International Journal of Current Engineering and Technology
ISSN 2277 - 4106

Available at http://inpressco.com/category/ijcet

Research Article

Protein Tertiary Structure Prediction using Data mining Techniques

Chandrayani Nikhil Rokde[Å*], Shital Satyen Telrandhe[Å] and Manali M. Kshisagar[Å]

[Å]Department of Information Technology, Dr.BabasahebAmbedkar College of Engineering Nagpur, Maharashtra, India

Accepted 20 November 2013, Available online 30 December 2013, **Vol.3, No.5 (December 2013)**

Abstract

Proteins are essential part of our life and participate in virtually every process within a cell. The understanding of protein structures is vital to determine the function of a protein. 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. Thus, protein structure prediction 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 computational methods for PSP is discussed and results are taken along with the help of parallel processing server.

Keywords: *Proteins,Protein structure prediction, Computational methods used in PSP,Parallel Processing.*

1. Introduction

Proteins are main building blocks of our Life. They are responsible for catalyzing 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. There are 20 different kinds of amino acid and each amino acid is identified by its side chain which determines the properties of amino acid. Amino acids are separated into four groups Non- polar Polar, Basic, Acidic, Polar and Non-Polar are again categorized under Hydrophobic (attracted towards water) and Hydrophilic (repelled by water). The combination of the properties that allow a specific protein to form into a certain structure is not completely known. There are many inherent properties that amino acids have that are involved in determining the structure of a protein.
One of the most important distinguishing factors of amino acids is their different tails which are also called the R Groups. Other factors play key roles in determining the final structure of a protein, these include: the energy level of the structure which needs to be low and stable and links between amino acids.

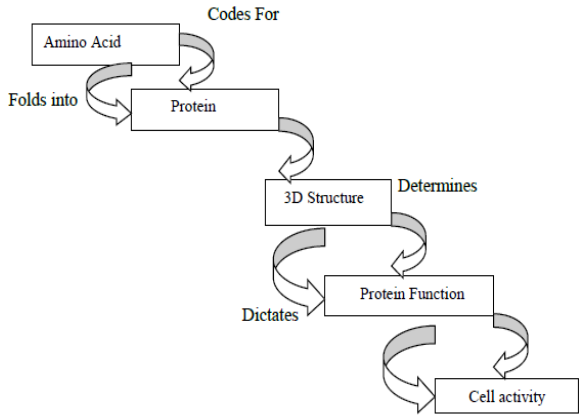
A protein does not exhibit a full biological activity until it folds into a three-dimensional structure. Information on the secondary and three dimensional(3D) structures of a protein is important for understanding its biological activity, because the shape and nature of the protein molecule surface account for the mechanisms of protein functions.

1.1 *Protein structure*

Formation of protein passes through different levels of structure. 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 .It is observed that due to the size, shape and charge of amino acid side chain ,each amino acid may fit better in one type of secondary structure than another.

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.

*Corresponding author: **Chandrayani Nikhil Rokde**



. .

Fig 1.1 shows Protein life cycle

1.2 *Protein Tertiary structure prediction*

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 . Although protein is constructed by the polymerization of only 20 different amino acidsinto 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. 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.

