International Journal of Recent Research and Review, Vol. XI, Issue 3, September 2018
ISSN 2277 – 8322

Protein Structure Determination and Prediction: A Review of Techniques

Sapna V M[1], Roshan Makam[2], V K Agrawal[3]
[1] *Research Scholar, Biotechnology/Computer Science Department, PES University, Bangalore-560085, India*
[2] *Professor, Biotechnology Department, PES University, Bangalore-560085, India*
[3] *Professor, Computer Science Department and Ex Director CORI, PES University, Bangalore-560085, India*
Email: [1]sapnavmk@gmail.com, [2]mvrmakam@pes.edu, [3]vk.agrawal@pes.edu

***Abstract*—Protein structure determination and prediction are important techniques in understanding the protein function. Although determination is laborious, expensive and cumbersome at times, computational techniques have been used to ease the structure prediction. Predicting the 3-D structure of a protein from its primary structure is possible with computational techniques, however there is no single computational method which can predict all the protein structures. This article explores the protein structure determination and prediction techniques that are available to explore the best method to be chosen for a given situation. Special mention is made on Critical Assessment of Structure Prediction (CASP) as well. The article concludes that computational methods have to be further explored for a single method to predict all the protein structures. The Protein structure prediction has been and will continue to be in the frontiers of research.**

***Keywords*—Protein structure determination, Levinthal's paradox, Protein crystallization, Computational biology, CASP**

I. INTRODUCTION

Knowing the protein 3-D structure from its primary structure [1] helps in applications such as medicine, agriculture, industry and never the less in life [2]. The body makes many different types of proteins, each of which has a specific role in metabolism. The shape of the protein molecule is what dictates its function [2]. Introduction of protein crystallization is a major outbreak in determining the structure of the protein [3]. Knowing the structure helps in elucidating the function of the protein. Techniques such as vapor diffusion, microbathe and microdialysis as well as specialized crystallization techniques like high throughput crystallization screening and protein engineering are employed to obtain the protein crystals. Once a pure protein is obtained, X-ray diffraction, NMR spectroscopy [4], and Electron microscopy are widely used techniques to determine the structure. But, the protein crystallization hinders the progress in determining its structure with these techniques as the crystallization slows down the protein determination and also the techniques are robust, time consuming and expensive, though quite accurate. Levinthal's paradox is a thought experiment; it acts as a self-reference in protein folding theory. In 1969, Cyrus Levinthal noted that, the molecule has an astronomical number of possible conformations due to huge number of degrees of freedom in an unfolded polypeptide chain. The process of protein folding takes in the order of milliseconds, Levinthal's paradox [5-6] suggests a contradiction between this folding timescale and the multitude of all possible conformations the system can explore. This leads to the assumption that the native state is determined by kinetics and energy considerations, such that it is the lowest kinetically accessible minimum, which is not necessarily the global energy minimum by Anfinsen's hypothesis [7]. A further development of this idea is that the energy surface must take the form of a 'funnel' to be biased towards the native state [8]. This being the reason and also due to huge availability and increase in the growth of biological data, the need of bioinformatics, computational techniques and tools are undoubtedly important to predict the protein structure. Application of computer technologies in the

field of biology is in demand especially in structural prediction of proteins as there is a notable gap between the sequences currently available in PDB and existing three-dimensional structures.PDB contains 141,842(10.07.2018) known structures [9a], which is remarkably less compared to the sequences available in UniProt/TrEMBL database which is 116,030,110 (10.07.2018) [9b]. Prediction of protein structure depends on the accuracy and complexity of the models used. The proteins tend to fold to their native structures according to the natural process and natures flawless algorithm, hence when a new protein is synthesized within the cell it spontaneously folds to its three-dimensional structure to perform its function. with. For the past four decades continuous work and efforts are put in, to study and determine the natural process and figure out the algorithm of nature. Several computational methods have been technologically advanced to understand this natural process of protein folding and predict the protein structure from their primary structure [10]. Broad classes of prediction techniques available move ahead in solving the structure of proteins. To name, the prediction methods are template based and template free modeling accompanied with high pace algorithms and tools to complete the job of predicting the protein structure. This field of computational biology is quite open to put in efforts to solve the problem of protein structure. To this date there is no single prediction tool that is able to predict all the protein structures.

II. PROTEIN STRUCTURE DETERMINATION

A. History

Determination of protein structure is a daunting task. Its history dates back to 1958 when British scientists John Kendrew and Max Perutz made a remarkable publication. Their work on the protein structure of very high-resolution value, oxygen storage protein myoglobin and then related to it the oxygen-transporting protein hemoglobin [11] was a remarkable one for which the two shared Nobel Prize in chemistry in 1962. Unlike the work done by Watson and Crick of introducing the structure of DNA that they had revealed five years earlier, there were many significant irregularities in the first structures deduced and was not neat and fine like DNA, and also Kendrew and Perutz together came up with vastly varying different shapes and features. These foretold the great variety in protein structure that researchers would work on to discover over the next decades, showing how in biology the molecules can take different roles. It has been a pretty long path since 1958 till date which has helped to understand the value of structure of protein and its determination. Proteins structure determination mainly involves, crystallization of the protein first and then is subjected to X-rays to determine its structure. Crystallization of proteins itself is a challenge especially for membrane proteins.

