We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900
Open access books available

185,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



9

Understanding Tools and Techniques in Protein Structure Prediction

Geraldine Sandana Mala John[1], Chellan Rose[1] and Satoru Takeuchi[2]
[1]*Central Leather Research Institute*
[2]*Factory of Takeuchi Nenshi ,Takenen*
[1]*India*
[2]*Japan*

1. Introduction

Protein structure prediction is an important area of protein science. Every protein has a primary structure, its sequence; a secondary structure, the helices and sheets; tertiary structure, the fold of the protein; and for some, the quaternary structure, multimeric formation of its polypeptide subunits. Protein structure has been experimented for the past several decades by physical and chemical methods. The dawn of protein sequencing began early in 1950s upon complete sequencing of insulin and then, ribonuclease. A key step towards the rapid increase in the number of sequenced proteins by 1980s was the development of automated sequencers followed by advances in mass spectrometry for structure identities. Structural knowledge is vital for complete understanding of life at the molecular level. An understanding of protein structure can lead to derivation of functions and mechanisms of their action. Bioinformatics is a novel approach in recent investigations on sequence analysis and structure prediction of proteins. With the advent of bioinformatics, it has been made possible to understand the relationship between amino acid sequence and three-dimensional structure in proteins. The central challenge of bioinformatics is the rationalization of the mass of sequence information not only to derive efficient means of storage and retrieval of sequence data, but also to design more analysis tools. Thus, there is a continual need to convert sequence information into biochemical and biophysical knowledge; to decipher the structural, functional and evolutionary clues encoded in the language of biological sequences (Attwood & Parry-Smith, 2003). Protein sequence information is stored in databases made available in the public domain to access, analyse and retrieve sequence and structural data. In general, protein databases may be classified as Primary and Secondary databases, composite protein pattern databases and structure classification databases. Primary and secondary databases address different aspects of protein analysis, because they store different levels of protein information. Primary databases are the central repositories of protein sequences, while secondary databases are based on the analysis of sequences of the primary ones. Composite protein pattern databases have emerged with a view to create a unified database of protein families. Protein structure classification databases have been established based on the structural similarities and common evolutionary origins of proteins. A number of tools are also

available for protein structure visualization and protein identification and characterization. Thus bioinformatics tools for protein analysis provide a wealth of information related to sequences and structures of proteins.

Use of computational tools is an essential kit for the biologist in this rapid pace of information technology. Eventually, tools and techniques for protein sequence analysis and further, the structure prediction, has become an integral study for protein biochemists. Random identification of protein structures based only on homology of proteins is by and large an ambiguous approach. Hence, a systematic analysis of the protein under study from its sequence annotations to its three-dimensional structure alignment is a feasible approach for the investigations of protein structure aided by computational networking and repositories available in the public domain. Thus, sequence data can be transformed to structural data by a line of database analyses. The identification of protein structures can be organized as a flow of information from protein characterization, primary structure analysis and prediction by database search; sequence alignment; secondary structure prediction; motifs, profiles, patterns and fingerprint search; modeling; fold structure analysis and prediction; protein structure visualization and analysis of structure classification databases to deposition of protein structures in the public domain. An identity of sequence similarity of query sequences with that of database sequences indicating homology derives the phylogenetic maps of the protein under consideration and reveals information on conserved patterns thereby predicting repeat folds among the proteins that have arisen from divergence or of convergence. Pattern recognition methods convey information on the characteristics of unique features of the protein as well as the identification of similar traits in other proteins.

However, it is noteworthy that identifying patterns and functions of proteins are still far from being perfect which are likely to result in false interpretations and assumptions. Hence, it is the expertise and the reasoning of the biologist to interpret protein and/or any sequence information in the light of physical and chemicals methods to determine structure predictions. The study of bioinformatics is an interdisciplinary approach which requires the skill sets of biologists, mathematicians, information analysts and software developers to design and develop computational methods for analysis of biological data. This is presumably the index of milestones in bioinformatics for a fruitful journey in the identification of protein structure. Hence, it can be correlated that bioinformatics is the hand tool in every biology laboratory for thorough investigations of proteins and their complements in establishing evolutionary hierarchy and in the identification of protein malfunctions by linking protein structure to its functions in health and disease, thereby opening possible avenues for genetic manipulations and undertake prophylactic measures.

2. Protein structure-an overview

Protein architecture is the fundamental basis of the living systems that coordinates the functional properties of cells to sustain life. Every metabolic action is dependent on a set (s) of proteins that function as chaperones, enzymes, cofactors, structural proteins etc. Hence, an understanding of protein structure is vital for implications in physiological and therapeutic investigations. Lesk (2001) and Whitford (2005) have provided much of the understanding on the structural aspects of proteins. Generally, proteins are made up of small units known as amino acids which form a polypeptide chain through formation of peptide bonds. Thus, amino acids are the building blocks of all proteins which are

characteristic for each type of the protein imparting specific functional attributes. There are 20 amino acids in nature that are of L-configuration that make up all kinds of proteins and are classified as aliphatic, aromatic, acidic, basic, hydroxylic, sulphur-containing and amidic amino acids. The discussion on the structure and chemical properties of amino acids is out of scope of this chapter and detailed information can be referred in most books covering protein structure. At the outset, we describe here the Primary, Secondary, Tertiary and Quaternary structures of a protein to enable keen insights of the structure prediction of proteins through bioinformatics. We also provide here the basic concepts of peptide bond and the Ramachandran plot that influence protein structure and conformation.

2.1 Primary structure

The primary structure of a protein resides in the linear order of the amino acid sequence along the polypeptide chain. Amino acids have been named in a three-letter code and in recent years, by a single letter code (Table 1) which is in current practice.

Amino acids	Three-letter code	Single letter code	Amino acids	Three-letter code	Single letter code
Alanine	Ala	A	Leucine	Leu	L
Arginine	Arg	R	Lysine	Lys	K
Asparagine	Asn	N	Methionine	Met	M
Aspartic acid	Asp	D	Phenylalanine	Phe	F
Cysteine	Cys	C	Proline	Pro	P
Glutamine	Gln	Q	Serine	Ser	S
Glutamic acid	Glu	E	Threonine	Thr	T
Glycine	Gly	G	Tryptophan	Trp	W
Histidine	His	H	Tyrosine	Tyr	Y
Isoleucine	Ile	I	Valine	Val	V

Table 1. Notations of amino acids in three-letter and single letter codes.

The amino acids that form the sequence are termed residues to denote the composition of a polypeptide. The primary sequence of a protein can therefore be visualized as a single letter code running from left to right with the left end constituting the N-terminal (amino group) of the first amino acid residue and the right end constituting the C-terminal (carboxylic acid group) of the last amino acid residue.

A particular amino acid residue of the amino acid sequence can therefore be identified by its position in the numerical sequence order. For example, a lysine residue can be identified as K6 when it appears in it 6^{th} position or a glycine residue as G3 when it appears in its 3^{rd} position and so on. The order of amino acid sequences is characteristic of a particular protein and of species and among protein families forming a conserved set of sequence in a region of the polypeptide(s). This sequential order determines the fold of a protein in achieving its native conformation and assigns the specific protein function. The primary sequence determination is therefore a significant criterion which defines the subsequent levels of the protein organization. An important aspect of the primary structure is that any mismatch of the sequence in a functional protein is often lethal to the cellular function carried out by the protein. This leads to several hereditary and metabolic defects such as in sickle cell anemia where the glutamic acid is replaced by valine in the 6^{th} position of the β-

chain of hemoglobin by a point mutation. The amino acid sequence of a protein is specified by the gene sequence by the process of transcription and translation.

2.2 Secondary structure

The secondary structure of a protein is the local conformation of the polypeptide chain or the spatial relationship of the amino acid residues which are placed close together in the primary sequence. This organizational level is found in globular proteins where three basic units of secondary structure are present, namely, the α-helix, β-strand and turns. Other secondary structures are based on these elements (Augen, 2004).

2.2.1 The α-helix

The right-handed α-helix is the most identifiable unit of secondary structure and the most common structural motif found in proteins with over 30% helix structure in globular proteins. In an α-helix, four or more consecutive amino acid residues adopt the same conformation resulting in a regular helical shape in the polypeptide backbone. This helix is stabilized by H-bonds between the main chain C=O group of one amino acid and the H-N group of the amino acid four residues further along the helix, forming a helix with 3.6 amino acid residues per helical turn resulting in a regular stable arrangement. The α-helix repeats itself every 0.54 nm along the helix axis i.e., the α-helix has a pitch of 0.54 nm. The radius of the helix is 0.23 nm with a translation distance per residue of 0.15 nm. The peptide planes are almost parallel with the helix axis and the dipoles within the helix are aligned. The α-helix arises from regular values adopted for φ (phi) and ψ (psi), the torsional or dihedral angles. The values of φ and ψ formed in the α-helix allow the backbone atoms to pack close together with few unfavorable contacts. This arrangement allows the H-bonding important for the stability of the helix structure. All the amino acids in the helix have negative φ and ψ angles, with ideal values of -57 and -47 respectively. It is important to note that proline does not form a helical structure due to the absence of an amide proton (NH) which is unable to form H-bond while the side chain covalently bonded to the N atom restricts backbone rotation.

2.2.2 The β strand

The second unit of protein secondary structure identified after the α-helix is the β strand which is an extended conformation when compared to the α-helix with 2 residues per turn and a translation distance of 0.34 nm leading to a pitch of nearly 0.7 nm in a regular β strand. A single β strand is not stable largely because of the limited number of local stabilizing interactions. When two or more β strands form additional H-bonding interactions, a stable sheet-like arrangement is created contributing to the overall stability of the β sheets. Adjacent strands can align in parallel or antiparallel arrangements and their orientations are established by the direction of the polypeptide chain from the N- to the C-terminal. Amino acid residues in the beta-conformation have negative φ and positive ψ angles with -139 and +135 angles respectively for parallel β sheets and -119 and +113 φ and ψ angles respectively for antiparallel β sheets. Polyamino acids in solution do not form β sheets and this hinders the study of their structures.