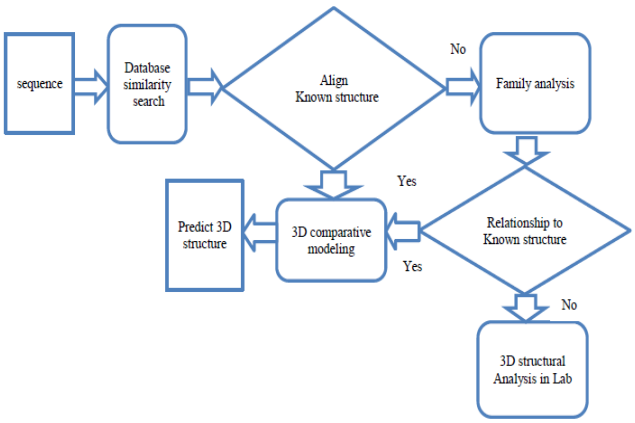


Fig 1.2 Flow chart for PSP

2. Methods Used in PSP

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. Homology modeling is for *easier targets. Accuracy of the prediction is 60%.Protein threading is carried out when sequence similarity with structure is Greater than 25%. Protein threading is for those targets with only fold-level homology found Protein threading is for harder targets(A Kelley *et al*, 2009). Accuracy of the prediction is 40%.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. This method is too slow and inaccurate and used for novel targets. Every two years, the performance of current methods is assessed in the CASP experiment stands for Critical Assessment of Techniques for Protein Structure Prediction.

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. 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 (Guido Bologna *et al*). Fold recognition requires a criterion to identify the best template for one target sequence.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:

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

Problem definition

Protein fold recognition methods attempt to recognize the suitable template from a structure template library for a query protein and generate an alignment between the query and the recognized template protein, from which the

structure of query protein can be predicted.Protein fold recognition using the protein threading technique has demonstrated a great success. There are four steps for the protein fold prediction for an amino acid sequence.
Step 1: Construct a protein structure template library
Step 2: Design a scoring function to measure the fitness between the target sequence and the template
Step 3: Design an efficient algorithm for searching over all the templates in the library
Step 4: Find the best alignment between the target sequence and the template by minimizing the scoring function

3. Implementation Methodologies

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 Alignment Search Tool (BLAST) were developed to perform rapid searches for sequence homologus in large sequence database(JamiaMilliaIslamia).These methods produce relatively accurate approximate sequence alignment by quickly finding sub-sequences in the databases. The two most popular databases for protein structure are the Protein Data Bank (PDB) and the NCBI Protein Database.

Protein p53 tumor suppressor is a flexible molecule composed of four identical protein chains.Flexible molecules are difficult to study by x-ray crystallography because they do not form orderly crystals, and if they do crystallize, The p53 protein is a phosphoprotein made of 393 amino acids. It consists of four units (or domains):

- A domain that activates transcription factors.
- A domain that recognizes specific DNA sequences (core domain).
- A domain that is responsible for the tetramerization of the protein.
- A domain that recognized damaged DNA, such as misaligned base pairs or single-stranded DNA.

3.1 Structure by parts

Most of the p53 mutations that cause cancer are found in the DNAbinding domain. The most common mutations are shown here, using PDB entry 1tup. This PDB entry includes three copies of the DNA-binding domain; only one (chain B in the file) is shown here. The mutations are found in and around the DNA-binding face of the protein.

Table 1 Predicted binding sites

Amino Acid	Residue	Contact	AV distance	JS divergence
CYS	176	25	0	0.77
HIS	179	25	0	0.72
CYS	238	25	0	0.81
CYS	242	25	0	0.77

The most common mutation changes arginine 248, colored red here. Notice how it snakes into the minor groove of the DNA (shown in blue and green), forming a strong stabilizing interaction. When mutated to another amino acid, this interaction is lost. Other key sites of mutation are shown in pink, including arginine residues 175, 249, 273 and 282, and glycine 245. Some of these contact the DNA directly, and others are involved in positioning other DNA-binding amino acids.

3.2 Algorithm Implementation

We have implemented algorithm known as Quasi Physical Algorithm there are two types of monomers: H (hydrophobic) and P (polar) ones. The polymer is modeled as a self-avoiding chain (amino acid sequences) on a regular lattice with repulsive or attractive interactions between neighboring nonbonded monomers.

We can imagine that all the balls are connected by a spring, and consider three types of forces: Fijp - the pulling force of spring between any two neighboring balls, Fijr- the repulsion forces between any two embedded balls and Fijg- the gravitational forces between any two H balls. Thus, at any time, composite force Fi decides the motion direction and velocity of ball i:

$$\sum Fi=\sum Fijp+\sum Fijr+\sum Fijg$$

Apparently, under the exertion of three types of forces, the H balls tend to congregate to form a center, and the P balls tend to layout peripherally.(Wang Gang,Liu*et al*, 2006) When the system reaches an equilibrium state, we get a good approximation to 3D protein structure .

