International Journal of Sciences
(ISSN 2305-3925)
http://www.ijSciences.com

Research Article
Volume 2, Issue June 2013

A Deep Glimpse into Protein Fold Recognition

Ahmed Sharaf Eldin[1], Taysir Hassan A. Soliman[2], Mohammed Ebrahim Marie[3], Marwa Mohamed M. Ghareeb[4]

[1]Information Systems Department, Faculty of Computers and Information, Helwan University, Cairo, Egypt
[2]Supervisor of Information Systems Department, Faculty of Computers and Information, Assiut University, Assiut, Egypt
[3]Information Systems Department, Faculty of Computers and Information, Helwan University, Cairo, Egypt
[4]Assistant Lecturer, Information Systems Department, Modern Academy, Cairo, Egypt

Abstract: The rapid growth in genomic and proteomic data causes a lot of challenges that are raised up and need powerful solutions. It is worth noting that UniProtKB/TrEMBL database Release 28-Nov-2012 contains 28,395,832 protein sequence entries, while the number of stored protein structures in Protein Data Bank (PDB, 4-12-2012) is 65,643. Thus, the need of extracting structural information through computational analysis of protein sequences has become very important, especially, the prediction of the fold of a query protein from its primary sequence has become very challenging. The traditional computational methods are not powerful enough to address theses challenges. Researchers have examined the use of a lot of techniques such as neural networks, Monte Carlo, support vector machine and data mining techniques. This paper puts a spot on this growing field and covers the main approaches and perspectives to handle this problem.

Keywords: Protein fold recognition, Neural network, Evolutionary algorithms, Meta Servers.

I. INTRODUCTION

Proteins transport oxygen to cells, prevent harmful infections, convert chemical energy into mechanical energy and perform many other important and beneficial biological processes. Proteins are also notable for their more deleterious effects; for instance, viruses are surrounded by protein shells that allow them to gain access to host cells. Knowing the protein's three-dimensional structure (i.e., the conformation or fold) is an important step towards understanding protein functions and brings many benefits like the ability to compose and invent drugs which interact with particular proteins, greater understanding of genetic defects, and improved therapies for diseases such as AIDS and malaria.

It is worth noting that UniProtKB/TrEMBL database Release 28-Nov-2012 [14] contains 28,395,832 protein sequence entries, while the number of stored protein structures in Protein Data Bank (PDB, 4-12-2012) [15] is 65,643. Thus, the need of extracting structural information through computational analysis of protein sequences has become very important and a lot of research has been conducted towards this goal in the late years. Especially, the prediction of the fold of a query protein from its primary sequence has become very challenging.

In the current work, different methods to face the problem of protein fold recognition are covered. This paper is organized as follows: section 2 illustrates the preliminaries of the protein structure. Section 3 clarifies the current categories of Protein fold recognition methods; section 4 demonstrates the different methods for protein fold recognition. Finally, section 5 concludes this paper.

II. PROTEIN STRUCTURE IN A GLANCE

Proteins are formed using the genetic code of the DNA. Three different processes are responsible for the inheritance of genetic information: **Replication**: a double stranded nucleic acid is duplicated to give identical copies, **Transcription**: a DNA segment that constitutes a gene is read and transcribed into a single stranded sequence of RNA. The RNA moves from the nucleus into the cytoplasm, **Translation**: the RNA

Marwa Mohamed M. Ghareeb (Correspondence)
mortamm@hotmail.com

sequence is translated into a sequence of amino acids as the protein is formed.

Proteins are constructed from a set of twenty naturally occurring amino acids. Amino acids are organic compounds, formed from carbon, hydrogen, nitrogen, oxygen, and sulfur. Amino acids form proteins by pending together in our cells by ribosomes, which are a cellular organelle that synthesizes polypeptide chains that will become proteins. All amino acids have a backbone and a residue. The backbone contains a nitrogen, denoted by **N** , followed by two carbons labeled the **α**-carbon, $\mathbf{C_\alpha}$, and the prime-carbon, **C'** , so the backbone atoms are ordered from left-to-right **N**- $\mathbf{C_\alpha}$- **C'** and it is the same for all the amino acids. The **N** atom has a hydrogen, **H**, attached to it, the **C'** atom has an oxygen, **O**, attached to it (known as the carbonyl **O**), and the $\mathbf{C_\alpha}$ atom has a hydrogen **H,** and the residue attached to it [1]. To form the protein, the amino acid backbones are linked together in a long chain to form peptide bonds between the **C'** of an amino acid and the N of the following amino acid.

The residue is a set of atoms attached to the central $\mathbf{C_\alpha}$. Amino acid is distinguished by its residue. The atoms of the residue are labeled using the Greek alphabet, beginning with the backbone's **α**-carbon and followed by the residue's **β**-carbon, γ-carbon, ϵ-carbon, etc as in Fig 1.