2.2.3 Turns

A turn is a secondary structural element where the polypeptide chain reverses its overall direction. It is a structural motif where the Cα atoms of two residues are separated by 1 to 5

peptide bonds and the torsional angles are not constant for all the residues in a turn. Many different conformations exist on the basis of the number of residues making up the turn and the dihedral angles associated with the central residues. Turns are classified according to the number of residues they contain namely, the α-turn, where the end residues are separated by 4 residues, β-turn, by 3 residues, γ-turn, by 2 residues, δ-turn, by one residue and π-turn, by a factor of 5 residues. A β-hairpin turn occurs between two H-bonded antiparallel beta strands in which the direction of the protein backbone reverses.

2.2.4 Loop

A loop occurs between 6 and 16 residues to form a compact globular shape of the protein which contain polar residues and hence, predominantly occur at the protein surface which contribute to the formation of active sites for ligand binding or catalytic activity. The loops connect the secondary structure elements of the polypeptide chain. Loop structures that are random are less stable and referred as random coils.

2.2.5 Coiled coil

A coiled coil is a structural motif in proteins in which 2-7 alpha helices are coiled together to form a repeated pattern of hypdrophobic and charged amino acid residues referred as heptad repeat. The tight packing in a coiled coil interface is due to van der Waal interactions between side chain groups. The coiled coil element is responsible for the amphipathic structures.

2.3 Tertiary structure

Tertiary structure is the global three-dimensional folding that results from interactions between elements of secondary structure. Tertiary structure of a protein therefore represents the folded conformation of a polypeptide chain in three-dimensional space, i.e., the spatial arrangement of amino acid residues widely separated in its primary structure. Interaction between the side chain groups is the predominant driver of the fold of the protein chain. These interactions which stabilize the tertiary structure arise from the formation of disulfide bridges, hydrophobic effects, charge-charge interactions, H-bonding and van der Waal interactions.

Disulfide bridges form between thiol (-SH) groups of two nearby cysteine residues. With reference to hydrophilic/hydrophobic interactions, water soluble proteins fold to expose hydrophilic side chains on the outer surface retaining the hydrophobic residues in the interior of the protein. Charge-charge interactions occur when a charged residue is paired with a neutralizing residue of opposite charge forming a salt bridge. H-bonding contributes to the overall stability of the tertiary structure or the folded state by stabilization of the secondary structure involving α-helices and parallel or antiparallel β sheets and of side chain groups of Tyr (Y), Thr (T), Ser (S), Gln (Q) and Asn (N). Van der Waal interactions are important in protein folding occurring between adjacent, uncharged and non-bonded atoms. A variety of post-translational modifications also contributes to the protein conformation such as conversion of proline to hydroxyproline that influences the tertiary structure of collagen molecule while glycosylation, carboxylation and methylation have little or no effects but which alter the chemical properties of the protein. Another important aspect in a protein fold is the activation of inactive proteins by small molecules such as cofactors, which are essential for native conformation formation.

The tertiary structure may be organized around more than one structural unit, known as domains which are folded sections of the protein representing structurally distinct units and the same interactions govern its stability and folding. Most domain structures exhibit specific functions independent of the rest of the protein architecture. Domain regions may be α-helices or β strands or mixed elements of both α-helices and β strands. Motifs are smaller structures, usually composed of few secondary elements that recur in many proteins and are rarely structurally independent. This feature or structural significance is important when considering the prediction of folded structure of an individual motif in context of the rest of a protein unlikely of the domain structure.

2.4 Quaternary structure

Many proteins involved in structural or metabolic or enzymatic functions are oligomeric proteins because they consist of more than a single polypeptide chains referred as subunits. The quaternary conformation of a protein arises from the interactions similar to tertiary structures, but is a result of interactions between the subunits which may be identical or nonidentical. Therefore, the quaternary structure refers to the noncovalent, stable association of the multiple subunits. A classic example of a protein that exhibits quaternary conformation is hemoglobin which consists of 4 polypeptide chains or subunits. The quaternary conformation of a protein allows the formation of catalytic or binding sites at the interface between subunits, which is not possible for monomeric proteins. Ligand or substrate binding causes a conformational change affecting the protein assembly for regulation of its biological activity such as the allosteric regulation in enzymes.

Thereby, the four conformations of a protein molecule define its architectural arrangement in a three-dimensional model which contribute to the functional attributes of the protein. This is represented in Figure 1 which is a common theme for most globular proteins.

2.5 The peptide bond

Amino acids are joined to each residue along the sequence by a linkage of the amino group of one residue with the carboxyl group of the next residue, known as the peptide bond (Figure 2).

The physical characteristics of the peptide bond impart the specific folding properties of the protein and this folding pattern of the polypeptide chain is described in terms of the angles of internal rotation around the bonds in the main chain. The N-Cα and Cα-C are single bonds and the internal rotations around these bonds are not restricted by the electronic structure of the bond, but, only by possible steric collisions in the conformations produced. An important characteristic of the peptide bond is the rigidity of the bond caused by its relatively short length, which imparts a partial double bond character. Hence, peptide bonds are characterized by a lack of rotational freedom. The double bond character of the peptide bond (Table 2) was first recognized by Linus Pauling who suggested that the peptide bond is rigid planar (Table 3) and hence exists as cis or trans isomer, with the trans isomer stable.

The entire conformation of the protein is described by these angles of internal rotation. Peptide bonds are invariably fixed at $\omega = 180^{o}$. The φ and ψ angles are limited by steric hindrance between amino acid side chains which reduce the number of allowed conformations for a polypeptide chain. The rigidity of the peptide bond limits the number of arrangements that could fit without distorting the bonds. Without this constraint, the peptide would be free to adopt many numbers of structures and no single consistent pattern

could exist. Therefore, by reducing the degrees of freedom, a well defined set of states of the protein could emerge. This is particularly significant because the proteins should indeed have a defined conformation to accomplish its physiological functions.

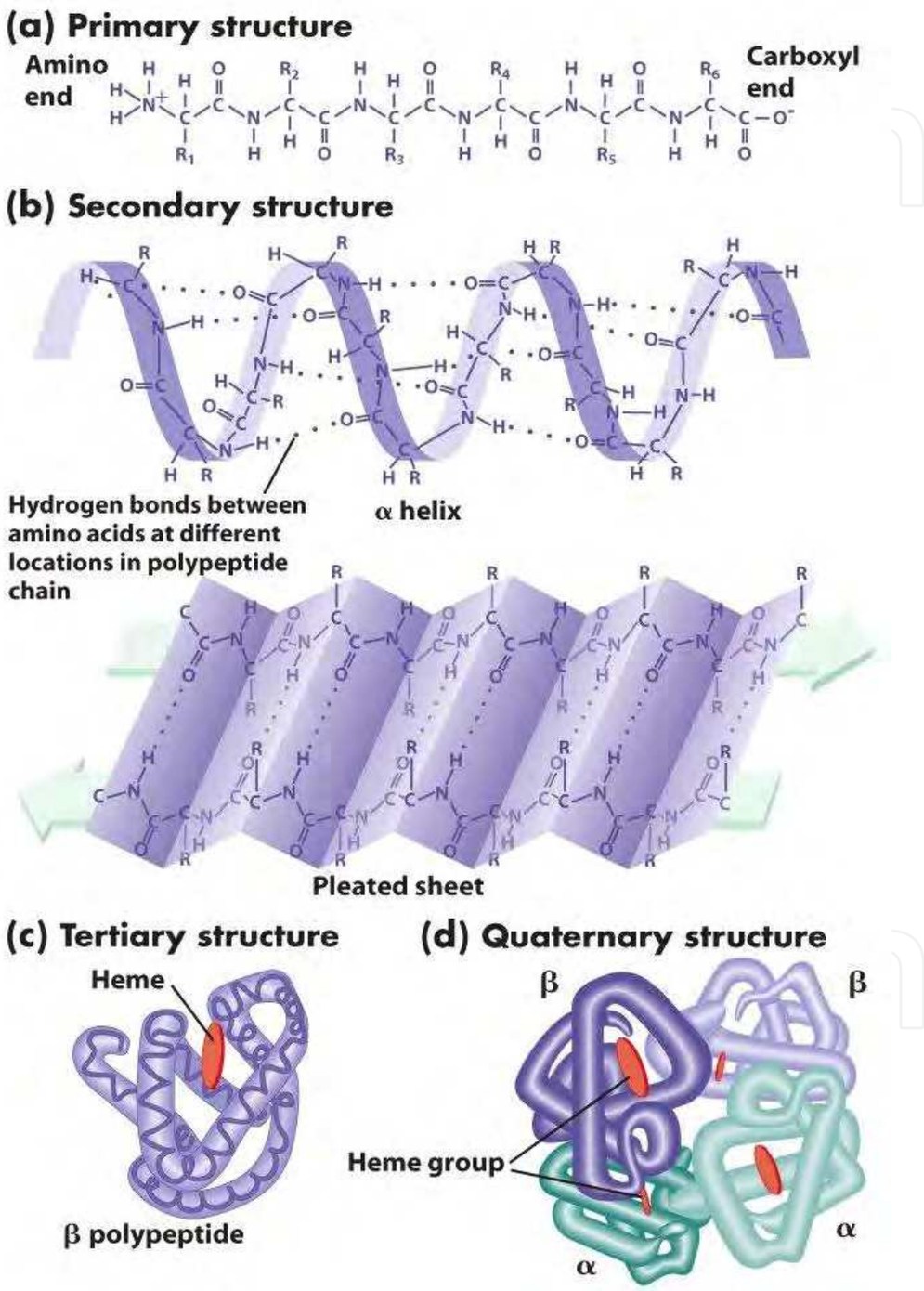


Fig. 1. Hierarchy levels of protein structure. The figure represents the different levels of hemoglobin structure.