B. Protein Crystallization

Crystallization of proteins is the process of formation of three-dimensional array of proteins. In nature some protein crystals have been observed [12] and also proteins when dissolved, tend to form crystals in the supersaturated solution. In such conditions, every single protein molecule can be seen as a pack of a repeating array, seized together by non-covalent interactions [13]. Later the observed crystals help in understanding the structural biology i.e. to know the molecular structure of the protein which in turn makes remarkable application in industry and field of biotechnology, most markedly for the study of X-ray crystallography. The prime purpose of crystallization is to have contamination free crystals at the same time aim to have large enough crystals to produce a diffraction pattern when exposed to X-rays, and then the protein's tertiary structure is revealed and studied with the diffraction pattern obtained. Crystallization of protein is indeed inherently difficult due the fragile nature of protein crystals. Protein crystallization is a challenging task. Various points need to be considered in the crystallization process like, the confines of the aqueous environment, complications in getting high-quality protein samples, also protein sample sensitivity, temperature, pH, ionic strength, and other factors. Protein crystallization is rarely predictable because they vary significantly in their physio-chemical characteristics. Determining and resolving suitable environment for crystallization

requires testing empirically numerous situations before an effective condition is found for the protein crystallization.

C. Methods of protein crystallization

Vapor diffusion method is the simplest and most common method used for protein crystallization. Here the purified protein along with precipitant and buffer are taken in a droplet and are allowed to equilibrate in a large reservoir containing precipitants and buffers in higher concentrations [13]. Primarily, the drops of protein solution comparatively have low precipitant and protein concentrations, but as the drop and reservoir equilibrate, the concentration of protein and precipitant increase in the drop [14]. There will be crystal growth in the drops, provided appropriate and precise crystallization solutions are used for a protein given. [13] [14]. Vapor diffusion is favoured as it lets for steady and gentle variations in protein and precipitant concentrations, this helps in growth of large and well-ordered crystals.

A microbatch generally involves dipping a very little protein volume droplets in oil. Oil is used because of low protein solution volume and in order to continue experimentation aqueously the evaporation must be subdued. For the experimentation several oils can be used, the preferred ones are paraffin oil and silicon oils (described by D'Arcy) which are liquid sealing agents [15]. Few other approaches existing for micro batching do not make use of liquid sealing agent but instead need a hands-on role of an expert to quickly place a tape or a film on a welled plate after the drop in the well is placed.

Microdialysis takes a better place in crystallization process. The major noteworthy part of this method is an advantage of a semi-permeable membrane as small molecules and ions easily pass through, but proteins and large polymers cannot. The system slowly moves toward supersaturation by forming a gradient of concentrations of solutes across the membrane and helps reach the system towards equilibrium, at which point crystals of protein may form. This method yields crystals by considering salts with higher concentration rate or some small membrane-permeable compounds which aid in decreasing the protein solubility. Occasionally it is possible to crystallize few proteins by dialysis salting in, dialyzing against pure water, removing solutes driving self-association and crystallization.

High throughput crystallization screening methods help update the experiments which are needed to figure out the numerous conditions which are required for the growth of crystals. Robots that can handle liquids can be made used to automate and set various crystallization experiments simultaneously [16]. Robotic crystallization systems use the same components that are used manually, but the experiments carry out the procedures quickly with large number of replicates. Every single experiment is under observations by a camera which detects the growth of the crystal [14].

D. Experimental Methods to detect protein structure

X-ray crystallography: It is most appropriate method and is of in Vitro type. Most of the structures in the PDB (Protein Data Bank) library have been determined with this outstanding technique [17]. Purification of protein and then its crystallization is the first key step of this method, followed by which the protein crystal is subjected to the intense X-rays beams [17]. The X-ray beams get diffracted by the protein crystal into characteristic pattern of spots. These spots or pattern produced are examined with the available methods to determine the phase of the X-ray wave in each spot to figure the electrons distribution in the protein. Then the resulting electron density map is considered, studied and interpreted to find the position of each atom. Like every technique X-ray crystallography has both pros and cons.

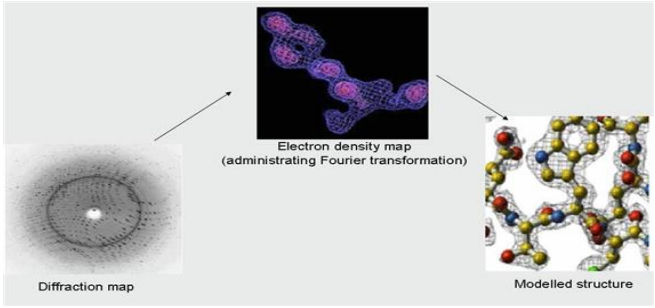

Fig.1. An image of determination of protein structure with X-ray crystallography
Dr. József Tőzsér, Dr. Tamás Emri, Dr. Éva Csősz, Dr. József Tőzsér (2011)

Advantages

- Get whole 3D structure by analysis of good crystallized material
- Single model is produced that is easy to interpret and visualize
- More mathematically direct image construction
- Availability of quality indicators (resolution, R-factor)
- Large molecules can be determined

Disadvantages

- Formation of stable crystals that diffract well
- Crystal formation can be time consuming and difficult
- Inability to examine solutions and the behavior of the molecules in solution
- There is no chance for direct determination of secondary structures
- Unnatural, non-physiological environment

NMR spectroscopy: It is one of the widely used technique to determine protein structure. It is quite accurate and in-vivo method [18] [19]. Here in this technique the protein is purified and is positioned in a strong magnetic field and then radio waves are made use to explore the results. A distinctive set of resonances observed are then studied and analysed to produce a list of atomic nuclei which fall close to each other, and to portray the atoms local conformation that are bonded together [19]. The list is later used to build the model of the protein that shows the position of each atom. Some experiments using, NMR spectroscopy [20] have been used to bring out properties regarding the folding pathway, however, complete knowledge of the folding process remains elusive. As such, computational models of proteins are positioned to give great insight into the physics behind this phenomenon. The technique currently is usually applicable to small and medium sized proteins as larger proteins boost up problems with overlapping peaks in the NMR spectra.

