International Journal of Computer Science and Informatics

Volume 4 | Issue 1 Article 4

July 2014

A COMPARATIVE STUDY OF PROTEIN TERTIARY STRUCTURE PREDICTION METHODS

CHANDRAYANI N.ROKDE
Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India-441110, chandrayanirokde@gmail.com

DR.MANALI KSHIRSAGAR
Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India-441110, Manali_kshirsagar@yahoo.com

A COMPARATIVE STUDY OF PROTEIN TERTIARY STRUCTURE PREDICTION METHODS

CHANDRAYANI N.ROKDE[1], DR.MANALI KSHIRSAGAR[2]

[1,2]Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India-441110
E-mail: chandrayanirokde@gmail.com, Manali_kshirsagar@yahoo.com

Abstract- Protein structure prediction (PSP) from amino acid sequence is one of the high focus problems in bioinformatics today. This is due to the fact that the biological function of the protein is determined by its three dimensional structure. The understanding of protein structures is vital to determine the function of a protein and its interaction with DNA, RNA and enzyme. Thus, protein structure is a fundamental area of computational biology. Its importance is intensed by large amounts of sequence data coming from PDB (Protein Data Bank) and the fact that experimentally methods such as X-ray crystallography or Nuclear Magnetic Resonance (NMR)which are used to determining protein structures remains very expensive and time consuming. In this paper, different types of protein structures and methods for its prediction are described.

Keywords*- protein structure, protein threading, Computational methods for Protein structure prediction.*

I. INTRODUCTION TO PROTEINS

Proteins are essential to biological processes. They are responsible for catalysing and regulating biochemical reactions, transporting molecules, and they form the basis of structures such as skin, hair, and tendon. The shape of protein is specified by its amino acid sequence. This protein sequence comprises a translation of the four-letter DNA alphabet into a 20-letter alphabet of native amino acids. Proteins differ in length (from 30 to over 30 000 amino acids). Protein function can be understood in terms of its structure. Indeed, the three-dimensional structure of a protein is closely related to its biological function. In general, a protein consists of a linear chain of a particular sequence of the 20 naturally occurring amino acids. Some of the functions of proteins include enzymes that catalyze biochemical reactions, structural or mechanical functions, maintaining cell shape, cell signaling, immune responses, cell adhesion, and regulating the cell. A protein does not exhibit a full biological activity until it folds into a three-dimensional structure. Information on the secondary and three dimensional(3D) structures of a protein is important for understanding its biological activity, because the shape and nature of the protein molecule surface account for the mechanisms of protein functions.

II. INTRODUCTION TO PROTEIN STRUCTURE

Protein formation passes through different levels of structure.[1] The primary structure of a protein is simply the linear arrangement, or sequence, of the amino acid residues that compose it. Secondary protein structure occurs when sequence of amino acid are linked by hydrogen bonds. The prediction consists of assigning regions of the amino acid sequence as likely alpha helices, beta strands. The main goal in prediction of secondary structure is to take primary structure (sequence)of protein.

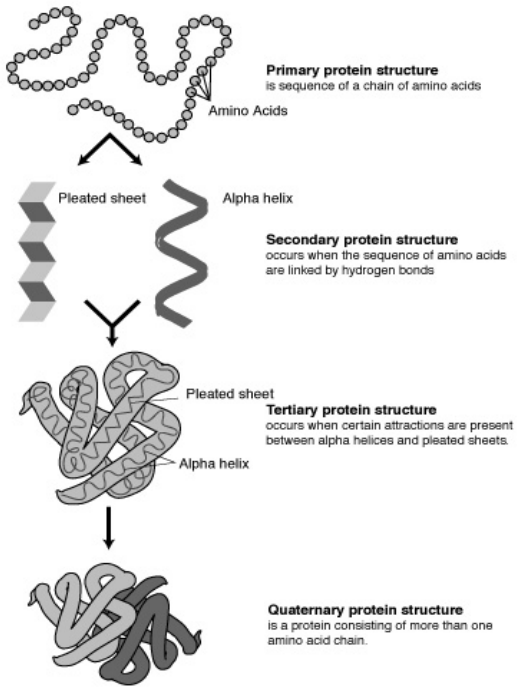

Fig1.Four level protein structure

Tertiary structure refers to the overall conformation of a polypeptide chain that is, the three-dimensional arrangement of all its amino acid residues. In contrast with secondary structures, which are stabilized by hydrogen bonds, tertiary structure is primarily stabilized by hydrophobic interactions between the non polar side chains, hydrogen bonds between polar side chains, and peptide bonds. These stabilizing forces hold elements of secondary structure, helices, strands, turns, and random coils compactly together.