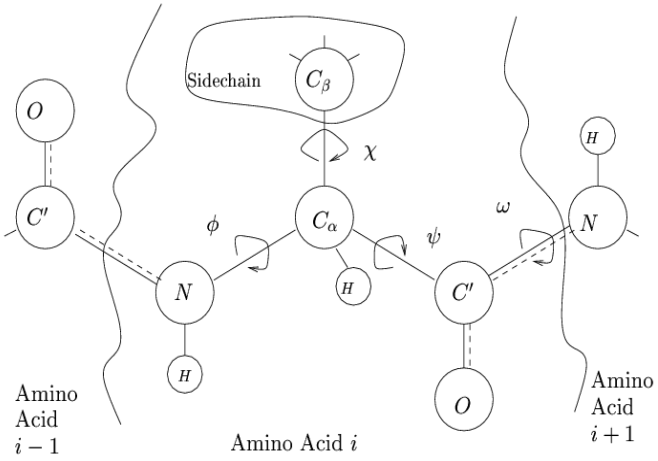

Fig 1 A typical tri-peptide sequence.

A. Covalent Bonding

The primary definer of protein structure is the covalent bonding. This bond has some parameters like the bond length, **L** which is measured as the distance between two atom's centers and is in angstroms, Å, the bond angle, **K** which is measured as the angle created by three atoms with two bonds between them and is in degrees, and the torsion angle, **θ** which is measured as the rotation of the bond about some axis and is in degrees [2] as seen in Fig. 2.

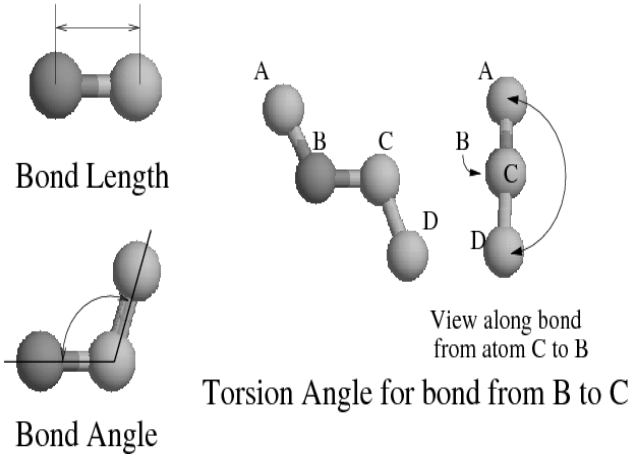

Fig 2 The covalent bonding parameters: the bond length, bond angle and torsion angle.

The values of the parameters depend upon the atoms forming the bond and the processes that went into creating the bond [2].The common covalent bonding types in proteins are single bonds, double bonds, peptide bonds and disulfide bridges.

Torsion angle is the main degree of freedom within the bonds. The three torsion angles related to the backbone bonds are: the rotational angle between the **N** and $\mathbf{C}_{\alpha}$ atoms which is called **Ǿ,** the rotational angle between the $\mathbf{C}_{\alpha}$ and **C'** atoms which is called **ψ**, and the rotational angle between the **C'** and the following amino acid's N which is called **w** (see Figure 1). **w** is a peptide bond and it is restricted to angles 180^{o} and 0^{o}. The **Ǿ** and **ψ** angles rotate much more freely [2], [3].

Beside the three torsion angles, there are rotamer angles associated with the torsion angles found in the single bonds of the amino acid sidechains. These angles begin with **x1**, the torsion angle of the **C'**-$\mathbf{C}_{\beta}$ bond, and continue on with **x2, x3**, etc. The protein's conformation is affected by changes in rotamer and backbone torsion angles. A change in a rotamer angle will only affect the location of a single sidechain's atoms, but a change in **Ǿ**, **ψ**, or **w** will affect the location of every backbone and sidechain atom following it. Thus a protein is a long chain that varies in shape primarily because of the rotation of each amino acid's **Ǿ** and **ψ** angles, and their choice of two possible **w** values [2].

B. Secondary Structure

A secondary structure is a repeating three-dimensional structure with a fixed bonding pattern. These structures are formed by a weak hydrogen bonding between atoms on different amino acids.

N atoms are known as donors where they share their **H** atoms with the other atoms. Acceptors, mostly **O**, are attracted to these donated **H** atoms because of their respective opposite charges

α-helix is considered the most common protein secondary structures. The hydrogen bonds in a **α**-helix occur between the **N** of the **i** amino acid and the carbonyl **O** of the **i+4** amino acid.

The other common secondary structures are **β**-sheets. **β**-sheets are constructed when a hydrogen bonds are formed between two relatively straight protein backbone segments that are lied parallel to each other as seen in Fig 3. The individual segments are referred to as **β**-strands. There are two types of **β**-sheets, anti-parallel and parallel.

β-sheets are typically anti-parallel where the directions of the two **β** strands run opposite to each other. The Hydrogen bonds are formed between the two **β**-strands where the $\mathbf{i}^{th}$ amino acid's carbonyl **O** attached with the $\mathbf{j}^{th}$ amino acid's backbone N, and between the $\mathbf{i}^{th}$ amino acid's **N** and the $\mathbf{j}^{th}$ amino acid's carbonyl O. Unlike the **α**-helices, the hydrogen bonding pattern between the two **β**-strands skips the **i+1** and **j-1** amino acids and continues with the **i+2** and **j-2** amino acids [2], [4].