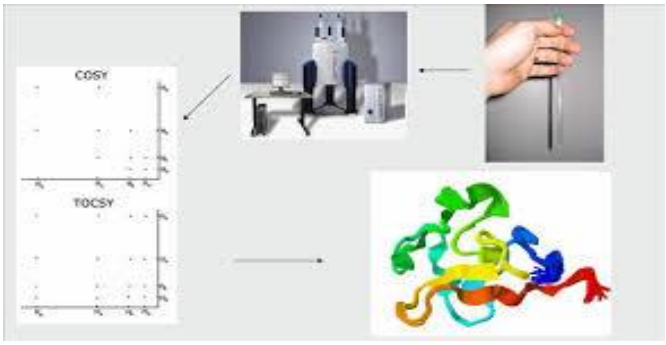

Fig.2. An image of determination of protein structure by NMR spectroscope
Dr. József Tőzsér, Dr. Tamás Emri, Dr. Éva Csősz, Dr. József Tőzsér (2011)

Advantages

- Provides valuable dynamics information
- Identifies precise motion of side-chains
- Help result in secondary structure from restricted experimental data
- Free from relics that is seen usually on crystallization
- Useful for protein-folding studies
- Closer to biological conditions in some respects

Disadvantages

- Danger of aggregation as it needs a concentrated solution
- Difficult for proteins of large size and hence restricted to small and medium protein size
- A weaker interpretation of the experimental data
- Produces an ensemble of possible structures rather than one model

Electron-microscopy: It is an imaging technology used to explore and carry work related to moderately larger objects, like large macromolecular complexes cellular organelles. The method uses individual particle reconstruction with low resolution and not more observable. To carry out the study the technique does not demand protein crystallization and also it requires less amount of sample material.

Advantages

- Allows the specimen observation that have not been stained or fixed in any way
- Viewing them in their native environment

Disadvantages

- Expensive
- The resolution of cryo-electron microscopy maps is not high enough

III. PROTEIN STRUCTURE PREDICTION

Having seen the pros and cons of structure determination it's clear that structure prediction is preferred because structure determination is quite tedious and difficult. The largest challenges being the cost, time and expertise. It is the structure which allows a protein to do the work it is built to do so. The protein structure has a major role as it provides a greater level of understanding of how it works and gives a way to create hypothesis about how to affect it, control it, or modify it.

Protein structure prediction can be classified into two major categories

- Prediction of secondary structure
- Prediction of tertiary (3-D) structure

A. Prediction of secondary structure

Protein secondary structure when predicted accurately helps in structure alignment [21] and gives the protein function information without the knowledge of its 3-D structure. Secondary structures are the major input for prediction algorithms of tertiary structures. Since the early 1970, many predictive algorithms have been developed and also have advanced significantly based on the knowledge of amino acid residue and its conformation as observed in the proteins crystal. Prediction of secondary structure initiated from statistics of single residues.

The Chou-Fasman (CF) [22] and the Garnier-Osguthorpe-Robson methods (GOR) [23] [24] have been widely considered for the protein secondary structure prediction. The Chou Fasman is an empirical technique [25] and is based on study of the relative frequencies of every single amino acid in secondary structures like helices, sheets and turns based on already prevailing structures documented by X ray crystallography, further these detected frequencies ,the set of probability parameters are inferred for the appearance of each amino acid in its secondary structure type, and then used to predict the probability of how the given amino acid sequence would form helix, a beta sheet, a beta strand or a turn in the protein [26]. The Garnier-Osguthorpe-Robson methods(GOR) is based on information theory method [27] in which the sequence of amino acid are observed ,studied and analysed to predict secondary structures like α helix, β sheet, turn or random coil at each position based on a window size which is of 17-amino acid residues per window .This method works based on probability parameters like Chou Fasman method but also considers conditional probability of the amino acid takes and the propensities of every amino acid to form secondary structures, but conditional probability of the amino acid to form a secondary structure provides that its direct neighbours have already formed that structure [28]. Later the statistics of residues blocks are presented to do the predictions [29]. These methods are based on Bayesian model [30] for predicting secondary structure where it considers the packing influence of residues on the structure determination, including those packed close in space but distant in sequence [31].

Machine learning methods play a very important role in the field of protein tertiary structure prediction. To name a few neural networks, nearest-neighbour techniques and hidden Markov models etc have been developed and are playing a remarkable role in this field of prediction in the recent years [32]. Many algorithms and methods tackle the problem of secondary structure prediction of proteins. Secondary structure prediction algorithms work well with the concepts based on neural network [33]. Neural network is a machine learning process. It is built with many intermediate layers which has interconnected nodes. In the prediction of secondary structure, the input is the sequence of amino acid and the output is the residue probability to adopt a specific structure. There are many hidden layers which are inter connected between the input and output layers, here the machine learning concept picturizes to adjust the weights of internal connections mathematically. The training of neural network is the first step. This is done by those sequences whose structures are already known, this helps to recognize the patterns of amino acids and their relationships with known structures.