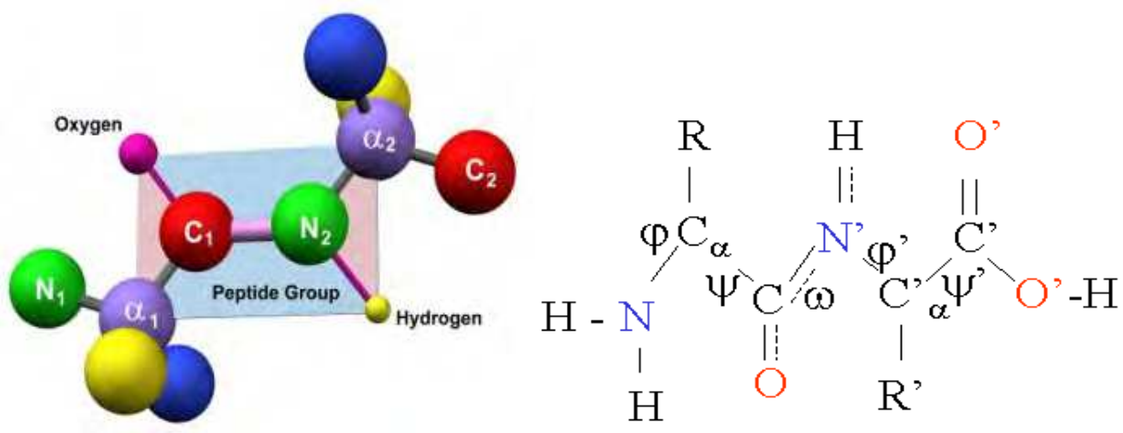


Fig. 2. The peptide bond structure. A. Ball and stick model, B. Torsional angles of the peptide structure

Bond nature	Length
C-N	1.47 Å
C=N	1.27 Å
C=O to NH	1.33 Å

Table 2. Bond character of the peptide bond.

Bond	Rotation	Torsional angle
NH to Cα	Free	Phi φ
Cα to C=O	Free	Psi ψ
C=O to NH (peptide bond)	Rigid planar	Omega ω

Table 3. Conformational angles of folding of polypeptide chain.

2.6 The Ramachandran plot

The peptide bond is planar as a result of resonance and its bond angle, ω has a value of 0 or 180˚. A peptide bond in the trans conformation (ω =180˚) is favoured over the cis arrangement (ω =0˚) by a factor of ~1000 because the preferential arrangement of non-bonded atoms lead to fewer repulsive interactions that otherwise decrease stability. In the cis peptide bond these non-bonded interactions increase due to the close proximity of side chains and Cα atoms with the preceeding residue and hence results in decreased stability relative to the trans state. Peptide bonds preceeding Proline are an exemption to this trend with a trans/cis ratio of ~4. The peptide bond is relatively rigid, but far greater motion is possible about the remaining backbone torsion angles. In the polypeptide backbone C-N-Cα-C defines the torsion angle φ whilst N-Cα-C-N defines ψ. In practice these angles are limited by unfavourable close contacts with neighbouring atoms and these steric constraints limit the conformational space that is sampled by the polypeptide chains. The allowed values for φ and ψ were first determined by G.N.Ramachandran using a 'Hard sphere

model' for the atoms and these values are indicated on a two-dimensional plot of φ against ψ that is called a Ramachandran plot.

In the Ramachandran plot shown in Figure 3 the freely available conformational space is shaded in green. This represents ideal geometry and is exhibited by regular strands or helices. Analysis of crystal structures determined to a resolution of <2.5 Å showed that over 80 percent of all residues are found in this region of the Ramachandran plot. The yellow region indicates areas that although less favourable can be formed with small deviations from the ideal angular values for φ and ψ. The yellow and green regions include 95 percent of all residues within a protein. Finally, the purple coloured region, although much less favourable will account for 98 percent of all residues in proteins. All other regions are effectively disallowed with the minor exception of a small region representing left handed helical structure. In total only 30 percent of the total conformational space is available suggesting that the polypeptide chain itself imposes severe restrictions. One exception to this rule is Glycine. Glycine lacks a Cβ atom and with just two hydrogen atoms attached to the Cα centre, this residue is able to sample a far greater proportion of the space represented in the Ramachandran plot. For glycine, this leads to a symmetric appearance for the allowed regions. As expected residues with large side chains are more likely to exhibit unfavourable, non-bonded interactions that limit the possible values of φ and ψ. In the Ramachandran plot the allowed regions are smaller for residues with large side chains such as phenylalanine, tryptophan, isoleucine and leucine when compared with the allowed regions for alanine.

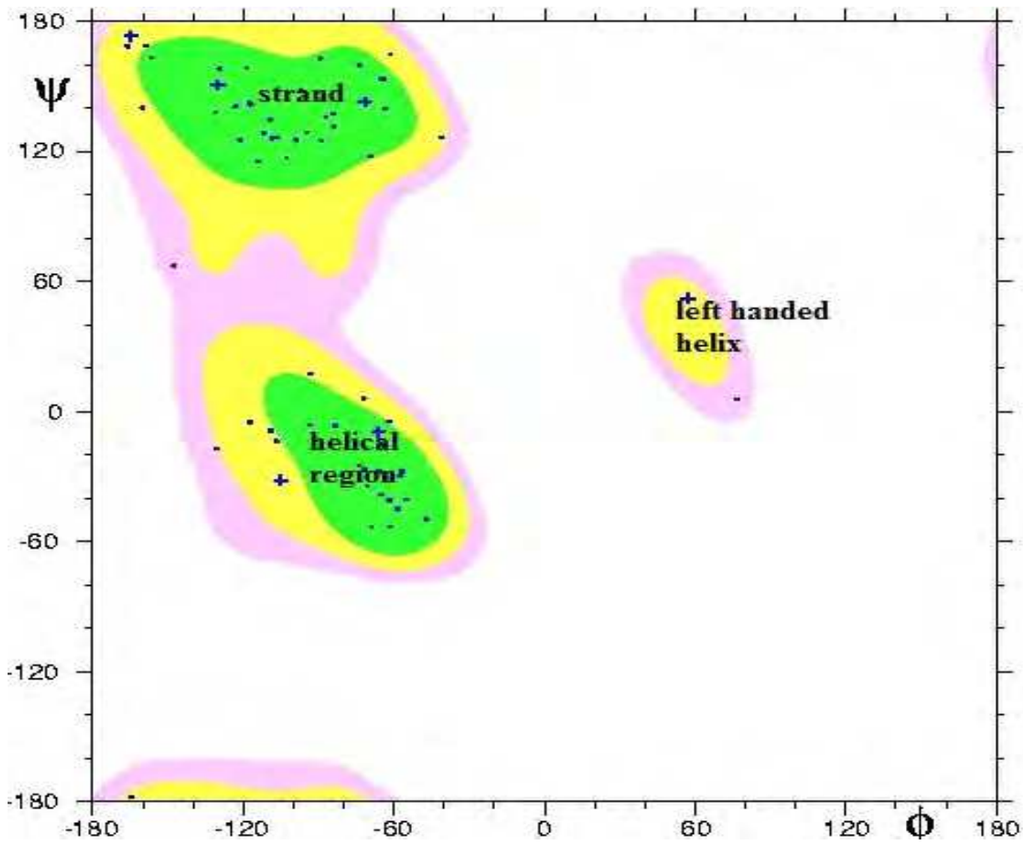


Fig. 3. Ramachandran plot showing allowed and disallowed conformations

3. The need for structural bioinformatics

Proteins are manifested in every aspect of biological activity/function. Many metabolic, cellular and structural events require the proper functioning of proteins in a cell. Any rupture of cellular function stems from the distortion or misfolding of proteins that prevents its normal function. Hence, protein science augments the advancements in genomic sciences to understand health and disease at the molecular level. Years back, an understanding of the structure of proteins, their interactions with other biomolecules, their roles within different biological systems have been made possible through molecular genetics and chemical methods and through biochemical pathways. This has taken years of intensive efforts and with the advent of modern techniques. The recent surge in bioinformatics has created a landmark in deciphering and decoding the gene and protein characteristics and functions. During the past decade, sequence information has been on a tremendous rise in contrast to the three-dimensional structural elucidation of proteins. This has resulted in the sequence/structure deficit of protein sequence and structure information. This can be estimated by the number of sequences available in sequence databases in contrast to the number of structures available in structure databases. A search for a protein sequence would generate hundreds of thousands of sequences while it would generate a few possible structures in a structure repository such as the Protein Data Bank (PDB) for the same protein query. This has prompted several consortia of groups to identify and deposit new protein structures through bioinformatics from the largely available protein structure prediction tools in the WWW.

Structure prediction has fascinated protein biochemists and the pioneering work of Margaret Dayhoff has contributed much to the understanding of protein structure through computational methods. She had developed the one-letter code for protein naming to reduce the complexity of the three-letter naming in the development of sequence information, storage and retrieval. She initiated the collection of protein sequences in the Atlas of Protein Sequence and Structure, a book collecting all known protein sequences that she published in 1965 which led to the development of Protein Information Resource database of protein sequences. In general, structure prediction is an attempt to predict the relative position of every protein atom in three-dimensional space using only its sequence information. Structural bioinformatics of the protein structure is based on a hierarchy of tools and techniques that identify the different levels of protein architecture (Figure 4). Many web tools for protein structure prediction have arisen to simplify the tasks of biochemists and bioinformaticians as well. Figure 5 provides a bird's eye view of the sequential steps in the identification/prediction of the protein structure.

4. Protein databases

Protein sequence information has been effectively dealt in a concerted approach by establishing, maintaining and disseminating databases, providing user-friendly software tools and develop state-of-the-art analysis tools to interpret structural data. Databases are central, shareable resources made available in public domain and represent convenient and efficient means of storing vast amount of information. Depending on the nature of the different levels of information, databases are classified into different types for the end user. This section describes the various databases for each of the nature of protein information that range from primary, composite, secondary and pattern databases. The different

databases address different aspects of protein information which enable the analyst to perform an effective structure prediction strategy (Mala & Takeuchi, 2008).

4.1 Primary protein databases

4.1.1 PIR

This is the Protein Information Resource developed as a Protein sequence database at the National Biomedical Research Foundation (NBRF) in the early 1960s and collaboratively by PIR-International since 1988. The consortia include the PIR at NBRF, JIPID the International Protein Information Database of Japan and MIPS the Martinsried Institute for Protein Sequences.

4.1.2 MIPS

The Martinsried Institute for Protein sequences collects and processes sequence data for PIR and can be accessed at its web server.

4.1.3 SWISS-PROT

This protein database was produced collaboratively by the Department of Medical Biochemistry at the University of Geneva and the EMBL (European Molecular Biology Laboratory). Since 1994, it moved to EMBL's UK outstation, the EBI (European Bioinformatics Institute) and in April 1998, it moved to Swiss Institute of Bioinformatics (SIB) and is maintained collaboratively by SIB and EBI/EMBL. It provides the description of the function of proteins, structure of its domains, post-translational modifications etc., is minimally redundant and is interlinked to many other resources.