A less common conformation is the parallel **β**-sheet where the two strands are running in the same direction. The backbone **N** of the **i** amino acid forms a hydrogen bond with the carbonyl **O** of the **j** amino acid, but the **i** carbonyl **O** forms a hydrogen bond with the **j+2** backbone **N**. Then the **i+2** backbones **N** will form a hydrogen bond with the **j+2** carbonyl **O** and so on.

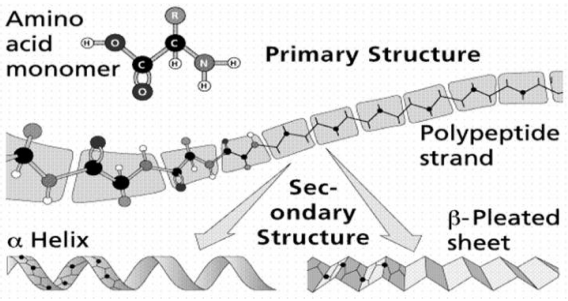

Fig 3 The most common secondary structures: alpha helices and beta sheets

Often, the secondary structures are organized in large groups called motifs. Common motifs include helix-helix, helix-loop-helix, and the Greek key motif, which is four adjacent anti-parallel **β** –sheets [2], [4].

C. Factors affect the 3D-Conformation

1) Hydrophobicity: Some of amino acids are polar. When they are exposed to the water, they interact with it. These amino acids are called hydrophilic like Thr, Tyr, Ser, Asp and Trp. Being exposed to water allows

their polar atoms to hydrogen bond with the surrounding water atoms. This is energetically beneficial to the protein. Other amino acids are called hydrophobic which are repelled from water. They are generally seen on the interior of the protein away from the surface and the surrounding solvent. These include Ala, Val, Pro, and Met. The hydrophobic and hydrophilic effect determines, to a large extent, how the protein will fold [5].

2) Protein Energetics: The hydrophobic and hydrophilic effect is one aspect of the protein's energetics. By energetics, the protein's folding process is referred to as an attempt to minimize its thermodynamic energy. Beside the hydrophobic and hydrophilic effect, the interactions such as hydrogen bonding and side-chain entropy also contribute to the protein's energy.

Side-chain entropy relates to the conformation of each amino acid's side-chain. Side-chains have a variety of configurations that they can appear in. As they become buried, this freedom is reduced and entropy decreases [5].

3) Energy Minimization: Proteins are constructed of atoms constrained by bonds and affected by forces exerted on them by surrounding atoms, both in the protein and in the surrounding solvent. These problems can be thought of as an optimization function. This energy function will contain parameters such as the covalent bonds between atoms, the hydrophobic effect, the hydrogen bonding, and other effects. This constitutes a very large complicated function that theoretically should be solvable [5], [6].

D. Tertiary structure

The process of protein folding results in a compact structure in which secondary structure elements are packed against each other in a stable configuration, often called a `fold'. Many elegant structures have evolved, including curved, barrel-like β-sheets, parallel bundles of helices, and propellor-like structures. Folds are also referred to as `topologies' since they can be thought of as sets of connected secondary structure elements [6].

Fold recognition and threading methods can be used to assign tertiary structures to protein sequences, even in the absence of clear homology. Fold recognition and threading methods aim to assign folds to target sequences that have very low sequence identity to known structures. Fold recognition methods work by comparing each target sequence against a library of potential fold templates using energy potentials and/or other similarity scoring methods. The template with the lowest energy score (or highest similarity score) is then assumed to best fit the fold of the target protein.
Although fold recognition and threading techniques will not yield equivalent results as those from X-ray crystallography, they are a comparatively fast and inexpensive way to a build a close approximation of a structure from a sequence, without the time and costs of experimental procedures.

III. CURRENT PROTEIN FOLD RECOGNITION CATEGORIES

In year 1994, The Protein Structure Prediction Center at Lawrence Livermore Laboratories has conducted a community wide experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP).

CASP goal is to help the advance of methods to identify protein structure from sequence. CASP provides the means of objective testing and evaluating of these methods via the process of blind prediction. CASP experiments provide the current state of protein structure prediction, the identifying of what the progress has been made, and highlighting where future effort may be most productively focused.

For the purpose of evaluating the different techniques, CASP divides the predictions into three different types: comparative modeling, **ab** initio predictions and fold recognition [7].

A. Comparative Modeling(CM)

Comparative models attempt to match the target's amino acid sequence with the amino acid sequences of proteins with known structures. Then the target protein's structure is assumed to be similar to that of the matched protein's known structure. This technique performs best when one can find a family of similar proteins. The main framework of these models is similar with significant variations at each step. The process starts with defining families or classes of folds and noting the sequences that produce those folds, then searching for sequences similar to our target using some alignment technique, such as SWISS-PROT [8]. The similar sequences will provide a template fold family for the target protein. Finally, the target protein is fitted to the template using the restraints inherent in the template.