Training being the crucial step, demands the optimization of weight functions to relate input to output appropriately. The satisfactorily trained network processes an unknown sequence; it puts on the rules learnt during training to recognize particular patterns of structure [33]. Like neural network approach, Hidden Markov model also made a remarkable entry in the secondary structure prediction. It is the statistical model composed of number of interconnected Markov chains with the capability to generate the probability value of an event by taking into account the influence from hidden variables. Mathematically, it calculates probability values of connected states among the Markov chains to find an optimal path within the network of states. It requires the training to obtain the probability values of state transitions. Later position scoring matrices were used to improve the results in secondary protein structure prediction [34]. The work of predicting the secondary structure with neural net and statistical methods had a great impact in the field [35]. This approach of neural networks and related algorithms gave better accuracy in the predicted results [36]. The nearest neighbour algorithms served the extraordinary remark in the field as well [37]. Algorithm based on local sequence homologies and sequence similarities supported very well in secondary structure prediction [38-39]. Nearest neighbour algorithms in combination with local multiple sequence alignment [40] [42] and local alignment respectively [41-42] did help in secondary prediction of protein structures [43]. Also, Multiple Sequence Alignment (MSA) [44] which is a sequence alignment of three or more biological sequences like proteins is one of the preferred methods for secondary structure prediction. The query sequences which form the input are expected to have evolutionary relationship. This concept shows that the query set share a connection and is descended from common ancestor. This piece of information is frequently used to measure sequence conservations of protein domains, secondary and tertiary structures. MSA generally denotes to the process of aligning such a sequence set, to do so many computational algorithms are used to produce and analyse the alignments [45]. Fragment Data base Mining (FDM) [46] laid a foundation with an advantage of prediction accuracy. Here the structural fragment database is being mined and quarried [47], and then uses the information of structures from the matching sequences fragments for the prediction of protein structure, however if fragments are not available the performance drops. The study of MSA and FDM indeed has put forth an advantage of prediction accuracy [48]. Secondary structure prediction of proteins supports other prediction problems like helps in finding out remote homologs. In fold recognition and structural clustering, secondary structure prediction is very important and most essential step.

B. Prediction of tertiary (3-D) structure

Prediction of tertiary, aims to predict the inborn or the 3-D structure of a protein [49 -50]. Physical methods being slow and expensive undoubtedly call for use of computational approaches [49]. Numerous approaches have been advanced in the attainment of predicting the tertiary structure [51-52]. All the structure prediction methods basically convey that there is a correlation between structure and its residue sequence [49]. Packing the secondary structure elements of the protein to form distinct domains or independent folding units is what yields the tertiary structure of a protein [53]. There is a huge gap present in the number of amino acid sequences to the number of known protein structures in the Protein Data Bank (PDB) [54] and this difference is continuously and quickly increasing. The basic steps in the traditional structure prediction are of the following order:

- Finding and understanding of template structure(s)
- There are two steps in generation of structure(s)-first the core structure comprising the secondary structure elements are generated using secondary structure prediction techniques and methods, followed by the non-conserved loops. Non-conserved loop structure prediction is difficult and is done with the help of loop structure prediction algorithms like Phyre Server, FREAD, JPred. (Although the secondary

structure elements are generated by the side chain conformations are not accurate)

- Side chain structure libraries like dynamic rotamer libraries and BetaSCPWeb are used to predict amino acid side chain conformations. The prediction of an overall model for the protein sequence of interest is completed with this step.
- To enhance the prediction accuracy of the model, the predicted models are further refined with the energy minimization algorithms like steepest decent, genetic algorithm, Monte Carlo, simulated annealing. Accuracy of a predicted model is measured in terms of RMSD between the a-carbon positions in the predicted and the real structure of the target sequence. Less than 1.0Å RMSD represent very good predictions.

Note: A target three-dimensional protein structure can be built from related known protein structures

called templates, if it shares statistically meaningful sequence similarity.

Protein tertiary structure prediction can be classified into two major categories [55]

Template Method (Knowledge based)

- Homology Modeling (Comparative modeling)
- Threading (Fold Recognition)
- Fragment based approach

Template-Free Method

- *Ab initio* Methods

C. Template Method (Knowledge based)

Homology Modelling Process (Comparative Modeling) is the simplest and most reliable. Proteins usually tend to fold into similar structures, when they have similar sequences [56] [57]. It involves building a 3-dimensional protein model for an unknown structure based on the sequence similarity to templates of the protein structure known. The basic principles of comparative modelling are built on the concept that a minor change in the sequence of amino acid usually will only give rise to a small change in the final overall structure [58-59]. The comparative modelling is one of the choice when >30% sequence identity exists [51]. One of the major problems is optimal template selection and alignment. Several templates or fragment recombination of proteins (consensus strategies) are used to build final protein models and is advantageous due to increase in chance of optimal template selection. The excellence and practicality of comparative models is directly proportional to the evolutionary distance between template and target. Main factors that influence homology modeling are correctness of alignment, to the extent structure is conserved between target and template and refining of models. Since the fold number is limited in nature and as the numbers of evolutionary related structures are made available, the prediction issues can be simplified [43]. The major steps for comparative modeling include fold assignment, alignment of template and target, model building and error corrections [60]. Fold assignment involves recognizing the resemblance between target and at least one known template structure. Later which the alignment of a target sequence and template is carried out followed by this the model is built based on alignment with template chosen. Then predicting errors in model built is the final step in the process of prediction. The magnitude of errors in the above steps can be decreased by improvising the algorithm, techniques considered for the same and by sampling a large number of sequences and structures of known proteins. SWISS- MODEL [61] Modeller,3D-JIGSAW [62] SCWR [63] are the prevalent tools and web servers meant for comparative modeling.

Advantages

- Finds the location of alpha carbon s of important residues inside the protein fold
- Helps to guide mutagenesis experiment
- Hypothesize structure function relation

Disadvantages

- Difficulty in modeling proteins with lower similarity (e.g. < 30% sequence identity)
- Model accuracy is a prime issue
- Need in optimizing the techniques of side chain modeling and loop modeling

- Need of improved optimizers and potential function

Threading (Fold Recognition) attempt to detect the fold that is well matched with a precise sequence of a query It is basically threading a specific sequence through all known folds and for each fold estimate the probability that the sequence can have that fold [64] [65]. The method takes an advantage of the extra material made available by 3-D structure. Instead of finding out how a sequence will fold, the method figures out and predict how well a fold will fit a sequence. The method improvises the protein folding problem effectively. Here, a protein configuration is matched to a library of known structures for highest compatibility. The amino acid sequence therefore has to be folded to the structure it is matched to and the compatibility of this match has to be evaluated. The compatibility function in this approach is typically an energy function [66] [67]. The fold recognition still needs to be made better in recognizing distant related sequence-structure pairs and sequence-structure alignment algorithm [68]. As of SCOP release 1.75 and CATH release v 3.5, identified numbers of folds are 1195 and 1313 respectively. The present multiple threading approach [69] uses probabilistic multiple threading algorithm for a target with multiple templates, Better alignments between target and template is the key and multiple template method can generate better models than single template method. The progress of effective threading algorithms to perceive distant structure templates has been a fundamental theme in this field [70]. With the limited numbers of folds, one can generate atomic level models and this makes the study of fold recognition even more interesting.

