Journal of Computer Science 4 (9): 768-776, 2008
ISSN 1549-3636

A Survey of Protein Fold Recognition Algorithms

Mohammed Said Abual-Rub and Rosni Abdullah
School of Computer Sciences, University Sains Malaysia, 11800 Penang, Malaysia

Abstract: Problem statement: Predicting the tertiary structure of proteins from their linear sequence is really a big challenge in biology. This challenge is related to the fact that the traditional computational methods are not powerful enough to search for the correct structure in the huge conformational space. This inadequate capability of the computational methods, however, is a major obstacle in facing this problem. Trying to solve the problem of the protein fold recognition, most of the researchers have examined the use of the protein threading technique. This problem is known as NP-hard; researchers have used various methods such as neural networks, Monte Carlo, support vector machine and genetic algorithms to solve it. Some researchers tried the use of the parallel evolutionary methods for protein fold recognition but it is less well known. **Approach:** We reviewed various algorithms that have been developed for protein structure prediction by threading and fold recognition. Moreover, we provided a survey of parallel evolutionary methods for protein fold recognition. **Results:** The findings of this survey showed that evolutionary methods can be used to resolve the protein fold recognition problem. **Conclusion:** There are two aspects of protein fold recognition problem: First is the computational difficulty and second is that current energy functions are still not accurate enough to calculate the free energy of a given conformation.

Key words: Protein fold recognition, protein threading, evolutionary methods, parallel evolutionary methods

INTRODUCTION

It is a great challenge for nowadays biologists to predict the three-dimensional structure of a protein from its linear sequence. Proteins, amino acid chains, are made up from 20 different amino acids that are folded into unique three-dimensional protein structures. These structures are determined by their sequence of amino acids.

In the mean time, there are two experimental methods available for determining the three-dimensional structure of a protein from its amino acid sequence: X-ray crystallography and Nuclear Magnetic Resonance (NMR). Unfortunately, these methods are not efficient enough and that is due to the fact that they are expensive and time-consuming. As a result, there is a bad need for a fast and reliable computational method to predict structures from protein sequences -especially since the number of completely-sequenced genomes is growing very fast.

Biologists have recognized that proteins could have similar structural folds even if they have no sequence similarity or functional similarity. In fact, the total number of structural folds in nature is very small compared to the number of known protein sequences. (Fold recognition methods try to recognize the structural fold of a protein from a structure template library, given its sequence information then generate an alignment between the query and the recognized template protein, from which the structure of query protein can be predicted). Fold recognition methods are so efficient especially in the following cases: First, when the sequence has little or no primary sequence similarity to any sequence with a known structure. Second, when some model from the structure library represents the true fold of the sequence.

Although there have been many tests and developments of the different fold recognition methods, researchers have found out two main 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. The size of the conformation space is huge. Both of Lathro[13] and Akutsu *et al*.[1] have argued that the protein threading problem is NP-complete and MAX-SNP-hard.

Many techniques; such as Monte Carlo, Molecular Dynamics, Neural Network and Genetic Algorithms,

Corresponding author: Mohammed Said Abual-Rub, School of Computer Sciences, University Sains Malaysia, 11800 Penang, Malaysia

have been used to face the computational difficulty. Moreover, researchers, namely Yadgari et al.[28], Liang and Wong[14], Krasnogor[12], Unger[24] have used evolutionary methods to solve the protein fold recognition. On the other hand, researchers like[2,5,8,19], have used some parallel methods to solve the problem.

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[9]. There are four steps for the threading technique for 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[9]

Protein threading: Aligning the query to the template is the main element (componont) of the protein threading problem. The second step is to figure out the best alignment among all possible alignments between the query and the template and that is by looking for an alignment that produces a proper score function Yanev[29].