These models suffer from several difficulties at each step. Protein families can be difficult to discern. Due to evolutionary effects within different species, proteins with divergent amino acid sequences may share a similar conformation. Also, there are many examples of proteins with a high degree of sequence similarity, but different folds. In addition, there is a large class of proteins for which no family can be found, and, hence, no model created [11].

B. Ab Initio Techniques

Ab initio or "new fold" (NF) models construct the protein using general principles-many of which are thermodynamic in nature. **Ab** initio folding methods build the 3D structure of a protein from its sequence without using any templates. Generally, it is assumed that a protein folds to a global minimum-energy conformation.

In order to find such a conformation, there are two approaches. In the first one, researchers simulate protein folding by doing standard molecular dynamic simulation with a physically reasonable potential function. This approach is computationally expensive. In addition, due to the inadequacies of current potential functions, the probability that a native state will be found at a global minimum-energy conformation is significantly reduced [9].

The second approach to do **ab** initio folding is the direct conformation space search where the successful prediction of the native structure of a protein requires both an efficient sampling of the conformational space and an energy function that recognizes the native conformation as the lowest in energy. However, exhaustive conformation space search is still formidable due to current computing speeds. To deal with that, researchers have attempted to reduce the search space by simplifying models or reducing the conformation space.

It has been observed that ab initio folding cannot perform consistently for all classes of proteins. In fact, ab initio folding totally fails for proteins longer than 150 residues [9].

C. Fold Recognition

Fold recognition (FR) or threading is a cousin to sequence homology. Instead of searching for significantly similar sequences and deducing the structure of the protein, Fold recognition methods try to recognize the structural fold of a protein by using a structure template library and the protein's sequence information then generate an alignment between the query and the recognized template from which the structure of query protein can be predicted [10]. Fold recognition techniques do not require similar sequences in the protein databank, just similar folds [12]. Due to phenomena of deletions, insertions, varying sequence length and others, there are thousands of possible ways to match a sequence to a template [13].

Fold recognition methods are so efficient when the sequence has little or no primary sequence similarity to any sequence with a known structure and when some model from the structure library represents the true fold of the sequence.

Current fold recognition methods suffer from many points: First, current energy functions are not precise enough to determine the free energy of a certain conformation; Second, there is no direct computational method that can recognize the conformation. Also, the size of the conformation space is huge. Protein threading problem is considered as NP-complete and MAX-SNP-hard [16], [17].

In recent years, the traditional boundaries between CM, NF, and FR have become blurred and the distinction between individual methods has become less clear. Sequence searching has become more powerful and arguably the traditional threading techniques which are based on physical energy potentials are becoming less popular. The term "fold recognition" is now often used to encompass all methods able to carry out template based modeling beyond the so-called "twilight zone" of sequence identity.

IV. PROTEIN FOLD RECOGNITION METHODS

There are many successful methods to face protein fold recognition problem.

1) Support Vector Machine (SVM): Xu has proposed a Support Vector Machine (SVM) regression approach to directly predict the alignment accuracy of a sequence template alignment [25]. The authors implemented experiments on a large-scale benchmark using their Support Vector Machine (SVM) regression approach. They argued that experimental results show that SVM regression method has much better performance in both sensitivity and specificity than the composition corrected Z-score method and SVM regression method also outperforms SVM classification method. In addition, SVM regression method enables the threading program to run faster than the composition-corrected Zscore method.

Sangjo Han et al. presented an alternative method for estimating the significance of the alignments [26]. A protein query is aligned to a template of length n in the fold library, and then this alignment is transformed into a feature vector of length n+1, which is then evaluated by Support Vector Machine (SVM). The output from SVM is converted to a posterior probability that a query sequence is related to a template, given SVM output. The new method outperforms PSI-BLAST and profile-profile alignment with Z - score scheme. The reason of that is related to the intermediate sequence search and its ability to recognize the essential features

among alignments of remotely related proteins.

W. Chmielnicki has presented a combined SVM-RDA classifier for the Protein fold recognition [27]. This model combines a well-known Support Vector Machine (SVM) classifier with Regularized Discriminant Analysis (RDA). It is used on a real world data set. The experiments showed that it outperforms the previously published methods.

2) Structural pattern-based methods**:** SPREK (Sequence structure Pattern-matching by Residue Environment Comparison) is a method for evaluation of protein models based on residue packing interactions developed by Taylor and Jonassen [24]. SPREK evaluates the register of a sequence on a structure based on the matching of structural patterns against a library derived from the protein structure databank. SPREK is a very straightforward approach. It is characterized by its simplicity; also there are no large tables of potentials or any large weight matrices. It did not discard structural information as occurs in the majority of methods. The major advantage of this method is its ability to operate using only the α-carbon atom positions.

