International Journal of Chemical Studies

P-ISSN: 2349–8528
E-ISSN: 2321–4902
IJCS 2019; 7(1): 652-659

Received: 15-11-2018
Accepted: 20-12-2018

Sikha Snehal
PhD Scholar, Department of Agricultural Biotechnology and Molecular Biology, Faculty of Basic Sciences and Humanities, Dr. Rajendra Prasad Central Agriculture University, PUSA, Samastipur, Bihar, India

.

Correspondence
Sikha Snehal
PhD Scholar, Department of Agricultural Biotechnology and Molecular Biology, Faculty of Basic Sciences and Humanities, Dr. Rajendra Prasad Central Agriculture University, PUSA, Samastipur, Bihar, India

Recent advances in science of protein structure prediction

Sikha Snehal

Abstract
Proteins form the very basis of life. Proteins regulate a range of activities in all known organisms, from replication of the genetic code to carrying oxygen, and are in general responsible for regulating the cellular machinery and subsequently, the phenotype of an organism. Proteins complete their task by three-dimensional tertiary and quaternary interactions between different substrates such as DNA and RNA, and other proteins. Thus knowing the structure of a protein is a prerequisite to gain a thorough understanding of the protein's function.

A major problem in structural bioinformatics is to conclude the three-dimensional (3-D) structure of a protein when only the sequence of amino acid residues is known. Predicting the three-dimensional structure of a protein that has no templates in the Protein Data Bank is a very hard and sometimes virtually intractable task. Over the last years, many computational methods, systems and algorithms have been developed with the purpose of solving this difficult problem. Nevertheless, the problem still challenges biologists, computer scientists, bioinformaticians, chemists and mathematicians since the complexity and high dimensionality of the protein conformational exploration space. Many computational methodologies and algorithms have been recommended as a solution to the 3-D Protein Structure Prediction (3-D-PSP) problem.

These approaches can be classified as following:
(a) First principle methods without any database information
(b) First principle methods with database information
(c) Fold recognition and threading methods
(d) Comparative modelling methods and sequence alignment strategies.

Deterministic optimization techniques, computational techniques, data mining and machine learning approaches are typically used in the construction of computational solutions for the PSP problem. This paper reviews the various recent advances in the science of protein structure prediction.

Keywords: science, protein structure, prediction

1. Introduction

1.1 Introduction to protein Structure and Representation

From a structural perspective, a protein is an ordered linear chain of building blocks known as amino acid residues. Every protein is defined by its unique sequence of amino acids. The sequence is very important as it causes the protein to fold into a specific three-dimensional shape. Predicting the folded structure of a protein only from its amino acid sequence remains a challenging problem in mathematical. The challenge arises due to the combinatorial explosion of plausible shapes each of which represent a local minimum of an intricate non-convex function of which the global minimum is sought. In nature, proteins characteristically present having 50 to 500 amino acid residues.

In nature there are 20 distinct protein genic amino acids, each one with its own chemical properties (including size, charge, polarity, hydrophobicity, i.e. the tendency to avoid water packing) (Lehninger *et al*., 2005) [4]. According to the polarity of the side-chain, amino acids differ in their hydrophilic or hydrophobic characters. The importance of the physical properties of the side-chains comes from the influence they have on the amino acid residues interactions in the 3-D structure. The allocations of the hydrophilic and hydrophobic amino acids are significant to determine the tertiary structure of the protein polypeptide.

A peptide is a molecule which is made of two or more amino acid residues bound by a chemical bond called the peptide bond. This peptide bond is formed when the carboxyl group of one residue reacts with the amino group of the other residue, thereby releasing a water molecule (H2 O). Two or more linked amino acid residues are referred to as a peptide and larger peptides are generally referred to as polypeptides or proteins. The peptide bond (C-N)

has a double bond and is not allowed rotation of the molecule around this bond. The rotation is only permitted around the bonds N-C¸ and C¸ -C.

These bonds are known as PHI (φ) and PSI (ψ) angles, respectively, and are free to rotate. This freedom is generally responsible for the conformation adopted by the polypeptide backbone. However, the rotational freedom around the (N-C¸) and (C¸ -C) angles is limited by steric hindrance between the side chain of the amino acid residue and the peptide backbone. As a result, the probable conformation of a given polypeptide is quite limited and depends on the amino acid chemical properties.

The peptide bond itself tends to be planar, with two allowed states: Trans, ω 180◦ (usually) and cis, ω 0◦ (rarely) (Branden and Tooze, 1998) [1]. The sequence of, and ω angles of all residues in a protein defines the backbone conformation or fold. The angles and can have any value between −180˚ and +180˚. However, some combinations are prohibited by steric interferences between atoms from the main-chain and atoms from the side-chain (two atoms cannot occupy the same space). The allowed and prohibited values for the torsion angles and are graphically demonstrated by the map of Sasisekharan–Ramakrishnan–Ramachandran, or simply Ramachandran map (Ramachandran and Sasisekharan, 1968) [10].