A query is defined as a sequence of amino acids of a protein. A template, however, is the three-dimensional coordinates of all atoms for each amino acid in the sequence which is known as a series of cores (such as α-helix, β-sheet), loops, links and turns. Threading a query against a template is to determine which basic folds the amino acids of the query can fit and then compute the free energy of the query[8]. The word threading implies that we drag the sequence step by step through each location on each template, but in fact we are searching for the best arrangement of the sequence on that template, as measured by some scoring function.

The protein threading process is shown in Fig. 1.

Threading is a difficult computational problem and has been described and proved to be NP-complete[13] and hence should be addressed by effective heuristics. Also it has been proved that the protein threading problem is MAX-SNP-hard, which means that it cannot be approximated to an arbitrary accuracy in polynomial time[1].

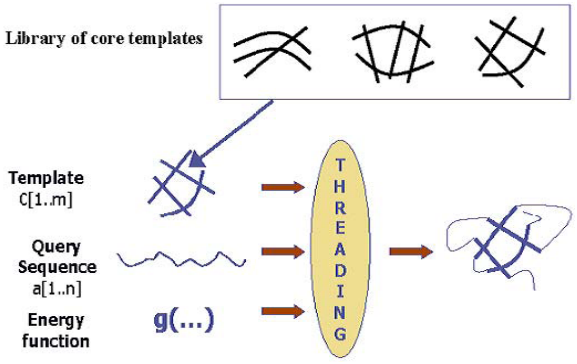

Fig. 1: Protein threading process

MATERIALS AND METHODS

Overview of protein fold recognition methods: Many researchers have tried different techniques; such as Molecular Dynamics, Monte Carlo, Genetic Algorithms and Neural Network in order to face the computational difficulty of protein fold recognition problem. The following part, however, discusses different successful approaches for protein fold recognition.

Neural network: Jones[10] introduces a new method for fold recognition. This method uses a traditional sequence alignment algorithm to produce alignments which are then evaluated by a method derived from threading techniques. Each threaded model is evaluated by a neural network in order to produce a single measure of confidence in the proposed prediction. So his study can be divided into three stages: (i) alignment of sequences, (ii) calculation of pair potential and salvation terms and (iii) evaluation of the alignment using a neural network. Jones[10] implemented GenTHREADER Protocol and GenTHREADER program. This method has been applied to the genome of Mycoplasma genitalium, his results shows that as many as 46% of the proteins derived from the predicted protein coding regions have a significant relationship to a protein of known structure. In some cases, only one domain of the protein can be predicted, giving a total coverage of 30% when calculated as a fraction of the number of amino acid residues in the whole proteome. The authors claim 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).

In terms of developing the method further, the authors claimed that this approach could easily be extended to take into account any number of input parameters and any sources of sequence-structure

information. We can note 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. Another note that it is able to find the correct answer for the vast majority of 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*[11]. have trained an artificial neural network model to predict compatibility of amino acid sequences with structural environment. They called their program TUNE (Threading Using Neural nEtwork). But their model is not trained to discriminate native protein structures. They tested their model 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. They used the protein structure classification CATH to select training and test sets. All the native structures in the decoy sets used for assessing ANN models.

Mcguffin and Jones[15] 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.

Nan Jiang *et al.*[17] proposed a new fold recognition model with mixed environment-specific substitution mapping (called MESSM) with three key features: (i) a structurally-derived substitution score is generated using neural networks. (ii) a mixed environment-specific substitution mapping is developed by combing the structural-derived substitution score with sequence profile from well-developed sequence substitution matrices. (iii) a support vector machine is employed to measure the significance of the sequence-structure alignment. They tested their model on two benchmark problems ;Wallner's Benchmark and Fischer's Benchmark, the model MESSM was found to lead to a good performance on protein fold recognition.

Bayesian networks: Raval *et al.*[20] present a Bayesian network approach for protein fold and superfamily recognition. The Bayesian network approach is a framework which combines graphical representation and probability theory, which includes, as a special case, hidden Markov models[20]. They introduced a novel implementation of a Bayesian network that can learn amino acid sequence, secondary structure and residue accessibility for proteins of known three-dimensional structure. They claimed that the cross validation experiments using Bayesian classification demonstrate that the Bayesian network model which incorporates structural information outperforms a hidden Markov model trained on amino acid sequences alone.