3) Neural network: Jones has presented GenTHREADER[18] as a new method for fold recognition. GenTHREADER can be divided into three stages: alignment of sequences, calculation of pair potential and salvation terms and evaluation of the alignment using a neural network. Jones claimed that the speed of this method, along with its sensitivity and low false-positive rate makes it ideal for automatically predicting the structure of all the proteins in a translated bacterial genome (proteome).

It is worth noting that GenTHREADER is able to produce structurally similar models for one-half of the targets, but significantly accurate sequence-structure alignments were produced for only one-third of the targets. Also, it can find the correct answer for the easy targets if a structurally similar fold was present in the server's fold libraries. However, among the hard targets it is able to produce similar models for only 40% of the cases, half of which had a significantly accurate sequence-structure alignment.

Kuang Lin et al has presented TUNE (Threading Using Neural nEtwork) [19]. TUNE uses an artificial neural network model to predict compatibility of amino acid sequences with structural environment.

But their model is not trained to discriminate native protein structures. TUNE is applied on the discrimination of protein decoy and native 3D structure, its performance is comparable to pseudo-energy functions with atom level structural description, better than the two functions with residue level structural descriptions.

Mcguffin and Jones [20] have improved and benchmarked GenTHREADER method; their improvements increase the number of remote homologies that can be detected with a low error rate which imply a higher reliability of score which also increase the quality of the models improved.

Thomas W. proposed protein fold class prediction using neural networks with tailored early-stopping [22]. This method consists of two stages: first the training patterns are used completely for gradient calculation and then they are split into a training and validation data set. This led to good generalizing neural networks. The experiments showed that standard feed-forward neural networks combined with an appropriate regularization scheme can classify the fold class of a protein given solely its primary and it outperformed the standard statistical approaches (like the nearest neighbor method etc and did not perform worse than Support Vector Machines (SVMs).

Nan Jiang et al. proposed MESSM which is a protein fold recognition model with mixed environment-specific substitution mapping [21]. It has three key features: a structurally-derived substitution score generated using neural networks, a mixed environment specific substitution mapping developed by combing the structural-derived substitution score with sequence profile from well-developed sequence substitution matrices, and a support vector machine employed to measure the significance of the sequence-structure alignment. MESSM is tested on two benchmark problems; Wallner's Benchmark and Fischer's Benchmark, and MESSM was found to lead to a good performance on protein fold recognition.

4) Evolutionary methods: Genetic algorithms were introduced early in 1992 [28]. Then, Unger and Moult have developed a genetic algorithm search procedure suitable for use in protein folding simulations [29]. Genetic algorithms are used to fold proteins on a two-dimensional square lattice in the HP model. A population of conformations of the polypeptide chain is used and then the mutation is used to change the conformations. Also, crossover is used in which parts of the polypeptide chain are interchanged between conformations. It was found that the genetic algorithm is dramatically superior to conventional Monte Carlo methods.

Yadgari et al. addressed the genetic algorithm paradigm used to perform sequence to structure alignments [30]. The sequence-structure pairs were

taken from a database of structural alignments where the sequence of one protein was threaded through the structure of the other. In this study, a proper representation is introduced where genetic operators can be effectively implemented. This representation consists of numbers usually zeros and ones or integer number when there is a sequence deletion; an example of representation is 11110011311 where 1 means a position of sequence on structure, 0 means structure deletion and any other number like 3 in the example, means sequence deletion. The effects of changing operators and parameters are explored and analyzed. The results of experiments indicate that the Genetic Algorithms method is a feasible and efficient approach for fold recognition.

Unger discussed the use of genetic algorithms to address the problem of protein structure prediction and protein alignments and introduced a general framework of how genetic algorithms can be used for protein structure prediction [31]. Using this framework, the significant studies that were published in recent years are discussed and compared. Unger suggested some improvements to be made to GA methods to improve performance. One obvious aspect is to improve the energy function. An interesting possibility to explore within the GA framework is to make a distinction between the fitness function and the energy function. Unger also introduced the use of explicit memory into the emerging substructure.

Modified Keep-Best is proposed by M.V.Judy and K.S.Ravichandran as an intermediate selection strategy for genetic algorithms and it is applied for protein folding problem [32]. The performance of the algorithm is tested for a set of six sequence-structure pairs. The effects of changing operators and parameters are explored and analyzed. The authors claimed that genetic algorithms threading is quite robust and is not overly dependent on the particular selection of parameter or operators.

5) Bayesian networks: Raval et al. has presented a Bayesian network approach for protein fold and superfamily recognition [23]. Bayesian network is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG), which includes, as a special case, hidden Markov models. This model is implemented to learn amino acid sequence, secondary structure and residue accessibility for proteins of known three dimensional structure. Raval argued that the cross validation experiments using Bayesian classification showed that the Bayesian network model which incorporates structural information outperforms a hidden Markov model trained on amino acid sequences alone.