Advantages

- Used when no suitable template structure can be found for homology-based modeling
- Accuracy better than comparative modeling

Disadvantages

- Threading methods seldom lead to the alignment quality that is needed for homology modeling.
- Less than 30% of the predicted first hits are true remote homologs (Predict Protein).

Fragment based approach to predict protein structure was first proposed by Bowie and Eisenberg in 1994 and is considered to be the most successful method to predict the tertiary structure of proteins [71]. The novel structures of protein were built by assembling fragments of short length obtained from known protein structures. The very basic need of this approach is the extent of existence of fragments that are similar in structure in the database of known structures for short fragments of a novel protein. Most fragment recognition methods existing rely on database-driven search stratagies to identify a concerned-candidate fragment, which are laborious and often hinders the possibility to trace longer fragments due to the limited databases size. It is difficult to alleviate the effect of noisy sequence based predicted features such as secondary structures on the quality of fragment. The widely used method to build the novel structures of protein is to assemble short fragments from the available structures. [72]. The fragment-based approach focuses mainly on fragment assembly methods used for protein structure prediction by using ROSETTA as a reference [73]. The concept of the fragment assembly strategy is that a local sequence (fragment) has a high probability for one or few specific local structures and that the complete structure depends on the assembly of the most likely resident(local)structures and non-local interactions between them. ROSETTA, implements the Bayes statistical theorem to predict the structure from the knowledge of the structure of short fragments. Further, ROSETTA is the most accurate fragment-based method which does extensive all atom refinement and is also used for homology modeling target as template refinement. Application "mix-and-match" methods or "Fragment assembly" methods are found to produce exceptionally good results in homology modeling which is based on templates, as well as in "*de novo*" folding which is template free in predicting the protein structure. The predicted models generated by the method of recombination of fragments extracted from known structures of protein are closer to the native target

protein structure. Fragment recombination method is considered to be the most successful in cases that protein exhibiting novel folds [74]. The approach of predicting the protein structure through the assembly of fragments is one of the finest ways.

Advantages

- Enhances the global optimization method that finds low-energy conformations
- Better choice due to non-availability of homologous structures
- Leads to more efficient and accurate models
- Reduces computational demand

Disadvantages

- Performance seems to be poor if the length of a target found greater than 100 residues,
- Huge run time
- Energy functions found to be Sub-optimal
- The major bottlenecks are the conformational sampling

D. *Template-Free Method (free modeling)*

Abinito method proceeds in predicting the structure entirely from scratch. The basic principle employed here is that the native protein structure is at the global free energy minimum [75] which is simulated by actual physical forces and potential of chemical interactions thereby large conformational space is reduced to only decoy fragments that obey the minimum free energy. Huge conformational space search for structures of proteins mainly the one that are particularly low in free energy for the sequence of amino acid are carried out. Here there is no use of information regarding alignments of sequence and no direct use of known structures [76-77]. The objective is to build empirical function that simulates the real physical forces and potentials of chemical contacts [78-79]. This approach relies much on the validity of Anfinsen's hypothesis [80] which states that the structure the protein forms in nature (the native structure) is the global minimum of the free energy and is determined only by amino acids sequence. As a consequence of this hypothesis, given an appropriate protein model with a free energy associated with each structure(conformation), global minimization of the free energy will yield the correct native state [81]. However, the free energy, which consists of potential energy and entropy, poses a complex modeling problem, therefore it has become common practice to model only the potential energy surface and successively correct for the entropy term [82]. *Ab initio* is a combination of knowledge based and physics-based approach to predict protein structures [83]. Many successful programs and servers for *ab initio* modeling are QUARK [84], I TASSER [85], ROBETTA server [86], ROSETTA@home [87], Bhageerath [88]. In ROSETTA method the native like conformations are yielded by assembling the fragments of short length of known proteins by Monte-Carlo strategy. For fragment insertion, a consecutive window of 3 or 9 residues is selected and torsion angles obtained from a fragment of known structure. A 3 or 9 residue window in the query are searched against all windows in a non-redundant database of protein structures composed of x-ray structures of <2.5 angstrom and <50% sequence identity. Fragments are selected using secondary structure prediction methods. Overall, a final fragment list for a query sequence is composed for every overlapping insertion window in the query. Fragment assembly occurs by Monte-Carlo search, followed by potential energy minimization which uses Monte-Carlo minimization and side chain optimization.

Advantages

- Very large search space
- Large success rate
- Fully automated

Disadvantages

- Enormous amount of computation
- Excessive running times

IV. IMPORTANCE OF MACHINE LEARNING IN STRUCTURE PREDICTION

Hidden Markov Models (HMM), neural networks and support vector machines are the three major supervised machine learning methods [89] and are unsupervised clustering methods employed for solving one, two, three and also the four-dimensional

structure prediction problems. Tertiary and quaternary protein structure prediction from primary sequence and many other problems related to structure prediction are intelligently handled by machine learning methods and certainly play a very important role in the arena of prediction of protein structures. The improvement in machine learning methods will find its importance in *ab initio* structure prediction problem in the foreseeable future [89]. The usefulness of the computationally predicted protein structures is strongly seen in biological research mainly in biomedicine which depends on accuracy of the prediction, in other words state-of-the-art algorithms [90].

