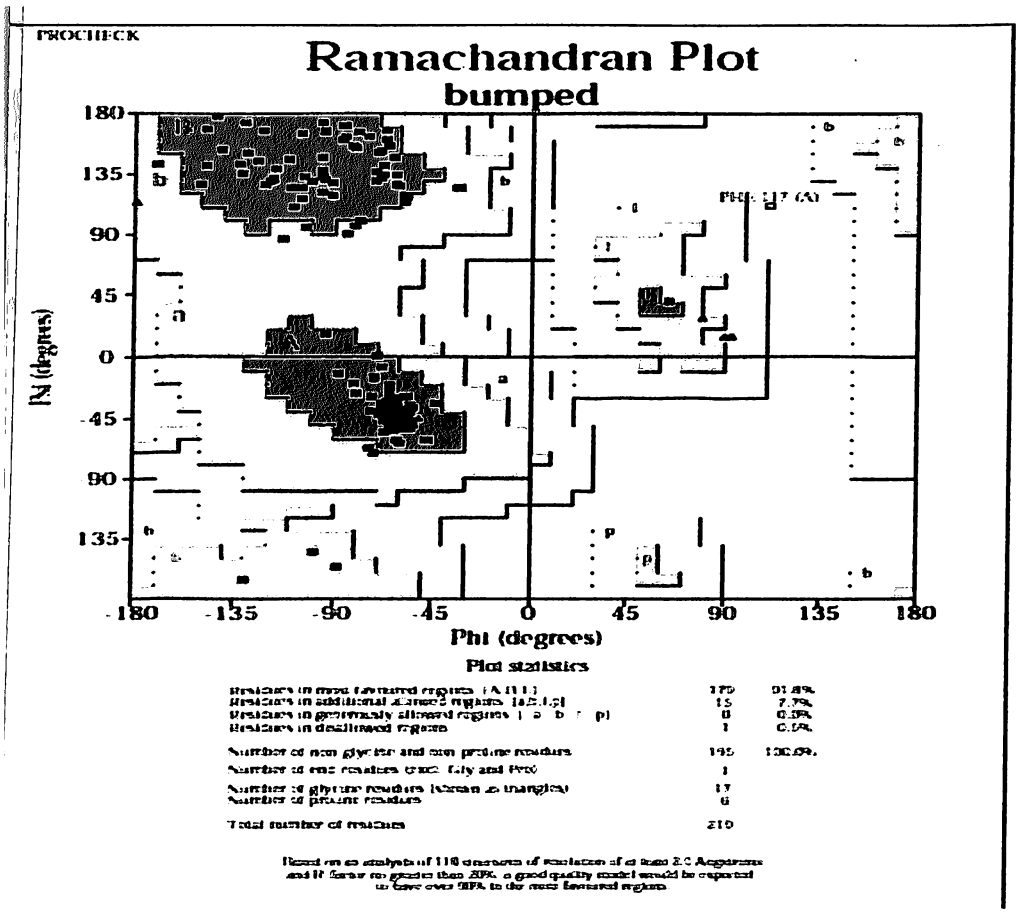
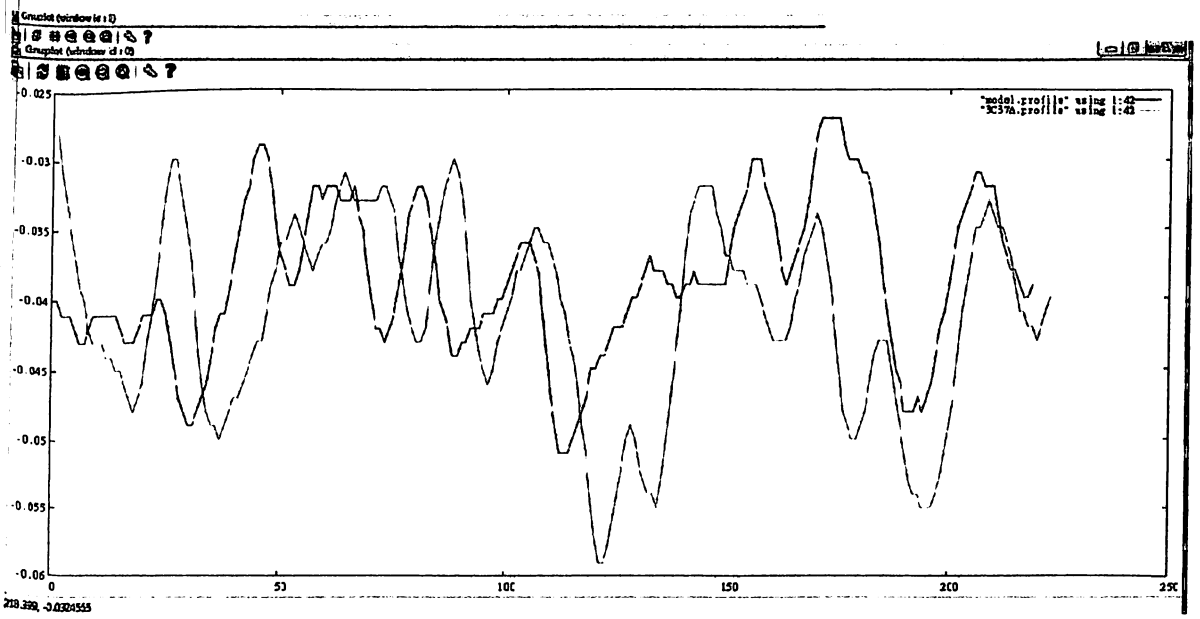