6) Monte Carlo methods: Monte Carlo methods are a class of computational algorithms that rely on repeated random sampling to compute their results. Monte Carlo methods are especially useful for simulating systems with many coupled degrees of freedom. Monte Carlo methods have traditionally been employed to address the protein folding problem. Traditional Monte Carlo and molecular-dynamics simulations tend to get caught in local minima, so the native structure cannot be located and the thermodynamic quantities cannot be estimated accurately. To resolve this problem, Liang and Wong proposed an Evolutionary Monte Carlo (EMC) approach for protein folding simulations [33]. EMC can be applied successfully for simulating the protein folding on simple lattice models and to finding the ground state of a protein. The authors claimed that EMC is faster than the genetic algorithm and the conventional Metropolis Monte.

7) Parallel Evolutionary Methods: Many researchers used Parallel Evolutionary Methods (PEM) for protein fold recognition. Parallel hybrid gas was introduced by Carpio et al. for three dimensional structure predictions of polypeptides [34]. The previous research of Carpio was insufficient to produce better fit conformers, so Carpio improved it in two substantial aspects. The first is a parallelization of the original algorithm to enrich the diversity of conformers in the population and the second a hybridization of the simple GA in order to process the atoms of the side chains. Carpio et al. claimed that a comparison of the best fit individual after the 500th generation obtained by the hybrid GA reveals more accurately the level of evolution of the process.

Nguyen et al. proposed a parallel hybrid genetic algorithm for solving the sum-of-pairs multiple protein sequence alignment problem [35]. They present a new GA-based method for more efficient multiple protein sequence alignment. A new chromosome representation and its corresponding genetic operators have been proposed. A multi-population GENITOR-type GA is combined with local search heuristics. It was then extended to run in parallel on a multiprocessor system for speeding up. The experimental results showed that the proposed method is superior to MSA, OMA and SAGS methods with regard to quality of solution and running time. It can be used to find multiple sequence alignment as well as testing cost functions.

MOfmGA was introduced by Day et al.[36] as parallel multi-objective implementation for protein structure prediction. The authors focused on tuning fmGA in an

attempt to improve the effectiveness and efficiency of the algorithm to address protein structure problem and to find better ways to identify secondary structures. Problem definition, protein model representation, mapping to algorithm domain, tool selection modifications and conducted experiments were discussed in this study. They claimed that their progress of using MOfmGA have been modified to scale its efficiency to 4.7 times.

8) Parallel evolution strategy: The Single Query Single Template Parallel ES Threading (SQSTPEST) method is proposed by Islam and Ngom[37] protein threading based on evolution strategy. They used two parallel approaches for fast threading. The parallelization is based on master-slave architecture. The method threads one query against one template. They used High Performance Computing environment, SHARCNET (Shared Hierarchical Academic Research Computing Network) as computing platform for experiment. The authors claimed that this method has obtained at least better results than current comparable approaches, as well as significant reduction in execution time.

EST is a novel evolution strategy for protein threading problem using evaluation strategy proposed by Alioune Ngom [38]. The author also proposed a parallel method for fast threading called parallel EST. Parallel EST was implemented on Grid-enabled platforms for High-Performance Computing. The author was only interested in determining the best alignment between a query and a template given an energy function so he was planning to use a better energy function than the one discussed in the study. Also, a threading score between a query and a template may not provide enough information about whether the template is the "correct" fold. That is, from the threading scores between a query and a pool of templates, so it is unknown whether the query's correct fold template is in the pool, or which is the correct fold even if it is there.

Probabilistic roadmap methods for motion planning are used in a new computational technique proposed by Thomas and Amato for studying protein folding [39]. This technique yielded an approximate map of a protein's potential energy landscape that contains thousands of feasible folding pathways. Thomas and Amato claimed that the other simulation techniques, such as molecular dynamics or Monte Carlo methods required many orders of magnitude more time to produce a single, partial, trajectory. They use STAPL method to easily parallelize their sequential code to obtain scalable speedups.

Wiese and Hendriks introduced a parallel evolutionary algorithm called PRnaPredict for RNA secondary structure prediction [40]. PRnaPredict is a fully parallel implementation of a coarse-grained distributed EA for RNA secondary structure prediction and is based on RnaPredict. Two sets of experiments were performed on five known structures from 3 RNA classes. The first determines the actual speedup and the second evaluates the performance of P-RnaPredict through comparison to mfold. The experiment results claimed that P-RnaPredict possess good prediction accuracy, especially on shorter sequences and P-RnaPredict succeeds in predicting structures with higher true positive base pair counts and lower false positives than mfold on specific sequences.

9) Consensus: From the analysis of the proceeding methods, it can be seen that each of these methods has its advantages and disadvantages and if different methodologies are combined by using consensus algorithms, a Meta server can be built with a more reliable prediction and more stable performance.

Pcons was the first fully automated Meta server that worked by collecting the outputs of six different publicly available protein fold recognition servers [41]. Pcons used a set of neural networks to predict the quality and accuracy of the collected models. Pcons was specifically trained to predict the quality of the final models. It allocates higher final scores to folds that were predicted by more than one server. All Meta servers made available since then work on a similar basis; they select their final answer from a set of results, using a consensus approach. The strength of Meta servers lies in the theory that mistakes in predicted models are likely to be random, while accurate models will occur at a frequency greater than random.