Structural pattern-based methods: Taylor and Jonassen[22] developed a method for evaluation of protein models based on residue packing interactions. Their method was described to evaluate the register of a sequence on a structure based on the matching of structural patterns against a library derived from the protein structure databank. The computer program that implemented the method is called SPREK (Sequence-structure Pattern-matching by Residue Environment Comparison). The authors claimed that the performance of SPREK on the decoy models was equivalent to those obtained with more complex approaches. Compared to previous methods, their approach is very straightforward. There are no large tables of potentials or any large weight matrices. Despite its simplicity, their method did not discard structural information as occurs in the majority of methods that consider only pairwise residue interactions. The authors maintained a description of the structure environment around a residue, including the sequential order of the residues in the environment and their secondary structure state. A major advantage of their method is its ability to operate using only the α-carbon atom positions.

Support Vector Machine (SVM): Xu[27] presented a Support Vector Machine (SVM) regression approach to directly predict the alignment accuracy of a sequence-template alignment. The authors implemented experiments on a large-scale benchmark using their Support Vector Machine (SVM) regression approach. They claimed 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 performs better than SVM classification method. In addition, SVM regression method enables the threading program to run faster than the composition-corrected Z-score method.

Sangjo Han *et al.*[6] resented an alternative method for estimating the significance of the alignments. The took a query of a protein and aligned it to a template of length n in the fold library, 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. According to their results, the new method gave significantly better performance than PSI-BLAST and profile-profile alignment with Z - score scheme. The authors claimed that the reason that SVM worked so well is related to the intermediate sequence search and its ability to recognize the essential features among alignments of remotely related proteins.

Evolutionary methods:

Genetic algorithms: The first study to introduce GAs to the field of protein structure prediction was that of[4]. They introduced GAs as a new tool to study proteins. Their research showed that the genetic algorithm simulation which mimicked important folding constraints as overall hydrophobic packaging and a propensity of the betaphilic residues for trans positions achieved a unique fold.

Unger and Moult[24] have developed a genetic algorithm search procedure suitable for use in protein folding simulations. They used GAs to fold proteins on a two-dimensional square lattice in the HP model. They maintained a population of conformations of the polypeptide chain and changed the conformations by mutation, in the form of conventional Monte Carlo steps and crossovers in which parts of the polypeptide chain are interchanged between conformations. For folding on a simple two-dimensional lattice it was found that the genetic algorithm is dramatically superior to conventional Monte Carlo methods.

Schulze-KremerS and TiedemannU[21] used a genetic algorithm to search energetically and structurally favorable conformations. They used a hybrid protein representation, three operators to manipulate the protein genes and a fitness function based on a simple force field.

Yadgari *et al*.[28] addressed the genetic algorithm paradigm used to perform sequence to structure alignments. The sequence-structure pairs in their research 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 has been discussed in which genetic operators can be effectively implemented. Their representation consists of numbers usually zeros and ones or integer number when there is a sequence deletion; an example of representation is 11110011311 (1 means a position of sequence on structure, 0 means structure deletion any other number, algorithms and implemented it for protein folding problem. They proposed a new intermediate selection step, which they called as Modified Keep-Best like 3 in the example, means sequence deletion). The authors claimed that the algorithm performance is tested for a set of six sequence-structure pairs. The effects of changing operators and parameters are explored and analyzed. The data they have presented indicate that the Genetic Algorithms method is a feasible and efficient approach for threading.

The authors claimed that genetic algorithms threading is quite robust and is not overly dependent on the particular selection of parameter or operators.

Unger[24] addressed the problem of protein structure prediction and protein alignments by using genetic algorithms. It is widely recognized that one of the major obstacles in addressing this question is that the "standard" computational approaches are not powerful enough to search for the correct structure in the huge conformational space. Genetic algorithms, a cooperative computational method, have been successful in many difficult computational tasks. Thus it is not surprising that in recent years several studies were performed to explore the possibility of using genetic algorithms to address the protein structure prediction problem.