Fig.29: Shows Validation report for model after removal of bumps.



Model. Profile-red, 3C37A.profile-green

Figure.30: Shows the comparative plot of target and template using gnu plot.

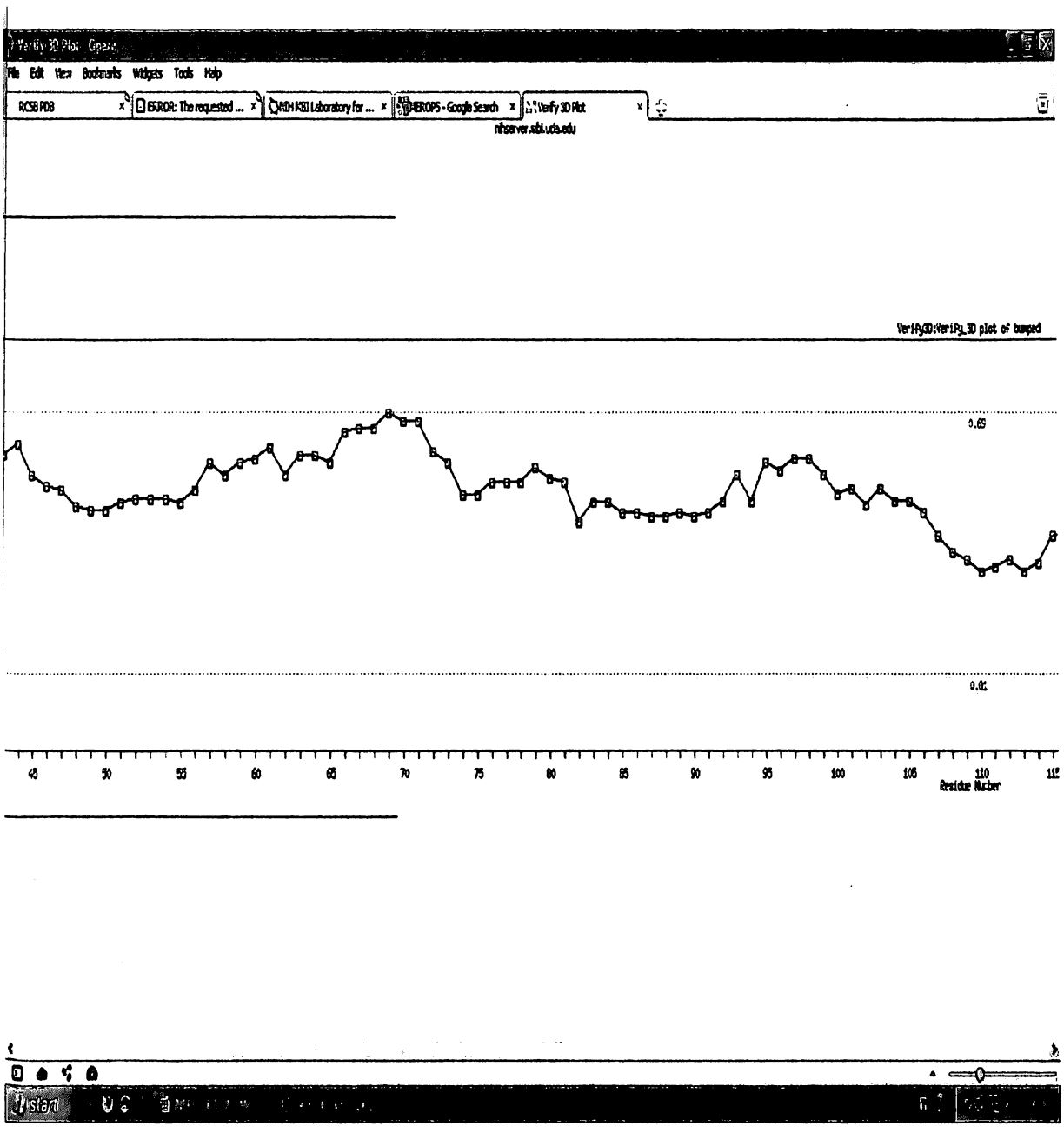


Fig.31: Shows the validation report of the model using veryfy3d.