V. CRITICAL ASSESSMENT OF STRUCTURE PREDICTION (CASP)

The main objective of Critical Assessment of Structure Prediction (CASP) is to obtain in depth knowledge and assessment of our present abilities in the area of structure prediction. It is community wide experiment. It is a biennial competition for which groups are challenged to predict the structure of protein provided by only the sequence of amino acid [91]. The obtained structures by simulation are compared to experimental results and are scored depending on the degree of structural agreement. CASP 10 which was held in Gaeta, Italy on December 9-12 (2012) has been ranked the best tools developed around the world for protein structure prediction. Predicting tertiary structure of protein from available sequence is one of the most substantial and still an open problem in the field of molecular biology. Measures taken up by the CASP and recent advancement in techniques and methodologies in predicting the protein structures with the aid of computational power will be of sure help to reduce the problems and challenges in structure prediction

VI. CONCLUSIONS

Although determination of protein structure by experimental methods is the better way to determine protein structures, computational methods have been gaining importance for structure prediction for reasons that have been discussed in this review. Use of computational approaches for structure prediction has its own benefit, it has to be supplemented with determination of experimental techniques. Currently, there are no computational methods that have replaced the experimental determination of proteins. The ultimate goal of computational methods is to obtain the function of proteins by structure determination. Most of the above-mentioned methods tackle this problem from different angles and starting points, so that the determination of the structure is a co-operative approach, then competition between them. Computational methods although are fast, economical and easy to predict protein structures still has limitations of not being able to predict all structures using a single computational approach. Hence, opportunities are abundant to fulfil the goal of protein structure and function prediction. Of lately, Artificial Intelligence can be used as a new approach to predict the protein structures. The rapid changes occurring in computational technology will one day be able to address this problem.

VII. ACKNOWLEDGEMENTS

The authors would like to acknowledge the management and staff of PES University. The authors are thankful to Mr. Subash Reddy, the librarian of PES University for supportive service to prepare the manuscript.

VIII. REFERENCES

[1] Lesk, A. M. (2001). *Introduction to protein architecture: the structural biology of proteins*. Oxford: Oxford University Press.

[2] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). The shape and structure of proteins.

[3] McPherson, A. (2004). Introduction to protein crystallization. *Methods*, *34*(3), 254-265.

[4] Brünger, A. T. (1997). X-ray crystallography and NMR reveal complementary views of structure and dynamics. *Nature structural biology*, *4*, 862-865.

[5] Levinthal, C. (1968). Are there pathways for protein folding? *Journal de chimie physique*, *65*, 44-45.

[6] Zwanzig, R., Szabo, A., & Bagchi, B. (1992). Levinthal's paradox. *Proceedings of the National Academy of Sciences*, *89*(1), 20-22.

[7] Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, *181*(4096), 223-230.

[8] Dill, K. A., & Chan, H. S. (1997). From Levinthal to pathways to funnels. *Nature Structural and Molecular Biology*, *4*(1), 10.

[9] a. https://www.rcsb.org/stats/growth/overall
b. https://www.uniprot.org/statistics/TrEMBL9

[10] Ramyachitra, D., & Veeralakshmi, V. (2014). Computational Analysis of Protein Structure Prediction and Folding. *Int. J. Comput. Sci. Inform. Technol. Secure*, *4*, 116-127.

[11] Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., & Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, *181*(4610), 662-666.

[12] Doye, J. P., & Poon, W. C. (2006). Protein crystallization in vivo. *Current opinion in colloid & interface science*, *11*(1), 40-46.

[13] Rhodes, G. (2010). *Crystallography made crystal clear: a guide for users of macromolecular models*. Elsevier.

[14] The Crystal Robot, *December 2000. Retrieved 2003-02-18*.

[15] Chayen, N. E., Shaw Stewart, P. D., Maeder, D. L., & Blow, D. M. (1990). An automated system for micro-batch protein crystallization and screening. *Journal of applied crystallography*, *23*(4), 297-302.

[16] Lin, Y. (2018). What's happened over the last five years with high-throughput protein crystallization screening?

[17] Drenth, J. (2007). *Principles of protein X-ray crystallography*. Springer Science & Business Media.

[18] Wüthrich, K. (2003). NMR studies of structure and function of biological macromolecules (Nobel Lecture). *Angewandte Chemie International Edition*, *42*(29), 3340-3363.

[19] Udgaonkar, J. B., & Baldwin, R. L. (1988). NMR evidence for an early framework intermediate on the folding pathway of ribonuclease A. *Nature*, *335*(6192), 694.

[20] Wales, D. (2003). *Energy landscapes: Applications to clusters, biomolecules and glasses*. Cambridge University Press.

[21] Krissinel, E., & Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographic a Section D: Biological Crystallography*, *60*(12), 2256-2268.

[22] Chou, P. Y., & Fasman, G. D. (1974). Prediction of protein conformation. *Biochemistry*, *13*(2), 222-245.

[23] Garnier, J., Osguthorpe, D. J., & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of molecular biology*, *120*(1), 97-120.

[24] Garnier, J., & Robson, B. (1989). The GOR method for predicting secondary structures in proteins. In *Prediction of protein structure and the principles of protein conformation* (pp. 417-465). Springer, Boston, MA.

[25] Chou, P. Y., & Fasman, G. D. (1978). Empirical predictions of protein conformation. *Annual review of biochemistry*, *47*(1), 251-276.

[26] Kabsch, W., & Sander, C. (1983). How good are predictions of protein secondary structure? *FEBS letters*, *155*(2), 179-182.