Libo Yu developed a consensus-based server for protein Fold Recognition [42]. A consensus-based server combines the outputs of several individual servers and generates better predictions than any individual server. Libo Yu proposed a Support Vector Machine (SVM) regression-based consensus method for protein fold recognition. SVM first extracts the features of a structural model by comparing the model to the other models produced by all the individual servers. Then, the SVM predicts the quality of each model. The experimental results from several data sets showed that the proposed consensus method, SVM regression outperforms any individual server.

Riccardo Lovsey describes the complete systematic development and benchmarking of an ensemble system for protein fold recognition, and examines the reasons behind the resultant improvements in performance [43]. Ensemble methods are learning algorithms that

construct a set of classifiers and then combine their individual decisions in some way to classify new examples. This ensemble system is designed to carry out a wide variety of fold recognition methods, searching a database of known structures. These methods include profile-profile, secondary structure, and structure-specific gapped alignment algorithms. These methods are optimized and tested using strictly selected protein data sets consisting of disparate subsets of the Structural Classification of Proteins (SCOP) database. Analyses showed that there is an increase in recognition accuracy due to the effect of 'noise filtering' by using multiple recognition algorithms in a consensus approach.

V. CONCLUSION

Many researches are conducted to invent new models for addressing the protein folding recognition problem. They rely on different supporting techniques like Genetic algorithms, support vector machines, Hidden Markov models, Multi-objective evolutionary algorithms, data integration techniques and ensemble classifiers. The accuracy, energy function, fitness score function, and the speed are all very significant factors when building Protein folding recognition model. There are two aspects of protein fold recognition problem: first is the computational difficulty and second is that the current energy functions are still not accurate enough to calculate the free energy of a given conformation. Computational difficulty can be solved by parallelization of one of the evolutionary methods so it can give a high performance. Also ensemble systems are considered one of the most powerful tools to recognize the correct fold.

REFERENCES

[1] Raymond Chang. Chemistry. McGraw Hill, 4 edition, 1991.

[2] Thomas E. Creighton. Proteins, Structures and molecular properties.W. H. Freeman and Company, New York, 2nd edition, 1993.

[3] G. N. Ramachandran and V. Sasiskharan. Conformation of polypeptides and proteins. Adv. Protein Chem., 23:283-437, 1968

[4] J. E. Wampler. Tutorial on peptide and protein structure. http://bmbiris.bmb.uga.edu/wampler/tutorial/.

[5] Carl Branden and John Tooze. Introduction to protein structure. Garland Publishing, Inc., New York and London, 1991.

[6] Robert Matthew MacCallum. "Computational Analysis of Protein Sequence and Structure," Ph.D. thesis, university College London, September 1997.

[7] Eaton E. Lattman. CASP4. Proteins: Structure, Function, and Genetics, 44(4):399, 2001.

[8] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Research, 28:45-48, 2000.

[9] Richard Bonneau and David Baker. Ab initio protein structure prediction: progress and prospects. Annual Review Biophysics Biomolecular Structure, 30:173{89, 2001.

[10] Richard H. Lathrop et al. Computational Methods in Molecular Biology, chapter 12, pages 227-283. Elsevier Press, Amsterdam, 1998.

[11] An-Suei Yang and Barry Honig. Sequence to structure alignment in comparative modeling using PrISM. PROTEINS: Structure, Function and Genetics Supplement, 3:66-72, 1999.

[12] David Baker. A surprising simplicity to protein folding. Nature, 405:39-42, 2000.

[13] Ken A Dill, S Banu Ozkan, M Scott Shell, and Thomas R Weikl, "The protein folding problem," Annual review of biophysics, 37:289–316, 2008.

[14] Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N. and Yeh, L.L. 'UniProt: the universal protein knowledge base' Nucleic Acids Research, Vol. 32, pp.D115–D119, 2004.

[15] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. 'The Protein Data Bank', Nucleic Acids Research, Vol. 28, pp.235–242, 2000.

[16] Akutsu, T. and S. Miyano, On the approximation of protein threading. Theoret. Comput. Sci., 210: 261-275. DOI: 10.1016/S0304-3975(98)00089-9, 1999.

[17] Lathrop, R., The protein threading problem with sequence amino acid interaction preferences is NP-complete. Protein Eng. Des. Select., 7: 1059-1068, 1994.

[18] Jones, D.T., "GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences," J. Mol. Biol., vol. 287, pp. 797-815, 1999.

[19] Lin, K., A.C.W. May and W.R. Taylor, "Threading Using Neural NETwork (TUNE): The measure of protein sequence-structure compatibility," Bioinformatics, vol. 18, pp. 1350-1357, 2002.

[20] Mcguffin, L.J. and D.T. Jones, "Improvement of the GenTHREADER method for genomic fold recognition," Bioinformatics, vol. 19, pp. 874-881, 2003.