4.1.4 TrEMBL

This database has been designed to allow rapid access to protein sequence data. TrEMBL refers to Translated EMBL and was created as a supplement to SWISS-PROT in 1996 to include translations of all coding sequences in EMBL.

4.1.5 NRL-3D

This database is a valuable resource produced by PIR from sequences extracted from the Brookhaven Protein Data Bank (PDB). The significance of this database is that it makes available the protein sequence information in the PDB for keyword interrogation and for similarity searches. It includes bibliographic references, MEDLINE cross-references, active site, secondary structure and binding site annotations.

4.2 Composite protein sequence databases

Composite databases have been created to simplify the sequence search for a protein query in a single compilation in context of the many different primary database searches, by merging a variety of different primary resources. These databases are non-redundant and render sequence searching much more efficient.

4.2.1 NRDB

Non-Redundant DataBase (NRDB) is the default database of the NCBI (National Center for Biotechnology Information) BLAST (Basic Local Alignment Search Tool) service and is a

composite of GenPept, PDB sequences, SWISS-PROT, SPupdate (weekly update of SWISS-PROT), PIR and GenPept update (daily updates of GenPept). It provides comprehensive up-to-date information and is non-identical rather than non-redundant, that is, it reiterates only identical sequence copies and hence results in artifacts.

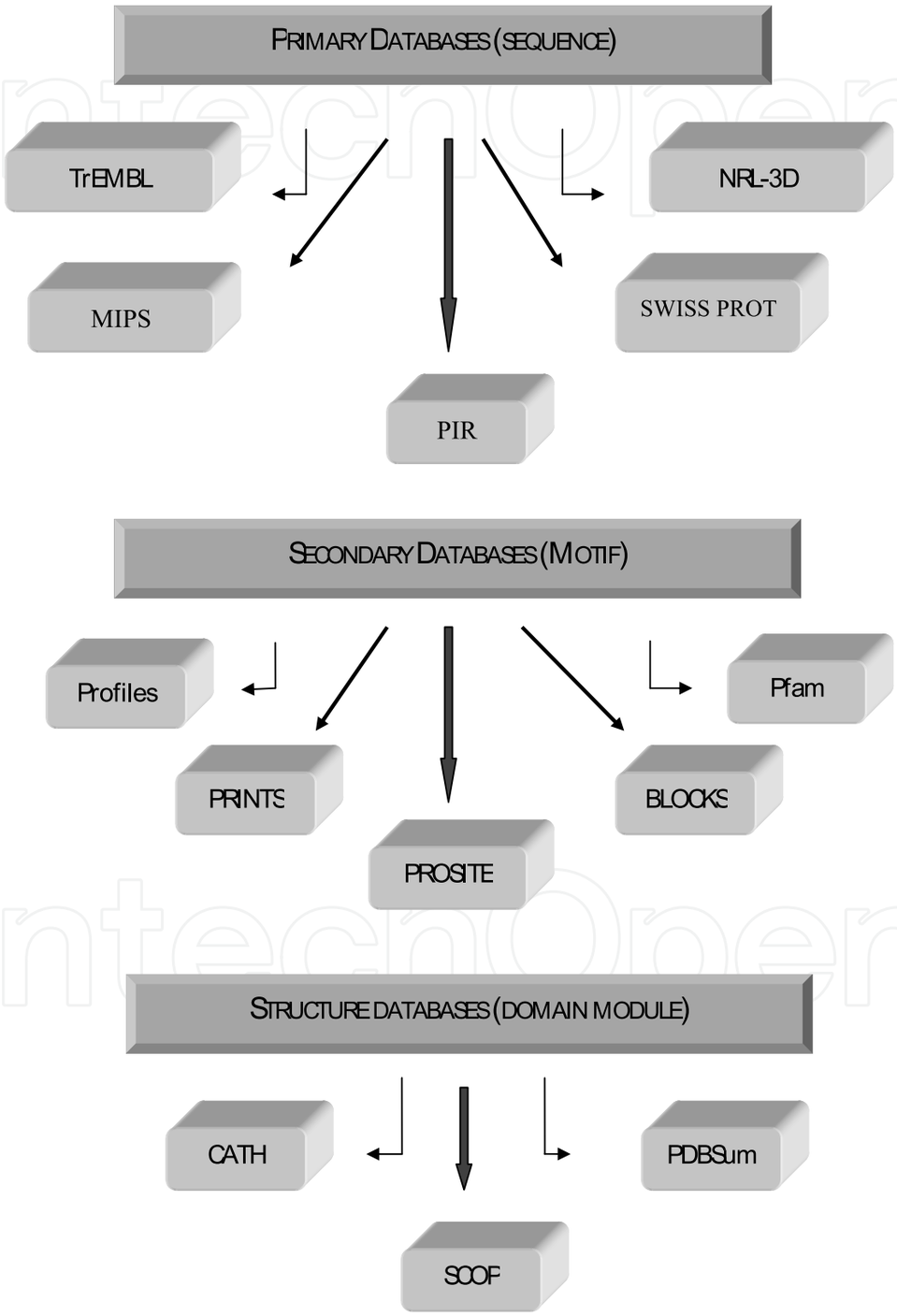


Fig. 4. Protein databases addressing different levels of protein structural information.

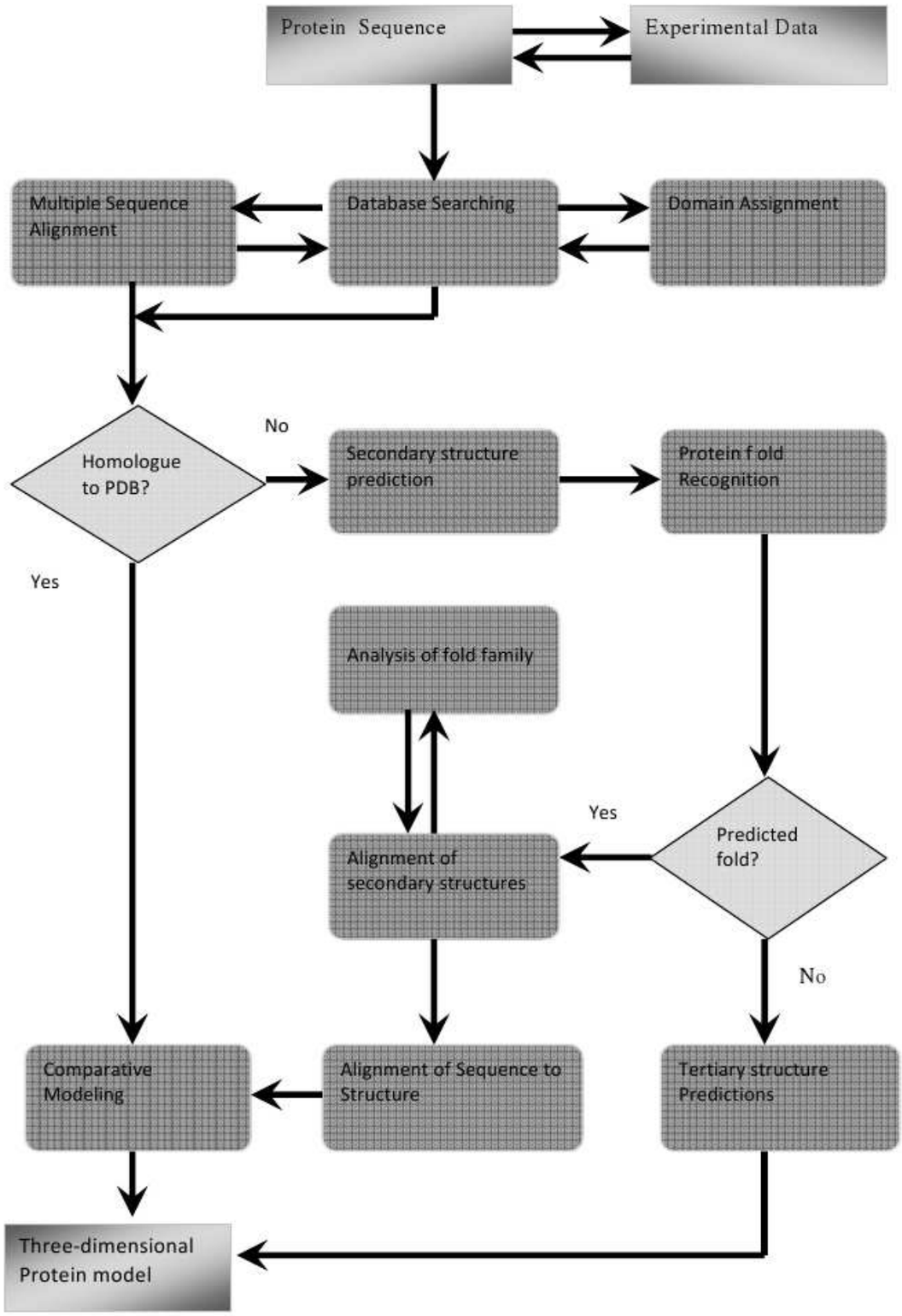


Fig. 5. Flow chart for protein structure modeling

4.2.2 OWL

This is a composite database of SWISS-PROT, PIR, GenBank and NRL-3D and is available from the UK EMBnet National Node and the UCL Specialist Node. It is a non-redundant database and is however not an updated resource but an efficient database for sequence comparisons.

4.2.3 MIPSX

This is a merged database produced at the Max-Planck Institute in Martinsried and reiterates unique copies of protein sequence search by removing identical sequences within or between them.

4.2.4 SWISS-PROT + TrEMBL

It is a combined resource of SWISS-PROT + TrEMBL at the EBI and is minimally redundant. It can be searched at the SRS sequence retrieval system on the EBI webserver.

4.3 Secondary databases

Secondary databases are a consequence of analyses of the sequences of the primary databases, mainly based from SWISS-PROT. Such databases augment the primary database searches, derived from multiple sequence information, by which an unknown query sequence can be searched against a library of patterns of conserved regions of sequence alignments which reflect some vital biological role, and based on these predefined characteristics of the patterns, the query protein can be assigned to a known family. However, secondary databases can never replace the primary sources but supplement the primary sequence search.