[27] Garnier, J., Gibrat, J. F., & Robson, B. (1996), GOR method for predicting protein secondary structure from amino acid sequence, m*ethods in enzymology* (Vol. 266, pp. 540-553), Academic Press

[28] Mount, D. W. Bioinformatics: sequence and genome analysis. 2004. *Bioinformatics: Sequence and Genome Analysis*

[29] Gibrat, J. F., Garnier, J., & Robson, B. (1987).

[30] Garnier, J., Osguthorpe, D. J., & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of molecular biology*, *120*(1), 97-120.

[31] Li, Q., Dahl, D. B., Vannucci, M., Joo, H., & Tsai, J. W. (2014). Bayesian model of protein primary sequence for secondary structure prediction. *PloS one*, *9*(10), e109832*Journal of molecular biology*, *198*(3), 425-443.

[32] Frishman, D., & Argos, P. (1997). Seventy-five percent accuracy in protein secondary structure prediction. *Proteins-Structure Function and Genetics*, *27*(3), 329-335.

[33] Holley, L. H., & Karplus, M. (1989). Protein secondary structure prediction with a neural network. *Proceedings of the National Academy of Sciences*, *86*(1), 152-156.

[34] Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices1. *Journal of molecular biology*, *292*(2), 195-202.

[35] Stolorz, P., Lapedes, A., & Xia, Y. (1992). Predicting protein secondary structure using neural net and statistical methods. *Journal of Molecular Biology*, *225*(2), 363-377.

[36] Rost, B., & Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins: Structure, Function, and Bioinformatics*, *19*(1), 55-72.

[37] Biou, V., Gibrat, J. F., Levin, J. M., Robson, B., & Garnier, J. (1988). Secondary structure prediction: combination of three different methods. *Protein Engineering, Design and Selection*, *2*(3), 185-191.

[38] Levin, J. M., & Garnier, J. (1988). Improvements in a secondary structure prediction method based on a search for local sequence homologies and its use as a model building tool. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, *955*(3), 283-295.

[39] Levin, J. M., Robson, B., & Garnier, J. (1986). An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS letters*, *205*(2), 303-308.

[40] Salamov, A. A., & Solovyev, V. V. (1995). Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments.

[41] Salamov, A. A., & Solovyev, V. V. (1997). Protein secondary structure prediction using local alignments1. *Journal of molecular biology*, *268*(1), 31-36.

[42] Salzberg, S., & Cost, S. (1992). Predicting protein secondary structure with a nearest-neighbor algorithm. *Journal of molecular biology*, *227*(2), 371-374.

[43] Yi, T. M., & Lander, E. S. (1993). Protein secondary structure prediction using nearest-neighbor methods. *Journal of molecular biology*, *232*(4), 1117-1129.

[44] MSA: Wang, L., & Jiang, T. (1994). On the complexity of multiple sequence alignment. *Journal of computational biology*, *1*(4), 337-348.

[45] Just, W. (2001). Computational complexity of multiple sequence alignment with SP-score. *Journal of computational biology*, *8*(6), 615-623.

[46] Cheng, H., Sen, T. Z., Jernigan, R. L., & Kloczkowski, A. (2007). Consensus data mining (CDM) protein secondary structure prediction server: combining GOR v and fragment database mining (FDM). *Bioinformatics*, *23*(19), 2628-2630.

[47] Cheng, H., Sen, T. Z., Kloczkowski, A., Margaritis, D., & Jernigan, R. L. (2005). Prediction of protein secondary structure by mining structural fragment database. *Polymer*, *46*(12), 4314-4321

[48] Cheng, H., Sen, T. Z., Jernigan, R. L., & Kloczkowski, A. (2009). Data Mining for Protein Secondary Structure Prediction. In *Data Mining in Crystallography* (pp. 135-167). Springer, Berlin, Heidelberg

[49] Sobha, K., Kanakaraju, C., & Yadav, K. S. K. (2008). Is protein structure prediction still an enigma? *African Journal of Biotechnology*, *7*(25).

[50] Zhang, Y. (2009). Protein structure prediction: when is it useful? *Current opinion in structural biology*, *19*(2), 145-155.

[51] Ginalski, K. (2006). Comparative modeling for protein structure prediction. *Current opinion in structural biology*, *16*(2), 172-177.

[52] Schwede, T., Kopp, J., Guex, N., & Peitsch, M. C. (2003). SWISS-MODEL: an automated protein homology-modeling server. *Nucleic acids research*, *31*(13), 3381-3385.

[53] Tusnady, G. E., & Simon, I. (1998). Principles governing amino acid composition of integral membrane proteins: application to topology prediction1. *Journal of molecular biology*, *283*(2), 489-506

[54] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. & Bourne, P. E. (2000). The Protein Data Bank Nucleic Acids R Toomulaesearch, 28, 235-242.

[55] Sliwoski, G., Kothiwale, S., Meiler, J., & Lowe, E. W. (2014). Computational methods in drug discovery. *Pharmacological reviews*, *66*(1), 334-395.

[56] Sanchez, R., & Šali, A. (1997). Evaluation of comparative protein structure modeling by MODELLER. *Proteins: Structure, Function, and Bioinformatics*, *29*(S1), 50-58.

[57] Guex, N., & Peitsch, M. C. (1997). Swiss model and the Swiss Pdb Viewer: an environment for comparative protein modeling. *electrophoresis*, *18*(15), 2714-2723.

[58] Wüthrich, K. (2003). NMR studies of structure and function of biological macromolecules (Nobel Lecture). *Angewandte Chemie International Edition*, *42*(29), 3340-3363.

[59] Lougher, M., Lücken, M., Machon, T., Malcomson, M., & Marsden, A. Computational modelling of protein folding

[60] Eswar, N., Webb, B., Marti Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. Y., & Sali, A. (2006). Comparative protein structure

modeling using Modeller. *Current protocols in bioinformatics*, *15*(1), 5-6.

[61] Arnold, K., Bordoli, L., Kopp, J., & Schwede, T. (2006). The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, *22*(2), 195-201.