3.3. Algorithm implementation on database

Algorithm is implemented on protein database downloadedfromUniProtKB/Swissprot.UniProtKB/Swiss-Prot is a high-quality, manually annotated, non-redundant protein sequence database. It combines information extracted from scientific literature and biocurator-evaluated computational analysis. The aim of UniProtKB/Swiss-Prot is to provide all known relevant information about a particular protein. Annotation is regularly reviewed to keep up with current scientific findings. The manual annotation of an entry involves detailed analysis of the protein sequence and of the scientific literature. Sequences from the same gene and the same species are merged into the same database entry.
Differences between sequences are identified, and their cause A range of sequence analysis tools is used in the annotation of UniProtKB/Swiss-Prot entries. Computer-predictions are manually evaluated, and relevant results selected for inclusion in the entry (A Kelley et al. 2009).These *predictions* include post-translational modifications, tranmembrane domains and topology, signal peptides, domain identification, and protein family classification. Our manual prediction procedure consists of the following components:
1. Pre-processing for identification of protein domains, identification and removal of signal peptides, and protein secondary structure prediction.

2. Collection of functional/structural information of a prediction target through various database searches;
3. Protein fold recognition for identification of native-like folds (Mohammed Said*et al*, 2008).
4. Prediction result validation through comparing predicted structures and collected structural and functional information for consistency check.

4. Results

Table 2 Values for different amino acidresidue

Sr.No	Amino Acid	Resi due	P.F	R.F	G.F
1	SER	96	-12.886	5.999	-31.892
2	SER	96	-12.537	10.798	-28.291
3	SER	96	-12.392	9.576	-29.156
4	SER	96	-12.151	4.597	-34.107
5	SER	96	-11.887	4.831	-29.067
6	SER	96	-11.669	5.957	-32.084
7	VAL	97	-11.435	6.72	-27.479
8	VAL	97	-11.391	11.646	-28.74
9	VAL	97	-11.172	7.063	-29.961
10	VAL	97	-11.097	5.335	-33.354
11	VAL	97	-11.034	6.06	-28.794
12	VAL	97	-10.93	9.389	-29.188
13	VAL	97	-10.779	6.442	-31.223
14	PRO	98	-10.489	6.432	-34.235
15	PRO	98	-10.307	8.291	-29.659