1.1 Proteins can be analysed at four levels (Lehninger *et al*., 2005) [4]

a) Primary structure,
b) Secondary structure,
c) Tertiary structure and
d) Quaternary structure.

This hierarchy facilitates the description and the understanding of proteins. However, it does not aim at describing precisely the physical laws that produce protein structures; it is an abstraction that aims at making protein structure studies more tractable.

The primary structure simply describes the sequence of amino acid residues in a linear order (Lehninger *et al*., 2005) [4]. Each amino acid residue binds to other amino acid residue through a peptide bond. The beginning of the primary structure corresponds to its N-terminal region and the end of its primary structure is the C-terminal region. Proteins are linear polymers that can assume several conformations.

The stable arrangement of amino acid residues of the polypeptide forms structural patterns (Lehninger *et al*., 2005) [4]. These structural patterns represent the secondary structure of a polypeptide. The secondary structure is defined by the presence of hydrogen bond patterns between the hydrogen atoms of the amino groups and the oxygen atoms of the carboxyl groups in the polypeptide chain. Regularity in the spatial conformation is maintained through these intermolecular interactions. There are two regular secondary structures: -helices (Pauling *et al*., 1951) [8, 9] and ˇ-sheets (Pauling and Corey, 1951) [8, 9]. There are also irregular conformations (coil and turns), but the ¸-helix and ˇ-sheets are the most stable and can be considered as the main elements present in 3-D protein structures.

The tertiary structure of a protein is represented by the distribution of secondary structures in a 3-D space. The three-dimensional shape assumed by a protein is also called native structure of the protein or functional structure. The native structure of a protein is formed by the variation of thermodynamic factors, i.e., covalent interactions, hydrogen bonds, hydrophobic interactions, electro-static interactions, van der Waals, and repulsive forces. In addition, the side-chains play an important role in creating the final structure of the polypeptide.

The tertiary structure of a protein allows the analysis and prediction of the function of the protein in the cell. It is possible to identify the active site, binding sites on a receptor, or a recombination site for the action of another protein (Lehninger *et al*., 2005) [4]. The tertiary structure of a protein is related to its topology (or fold). The topology of a protein is given by the type of succession of secondary structures that are connected to and from the shape in which these structures are organized in a 3-D space.

A protein may have different polypeptide chains (or subunits) forming a quaternary structure. The quaternary structure of a protein is the arrangement of various tertiary structures. This structure is maintained by the same forces that determine the secondary and tertiary structures (hydrogen bonds, hydrophobic interactions, hydrophilic interactions) (Lehninger *et al*., 2005) [4].

2. Protein Structure prediction methods – CASP & latest modifications

The prediction of the 3-D structure of polypeptides based only on the amino acid sequence (primary structure) is a problem that has, over the last decades, challenged biochemists, biologists, computer scientists and mathematicians. The Protein Structure Prediction Problem is one of the main research problems in Structural Bioinformatics. The main challenge is to understand how the information encoded in the linear sequence of amino acid residues is translated into the 3-D structure, and from this acquired knowledge, to develop computational methodologies that can correctly predict the native structure of a protein molecule. Many methods and algorithms have been proposed, tested and analysed over the years as a solution to this complex problem.

2.1 Floudas (Floudas *et al*., 2006) [2] classifies the computational methods for protein structure prediction into four groups

1. First principle methods without database information
2. First principle methods with database information
3. Fold recognition and threading methods
4. Comparative modelling methods and sequence alignment strategies

Regardless of the group, all developed 3-D protein structure prediction methods have to be tested for the ability to predict new protein structures. Every other year since 1994 a worldwide experiment called CASP (Critical Assessment of Structure Prediction) is performed to test protein structure prediction methods. Structural biologists who are about to publish a structure are asked to submit the corresponding sequence for structure prediction. The predictions are then compared with the newly experimentally determined structures (by NMR or X-ray crystallography methods). CASP allows research groups with an opportunity to objectively test their structure prediction methods and provides an independent assessment of the state-of-the art protein structure modelling. The CASP competition involves a large number of research groups using a variety of methods from the four groups listed above.

The most significant progress in last CASP was identified by template-based modelling methods (methods that use database information) (Huang *et al*., 2014) [3].

There was evidence of improved accuracy for targets of mid-

range difficulty, probably due to improved methods that combine information from multiple templates. The major remaining challenge in this class of methods is the development of better methods for template production and identification; accurate structures for those regions are not easily derived from an obvious template.