Because the stabilizing interactions are weak, however, the tertiary structure of a protein is not rigidly fixed but undergoes continual and minute fluctuation. This variation in structure has important consequences in the function and regulation of proteins. The final level of protein structure is quaternary structure, which refers to when more than one protein come together to form a complex.[2].The four level protein structure is shown below.

III. PROTEIN TERTIARY STRUCTURE PREDICTION METHODS

Protein structure prediction is the prediction of the three-dimensional structure of a protein from its amino acid sequence thus all activities of proteins are depends upon its three dimensional structure. Structure prediction is fundamentally different from the inverse problem of protein design. The three-dimensional structure of a protein is determined by the network of covalent and non-covalent interactions.[3]Although protein is constructed by the polymerization of only 20 different amino acids into linear chains, proteins carry out an incredible array of diverse tasks. A protein chain folds into a unique shape that is stabilized by noncovalent interactions between regions in the linear sequence of amino acids. This spatial organization of a protein its shape in three dimensions is a key to understanding its function. Only when a protein is in its correct three-dimensional structure, or conformation, is it able to function efficiently.[4] A key concept in understanding how proteins work is that function is derived from three-dimensional structure, and three-dimensional structure is specified by amino acid sequence.

There are three main strategies for solving the PSP(Protein structure prediction) problem: homology (comparative) techniques, protein threading (fold recognition), and Ab initio (de novo) techniques. Homology modeling is a knowledge-based approach, given a sequence database, use multiple sequence alignment on this database to identify structurally conserved regions and construct structure backbone and loops based on these regions, restore side-chains and refine through energy minimization. Protein threading prediction can used for protein structure prediction when

1) The target protein does not share a high sequence similarity with any protein in PDB (protein data bank)

2) The target protein shares a similar structure with some proteins in PDB.Homology modeling predicts the structure for a target by identifying some homologous protein from PDB.Two homologous proteins usually shares similar sequences and similar structures. Therefore homology modeling detects whether two proteins are homologous by aligning their sequences. Compared to homology modeling, which only consider sequence similarity between target and template, protein threading makes use of structural information encoded in template to improve prediction accuracy including the use of secondary structure, solvent accessibility and pair wise interaction. In order to generate good sequence template alignment homology modeling usually requires that the target and template shares at least 25% sequence identification. Protein threading can get beyond this limitation and sometimes can align the target and template very well even when their sequence identify is well below 25%.Ab initio folding predicts the structure for a target without using any complete protein structure in PDB as template. Following table shows the difference between three computational methods.

Homology Modeling	Protein Threading	Ab Initio Method
Carried out when sequence similarity with structure is greater than 35%.	Carried out when sequence similarity with structure is Greater than 25%.	Carried Out When no suitable structure templates can be found.
Homology modeling is for those targets which have homologous proteins with known structure	Protein threading is for those targets with only fold-level homology found	The goal of Ab initio protein structure prediction is to predict a protein's structure accurately by focusing on the chemical and physical properties of the amino acid sequence making up the mature protein
Homology modeling is for "easier" targets. Accuracy :60%	Protein threading is for "harder" targets. Accuracy:40%	This method is too slow and inaccurate and used

		for novel targets.

Table 1 Differences between three methods

Every two years, the performance of current methods is assessed in the CASP experiment stands for Critical Assessment of Techniques for Protein Structure Prediction.