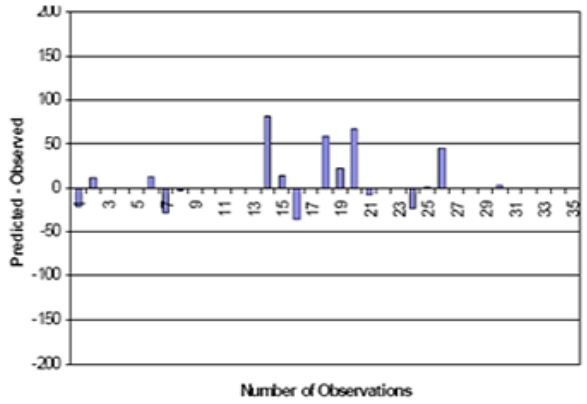


Figure 1 Graph showing predicted protein structure vs number of observation

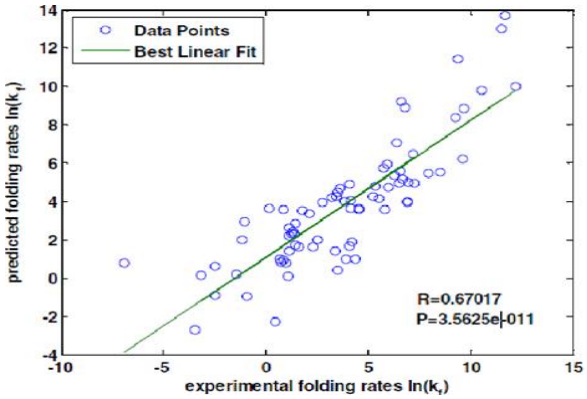


Figure 2 Graph showing predicted folding rates vs. experimental folding rates

If we define the computation of one Fij (and Fji) as basic operation, and assume that it take unit time, the run time of algorithm 1 .where 1 is the number of iterations. The experiments show that the solutions produced by our algorithm have lower energy than those produced by other methods. But the algorithm needs very big 1 (generally hundreds of millions) to get good result. For long amino acid sequences, the run time of this algorithm is not acceptable. So we consider parallelizing the algorithm using parallization techniques (www.openmp.org).

5. Parallel Approach

Parallel processing is the simultaneous use of more than one CPU or processor cores to execute a program .Motivations for parallel processing Higher speed or solving problem faster Higher computational power. We have applied OpenMp programming to parallelize the quasi physical algorithm (R. Eigenman*et al*, 2001).
OpenMP is widely accepted standard API for writing shared memory parallel applications in c. It consists of compiler directives, runtime routines and various environment variables whose specifications are maintained by OpenMP Architecture Review Board. It is basically based on fork-join model. As multicore machines and multithreading processors spread in the marketplace, it might be increasingly used to create programs for uniprocessor computers also. OpenMP is not a new programming language, rather, it is notation that can be added to a sequential program in FORTRAN, C, or C++ to describe how the work is to be shared among threads that

	On Dual Core	On Quad core	On Six core
Tseq	5.11899	3.911532sec	0.02048
Tpar	3.08569	1.006332sec	0.00501
Speedup	1.65sec	3.93sec	**4.08sec**

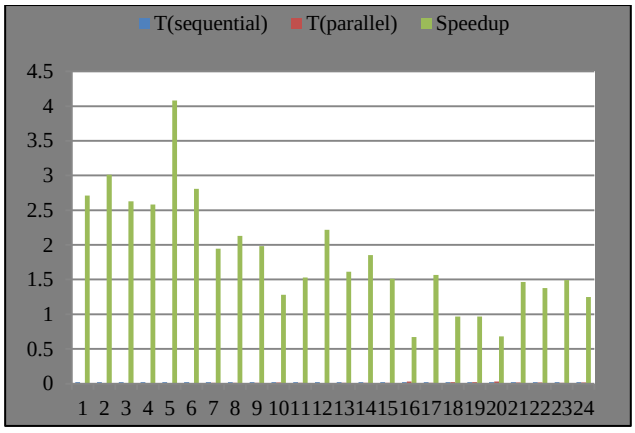


Figure 3 graph showing timings for no.of threads

will execute on different processors or cores and to order accesses to shared data as needed. The appropriate insertion of OpenMP features into a sequential program will allow many, perhaps most applications to benefit from shared-memory parallel architectures—often with minimal

modification to the code.The basic idea for OpenMP is based on threads that are a runtime entity that is able to independently execute a stream of instructions. OpenMP builds on a large body of work that supports the specification of programs for execution by a collection of cooperating threads. Threads running simultaneously on multiple processors or cores may work concurrently to execute a parallel program. We have used OpenMp programming for parallelization of algorithm, for parallelization of an algorithm.

As HP Z600 has 12 cores it is having 24 thread,Results for the 24 threads are taken by calculating Time which is measured by: QueryPerformanceCounter(&start).And Speed of the algorithm is calculated as time required for executing the sequential code divided by time required for executing parallel code.