4.3.1 Prosite

It is the first secondary database and consists of entries describing the protein families, domains and functional sites as well as amino acid patterns, signatures, and profiles. This database was created in 1988 and is manually curated by a team of the Swiss Institute of Bioinformatics and tightly integrated into Swiss-Prot protein annotation.

4.3.2 Prints

This is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterize a protein family by iterative scanning of a SWISS-PROT/TrEMBL composite. Usually the motifs do not overlap, but are separated along a sequence, though they may be contiguous in 3D-space. Fingerprints can encode protein folds and functionalities more flexibly and powerfully than can single motifs. PRINTS can be accessed by Accession number, PRINTS code, database code, text, sequence, title, number of motifs, author or query language.

4.3.3 Blocks

Blocks are multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins. The blocks for the Blocks database are made automatically by looking for the most highly conserved regions in groups of proteins documented in InterPro. Results are reported in a multiple sequence alignment format without calibration and in the standard Block format for searching.

4.3.4 Profiles

In the motif-based approach of protein family characterization, it is probable that variable regions between conserved motifs also contain valuable sequence information. Profiles indicate where the insertions and deletions are allowed in the complete sequence alignment and provide a sensitive means of detecting distant sequence relationships.

4.3.5 Pfam

The Pfam database contains information about protein domains and families. For each entry a protein sequence alignment and a hidden Markov model is stored. These hidden Markov models can be used to search sequence databases. For each family in Pfam it is possible to look at multiple alignments, view protein domain architectures, examine species distribution, follow links to other databases and view known protein structures.

4.3.6 Identify

This resource is derived from BLOCKS and PRINTS and its search software eMOTIF is based on the generation of consensus expressions from conserved regions of sequence alignments. It can be accessed via the protein function webserver from the Department of Biochemistry at Stanford University.

4.4 Structure classification databases

Many proteins share structural similarities, reflecting common evolutionary origins. It can therefore be presumed that when the functions of proteins are conserved, the structural elements of active site residues may also be conserved giving rise to different fold families. Thus structure classification databases have evolved to better understand sequence/structure relationships. Important protein structure classification schemes are the CATH (Class, Architecture, Topology, Homology), SCOP (Structural Classification of Proteins) databases which will be dealt in detail in Section 9 of this Chapter.

4.5 Weblinks for protein databases

PIR	http://pir.georgetown.edu/
SWISS-PROT	http://expasy.org/sprot/
PROSITE	http://expasy.org/prosite/
PRINTS	http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php
Pfam	http://pfam.sanger.ac.uk/
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/
CATH	http://www.cathdb.info/

5. Sequence alignment

Two or more sequences share sequence similarities when they are homologous and share an ancestral sequence due to molecular evolution. Homology arises when the sequences share a common ancestor although similarity does not necessarily reflect homology below a certain threshold. When sequences exhibit similarities, it is likely that they will exhibit similarity of structures as well as biological functions, which enable to make predictions. This is the ultimate aim of sequence databases which requires the use of search tools that searches the sequences in the entire database against the new sequence or the query that has

been input by the user. Multiple alignments of protein sequences help to demonstrate homology which would otherwise have been considered non-significant in a pairwise alignment. In contrast to the homology of sequences over the entire length, it is also desirable to restrict homology to a limited region of the sequences. This is achieved by using a local alignment search tool, more commonly, the BLAST tool at NCBI. Multiple alignment tools are provided by EBI known as ClustalW program, most widely used with default and editable options in performing a multiple alignment (Figure 6).

BLAST is a heuristic method to find the highest scoring locally optimal alignments between a query sequence and a database sequence. It has been designed for fast database searching with minimal sacrifice of sensitivity and finds patches of local similarity, rather than a global fit. This tool works on statistics of ungapped sequence alignments and uses a substitution matrix in all phases of sequence searches. The use of filters reduces the artifacts in the databases. The BLAST algorithm works in a three-step process- the preprocessing of the query, generation of hits and extension of the hits. For a protein query, one can perform the standard BLASTP (a protein query vs. a protein database), TBLASTN (a protein query vs. six-frame translation of nucleotide sequences in the database), pairwise BLAST (between the first protein query sequence vs. the second protein sequence), PHI-BLAST (Pattern hit initiated BLAST which locates other protein sequences that contain both the regular expression pattern and the homologous sequences to a protein query) and the PSI-BLAST (Position specific iterated BLAST for finding protein families to determine domain identification and fold assignment).

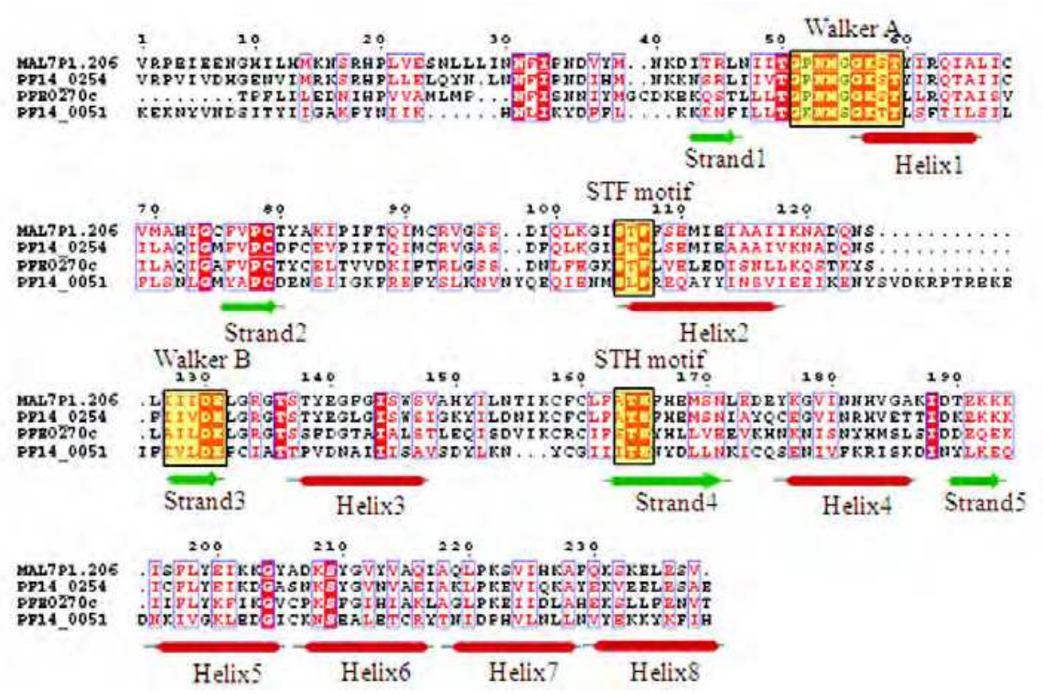


Fig. 6. Multiple sequence alignment of P-loop NTPase domain of ***P. falciparum*** MutS proteins**.** Conserved residues are in solid red and characteristic motifs are boxed in black and shaded in yellow. The corresponding secondary structure is shown below the alignment. Red cylinder represents helices and green arrows represent β-strands.

Comparing each and every sequence to every other sequence is an impractical means to obtain sequence similarity data. Often it is desirable to compare sequence sets of a given protein among its species and this is accomplished by a multiple sequent alignment by comparing all similar sequences in a single compilation, where, the sequences are aligned on top of each other, so that a co-ordinate system is set up. Each row corresponds to the sequence of a protein and each column is the same position in each sequence. Gaps are shown by dash '-' or dot '.' character. CLUSTALW is a standard program and W represents a specific version of the program. This program computes the pairwise alignments for all against for all sequences and the similarities are stored in a matrix. It converts the sequence similarity matrix values to distant measures, reflecting evolutionary distance between each pair of the sequences. It constructs a tree using neighbour-joining clustering algorithm and it progressively aligns the sequences/alignments together into each branch point of the tree. Clustal accepts alignments in several formats as: EMBL/SWISS-PROT, NBRF/PIR, GCG/MSF and its own format. There are 50 residues per line with one blank after 10 residues.

6. Protein data bank (PDB)

The Protein Data Bank (PDB) is the collection of structures and structural data of proteins, nucleic acids and other biological macromolecules. It was established in 1971 as a repository for the 3-D structural data at the Brookhaven National Laboratory, New York, and is available freely in the public domain. It is a key resource in the area of structural biology and structural genomics. PDB structures are deposited by researchers worldwide derived typically from X-ray crystallography, NMR spectroscopy, cryoelectron microscopy and theoretical modeling. PDB therefore serves as a platform to collect, organize and distribute structural information. Since 1998, PDB is an International Organization, managed by the Research Collaboratory for Structural Bioinformatics (RCSB) which facilitates the use and analysis of structural data in biological research. The PDB is overseen by an organization called the Worldwide Protein Data Bank, wwPDB. The founding members are PDBe (Europe), RCSB (USA) and PDBj (Japan). The BMRB (Biological Magnetic Resonance DataBank) joined in 2006. Each of the four members can act as deposition, data processing and distribution centres for PDB data. The data processing refers to the fact that wwPDB staff review and annotate each submitted entry. The data are then automatically checked for plausibility. The PDB website and its 'mirrors' permit retrieval of entries in computer-readable form (Kothekar, 2004).

6.1 PDB search

The PDB can be accessed at its homepage in the WWW (http://www rcsb.org/pdb/home/home.do) and several ways are available for search analysis using PDB identification code (PDB ID), searching the text found in PDB files (SearchLite), searching against specific fields of information such as deposition date or author (SearchFields), by searching the status of an entry on hold or released (StatusSearch) and by iterating on a previous search.

6.2 PDB structure

The PDB archive contains atomic coordinates, bibliographic citations, primary and secondary structure information, crystallographic structure factors and NMR experimental data. There are various options to view, download and search for structural neighbours. A

set of coordinates deposited with the PDB is subjected to a set of standard stereochemical checks and translated into a standard entry format. Each PDB entry is assigned an identifier with the first character denoting its version number.

7. Structure prediction methods

Structure prediction is an important aspect in modern biology which helps in the understanding of the functions and mechanisms of the protein macromolecule in medicine, pharmacology and biotechnology. In view of the complexity of the elucidation of protein structure by experimental means, it is now possible to use bioinformatics approaches for predictions of the protein structure. A number of software programs are available for structure predictions and the reasoning of the biologist to assess the suitability of the tools for the nature of the protein whose structure is to be determined is critical. The present methods for protein structure prediction include homology or comparative modeling, fold recognition or threading and *ab initio* or the *de novo* structure predictions for the appropriate proteins (Westhead et al., 2003). The basic approaches of these methods are discussed.