[21] Jiang, N., W.X. Wu and I. Mitchell, "Protein fold recognition using neural networks and support vector machines," Proceeding of the 6th International Conference on Intelligent Data Engineering and Automated Learning-IDEAL, July 6-8, 2005.

[22] W. Thomas, C Igel, Jutta Gebert, "Protein Fold Class Prediction Using Neural Networks with Tailored Early-Stopping," in International Joint Conference on Neural Networks IJCNN, 2004.

[23] Raval, A., Z. Ghahramani and D.L. Wild, "ABayesian network model for protein fold and remote homologue recognition," Bioinformatics, vol.18, pp.788-801. 2002.

[24] Taylor, W.R. and I. Jonassen, "A structural pattern-based method for protein fold recognition. Proteins," Struct. Funct. Bioinformatics, vol. 56, pp. 222-234. 2004.

[25] Xu, J., "Fold recognition by predicted alignment accuracy," IEEE/ACM Trans. Comput. Biol. Bioinform., 2: 157-165, 2005.

[26] Sangjo Han, Byung-chul Lee, Seung Taek Yu, Chan-seok Jeong, Soyoung Lee and Dongsup Kim, "Fold recognition by combining profile-profile alignment and support vector machine," Bioinformatics, vol.21, pp. 2667-2673. 2005.

[27] W. Chmielnicki, K. Stapor, "Protein fold recognition with combined SVM-RDA classifier," in: Proceedings of the HAIS 2010, Part I, LNAI, vol. 6076, pp. 162-169. 2010.

[28] Dandekar, T. and P. Argos, "Potential of genetic algorithms in protein folding and protein engineering simulations," Protein Eng. Des. Select., vol.5, pp. 637-645. 1992.

[29] Yanev, N. and R. Andonoy, "Solving the protein threading problem in parallel," Proceedings of the 17th International Symposium on Parallel and Distributed Processing, Apr. 22-26, IEEE Computer Society, Washington, DC., USA., pp: 157, 2003.

[30] Yadgari, J., A. Amir and R. Uunger, "Genetic threading," Constraints, 6: 271-292, 2001.

[31] Unger, R., "The genetic algorithm approach to protein structure prediction," Struct. Bond., vol.110, pp.153-175, 2004.

[32] Judy, M.V. and K.S. Ravichandran. "A solution to protein

folding problem using a genetic algorithm with modified keep best reproduction strategy,"Proceeding of IEEE Congress on Evolution Computation, Sep. 25-28, IEEE Xplore Press, Singapore, pp. 4776-4780, 2007.
[33] Liang, F. and W.H. Wong, "Evolutionary monte carlo for protein folding simulations," J.Chemi. Phy., vol. 115, pp. 3374-3380, 2001.
[34] Carpio, C.A.D., Sasaki, S.I., L. Baranyi and H. Okada, "A parallel hybrid GA for peptide 3D structure prediction," Proceedings of the Workshop on Genome Informatics, Dec. 11-12, Universal Academy Press, Tokyo, 1995.
[35] Nguyen, D.H., Yoshihara, I. Yamamori, K. and Yasunaga, M. "Aligning multiple protein sequences by parallel hybrid genetic algorithm," Genome Inform., vol.13, pp. 123-132, 2002.
[36] Day, R.O., G.B. Lamont and R. Pachter, "Protein structure prediction by applying an evolutionary algorithm," Proceedings of the International Symposium on Parallel and Distributed Processing, Nice, France, pp. 155-162, Apr. 22-26, 2003.
[37] Islam, R. and A. Ngom, "Parallel evolution strategy for protein threading," Proceedings of the 25th International Conference on Chilean Computer Science Society, IEEE Computer Society, Washington, DC., USA., pp. 74, Nov, 07-11, 2005.
[38] Alione, N., "Parallel evolution strategy on grids for the protein threading problem," J. Parallel Distributed Computing, vol. 66, pp. 1489-1502, 2006.
[39] Thomas, S. and N.M. Amato, "Parallel protein folding with STAPL," Proceedings of the 18th International Parallel and Distributed Processing Symposium, IEEE Computer Society, Washington, DC., USA., pp. 189, Apr. 26-30, 2004.
[40] Wiese, K.C and A. Hendriks, "A detailed analysis of parallel speedup in P-RnaPredict-an evolutionary algorithm for RNA secondary structure prediction," Proceeding of the IEEE Congress on Evolutionary Computation, IEEE Computer Society, Washington, DC., USA., pp. 2323-2330. July 16-21, 2006.
[41] Lundstrom, J., Rychlewski, L., Bujnicki, J. & Elofsson, A., "Pcons: a neural network based consensus predictor that improves fold recognition," Protein Sci.,vol.10, pp. 2354–2362, 2001.
[42] Xu J, Yu L, Li M,"Consensus fold recognition by predicted model quality," Asian-Pacific Bioinformatics Conference (APBC) , 105-116. 2005.
[43] Riccardo Lovsey,"Development of an Enhanced Fold Recognition Ensemble System for Protein Structure Prediction," Ph.D. thesis, University of London, London, England, September 2006.