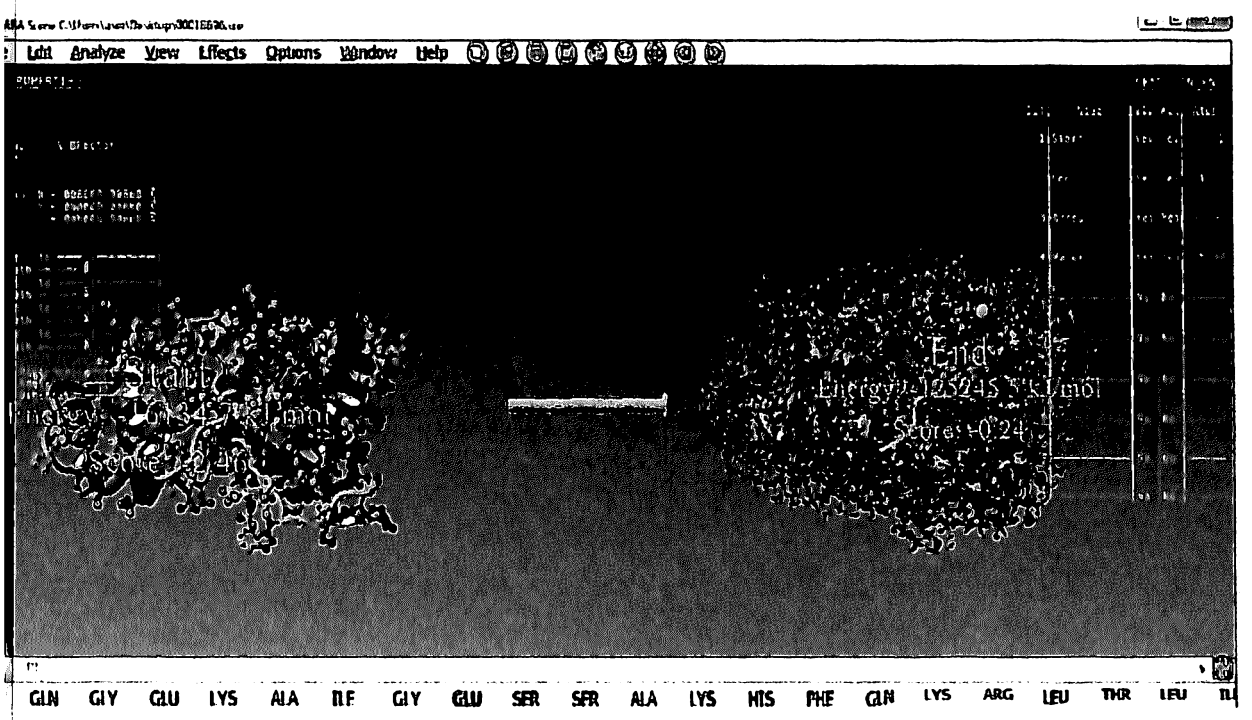


Fig.32: Shows the result for energy minimization