7.1 Homology or comparative modeling

This method is based on the consideration that sequences that are homologous by at least 25% over an alignment of 80 residues adopt the same structure while sequences falling below a 20% sequence identity can have very different structure. An important consideration is that tertiary structures of proteins are more conserved than their amino acid sequences. This is especially significant if a protein is similar but has been diverged; it could still possess the same overall structure. If a sequence of unknown structure (the target or query) can be aligned with one or more sequences of known structure (the template) that maps residues in the query sequence to residues in the template sequence, then, it produces a structural model of the target. Thus, homology modeling of a protein refers to constructing an atomic-resolution model of the target protein from its amino acid sequence and an experimental three-dimensional structure of the template. Homology models can be useful to derive qualitative conclusions about the biochemistry of the query sequence, about why certain residues are conserved. The spatial arrangement of conserved residues may suggest whether a particular residue is conserved to stabilize the folding, to participate in binding some small molecule, or to foster association with another protein or nucleic acid. Homology modeling can produce high-quality structural models when the target and template are closely related. The homology modeling procedure can be broken down into four sequential steps: template selection, target-template alignment, model construction, and model assessment. Figure 7 describes the sequence for homology modeling of a query protein.

The first critical step is to locate possible template structures using standard sequence similarity search methods such as BLAST for which the structures are experimentally known by experimental methods such as by X-ray crystallography or NMR spectroscopy and is available in the database. One of the limitations of homology modeling is the lack of a template structure for most proteins which is hoped to be available in the next 10-15 years with the advancements in structural genomics. When the template structure has been obtained, it is now essential to align the sequences with the target sequences by using a multiple alignment tool. When the target and template sequences closely match with high percentage identities, then, a good model is generated. The alignment should be generally checked for conserved key structural and functional residues to prevent obvious alignment

errors when there is a high percentage identity. Given a template and an alignment, the information contained therein must be used to generate a three-dimensional structural model of the target, represented as a set of Cartesian coordinates for each atom in the protein. Three major classes of model generation methods have been proposed-fragment assembly, segment matching and satisfaction of spatial restraints. Regions of the target sequence that are not aligned to a template are modeled by loop modeling. The coordinates of unmatched sections determined by loop modeling programs are generally much less accurate particularly if the loop is longer than 10 residues. Homology models without reference to the true target structure are assessed by statistical potentials or physics-based energy calculations which produce an estimate of the energy for the model being assessed. The assessment of homology models' accuracy when the experimental structure is known is direct, using the root-mean-square deviation (RMSD) metric to measure the mean distance between the corresponding atoms in the two structures after they have been superimposed.

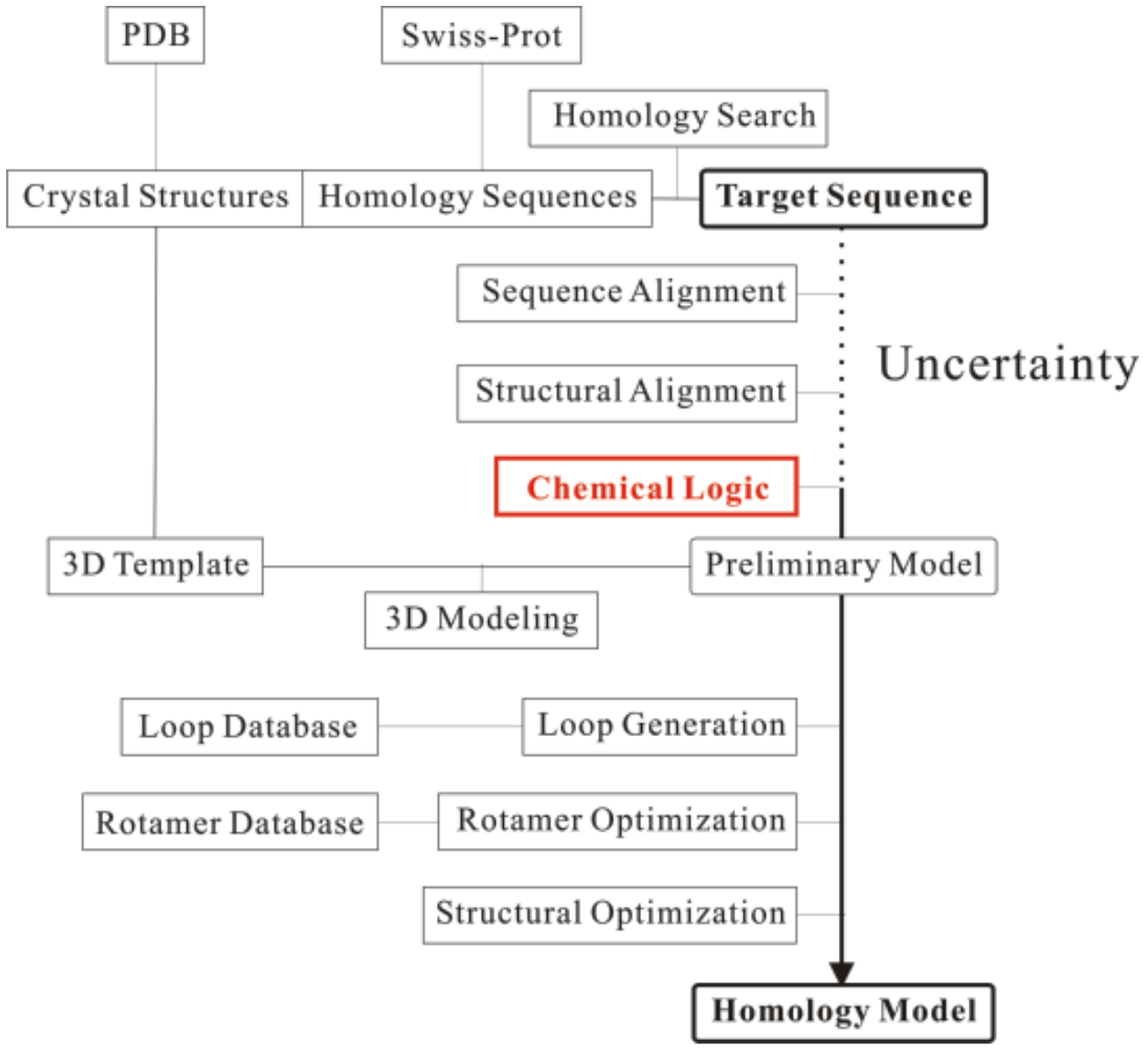


Fig. 7. Flow chart to derive protein structure by homology modeling

A number of free and commercial softwares are available in the WWW. SWISS-MODEL is a fully automated protein structure homology modeling server accessible via the ExPASy webserver or the SWISS-PDBVIEWER. It searches for suitable templates, checks sequence identity with targets, generates models and calculates energy minimization. MODELLER is another program for homology modeling. An alignment of the sequence to be modeled is to be provided and it automatically calculates a model with known related structures by satisfaction of spatial restraints. Table 4 lists some bioinformatics tools used for Homology modeling.

Web tool	**Method**
CABS	Reduced modeling tool
MODELLER	Satisfaction of spatial restraints
ROSETTA	Rosetta homology modeling
SWISS-MODEL	Local similarity / fragment assembly
TIP-STRUCTFAST	Automated comparative modeling
WHATIF	Position specific rotamers

Table 4. Homology modeling tools

7.2 Fold recognition or threading

The basic concept of threading was a result of the observation that a large percentage of proteins adopt one of a limited number folds; 10 different folds account for 50% of the known structural similarities between protein superfamilies. Rather than finding the correct structure for a given protein for all possible conformations, the correct structure is likely to have already been observed and stored in a database. In cases where the target protein shares significant sequence similarity to a protein of known 3-D structure, the fold recognition is made simple just by sequence comparison to identify the correct fold. The method of threading is thus used to detect structural similarities that are not accompanied by sequence similarity. Therefore, when a protein displays less than 25% sequence similarity to that of a template, the threading method can be used to predict its structure. This is unlike the homology modeling where sequence similarity is sufficient to guarantee similarity in structure. It is also evident that structures are conserved than sequences during evolutionary processes. Fold recognition method detects such distant relationships by searching through a library of known protein structures known as the fold library. Threading works by using statistical knowledge of the relationship between the structures deposited in the PDB and the sequence of the target protein. The prediction is made by "threading" or aligning each amino acid in the target sequence to a position in the template structure, and evaluating how well the target fits the template. After the best-fit template is selected, the structural model of the sequence is built based on the alignment with the chosen template. The flowchart of threading follows from the selection of protein structures from protein structure databases such as PDB and SCOP by eliminating structures with high sequence similarities, designing a good scoring function to measure the fitness between target sequences and templates based on the knowledge of the known relationships between the structures and the sequences, aligning the target sequence with each of the structure templates by optimizing the designed scoring function, selecting the threading alignment that is statistically most probable as the threading prediction and constructing a structure model for the target by placing the backbone atoms of the target sequence at their aligned

backbone positions of the selected structural template. Fold recognition methods can be broadly divided into those that derive a 1-D profile for each structure in the fold library and align the target sequence to these profiles, and those that consider the full 3-D structure of the protein template. Fold recognition methods are widely used and effective because there are a strictly limited number of different protein folds in nature, mostly as a result of evolution and also due to constraints imposed by the basic physics and chemistry of polypeptide chains, which authenticate the derived protein structure by this method. Homology modeling and threading are both template-based methods but the protein templates that they target are very much different. However, this method suffers from its limitations such as the weak fold recognition and domain problem in proteins with multiple domains. Table 5 lists a few of the bioinformatics tools used in Threading .