CASP9 and CASP10 did not reveal much progress in Free Modelling methods (first principle methods without database information) among the methods that have been tested, I-Tasser presented a significant improvement in its pre-dictions. This improvement happened because I-Tasser incorporates two components: REMO and FG-MD (Li and Zhang, 2011) [5]. REMO is a method for atomic structure construction and improvement of hydrogen-bonding network and FG-MD is fragment-guided molecular dynamics based method that uses constrained molecular dynamics simulation to adjust the position of each atom in the protein.

Each of the four classes of protein structure prediction methods that will be detailed below have some limitations. The analysis of CASP9 experiments reveals that the best results are achieved by methods which combine principles of the four groups of methods. First principle methods without database information have limitations with respect to the size of the conformational search space.

It is not possible to simulate, in plausible time, all folding process of long sequences of amino acid residues. Methods that use fragments still have two major limitations: the first one is related to the challenge of dealing with large conformational search spaces caused by different combination of such fragments; the second refers to the challenge of reducing the potential energy in regions where combinations of fragments occur.

Despite the high quality predictions, comparative modelling and fold recognition also have some limitations such as the inability to perform prediction of new folds. This is explained by the fact that these methodologies can only predict structures of protein sequences which are similar or nearly identical to other protein sequences of known structures in the PDB. Another limitation is that it is not possible to study the folding process of the protein, i.e., the path that an unfolded protein traverses to the functional state (native state).

3. First principle methods without database information

Ab initio methods, the first principle methods without database information, are founded on thermodynamics and based on the fact that the native structure of a protein corresponds to the global minimum of its free energy. Ab initio structure prediction methods aim at predicting the native conformation of a protein considering only the amino acid sequence defines "Ab initio folding" as the class of methods that are based on potential energy functions that describe the physics of a current conformational state and where only this potential function is used to search the native structure of the polypeptide.

In pure Ab initio methods the use of structural templates from a database such as the PDB is not allowed. The structural information from determined structures is only used in the parameterization of empirical all-atoms potentials used in force-fields (potential energy functions). Ab initio protein folding is considered a global optimization problem where the goal is to identify the values of a variable set (torsion angles, position of all atoms or a specific set of atoms in the protein structure) that describe the minimum energy of the polypeptide conformation.

Ab initio methods simulate the protein conformational space using an energy function, which describes the internal energy of the protein and its interactions with the environment in which it is inserted. The goal is to find a global minimum of free energy that corresponds to the native or functional state of the protein. Ab initio methods can predict new folds because they are not limited to templates from the PDB. However, these methods have some limitations with respect to the size of the conformational search space. This problem is frequently referred to by many authors as the Levinthal's paradox (Zwanzig *et al*., 1991) following studies carried out by Cyrus Levinthal in 1968.

In his experiments, Levinthal noted that due to the very large number of degrees of freedom in an unfolded polypeptide chain, a protein molecule has an enormous number of possible conformations (thus rendering a NP-Complete problem).

3.1 In general an AB initio method requires three elements:

a) A geometric representation of the protein chain,
b) A potential function and
c) An energy surface searching technique.

3.1.1 Geometric representation: This representation corresponds to the way that computationally we will represent the structure of a protein. The most detailed representations include all atoms of the protein and the surrounding solvent molecules (for example, H2 O). Using all atoms to represent the protein is computationally expensive. Such representations can be simplified in a number of ways: the all-atom model of both the protein and the solvent environment (explicit solvent) is usually replaced by employing an united atom model, where the solvent is modeled by potential fields of various descriptions (implicit solvent).

In general, the united-atom model is frequently used to reduce the computational cost. In this model, explicit hydrogen atoms-with the exception of those that have the capability to participate in hydrogen bonds – are eliminated. Virtual-atoms can also be used to represent one residue and reduce the computational cost. In turn, Rotamers can also be used to represent a limited set of conformations that side-chains can adopt in the polypeptide structure.

Almost all Ab initio folding methods use some form of simplified geometry model, in which single virtual atoms of the model represent a number of atoms in the all-atom model. The geometric representation is one of the most important elements of an Ab initio method and is directly related to the reduction or increase of the associated computational complexity. An all-atom model can demand enormous computational effort during a simulation. On the other hand, simplified representation models can preserve the main structure characteristics and reduce the computational time demanded by a protein folding simulation.

3.1.2 Potential functions: The second element of an Ab initio method is a potential energy function. Potential energy functions are used in Molecular Mechanics (MM) simulations, protein design (Li *et al*., 2013) [6] and protein structure prediction.