IV. PROTEIN THREADING (FOLD RECOGNITION)

Proteins fold due to hydrophobic effect, Vander Waals interactions, electrostatic forces, and Hydrogen bonding. Protein threading, also known as fold recognition, is a method of protein modelling (i.e. computational protein structure prediction) which is used to model those proteins which have the same fold as proteins of known structures, but do not have homologous proteins with known structure. Threading is similar to homology modeling. But, instead of finding similar sequences to deduce the native conformation of the target protein, threading assumes that the target structure is similar to another existing structure, [5][6]which should be searched for. Protein threading consists of five main components:

1) A library of template structures
2) Representation of templates and targets
3)Objective functioning measuring quality of Sequence template alignment
4) An algorithm finding best sequence-template alignment
5) One method selecting best template based on all the sequence-template alignment. A library of template structure is set of representative structures selected from the PDB, in order to save computing time, among all the highly similar protein sequences, only one is kept in the template library .To construct a library of template structures, we can cluster all the proteins in the PDB into several thousand groups and then choose one representative from each group as a structural template. Threading improves the sequence alignment sensitivity by introducing structural information into the alignment, where the structural information refers to the secondary or tertiary structural features of proteins.

V. FOLD RECOGNITION

PROTEIN folding is the process by which a protein assumes its 3D structure. All protein molecules are endowed with a primary structure consisting of the polypeptide chain[7]. Fold recognition requires a criterion to identify the best template for one target sequence.

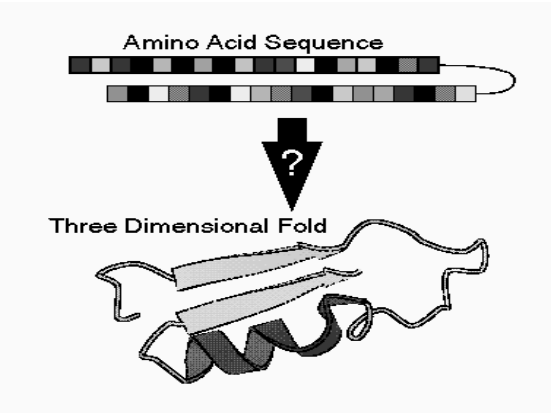

Figure 2. Protein folding

To build the three-dimensional structure of a target sequence, a high quality sequence-template alignment is indispensable. A high-quality sequence-template alignment cannot be obtained easily unless the structures of both proteins are available. The sequence-template alignment score cannot be directly used to rank the templates due to the bias introduced by the residue composition and the number of alternative sequence-template alignments. So far, there are two strategies used by the The protein fold-recognition approach to structure prediction aims to identify the known structural framework (i.e. the backbone of an experimentally determined protein structure) that accommodates the target protein sequence in the best way. Typically, a fold-recognition program comprises four components:

(1) The representation of the template structures (usually corresponding to proteins from the Protein Data Bank database),
(2) The evaluation of the compatibility between the target sequence and a template fold,
(3) The algorithm to compute the optimal alignment between the target sequence and the template structure, and (4) the way the ranking is computed and the statistical significance is estimated [8].

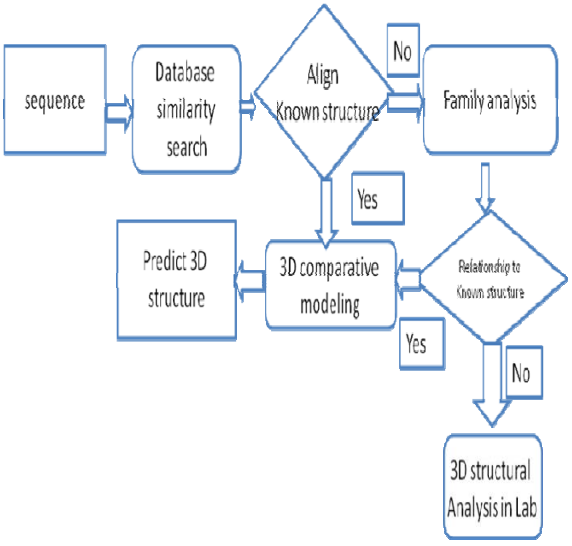

Figure.3 Flow chart showing Procedure for predicting a protein structure from its amino acid sequence.

VI. AB INITIO METHOD

The goal of Ab initio protein structure prediction is to predict a protein's structure accurately by focusing on the chemical and physical properties of the amino acid sequence making up the mature protein. These methods rely on the fact that the folded protein is in a state of lowest free energy and they attempt to compute that lowest energy information based on possible interactions between the residues comprising the sequence. This is a very compute-intensive problem and the assumptions that are made preclude accurate prediction.