Table 4 Timing for number of threads

No. of Threads	T(sequential)	T(parallel)	Speedup
1	0.020478	0.007534	2.71
2	0.020478	0.006804	3.01
3	*0.020478*	*0.00779*	*2.629*
4	0.020478	0.007938	2.58
5	0.020478	0.00501	**4.0836**
6	0.020478	0.007299	2.806
7	0.020478	0.010526	1.945
8	0.020478	0.009622	2.128
9	0.020478	0.010342	1.98
10	0.020478	0.015986	1.281
11	0.020478	0.013383	1.53
12	0.020478	0.009239	2.216
13	0.020478	0.012676	1.615
14	0.020478	0.011038	1.855
15	0.020478	0.013569	1.509
16	0.020478	0.030526	0.671
17	0.020478	0.013063	1.568
18	0.020478	0.021186	0.967
19	0.020478	0.021153	0.968
20	0.020478	0.030115	0.68
21	0.020478	0.01398	1.465
22	0.020478	0.014869	1.377
23	0.020478	0.0137	1.495
24	0.020478	0.016376	1.25

Table 3 Number of observation on dual core, quad core and six core machine

6. Conclusions

Thus we observed that maximum speedup obtained for Quasi Physical Algorithm is at thread no.5 is 4.0836
The main idea of the parallel algorithm is that using coarse grained data composition strategy to partition tasks, and exchanging data between processes periodically to minimum communication overhead. Our approach starts with a pair of sequences in the set and uses the local alignment results of the two sequences to construct an initial step. It then progressively processes the remaining Sequences. Experimental results show that this approach can achieve comparable accuracy on sequences.

Accuracy: The predicted accuracy of quasi physical algorithm is 61.85 which is 0.85 higher than the existing methods.

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

YunlingLan Tao (2008), Protein Structure Prediction based on An Improved Genetic Algorithm , 2978-1-4244-1748-3/08/, *IEEE*

Rafiqul Islam and AliouneNgom (2005), Parallel Evolution Strategy for Protein Threading Proceedings of the *XXV International Conference of the Chilean Computer Science Society* (SCCC'05) 0-7695-2491-5/05.

A Kelley & Michael J E Sternberg (Feb 2009), Protein structure prediction on the Web: a case study using the Phyreserver Lawrence -26 9; doi:10.1038/nprot.2009.2

Guido Bologna, Ron D. Appel , AComparison Study on Protein Fold Recognition Proceedings of the 9th *International Conference* on *Neural Information Processing (ICONIP'OZ)* , Vol. 5

Colony AlgorithmHeshamAwadh A. Bahamish, Rosni Abdullah 978-0-7695-3648-4/09, *IEEE*

Khalid Raza , Application of Data Mining In Bioinformatics *Centre for Theoretical Physics,Jamia Millia Islamia, New Delhi-110025, India* Vol 1 No 2, 114-118

Wang Gang,LiuXiaoguang, Liu Jing (2006), Parallel Algorithm for Protein Folds Prediction 1-4244-060 *IEEE.*

Mohammed Said Abual-Rub and RosniAbdullah (2008), A Survey of Protein Fold Recognition Algorithms Journal of Computer Science 4 (9): 768-776, 2008 ISSN 1549-3636,*Science Publications*

A Kelley & Michael J E Sternberg (2009), Protein structure prediction on the Web: a case study using the Phyre server Lawrence 26, doi:10.1038/nprot.2009.2

Mohammed Said Abual-Rub and RosniAbdullah (2008), A Survey of Protein Fold Recognition Algorithms Journal of Computer Science 4 (9): 768-776, 2008 ISSN 1549-3636, *Science Publications*

Mount, David W. (2004), Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor, NY: *Cold Spring Harbor Laboratory Press.*

Hongyu Zhang, Celera Genomics, Rockville, Maryland ,Protein Tertiary Structures Prediction from Amino Acid Sequences, USA

OpenMP home page (http://www.openmp.org) Using OpenMPBy Barbara Chapman, Gabriele Jost& Ruud Van Der Pas. *MIT Press*

R. Eigenman, Michael J. Voss(Eds*): OpenMP Shared Memory Parallel Programming*. Springer LNCS 2104, Berlin, 2001 ISBN 3-540-42346-X