In this study, a general framework of how genetic algorithms can be used for protein structure prediction was described. Using this framework, the significant studies that were published in recent years are discussed and compared. Applications of genetic algorithms to the related question of protein alignments are also mentioned. The rational of why genetic algorithms are suitable for protein structure prediction is presented.

The author claimed that GAs are efficient general search algorithms and as such are appropriate for any optimization problem, including problems related to protein folding. The author 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. In this way it might be possible to emphasize different aspects of the fitness function in different stages of folding. Another possibility is to introduce explicit memory into the emerging substructure, such that substructures that have been advantageous to the structures that accommodate them will get more level of immunity form changes.

M.V.Judy and K.S.Ravichandran[16] proposed a new intermediate selection strategy for genetic Reproduction (MKBR) to overcome the problem that the parents may be worse than the children as it is known in GA in practice. The new selection method

ensures that new genetic information is entered into the gene pool, as well as good previous genetic material is being preserved.They have demonstrated the superiority of modified keep-best reproduction on several instances of the proteinfolding problem, which not only finds the optimum solution, but also finds them faster than the standard generational replacement schemes.

Evolutionary monte carlo: Monte Carlo methods have traditionally been employed to address the protein folding problem. Monte Carlo algorithms based on minimizing the energy function, through a path that does not necessarily follow the natural folding pathway. The GA approach incorporates many Monte Carlo concepts[24].

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[14]. To resolve this problem, Liang and Wong[14] proposed an Evolutionary Monte Carlo (EMC) approach for protein folding simulations. They demonstrated that EMC can be applied successfully to simulations of protein folding on simple lattice models and to finding the ground state of a protein.

The authors claimed that in all cases, EMC is faster than the genetic algorithm and the conventional Metropolis Monte Carlo and in several cases it finds new lower energy states. The authors also proposed one method for the use of secondary structure in protein folding; their numerical results showed that it is extremely superior to other methods in finding the ground state of a protein. But, the authors just have considered only 2D HP models and they claim that the extension to 3D HP and real protein models is straightforward.

Parallel Evolutionary Methods (PEM) for protein fold recognition: Many researchers used parallel methods to solve the protein fold recognition problem in recent studies. While some researchers also used parallel methods to solve RNA sequence problem. There are three domains of biological sequences, namely DNA, RNA and protein. Some research mainly deals with the alignment in one domain. However, the method can be easily extended to deal with other domains. So in the following part, some parallel evolutionary methods for biological structure prediction will be discussed.

Parallel genetic algorithms:

Parallel hybrid gas: Carpio *et al.*[3] were the first to present a parallel hybrid genetic algorithm for three dimensional structure predictions of polypeptides. Their previous research based on a simple genetic algorithm was insufficient to produce better fit conformers, so they have proposed an improvement 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*.[3] 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.

In 2002, Nguyen *et al*.[19] proposed a parallel hybrid genetic algorithm for solving the sum-of-pairs multiple protein sequence alignment problem. They present a new GA-based method for more efficient multiple protein sequence alignment.

It is well known that the sum-of-pairs multiple sequence alignment problem can be exactly solved by the dynamic programming algorithm. However, this algorithm requires a running time which grows exponentially in proportion to the size of the problem. The majority of multiple sequence alignment heuristics is now carried out using the progressive approach (e.g., CLUSTALW, MULTAL, T-COFFEE), however, its main disadvantage is the local minimum problem. This study presented 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 authors claimed that experimental results of benchmarks from the BAliBASE 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 to find multiple sequence alignment as well as testing cost functions.