There are two categories of potentials: MM potentials and protein structure-derived potential functions (scoring functions). The first category aims at modeling the forces that determine protein con-formations using physically-based parameterized functional forms from small molecule data or in vacuo quantum mechanics calculations. The second

category is empirically derived from experimental structures from the PDB. These two classes of potentials represent the forces that determine the macromolecular conformation: solvation, electrostatic, van der Waals interactions, covalent bonds, angles, torsions.

The main advantage of using a knowledge-based energy function is that it can model any behaviour observed in known protein crystal structures, even when there is no good physical understanding of their behaviour. The disadvantage is that these functions cannot predict new behaviours absent in the training set obtained from the experimental database.

A potential energy function incorporates two types of terms: bonded and non-bonded. The bonded terms (bonds, angles and torsions) are covalently linked. The bonded terms constrain bond lengths and angles near their equilibrium values. The bonded terms also include a torsional potential (torsion) that models the periodic energy barriers encountered during bond rotation. The non-bonded potential includes: ionic bonds, hydrophobic inter-actions, hydrogen bonds, van der Waals forces, and dipole–dipole bonds. There is a great number of potential energy functions used in computational molecular biology. AMBER, CHARMM and ECEPP are the most widely used potential energy functions in 3-D PSP and Protein Folding problems.

3.1.3 Energy surface search techniques: methods for Ab initio pre-diction include Molecular Dynamics simulations of proteins and protein-substrate complexes; Monte Carlo simulations that do not use forces but rather compare energies, via the use of Boltz-mann probabilities. Genetic Algorithms which are based on populations of solutions by iterative cycles of operations and try to improve on the sampling and the convergence of Monte Carlo approaches and exhaustive and semi-exhaustive lattice-based studies which are based on using a crude/approximate fold representation (such as two residues per lattice point) and then exploring all or large amounts of the conformational space given the crude representation.

There are many computational packages used in Ab initio protein structure simulations. These simulation packages are frequently used in the protein folding problem and in other molecular modelling problems such as molecular docking, which predicts the preferred orientation of a molecule with respect to another molecule when bound to each other to form a stable complex. There are also Ab initio algorithms developed specifically for the 3-D PSP

Some of these are

AMBER- Assisted Model Building with Energy Refinement
CHARMM- Chemistry at HARvard Molecular Mechanics
GROMACS- Groningen Machine for Chemical Simulation
TINKER- Software Tools for Molecular Design
LINUS -Local Independent Nucleated Units of Structure

4. First principle methods with database information

In first principle methods with database information general rules of protein structures are extracted from protein databases and used to build starting point 3-D protein structures. These methods do not compare a target sequence to a known structure, but they compare fragments, i.e. short amino acid sub-sequences of a target fragment against fragments of known protein structures.

This arises from the observation that when a new fold is discovered, it is composed of common structural motifs or fragments from super-secondary structures of proteins with known structures. Thus, if there are protein fragments that fold into similar structures, then this information or these fragments can be used to construct 3-D structural models of proteins. This is the essence of the methods based on fragments. The conformation of a protein is seen as a set of various fragments of amino acid sequences representing various structural motifs that are combined to form a 3-D protein structure. When homologue fragments are identified they are assembled into a structure through scoring functions and optimization algorithms.

The fragments are assembled through a fragment assembly procedure with the purpose of finding the structure with the lowest potential energy. When finding polypeptide structures with the lowest energy potential, these methods are similar to ab initio methods. However, they cannot be classified as ab initio methods because they use database information to predict the structure of polypeptides. Fragment-based methods are based on the premise that local interactions can define local structures in proteins. Local structures present in known protein structures are used in order to predict the structure of a target amino acid sequence. When appropriate fragments have been identified, compact structures can be assembled by randomly combining fragments using, for example, a simulated annealing approach.

Similar local sequences do not always present the same 3-D structure. This occurs because in a 3-D structure a large number of physicochemical interactions are present; such interactions con-tribute not only to the stability of the global structure, but also to the configuration of the secondary structures. Thus, fragment-based methods cannot fragment the target amino acid sequence, search database template fragments, get their information and combine these fragments without any combination criterion. Non-covalent interactions between atoms of different regions of the molecule influence the formation of local structures.

Fragment-based methods need to establish a relationship criterion between the fragments so that they can determine the fragments with higher probability of insertion during the prediction of the final structure. In this sense, scoring functions are frequently used. The fitness of a conformation can be assessed with scoring functions derived from conformational statistics of known proteins (Floudas *et al*., 2006) [2].

Usually, given the complete sequence of amino acids in a protein, the fragment-based method are composed of five distinct stages where:

1. It divides the target sequence into fragments;
2. It carries out the search for similar sequences from each frag-ment, in a database of known structures;
3. It classifies the fragments (scoring);
4. It constructs the three-dimensional structure from the fragment template using a combination technique;
5. Finally, it refines the conformation.