[62] Bates, P. A., Kelley, L. A., MacCallum, R. M., & Sternberg, M. J. (2001). Enhancement of protein modeling by human intervention in applying the automatic programs 3DJIGSAW and 3DPSSM. *Proteins: Structure, Function, and Bioinformatics*, *45*(S5), 39-46.

[63] Krivov, G. G., Shapovalov, M. V., & Dunbrack, R. L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. *Proteins: Structure, Function, and Bioinformatics*, *77*(4), 778-795.

[64] Skolnick, J., & Kihara, D. (2001). Defrosting the frozen approximation: PROSPECTOR—a new approach to threading. *Proteins: Structure, Function, and Bioinformatics*, *42*(3), 319-331.

[65] Panchenko, A. R., Marchler-Bauer, A., & Bryant, S. H. (2000). Combination of threading potentials and sequence profiles improves fold recognition1. *Journal of Molecular Biology*, *296*(5), 1319-1331.

[66] Fischer, D. A. N. I. E. L., Rice, D. A. N. N. Y., Bowie, J. U., & Eisenberg, D. A. V. I. D. (1996). Assigning amino acid sequences to 3-dimensional protein folds. *The FASEB journal*, *10*(1), 126-136.

[67] Lewis, P. N., & Scheraga, H. A. (1971). Predictions of structural homologies in cytochrome c proteins. *Archives of biochemistry and biophysics*, *144*(2), 576-583.

[68] Fischer, D. A. N. I. E. L., Rice, D. A. N. N. Y., Bowie, J. U., & Eisenberg, D. A. V. I. D. (1996). Assigning amino acid sequences to 3-dimensional protein folds. *The FASEB journal*, *10*(1), 126-136

[69] Peng, J., & Xu, J. (2011). A multiple template approach to protein threading. *Proteins: Structure, Function, and Bioinformatics*, *79*(6), 1930-1939.

[70] Wu, S., & Zhang, Y. (2009). Protein structure prediction. In *Bioinformatics* (pp. 225-242). Springer, New York, NY

[71] Donald Voet, Judith Voet., Biochemistry, Chapter 8,3rd edition, 219-275,2004

[72] Lee, J., Kim, S. Y., & Lee, J. (2005). Protein structure prediction based on fragment assembly and parameter optimization. *Biophysical chemistry*, *115*(2-3), 209-214.

[73] Schulz, R. (2007). Protein Structure Prediction., *Proteins*

[74] Gajda, M. J., Pawlowski, M., & Bujnicki, J. M. (2011). Protein structure prediction: From recognition of matches with known structures to recombination of fragments. In *Multiscale Approaches to Protein Modeling* (pp. 231-254). Springer, New York, NY.

[75] Simoncini, D., Berenger, F., Shrestha, R., & Zhang, K. Y. (2012). A probabilistic fragment-based protein structure prediction algorithm. *PloS one*, *7*(7), e38799.

[76] Kolinski, A., & Skolnick, J. (1998). Assembly of protein structure from sparse experimental data: an efficient Monte Carlo model. *Proteins: Structure, Function, and Bioinformatics*, *32*(4), 475-494.

[77] Ortiz, A. R., Kolinski, A., & Skolnick, J. (1998). Fold assembly of small proteins using Monte Carlo simulations driven by restraints derived from multiple sequence alignments 1. *Journal of molecular biology*, *277*(2), 419-448.

[78] Simons, K. T., Strauss, C., & Baker, D. (2001). Prospects for ab initio protein structural genomics1. *Journal of molecular biology*, *306*(5), 1191-1199.

[79] Aszodi, A., Gradwell, M. J., & Taylor, W. R. (1995). Global fold determination from a small number of distance restraints. *Journal of molecular biology*, *251*(2), 308-326.

[80] Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, *181*(4096), 223-230.

[81] Floudas, C. A., Fung, H. K., McAllister, S. R., Mönnigmann, M., & Rajgaria, R. (2006). Advances in protein structure prediction and de novo protein design: A review. *Chemical Engineering Science*, *61*(3), 966-988.

[82] Floudas, C. A., Klepeis, J. L., & Pardalos, P. M. (1999). Global optimization approaches in protein folding and peptide docking. *DIMACS series in discrete mathematics and theoretical computer science*, *47*, 141-171.

[83] Zhang, Y. (2008). Progress and challenges in protein structure prediction. *Current opinion in structural biology*, *18*(3), 342-348.

[84] Xu, D., & Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge based force field. *Proteins: Structure, Function, and Bioinformatics*, *80*(7), 1715-1735.

[85] Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein

structure and function prediction. *Nature protocols*, *5*(4), 725

[86] Kim, D. E., Chivian, D., & Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. *Nucleic acids research*, *32*(suppl_2), W526-W531.

[87] Rohl, C. A., Strauss, C. E., Misura, K. M., & Baker, D. (2004). Protein structure prediction using Rosetta. In *Methods in enzymology* (Vol. 383, pp. 66-93). Academic Press.

[88] Jayaram, B., Bhushan, K., Shenoy, S. R., Narang, P., Bose, S., Agrawal, P., & Pandey, V. (2006). Bhageerath: an energy based web enabled computer software suite for limiting the search space of tertiary structures of small globular proteins. *Nucleic acids research*, *34*(21), 6195-6204.

[89] Cheng, J., Tegge, A. N., & Baldi, P. (2008). Machine learning methods for protein structure prediction. *IEEE reviews in biomedical engineering*, *1*, 41-49.

[90] Du, P., Andrec, M., & Levy, R. M. (2003). Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? An update. *Protein Engineering*, *16*(6), 407-414.

[91] Lougher, M., Lücken, M., Machon, T., Malcomson, M., & Marsden, A. Computational modelling of protein folding *38.*