Island parallel gas: Anbarasu *et al*.[2] presented an evolution-based approach for solving multiple molecular sequence alignment. The approach is based on the island Parallel Genetic Algorithm (iPGA) that relies on the fitness distribution over the population of alignments. The algorithm searches for an alignment among the independent isolated evolving populations by optimizing weighted sum of pairs-objective function which measures the alignment quality.

Some of the most widely used multiple molecular sequence alignment packages like ClustalW, Mutal and Pileup are based on dynamic programming. They have advantages of being fast and simple as well as

reasonably sensitive, but their main drawback is the local minimum problem. In this study, the authors describe an iPGA strategy that runs on a distributed network of workstations.
Their parallel approach was implemented on PARAM 10000; a parallel machine developed at the Center of Development of Advanced Computing, Pune and is shown to consistently perform better than the sequential genetic algorithm. The algorithm generated alignments that were qualitatively better than an alternative method, ClustalW.

Multi-objective fmGA: Previous research using the Simple Genetic Algorithm (GA), messy GA (mGA), fast messy GA (fmGA) and Linkage Learning GA (LLGA) has made progress on this problem. However, past research used off-the-shelf software such as GENOCOP, GENESIS and mGA[5]. Day *et al.*[5] presented a modified fmGA as multi-objective implementation of the fmGA (MOfmGA) and a farming model for the parallel fmGA for protein structure prediction. The authors focused on tuning fmGA in an attempt to improve the effectiveness and efficiency of the algorithm in solving a protein structure and in finding 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 a serial run time and computational results support their hypothesis that the MO version provides more acceptable results.

Parallel evolution strategy: Islam and Ngom[8] proposed an evolution strategy for protein threading and also developed two parallel approaches for fast threading based on an evolution strategy for protein threading. The parallelization is based on master-slave architecture. Their novel approach, for protein threading, is based on evolution strategy. The Single Query Single Template Parallel ES Threading (SQST-PEST) method threads one query against one template; The Single Query Multiple Templates Parallel ES Threading (SQMT-PEST) method threads one query against a set of templates. The parallelization is based on master-slave architecture.

They used High Performance Computing environment, SHARCNET (Shared Hierarchical Academic Research Computing Network) as computing platform for experiment.

The authors claimed that the two parallel approaches have obtained at least better results than current comparable approaches, as well as significant reduction in execution time, but they did not explain what they mean by "at least better".

Alioune Ngom[18] proposed a novel evolution strategy for the solution of the protein threading problem using evaluation strategy called EST. The author showed that with recombination, his EST algorithm gave much better results, both in energy and threading time, than an existing GA-based method. Without recombination, EST is comparable to the GA-based approach but much faster. He also proposed a parallel method for fast threading, his 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 planing 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, we generally cannot tell if the query's correct fold template is in the pool, nor can we always tell which is the correct fold even if it is there.

RnaPredict approach: Wiese and Hendriks[26] presented a parallel evolutionary algorithm called P-RnaPredict for RNA secondary structure prediction. P-RnaPredict is a fully parallel implementation of a coarse-grained distributed EA for RNA secondary structure prediction and is based on RnaPredict, a serial EA for the same purpose which encodes RNA secondary structures in permutations and includes two stacking-energy based thermodynamic models. 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 authors stated the results that P-RnaPredict was shown to 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.

Probabilistic roadmap methods: Thomas and Amato[23] introduced a new computational technique for studying protein folding that was based on probabilistic roadmap methods for motion planning. They claimed that their technique yielded an approximate map of a

protein's potential energy landscape that contains thousands of feasible folding pathways and they had validated their method against known experimental results. 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 reported their experiments parallelizing their method using STAPL, that is being developed in the Parasol Lab at Texas A&M. With STAPL, they were able to easily parallelize their sequential code to obtain scalable speedups.

RESULTS AND DISCUSSION

Many researchers used evolutionary methods to solve protein fold recognition problem and their results were promising as shown in Table 1, for the time problem many researchers tried to parallelize the problem and also got promising results as shown in Table 2.