As first principle method without database information, fragment-based methods offer advantages over the other classes of prediction methods. The first advantage refers to the ability of predicting new folds, which cannot be achieved by methods based on homology modelling. In comparison with Ab initio methods, fragment-based methods take advantage of the reduction of the conformational search space.

This reduction is due to the fact that in a simple replacement of a fragment in the target protein, this fragment moves from one region of a protein which has a structure with minimum

potential energy. However, despite reducing the conformational search space, the methods that use fragments still have two major limitations.

The first one is related to the challenge of dealing with large conformational search spaces caused by different combination of such fragments. The second one refers to the challenge of reducing the potential energy in regions where combination of fragments occur. Fragment-based methods produced very positive results in the CASP experiments (Moult *et al*., 2014) [7].

5. Fold recognition and threading methods

Fold recognition methods are motivated by the notion that structure is more evolutionary preserved than sequence, i.e., proteins with no apparent sequence similarity could have similar folds. Several studies in the last years have indicated that the number of protein structural folds in nature is limited. Today, for example, there are approximately ten different folds in fifty percent of the proteins with known structure.

The general goal of 3-D protein structure prediction by threading methods is to fit a protein sequence correctly against a structural model. During this procedure the target amino acid sequence is placed, following their sequential order, into structural positions of a template 3-D structure in an optimal way.

5.1 This involves two basic procedures:

a) Selecting a structural model from a library of models
b) Finding the correct replacement between the target sequences against the structural models in the space of possible sequence-structure alignments.

Threading methods use structural information such as residue-residue contact patterns, secondary structure and solvent accessibility, and after identifying the structural similarities, which cannot be detected solely by the similarities between the amino acid sequences, the predicted structural models are constructed.

In threading methods for the 3-D PSP problem it is necessary to solve the problem of sequence-structure replacement, where, given a solved structure T = t1, t2, . . ., tn and a target sequence S = s1, s2, . . ., sm the main goal is to find the best match between S and T. Threading methods use known 3-D protein structures as templates for sequences of unknown structures. Threading methods try to identify templates with similar fold with or without direct evolutionary relations (analogue). Homologue proteins are the result of divergent evolution and often share a common function. Analogue proteins do not have a common ancestor and generally do not have a common function. In both cases the proteins share a common three-dimensional structure without a significant sequence similarity. Comparative modeling usually employs sequence-sequence comparison while threading usually exploits structure information to assist alignment. Com-pared to first principle methods without database information (Ab initio), threading methods seek to optimize a potential energy function (an objective or scored function) measuring the fitting quality of a sequence in a particular 3-D configuration. This measure will be assessed using statistical or energetic measurements for the over-all likelihood of the target amino acid sequence adopting one of the available structural folds.

In a general form fold recognition methods can be divided in two group: profile-based and pair potentials-based. On the first group the information of the structural database containing potential target structures is represented in a linear form or profile. In this case the target protein is matched in turn with this profile. The second group uses pair potentials which score the propensity of two residues being at a certain distance.

5.2 A threading method typically consists of three components:

1) Construction of a library of potential folds or structural templates.
2) A scoring schema to evaluate any particular placement of a target sequence into each fold.
3) A method to search over the vast space of possible replacements between each sequence and each fold for the best set that gives the best total score. Next, we detail these four components.

5.3 Construction of a library of potential folds or structural tem-plates: the library of folds is constructed from known native protein structures derived, for example, from the PDB. Usually, the 3-D coordinates of a protein structure are reduced to more abstract representations. Structural core elements are defined by the secondary structure elements: ˇ-sheet, ˛-helix, left handed helix, coil, strands. Frequently, side-chain information is removed. What remains is a backbone template of blank or empty amino acid positions

5.4 A scoring scheme to evaluate any particular placement of a sequence into each fold: the scoring functions are usually a list of statistical references of each amino acid residue to each structural or fold environment. These functions describe how favorable a replacement of a query sequence and a template structure are. Most threading methods do not use physical full-atom free energy function as used by first principle methods without database information. Most threading objective energy functions are determined empirically by statistical analysis of 3-D data obtained from the PDB. These functions are referred to in general as knowledge-based functions and are used in both profile-based and pair potentials-based methods.

5.5 A method to search over the vast space of possible replacements: the use of an algorithm to identify the optimal sequence-structure replacement is essential in a threading method. The main task is to identify the global best score and the optimal fitting/threading. There are at least two main approaches to the sequence-structure replacement: (1) 3-D profile methods and (2) contact potentials Today most threading methods fall into category 2 above.