Web tool	**Method**
PSI-BLAST	Iterated sequence alignment for fold identification
3D-PSSM	3D-1D sequence profiling
SUPERFAMILY	Hidden Markov model
GenTHREADER	Sequence profile and predicted secondary structure
LOOPP	Multiple methods

Table 5. Tools for Threading method

7.3 *Ab initio* or *De novo* structure prediction

Proteins fold to attain a state of minimum thermodynamic free energy as in all physicochemical systems. This is exploited to predict the structure conformation of the protein by *ab initio* methods. Thus, this method does not require a template structure but attempts to predict tertiary structure from the sequences that govern protein folding. Therefore it uses the principles of theoretical calculations in statistical thermodynamics and quantum mechanics. The different *ab initio* methods are Molecular dynamics simulations, Monte Carlo simulations, Genetic algorithm simulations and lattice models. However, this method is not in practice when compared to homology modeling or fold recognition due to its complexity in its approach.

Table 6 indicates some of the *ab initio* structure prediction methods.

Web tool	**Method**
ROSETTA	Rosetta homology modeling and *ab initio* fragment assembly
Rosetta@ Home	Distributed-computing implementation of Rosetta algorithm
CABS	Reduced modeling tool

Table 6. *Ab initio* programs

7.4 Strategies in protein structure prediction

A set of guidelines can be followed to devise a protein structure prediction strategy. The first step in the structure prediction of the protein can be to identify the features that the protein can possess that can be examined by sequence alignment. The presence of coiled coils could be tested. A prior analysis of the target sequence with Interpro can reveal an overall domain structure. Comparative model is more suited in terms of accuracy; although it is possible only for a minority of the proteins. Fold recognition methods detect

evolutionary relationships inclusive of the consequence of divergence, however with lower accuracies.

8. Secondary structure prediction

The secondary structure of a protein refers to a consecutive fragment in its sequence that corresponds to a local region showing distinct geometrical features. These structural elements form during the early stages of the folding process. Knowledge of protein secondary structural regions along the protein sequence is a prerequisite to model the folding process or its kinetics. The ability to predict the secondary structure is a critical aspect in the structure prediction of a protein. Therefore, it is possible to recognize the three-dimensional topology by comparing the successfully predicted secondary structural elements of a query protein with the database of known topologies. Recently, it has been reported that helices and strand structures are maintained by evolution and the formation of regular secondary structure is an intrinsic feature of random amino acid sequences (Schaefer et al., 2010). Many methods are based on secondary structure propensity which reflects the preference of a residue for a particular secondary structure. Early methods were the Chou-Fasman method and the GOR method, while, predictions from multiple-aligned sequences are the Neural network methods which are based on statistical analysis. Other methods include Machine learning methods and Lim's and Cohen's methods.

8.1 Chou-Fasman method

It is a statistical prediction method based on calculation of statistical propensities of each residue forming either α-helix or β-strand. These propensities are used to classify the residues into six classes depending on their likelihood of forming an α-helix, and six classes depending on their likelihood of forming a β-strand. The class designations are used to find areas of probable α-helix and β-strands in the protein sequences to be predicted. The probable areas are then modified by a series of rules to produce the final prediction. This method is somewhat arbitrary and does not relate to chemical or physical theories. An improved version of this method for protein secondary structure prediction has been developed by Chen et al. (2006).

8.2 GOR method

The GOR (Garnier-Osguthorpe-Robson) method is based on statistical principles and is well-defined. It is based on the idea of treating the primary sequence and the sequence of secondary structure as two messages related by a translation process, which is then examined by using information theory. Structure prediction depends on measuring the amount of information the residues carry about their secondary structure and other residues secondary structure. Also theoretically complex, it is simple in practice.

8.3 PHD

This method uses a two-layered neural network method for sequence-to-structure prediction. The input of this network is a frame of 13 consecutive residues. Each residue is represented by the frequencies in the column of multiple sequence alignment which corresponds to that residue. The residues in the homologous proteins that correspond to the residue in the query protein are selected and frequencies of each type of residues are

calculated and input to the network. This means each residue introduces 20 inputs to the neural network. Also, one more input is used for each residue in the frame for the cases that the frame extends over the N or C terminus of the protein. One final input is added for each residue called the conservation weight. This weight represents the quality of a multiple sequence alignment. So every residue is represented by 20+1+1=22 inputs, thus the sequence-to-structure network has 13x22 input modes. The output of this network is 3 weights, one for each of the helix, strand and loop states. The structure-to-structure prediction part of the algorithm is also implemented as a two-layered feed-forward network (Singh et al., 2008).

8.4 Machine learning methods

The first full-scale application of machine learning to secondary structure prediction occurred with the development of a learning algorithm PROMIS. Since then, more powerful machine learning methods known as inductive logic programming (ILP) have been developed. ILP method is specifically designed to learn structural relationships between objects and is more advantageous for secondary structure prediction, using the database of known structures.

8.5

Lim's method and Cohen's method are based on physicochemical properties to encode structural knowledge of proteins.

8.6

Multiple sequence alignments significantly improve secondary structure prediction and reveal patterns of conservation as a result of evolution. A residue with a high propensity for a particular secondary structure in one sequence may have occurred by chance, but if it is part of a conserved column, in which all residues have high propensity for that type of secondary structure, then it provides predictive evidence. Multiple alignments can also reveal subtle patterns of conservation. Like, for example, a large proportion of α-helices in globular proteins are amphipathic, containing hydrophobic and hydrophilic residues associated with periodic patterns of sequences. The appearance of such conserved patterns is therefore predictive of α-helical structure.

8.7

Secondary structure prediction tools are the Jpred which is a neural network assignment method, PREDATOR which is a knowledge-based databse comparison method and Predict protein which is a profile-based neural network.

9. Structural classification

Protein structure is more conserved than its sequences. Hence, there is a need for classification of protein structures for management of protein structures deposited in databases to reflect both structural and evolutionary relatedness. Protein classification is based on a hierarchy of levels which assign the proteins to family, superfamily and fold depicting clear evolutionary relationship, probable common evolutionary origin and major

structural similarity respectively. Methods of protein structure classification rely on the sequence comparison methods and the structure comparison methods.
CATH and SCOP are the major hierarchical structure classification databases available at: http://www.cathdb.info/ and http://scop.mrc-lmb.cam.ac.uk/scop/ in the www.

9.1 CATH

This database classifies proteins based on its Class, Architecture, Topology and Homology. Class is determined by secondary structure and packing within the folded protein. Three major classes are recognized: all alpha, all beta and alphabeta, while the fourth class is composed of proteins with low secondary structure content. Architecture represents the overall shape of the domain as a function of the orientations of individual secondary structures. This level is assigned using a simple description of the secondary structure arrangement. The Topology level groups proteins into fold families depending on both the overall shape and connectivity of secondary structures. The Homologous superfamily level groups together protein domains that share a common ancestor. Structures within this level are further clustered according to their level of sequence identity.

9.2 SCOP

This database represents the Structural Classification of Proteins, a valuable resource for comparing and classifying new structures. It is designed to provide a comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known, which includes all entries in the PDB. The database is available as a set of tightly linked hypertext documents for accessibility. This classification has been constructed manually by visual inspection and comparison of structures.

10. Structure visualization

Structure visualization enables identification and manipulation of structural features in the three-dimensional view of protein macromolecules. Several programs have been developed to view structural data. Rasmol is one of the most popular tools for protein structure visualization developed by Roger Sayle which reads molecular structure files from PDB. Chemscape Chime and Protein explorer work as plug-ins to allow structure visualization in the web browser. Cn3-D is a helper application that allows viewing of 3-D structures and sequence-structure or structure-structure alignments for NCBI database. Swiss-PdbViewer provides an interface to analyze several proteins at the same time, which can be superimposed in order to deduce structural alignments and compare their active sites. It is tightly linked to Swiss-Model. PDBsum is a database which provides a largely pictorial summary of the key information on each macromolecular structure from the PDB. Table 7 lists some of the databases used for protein structure visualization.

11. Web tools in protein structure prediction

There exists unlimited information in the WWW for determination of protein structure prediction due to developments of webservers to analyse and interpret structural data. Webservers are developed and maintained by the Organizations for free availability or on commercial purposes. With the recent revolutions in bioinformatics, new software tools have been designed to meet updated protein information. This section is therefore intended to describe some of the webservers for obtaining structural information.

Program	Function
RasMol	3-dimensional visualization
Cn3-D	3-dimensional visualization, linked to sequence alignments
Chime	3-dimensional visualization
TOPS	Visualization of protein folding topologies
DSSP	Finds secondary structure elements in an input structure
Surfnet	Visualization of protein surface
PROCHECK	Checks stereochemical quality of protein structures
PROMOTIF	Analyses protein structural motifs

Table 7. Protein visualization tools

11.1 ExPASy

The ExPASy (Expert Protein Analysis System) is a proteomics server of the Swiss Institute of Bioinformatics (SIB) which analyzes protein sequences and structures and functions in collaboration with the European Bioinformatics Institute. The ExPASy server is a repertoire of tools for the many different types of protein analysis. These tools can be accessed at: http://expasy.org/tools/ and retrieve information on protein identification and characterization using mass spectrometric data, primary structure analysis, pattern and profile searches, secondary structure prediction, tertiary sequence analysis and tertiary structure prediction as well as quaternary structure analysis, molecular modeling and visualization.

Table 8 lists some of the Protein identification and characterization programs (Mala & Takeuchi, 2008).

Table 9 lists the protein structure prediction programs (Rastogi et al., 2004).

Program	Function
AACompIdent	Identification of amino acid composition
TagIdent	Identification of proteins using mass spectrometric data
PeptIdent	Identification of proteins using peptide mass fingerprint data
MultiIdent	Identification of proteins using pI, MW, amino acid composition
Propsearch	Find putative protein family
PepSea	Identification of protein by peptide mapping or peptide sequencing
FindPept	Identification of peptides resulting from unspecific cleavage of proteins
TMAP; TMHMM	Prediction of transmembrane helices

Table 8. Tools in protein identification and characterization

Program	Function
ProtParam	Physico-chemical parameters of a protein sequence
HeliQuest	A webserver to screen sequences with specific alpha-helical properties
Rep	Searches a protein sequence for repeats
Paircoil	Prediction of coiled coil in proteins
PepDraw	Peptide primary structure drawing
Jpred	A consensus method for protein secondary structure prediction
PredictProtein	A webserver from Columbia University for secondary structure prediction
PSIpred	Various protein structure prediction methods
SWISS-MODEL	An automated knowledge-based protein modeling server
LOOPP	Sequence to sequence, sequence to structure, and structure to structure alignment
Rosetta	Prediction of protein structure from sequence
MakeMultimer	Reconstruction of multimeric molecules present in crystals
Swiss-PdbViewer	A program to display, analyse and superimpose protein 3D structures

Table 9. Tools in Protein structure prediction

11.2 Predict protein

PredictProtein is a webserver available at http://www.predictprotein.org/ and works on the profile-based neural network method. It integrates feature prediction for secondary structure, solvent accessibility, transmembrane helices, globular regions, coiled-coil regions, structural switch regions, B-values, disorder regions, intra-residue contacts, protein-protein and protein-DNA binding sites, sub-cellular localization, domain boundaries, beta-barrels, cysteine bonds, metal binding sites and disulphide bridges. PredictProtein caches the prediction for each new query sequence it processes for quick and easy retrieval. Currently the PredictProtein cache contains 4,136,476 annotated proteins.