Table 1: protein fold recognition approaches based on technique used

Method	Paper	Major contribution
Neural network	GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences[19]	A neural network method for fold recognition
	Improvement of the GenTHREADER method for genomic fold recognition[15,19]	Impovement of the GenTHREADER
	Protein fold recognition using neural networks and support vector machines[17]	A new fold recognition model with Mixed Environment Specific Substitution Mapping (MESSM)
Bayesian networks	A Bayesian network model for protein fold and remote recognition[20]	A Bayesian network approach for protein fold and superfamily recognition
Structural pattern-based method	A structural pattern-based method for protein fold recognition[22]	A method (SPREK) was developed for the evaluation of protein models based on residue packing interactions
Support Vector Machine (SVM)	Fold recognition by predicted alignment accuracy[27]	A SVM regression approach for protein fold recognition
	Fold recognition by combining profile- profile alignment and support vector machine[6]	An alternative method for estimating the significance of the alignments evaluated by SVM
	Protein Fold Recognition Using Networks and SVMs[17]	A new fold recognition model with mixed environment specific substitution mapping (MESSM)
Genetic algorithms	Potential of genetic algorithms in protein folding and protein engineering simulations[4]	Genetic algorithms for protein folding and protein engineering simulations
	Genetic algorithms for protein folding simulations[25]	Genetic algorithms to fold proteins on a two- dimensional square lattice in the HP model
	Genetic threading[28]	A genetic algorithm paradigm for protein threading
	The genetic algorithm approach to protein structure prediction[24]	A framework of genetic algorithms for protein structure prediction.
	A Solution to Protein Folding Problem Using a Genetic Algorithm with Modified Keep Best Reproduction [16]	A new intermediate selection strategy for genetic algorithms and implemented it for protein folding problem.
Monte Carlo	Evolutionary Monte Carlo for protein folding simulations[14]	Evolutionary Monte Carlo approach for protein folding simulations

Table 2: Parallel evolutionary methods for protein fold recognition

Method	Paper	Major contribution
Parallel genetic algorithms	A parallel hybrid GA for peptide 3-D structure prediction[3]	A parallel hybrid GA for peptide 3-D structure prediction
	Multiple molecular sequence alignment by island parallel genetic algorithm[2]	Island parallel genetic algorithm for multiple molecular sequence alignment
	Aligning multiple protein sequences by parallel hybrid genetic algorithm[19]	A parallel hybrid genetic algorithm for solving the sum-of-pairs multiple protein sequence alignment
Multi-objective Evolutionary Approach	Protein structure prediction by applying an evolutionary algorithm	Multiobjective implementation of the fmGA (MOfmGA) and a farming model for the parallel fmGA
	Multi-class protein fold recognition using multi-objective evolutionary algorithms[5]	A Multi-Objective Feature Analysis and Selection Algorithm (MOFASA) for protein fold recognition
Parallel strategy evolution	Parallel evolution strategy for protein threading[8]	A novel approach based on evolution strategy for protein threading
	Parallel evolution strategy on grids for the protein threading problem[18]	A novel evolution strategy for the protein threading problem using evaluation strategy EST.
P-Rna predict Approach	A detailed analysis of parallel speedup in P-RnaPredict-an evolutionary algorithm for RNA secondary structure prediction[26]	A parallel evolutionary algorithm for RNA secondary structure prediction A new computational technique for studying protein folding that is based on probabilistic roadmap methods for motion planning
Probabilistic roadmap methods	Parallel protein folding with STAPL[23]	

CONCLUSION

This survey has shown that evolutionary methods can be used to resolve the protein fold recognition problem. We can see that there are two aspects of protein fold recognition problem: first is the computational difficulty and second is that current energy functions are still not accurate enough to calculate the free energy of a given conformation

However, the computational difficulty can be solved by parallelization of one of the evolutionary methods so it can give a high performance. An efficient parallel evolutionary method with an accurate energy function will be a good idea for my future research.