6. Comparative modelling methods and sequence alignment strategies

In comparative modeling a target sequence of amino acid residues (target protein) is aligned against the amino acid sequence of another protein with known structure (template protein) and stored in the PDB. If the target sequence is similar to the sequence of the template protein, the structural information obtained from the known structure is used for modeling the target protein. The main idea of this kind of method is to construct an atomic-resolution model of the target protein from its amino acid sequence and an experimental 3-D structure of a related homologous protein. Comparative modeling can be applied whenever it is possible to detect an evolutionary relation-ship between the target protein and the template protein of which the 3D structure is known. The evolutionary relationship between proteins is a fundamental factor in comparative modeling methods and the

target protein can be modeled from homologous proteins with 3-D structures determined experimentally. The structure of these proteins is similar in the sense that amino acid residues with identical physico-chemical properties occupy the same position in homologous proteins

6.1 Steps in model production

The homology modeling procedure can be broken down into four sequential steps: template selection, target-template alignment, model construction, and model assessment. The first two steps are often essentially performed together, as the most common methods of identifying templates rely on the production of sequence alignments; however, these alignments may not be of sufficient quality because database search techniques prioritize speed over alignment quality. These processes can be performed iteratively to improve the quality of the final model, although quality assessments that are not dependent on the true target structure are still under development process.

Optimizing the speed and accuracy of these steps for use in large-scale automated structure prediction is a key component of structural genomics initiatives, partly because the resulting volume of data will be too large to process manually and partly because the goal of structural genomics requires providing models of reasonable quality to researchers who are not themselves structure prediction experts.

6.2 Template selection and sequence alignment

The critical first step in homology modeling is the identification of the best template structure, if indeed any are available. The simplest method of template identification relies on serial pairwise sequence alignments aided by database search techniques such as FASTA and BLAST. More sensitive methods based on multiple sequence alignment – of which PSI-BLAST is the most common example – iteratively update their position-specific scoring matrix to successively identify more distantly related homologs. This family of methods has been shown to produce a larger number of potential templates and to identify better templates for sequences that have only distant relationships to any solved structure.

Protein threading, also known as fold recognition or 3D-1D alignment, can also be used as a search technique for identifying templates to be used in traditional homology modeling methods. Recent CASP experiments indicate that some protein threading methods such as RaptorX indeed are more sensitive than purely sequence(profile)-based methods when only distantly-related templates are available for the proteins under prediction. When performing a BLAST search, a reliable first approach is to identify hits with a sufficiently low E-value, which are considered sufficiently close in evolution to make a reliable homology model.

Other factors may tip the balance in marginal cases; for example, the template may have a function similar to that of the query sequence, or it may belong to a homologous operon. However, a template with a poor E-value should generally not be chosen, even if it is the only one available, since it may well have a wrong structure, leading to the production of a misguided model. A better approach is to submit the primary sequence to fold-recognition servers [9] or, better still, consensus meta-servers which improve upon individual fold-recognition servers by identifying similarities (consensus) among independent predictions. Often several candidate template structures are identified by these approaches. Although some methods can generate hybrid models with better accuracy from multiple templates, most methods rely on a single template. Therefore, choosing the best template from among the candidates is a key step, and can affect the final accuracy of the structure significantly.

This choice is guided by several factors, such as the similarity of the query and template sequences, of their functions, and of the predicted query and observed template secondary structures. Perhaps most importantly, the coverage of the aligned regions: the fraction of the query sequence structure that can be predicted from the template, and the plausibility of the resulting model. Thus, sometimes several homology models are produced for a single query sequence, with the most likely candidate chosen only in the final step.

6.3 Model generation

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.

6.4 Fragment assembly

The original method of homology modeling relied on the assembly of a complete model from conserved structural fragments identified in closely related solved structures. For example, a modeling study of serine proteases in mammals identified a sharp distinction between "core" structural regions conserved in all experimental structures in the class, and variable regions typically located in the loops where the majority of the sequence differences were localized. Thus unsolved proteins could be modeled by first constructing the conserved core and then substituting variable regions from other proteins in the set of solved structures. Current implementations of this method differ mainly in the way they deal with regions that are not conserved or that lack a template. The variable regions are often constructed with the help of fragment libraries.

6.5 Segment matching

The segment-matching method divides the target into a series of short segments, each of which is matched to its own template fitted from the Protein Data Bank. Thus, sequence alignment is done over segments rather than over the entire protein. Selection of the template for each segment is based on sequence similarity, comparisons of alpha carbon coordinates, and predicted steric conflicts arising from the van der Waals radii of the divergent atoms between target and template.