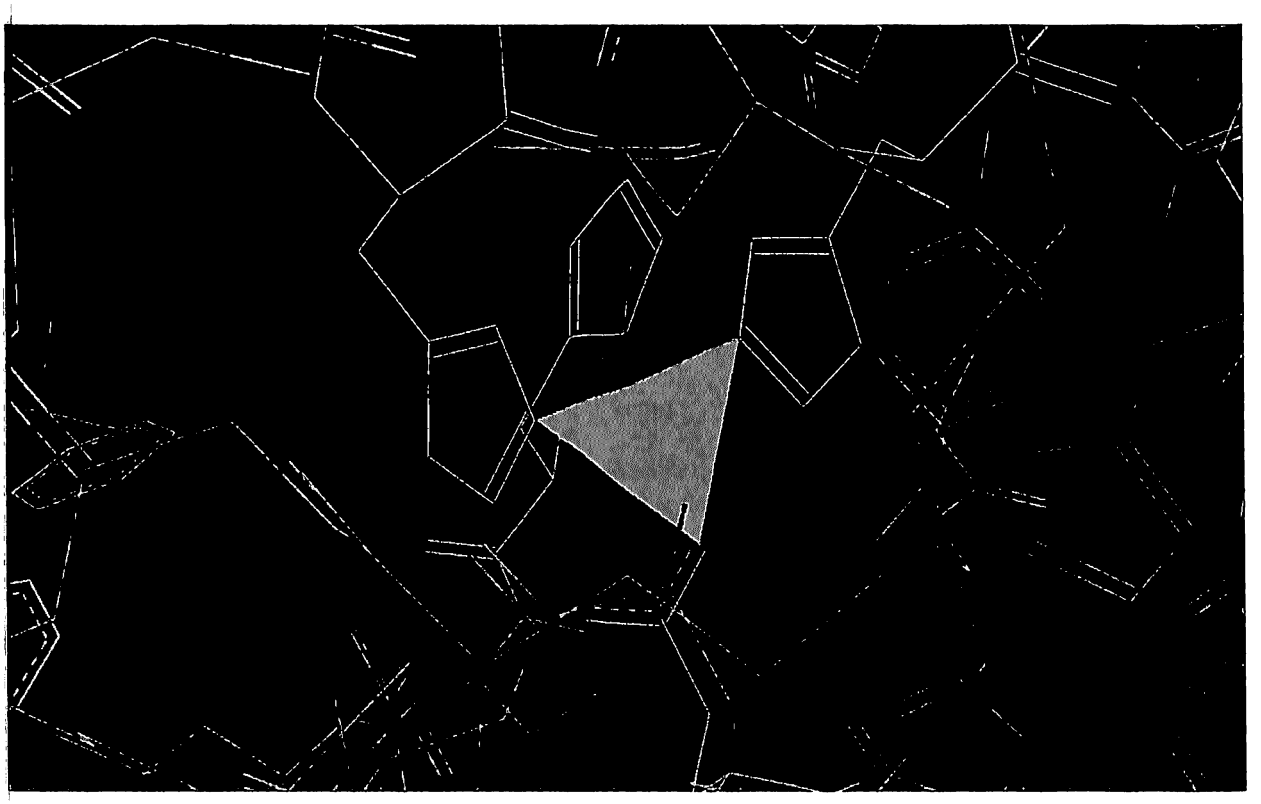


Fig.33: Shows the metalbinding site in the template sequence.