VII. PROTEIN SEQUENCE DATABASE

The large amount of protein sequence information and experimentally determined structure information and structural classification information is stored in publically available databases such as Universal protein resource(UniProt) is most comprehensive warehouse containing information about protein sequence and National centre for Biotechnology and information (NCBI) also provide non-redundant database for protein sequences using sequences from wide variety of sources.

VIII. SOFTWARES USED IN PSP

As of today, hundreds of servers and tools are widely available for protein structure prediction. For protein threading, methods such as FASTA and Basic Local Allignment Search Tool (BLAST) were developed to perform rapid searches for sequence homologus in large sequence database.These methods produce relatively accurate approximate sequence alignment by quickly finding sub-sequnces in the databases. Most of the methods are based on the dynamic algorithm, but the key difference is the scoring strategy, in most threading algorithms the score functions include the structure information in addition to the sequence information[9].Another important threading method is the Profile Hidden Markov Model method which uses probability distribution model.Some threading method also use the structure–structure match scores to evaluate the alignment between the target and the template.Software which are available for protein threading are RAPTOR,HHpred,Phyre and MUSTER[10]. Some notable tools that are used for Ab inito protein structure prediction are Jpred and PSIPred which both use a position specific scoring matrix (PSSM), which contains position values based on similar sequences that is retrieved from PSI-BLAST, a sequence alignment tool that generates PSSMs. Jpred's newest revision, Jpred3, incorporates Jnet 2.0 the highest prediction accuracy for secondary structure at greater than 81%. The two most popular databases for protein structure are the Protein Data Bank (PDB) and the NCBI Protein Database. Accuracy of tertiary structure prediction is usually measured by comparing the coordinates for correct and predicted structures using root mean square deviation(R.M.S).

IX. CONCLUSIONS

The major difference between homology modeling and protein threading is that besides sequence information, protein threading can make use of secondary structure and solvent accessibility to improve both alignment accuracy and fold recognition rate. The biggest obstacle to improving prediction tools in general is still the slow pace of experimental advancements in biological and biochemical research still; new protein structures are constantly being determined, increasing the data available to refine protein structure prediction methods, which will eventually lead to a breakthrough in the field to be done.

REFERENCES

[1] Gutachter:Prof. Dr. Martin Vingron, Walaa Fathy Ahmed Walid Gomaa" "Approaches to protein structures" 978-1-4577-0476-5/112011 IEEE.

[2] Marco Vassura, Luciano Margara, Pietro Di Lena, Filippo Medri, Piero Fariselli, and Rita Casadio," Reconstruction of 3D Structures from Protein Contact Maps", VOL. 5, NO. 3, JULY-SEPTEMBER 2008

[3] Maciej Kicinski," AB INITIO PROTEIN STRUCTURE PREDICTION ALGORITHMS" (2011). Master's Projects. Paper 165.

[4] Hongyu Zhang "Protein Tertiary Structures:Prediction from Amino Acid" Sequences ENCYCLOPEDIA OF LIFE SCIENCES / & 2002 Macmillan Publishers Ltd.

[5] D.T. Jones, "THREADER: protein sequence threading by double dynamic Programming," In Computational Methods in Biology (ed. S. Salzberg,D. Searl, and S. Kasif), Amsterdam: Elsevier Science, 1998.

[6] J. Skolnick, D. Kihara, and Y. Zhang, "Development and large scale benchmark testing of the PROSPECTOR 3 threading algorithm," Proteins, vol. 56, pp. 502-518, 2004.

[7] Daisuke Kihara, Hui Lu,Andrzej Kolinski , and Jeffrey Skolnick(2001) "TOUCHSTONE: An ab initio protein structure prediction method that uses threading based tertiary restraints"

[8] I.Cymerman, M. Feder, M. PawŁowski, M.A. Kurowski,J.M. Bujnicki "Computational Methods for Protein Structure Prediction and Fold recognition" Nucleic Acids and Molecular Biology,Vol. 15 Springer-Verlag Berlin Heidelberg 2004.

[9] Mount, David W. Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press 2004.

[10] http://www.biophysics.org/blot/seq_empirical.html