6.6 Satisfaction of spatial restraints

The most common current homology modeling method takes its inspiration from calculations required to construct a three-dimensional structure from data generated by NMR spectroscopy. One or more target-template alignments are used to construct a set of geometrical criteria that are then converted to probability density functions for each restraint. Restraints applied to the main protein internal coordinates – protein backbone distances and dihedral angles – serve as the basis for a global optimization procedure that originally used conjugate gradient energy minimization to iteratively refine the positions of all heavy atoms in the protein.

This method had been dramatically expanded to apply specifically to loop modeling, which can be extremely difficult due to the high flexibility of loops in proteins in aqueous solution. A more recent expansion applies the

spatial-restraint model to electron density maps derived from cryoelectron microscopy studies, which provide low-resolution information that is not usually itself sufficient to generate atomic-resolution structural models. To address the problem of inaccuracies in initial target-template sequence alignment, an iterative procedure has also been introduced to refine the alignment on the basis of the initial structural fit. The most commonly used software in spatial restraint-based modeling is MODELLER and a database called ModBase has been established for reliable models generated with it.

6.7 Loop modeling

Regions of the target sequence that are not aligned to a template are modeled by loop modeling; they are the most susceptible to major modeling errors and occur with higher frequency when the target and template have low sequence identity. The coordinates of unmatched sections determined by loop modeling programs are generally much less accurate than those obtained from simply copying the coordinates of a known structure, particularly if the loop is longer than 10 residues. The first two side chain dihedral angles ($\chi1$ and $\chi2$) can usually be estimated within 30° for an accurate backbone structure; however, the later dihedral angles found in longer side chains such as lysine and arginine are notoriously difficult to predict. Moreover, small errors in $\chi1$ (and, to a lesser extent, in $\chi2$) can cause relatively large errors in the positions of the atoms at the terminus of side chain; such atoms often have a functional importance, particularly when located near the active site.

7. List of protein structure prediction software

This **list of protein structure prediction software** summarizes commonly used software tools in protein structure prediction, including homology modeling, protein threading, *ab initio* methods, secondary structure prediction, and transmembrane helix and signal peptide prediction.

7.1 Software highlight

7.1.1 I-TASSER is the best server for protein structure prediction according to the 2006-2012 **CASP** experiments (CASP7, **CASP8,** CASP9, **CASP10,** and CASP11)**.** The standalone I-TASSER package is freely available for download.

7.1.2 HHpred was the leading server for template-based protein structure prediction in the 2010 CASP9 experiment. It has a median response time of a few minutes instead of days like other top-performing servers. HHpred is often used for remote homology detection and homology-based function prediction. It runs with the free, open-source software package HH-suite for fast sequence searching, protein threading and remote homology detection.

7.1.3 RaptorX excels at aligning hard targets according to the 2010 CASP9 experiments. RaptorX generates the significantly better alignments for the hardest 50 CASP9 template-based modeling targets than other servers including those using consensus and refinement methods. The RaptorX server is available at server

7.1.4 MODELLER is a popular software tool for producing homology models by satisfaction of spatial restraints using methodology derived from NMR spectroscopy data processing. The Mod Web comparative protein structure modeling web-server uses primarily MODELLER for automatic comparative modeling.

7.1.5 Geno3D is webserver for producing homology models by satisfaction of spatial restraints using methodology derived from NMR data processing. webserver

7.1.6 Swiss-Model provides an automated web server for protein structure homology modeling.

7.1.7 bioinfo-pl and Robetta widely used servers for protein structure prediction. SPARKSx is one of the top performing servers in the CASP focused on the remote fold recognition.

7.1.8 PEP-FOLD is a *de novo* approach aimed at predicting peptide structures from amino acid sequences, based on a HMM structural alphabet.

7.1.9 Phyre and Phyre2 are amongst the top performing servers in the CASP international blind trials of structure prediction in homology modelling and remote fold recognition, and are designed with an emphasis on ease of use for non-experts.

7.1.10 RAPTOR (software) **is** a protein threading software that is based on integer programming. The basic algorithm for threading is described in Bowie (1991) and is fairly straightforward to implement.

7.1.11 QUARK is an algorithm developed for *ab initio* protein structure modelling.

7.1.12 Abalone is a Molecular Dynamics program for folding simulations with explicit or implicit water models.

7.1.13 TIP is a knowledgebase of STRUCTFAST models and precomputed similarity relationships between sequences, structures, and binding sites. Several distributed computing projects concerning protein structure prediction have also been implemented, such as the Folding@home, Rosetta@home, Human Proteome Folding Project, Predictor@home, and TANPAKU.

7.1.14 CABS-FOLD is a server that provides tools for protein structure prediction from sequence only (de novo modelling) and also using alternative templates (consensus modelling).