DISCUSSION

The protein sequence of peptidase M48 Ste24p of *cyanothece sp.* of strain PCC 7822 was taken for blast in NCBI server, to find out the homologous sequences. The primary analysis of this structure shows in the protparam shows that Aliphatic index: 84.89 which show the good result, Grand average of hydropathicity (GRAVY): -0.542 shows hydrophobicity. On the basis of identity approximately 27%-48%, e-values, 65 number of different organism for peptidase were selected for evolutionary studies. Different species of cyanothece genera was taken in MEGA using clustal W algorithm. The tree was constructed using neighboring joining algorithm in MEGA 5.0; the result is displayed in figure 14. This tree depicts that the total 65 number of species were grouped in to seven different clads. Then the tree was extracted to the figtree environment and then the tree was compiled for a good presentation displayed in the figure 16. Here from the figure number 16 it has been observed that among seven number of clads of 65 numbers of species of cyanothece genera. In figure 16 the sky colour clad includes the species, 3C37 A, Geobacter sulfurreducens, Stigmatella aurantiaca, Candidatus koribacter versatilis, Acidobacterium sp. , Microcystis aerugenosa, Oscillatoria sp., Microcoleus vaginatus, Nostoc azollae, Raphidiopsis brookii, Cylindrospermopsis raciborskii.Brown colour clad includes the species Nitrosomonas eutropha, gallilonella capsiferriform, methylovorus sp., Aromatoleum aromaticum, Bordetella petrii, Ralstonia solanaceanum, Ralstonia picketti, Halorhodospira halophilla, delta proteobacterium, Desulphobacca acetoxidans, Desulphatibacillum alkalivorans, Syntrophus acidinophicus, syntrophobacter fumaroxidans, Desulphovibrio salexigens, Desulphovibrio fructosovorans, Desulphovibrio sp., Desulphovibrio vulgaris, bilophila wadswortha.Pink colour clad includes the species Desulfobacterium autotrophicum, obulbus propionicus, Desulfomicrobium baculatum, Dechloromonas aromatica, Desulfaculus baaarsii, Alcalilimnicola ehrlichii, Gamm proteo bacterium, Sulfuricurvum kujiense, Oxalobacter fermigenes, Candidatus poribacteria sp., C10 bacterium.Green colour clad includes the species Weeksella virosa, Runella slithyformis, Oscillatoria sp., Planctomyces maris, Alpha proteobacterium, Dyadobacter fermetans, Syderoxydans lithotrophicus.Red colour clad includes the species Sulfurihydrogenibium sp., Sulfurihydrogenibium yellowstonense, Sulfurihydrogenibium azonense,Thermodesulfovibrio yellowstonii, Aquifex aeolisus, Thermodesulfobacterium sp.Blue colour clad includes the species Nospoc sp.,Anabaena vaniabilis, Nodularia spumigena, Cyanothece sp. PCC 7822,Lyngbya majusculla, Microcoleus