11.3 Rasmol

The software developed by Roger Sayle displays a three-dimensional image of a structure in the standard structural database. The image can be rotated by using a mouse to produce different views and displayed in various formats such as wireframe, space filling, ball and stick and cartoon formats, which give clear displays of secondary structure elements. The user can choose between various colour schemes and even use customized colours. There are flexible ways of selecting parts of structures to enable highlighting with a different display format. Figure 8 shows the different displays of a protein structure viewed in Rasmol.

11.4 DOMO and PROF_PAT

A new database of aligned protein domains known as DOMO has been developed by Gracy & Argos (1998). DOMO can be accessed through the sequence retrieval system (SRS). A form-based query manager allows retrieval of familial domain alignments by identifiers, sequence accession numbers or keywords. The DOMO sequence analysis provides a simple tool for determining domain arrangements, evolutionary relationships and key amino acid residues in a query protein sequence. PROF_PAT 1.3 is an updated database of patterns to detect local similarities, containing patterns of more than 13,000 groups of related proteins in a format similar to that of PROSITE.

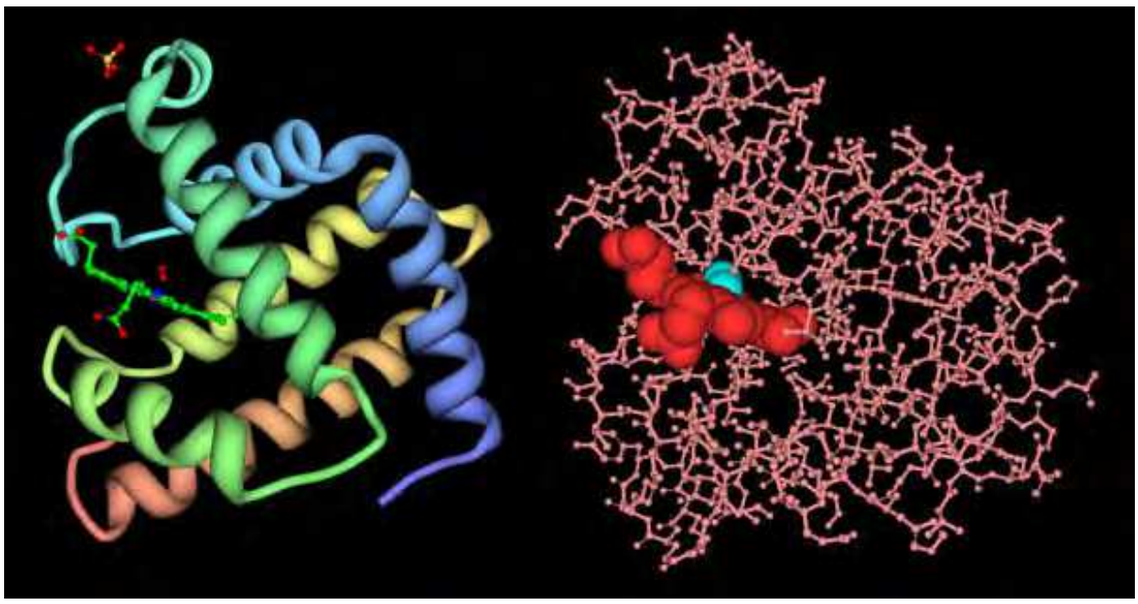


Fig. 8. Ribbon (A) and Ball stick (B) models of myoglobin viewed in Rasmol

12. CASP

CASP refers to Critical Assessment of protein Structure Prediction experimental methods to establish the current state of the art in protein structure prediction with identification of the progress made so far and highlight future efforts to be focused. CASP1 was initiated in 1994 and has been upgraded every two years. The recent method is CASP9 for the year 2010. CASP provides research groups with an opportunity to objectively test their structure prediction methods and delivers an independent assessment of the state of the art in protein structure modeling to the research community and software users. Prediction methods are assessed on the basis of the analysis of a large number of blind predictions of protein structure. The CASP results are published in special supplement issues of the scientific journal Proteins, all of which are accessible through the CASP website. The earlier version of CASP, CASP8 has been described byMoult et al. (2009) in Proteins.

13. Conclusion

Protein sequence information can be retrieved and analysed from databases that encompass much of the available sequence and structure data. On the other hand, it is of significant interest that a researcher be able to submit the sequence information for the protein investigated by him/her. Submission of sequences in any of the databases is transferred by FTP to the other databases for synchronized database management. The online submission tools provide a simple user interface and are maintained and curated on a daily basis. The vast sequence information available in the WWW requires potential search engines for data retrieval such as the Entrez from NCBI and SRS (sequence retrieval system) from EBI, which allow text-based searching of a number of linked databases. Thus, there is a continual need for sequence information and data retrieval in view of the sequence/structure deficit and also to provide links to the identification of protein biomarkers in health and disease which requires structural information. This Chapter therefore provides comprehensive information

to the reader on the application-based insights of protein structure prediction using bioinformatics approaches.

14. Acknowledgement

The authors thank Dr. A.B.Mandal, Director, CLRI, Chennai, India, for his kind permission to publish this work. The financial assistance extended by the Council of Scientific and Industrial Research (CSIR), New Delhi, India, to Dr. J.Geraldine Sandana Mala is gratefully acknowledged. The financial support of Mr. Kikuji Takeuchi and Mr. Naomi Takeuchi of Takenen, Japan for publication of this Chapter is also sincerely acknowledged.

15. References

Attwood, T.K. & Parry-Smith, D.J. (2003) *Introduction to Bioinformatics*, Fourth Indian reprint, Pearson Education Ltd., ISBN 81-7808-507-0, India

Augen, J. (2004) *Bioinformatics in the Post-genomic era*, First edition, Pearson Education Inc., ISBN 0-321-17386-4, USA

Chen, H., Gu, F. & Huang, Z. (2006) Improved Chou-Fasman method for protein secondary structure prediction. *BMC Bioinformatics*, 7, pp.S14

Gracy, J. & Argos, P. (1998) DOMO: a new database of aligned protein domains. *TIBS*, 23, pp. 495–497

Kothekar, V. (2004) *Introduction to Bioinformatics*, First edition, Dhruv Publications, ISBN 81-8240-006-6, India

Lesk, A.M. (2001) *Introduction to Protein architecture*, First edition, Oxford University Press, ISBN 0-19-850474-8, USA

Mala, J.G.S. & Takeuchi, S. (2008) Understanding structural features of microbial lipases-an overview. *Analytical Chemistry Insights*, 3, pp. 9-19

Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B. & Tramontano, A. (2009) Critical assessment of methods of protein structure prediction—Round VIII. *Proteins*, 77, pp.1-4

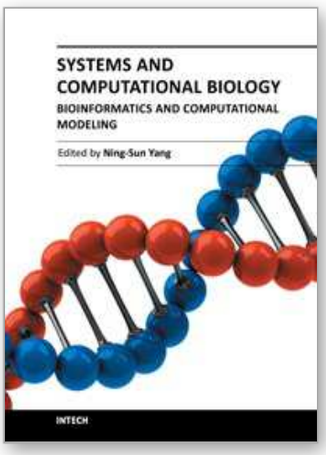
Rastogi, S.C., Mendiratta, N. & Rastogi, P. (2004) *Bioinformatics: Methods and Applications*, First edition, Prentice-Hall of India Pvt. Ltd., ISBN 81-203-2582-6, India

Schaefer, C., Schlessinger, A. & Rost, B. (2010) Protein secondary structure appears to be robust under *in silico* evolution while protein disorder appears not to be. *Bioinformatics*, 26, pp. 625-631

Singh, M., Sandhu, P.S. & Kaur, R.K. (2008) Protein secondary structure prediction. *World Academy of Science, Engineering and Technology*, 42, pp. 458-461

Westhead, D.R., Parish, J.H. & Twyman, R.M. (2003) *Instant Notes Bioinformatics*, First Indian edition, Viva Books Pvt.Ltd., ISBN 81-7649-419-4, India

Whitford, D. (2005) *Proteins structure and function*, John Wiley & Sons, ISBN 0-471-49893-9 HB, ISBN 0-471-49894-7PB, England



Systems and Computational Biology - Bioinformatics and Computational Modeling

Edited by Prof. Ning-Sun Yang

ISBN 978-953-307-875-5
Hard cover, 334 pages
Publisher InTech
Published online 12, September, 2011
Published in print edition September, 2011

Whereas some â€œmicroarrayâ€ or â€œbioinformaticsâ€ scientists among us may have been criticized as doing â€œcataloging researchâ€, the majority of us believe that we are sincerely exploring new scientific and technological systems to benefit human health, human food and animal feed production, and environmental protections. Indeed, we are humbled by the complexity, extent and beauty of cross-talks in various biological systems; on the other hand, we are becoming more educated and are able to start addressing honestly and skillfully the various important issues concerning translational medicine, global agriculture, and the environment. The two volumes of this book present a series of high-quality research or review articles in a timely fashion to this emerging research field of our scientific community.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Geraldine Sandana Mala John, Chellan Rose and Satoru Takeuchi (2011). Understanding Tools and Techniques in Protein Structure Prediction, Systems and Computational Biology - Bioinformatics and Computational Modeling, Prof. Ning-Sun Yang (Ed.), ISBN: 978-953-307-875-5, InTech, Available from: http://www.intechopen.com/books/systems-and-computational-biology-bioinformatics-and-computational-modeling/understanding-tools-and-techniques-in-protein-structure-prediction

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2011 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License, which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.

IntechOpen

IntechOpen