7.1.15 Bhageerath is another Ab-initio modelling server.

7.1.16 Foldit program seeks to investigate the pattern-recognition and puzzle-solving abilities inherent to the human mind in order to create more successful computer protein structure prediction software.

7.1.17 BBSP (Building Blocks Structure Predictor) is a program that makes use of Hybrid template-based approaches, which associate fragment conformations for the sequence and detect distant fold similarities based on the fragment similarities Computational approaches provide a fast alternative route to antibody structure prediction. Recently developed antibody F_V region high resolution structure prediction algorithms, like Rosetta Antibody, have been shown to generate high resolution homology models which have been used for successful docking

8. Conclusion

The classification of the prediction methods into four classes, (1) first principle methods without database information; (2) first principle methods with database information; (3) fold recognition and threading methods and (4) comparative modelling methods and sequence alignment strategies – gives a more general view about which methods can be used in the prediction, how experimental data can be used in the prediction tasks, and how a protein conformation can be represented in terms of physical and chemical laws (in the protein folding process).

Knowledge-based methods are limited to experimental data, e.g., homology modelling can only predict structures of protein sequences which are similar or nearly identical to other protein sequences of known structure. Fold recognition via threading is limited to the fold library derived from the PDB structure database. Ab initio methods can obtain new structures with novel folds. However, the complexity and high dimensionality of the conformational search space even for a small protein molecule still makes the problem intractable.

Over the last years, probably the most important results in this field were produced by hybrid methods such as the ones based on first principles with database information. Such hybrid methods combine the accuracy of knowledge-based methods with a more realistic, force field-based, physicochemical description of a protein. The last results presented in the CASP competition corroborate this statement. ROSETTA, FRAGFOLD, I-TASSER and LINUS all belong to this class of methods. ROSETTA and I-TASSER have been the most successful predictors over the last years according to data from the CASP experiments. In the last CASP, the bioinformatics com-munity focused on the problem of predicting the local and global regions of the 3-D model when experimental structural data are not available. Machine learning techniques, statistical potentials, physical energy functions have been applied in order to find accurate structures.

Finally, protein structure prediction is a very difficult problem and further research remains to be done. The development of new strategies, the adaptation and investigation of new methods and the combination of existing and state-of-the-art computational meth-ods and techniques to the 3-D PSP problem are clearly needed. Understanding how experimental data can be better used in combination with Ab initio techniques is another open research question. In summary, there are several research opportunities and avenues to be explored in this field, with relevant multidisciplinary applications in computer science, bioinformatics, chemistry, biochemistry, and the medical sciences.

9. References

1. http://www.iscb.org/cms_addon/conferences/ismb2000/tutorials/samudrala.html (Accessed on 3/10/2016)
2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839925/ (Accessed on 3/10/2016)
3. https://en.wikipedia.org/w/index.php?title=List_of_protein_structure_prediction_software & printable = yes (Accessed on 3/10/2016)
4. https://www.expasy.org/proteomics/protein_structure (Accessed on 2/10/2016)
5. https://www.predictprotein.org/ (Accessed on 2/10/2016)
6. Branden C, Tooze J. Introduction to Protein Structure, 2nd ed. Garland Publishing Inc, New York, 1998.
7. Floudas C, Fung H, McAllister S, Moennigmann M, Rajgaria R. Advances in protein structure prediction and de novo protein design: a review. Chem. Eng. Sci. 2006; 61(3):966.
8. Huang YJ, Mao B, Aramini JM, Montelione GT. Assessment of template-based protein structure predictions in casp10. Proteins: Struct. Funct. Bioinf. 2014; 82:43-56.
9. Lehninger A, Nelson D, Cox M. Principles of Biochemistry, 4th ed. W.H. Freeman, New York, 2005
10. Li Y, Zhang Y. Atomic-level protein structure refinement using fragment guided molecular dynamics conformation sampling. Structure 2011; 19(12):1784.
11. Li Z, Yang Y, Zhan J, Dai L, Zhou Y. Energy functions in de novo protein design: current challenges and future prospects. Annu. Rev. Biophys. 2013; 42(1):315–335
12. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (casp) – round x. Proteins: Struct. Funct. Bioinf. 2014; 82:1-6.
13. Pauling L, Corey R. The pleated sheet, a new layer configuration of polypep-tide chains. Proc. Natl. Acad. Sci. U. S. A. 1951; 37(5):251.
14. Pauling L, Corey R, Branson H. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl. Acad. Sci. U. S. A. 1951; 37(4):205.
15. Ramachandran G, Sasisekharan V. Conformation of polypeptides and pro-teins. Adv. Protein Chem. 1968; 23:238.