chthonoplastes, Crocophaera watsonii, Acaryochloris marina,Trichodesmium erythraceum, Lyngbya sp., Synchoccus sp., Thermosynechococcus elongatus.From the fig-16, it is very clear that the *cyanothece sp.* M48 Ste24p PCC 7822 is closely related to Lyngbya majuscula and Nodularia spumigena species.Then the 65 sequences were taken to bioedit for finding conserved region and the region is displayed in fig.17.It is found that one region is conserved in all the species that is from one 1-252.The consesus sequence is displayed in fig.22.So it might happen that due to other position changes occurred due to speciation during evolutionary process and which results the existances of different species. The secondary structure of the target sequence *Cyanothece sp.* was predicted using different tools namely SOPMA, GOR4, PSIPred. The result was shown in fig-18 and fig-19.The Table no.-1 depicts the analysis / comparision of secondary structural element found in the target protein .It has been observed that approximate 49-54% of Alpha helix and very less i.e. less than 5% Beta sheets were found, this implies mostly our target sequence is having Alpha helical structure. The secondary structure in PSIPred result given in fig-20 , Also predicts that the target protein has maximum Alpha helix ,where it attempts the secondary structure. After blast against the PDB database the protein 3C37 A was found 32% identical with the target sequence having least e-value 4e-25 and having length 253 aminoacid.Hence this structure was selected as template.Then pairwise alignment was performed between the target sequence and the template structural sequence and the result is shown in fig-22.From the result it is observed that the target sequence is properly aligned with the template sequence from the region 366-595 amino acid residue. So for proper structure construction of the target it was edited .Again the alignment was performed in the modeller environment and the both result was compared shown in fig -23.It was observed that loop region was found in the target from 166 to 179 no residue .Table no-2 shows the five top selected model on the basis of dope score out of 100 of models of our target again from this table, it has been cleared that the model no target.B99990054 having lowest dope score. Again target.B99990054 was taken as our target .The fig-26 displays superimposition of the retrieved model and template using DS 3.0 visualiser.From the picture it is very clear about the loop region that was mentioned. Then the predicted structure was processed for loop refinement using modloop server and after loop refinement again superimposition was done between ~~target~~ modloop output and ~~template~~ Modeller output and visualized in DS 3.0 visualizer which shown in fig.27.Then the predicted structure after loop refinement was processed for evaluation using procheck server displayed in fig.30. from the result it is very

cleared that the percentage of aminoacid in most favoured region is 91.8%.Fig.32 displays a comparative graph between predicted structure and template structure.from the picture it is found that most of the portion of the target and template structure is coinciding and less portion is found or marked as deviation and this graph is plotted using Gnu plot.The refined structure was processed for energy minimization was done of the predicted structure using YASARA server. The result is shown in fig.34.From the result of energy minimization it was seen that the energy of the protein minimized perfectly that is from -16134.7 kJ/mol to -125245.5 kJ/mol. Also the Z-score increased from -2.46 to -0.24. So it was cleared that the protein molecule became more stable by doing energy minimization.

SUMMARY

SUMMARY

The current research is intended to perform evolutionary study and structure prediction of peptidase M48 Ste24p [*cyanothece sp.*PCC 7822]. The sequence was retrieved from NCBI database having accession number YP_003888914.1 and verified in swissprot database having accession number EOUH40. After finding homologous species in Blastp, against nr database, and 64 different species of peptidase family were taken. It was further processed for phylogenetic analysis and tree construction was done in MEGA using NJ algorithm. The constructed tree was compiled and the result shows that tree was compile and the result shows that the tree was clustered in to 6 no. of clads and the target is closely evolutionary related with lyngbya majuscula and nodularia spumigena species. From the conserved site prediction of all the metallopeptidase family, it has been observed that the one conserved site from 1-252 residue position. This show significant similarity between different species and divergent regions for which speciation occurred during evolution. Again secondary structure prediction results shows most of the region of the target protein is alpha helical in nature. This shows the stability of the core region of the protein. From the tertiary structure prediction result after loop refinement confirmed by ramachandran plot and verify 3d and concluded the verified model is quite accurate.Then the metal binding region was found in the position His106, His110, Glu162, His208 in the template structure, it was found that the metal binding site is conserved in the target sequence. This implies the accuracy of the structure prediction and binding site determination in this work. Hence the work can be further extended to study structure and functionality of the protein in more detail.This would help the breeder in future for incorporating the particular gene in a new variety of rice plant. And this work will definite provide a boost to enhance the productivity of the rice plant.