ACKNOWLEDGEMENT

This research has been funded by Universiti Sains Malaysia Fundamental Research Grant Scheme (FRGS) for "Parallel Conformational Search Algorithm for Protein Tertiary Structure Prediction Using Honey Bee Colony Optimization" [203/pkomp/671184].

REFERENCES

1. Akutsu, T. and S. Miyano, 1999. On the approximation of protein threading. Theoret. Comput. Sci., 210: 261-275. DOI: 10.1016/S0304-3975(98)00089-9.
2. Anbarasu, L.A., P. Narayanasamy and V. Sundararajan, 2000. Multiple molecular sequence alignment by island parallel genetic algorithm. Curr. Sci., 78: 858-863. http://www.ias.ac.in/currsci/apr102000/researcharticles2.pdf.
3. Carpio, C.A.D., Sasaki, S.I., L. Baranyi and H. Okada, 1995. A parallel hybrid GA for peptide 3D structure prediction. Proceedings of the Workshop on Genome Informatics, Dec. 11-12, Universal Academy Press, Tokyo. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.35.9052.
4. Dandekar, T. and P. Argos, 1992. Potential of genetic algorithms in protein folding and protein engineering simulations. Protein Eng. Des. Select., 5: 637-645. http://peds.oxfordjournals.org/cgi/content/abstract/5/7/637.
5. Day, R.O., G.B. Lamont and R. Pachter, 2003. Protein structure prediction by applying an evolutionary algorithm. Proceedings of the International Symposium on Parallel and Distributed Processing, Apr. 22-26, Nice, France, pp: 155-162. http://www2.computer.org/portal/web/csdl/doi/10.1109/IPDPS.2003.1213291.
6. Sangjo Han, Byung-chul Lee, Seung Taek Yu, Chan-seok Jeong, Soyoung Lee and Dongsup Kim, 2005. Fold recognition by combining profile-profile alignment and support vector machine. Bioinformatics, 21: 2667-2673. DOI: 10.1093/bioinformatics/bti384.
7. Fischer, D., L. Rychlewski, R.L. Dunbrack, A.R. Ortiz and A. Elofsson, 2003. CAFASP3: The third critical assessment of fully automated structure prediction methods. Proteins, 53: 503-516. DOI: 10.1002/prot.10538.
8. Islam, R. and A. Ngom, 2005. Parallel evolution strategy for protein threading. Proceedings of the 25th International Conference on Chilean Computer Science Society, Nov, 07-11, IEEE Computer Society, Washington, DC., USA., pp: 74. http://portal.acm.org/citation.cfm?id=1115322.
9. Moult, J.J., F. Fidelis, A. Zemla and T. Hubbard, 2003. Critical assessment of methods on protein structure prediction (CASP)-round V. Proteins, 53: 334-339. http://www.ncbi.nlm.nih.gov/pubmed/14579322.
10. Jones, D.T., 1999. GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. J. Mol. Biol., 287: 797-815. DOI: 10.1006/jmbi.1999.2583.
11. Lin, K., A.C.W. May and W.R. Taylor, 2002. Threading Using Neural NETwork (TUNE): The measure of protein sequence-structure compatibility. Bioinformatics, 18: 1350-1357. http://bioinformatics.oxfordjournals.org/cgi/content/abstract/18/10/1350.
12. Krasnogor, N., B.P. Blackburne, E.K. Burke and J.D. Hirst, 2002. Multimeme algorithms for protein structure prediction. Proceedings of the International Conference on Parallel Problem Solving from Nature, Sep. 07-11, Springer, Berlin, London UK., pp: 769-778. http://portal.acm.org/citation.cfm?id=645826.669429.
13. Lathrop, R., 1994. The protein threading problem with sequence amino acid interaction preferences is NP-complete. Protein Eng. Des. Select., 7: 1059-1068. http://peds.oxfordjournals.org/cgi/content/abstract/7/9/1059.
14. Liang, F. and W.H. Wong, 2001. Evolutionary monte carlo for protein folding simulations. J. Chemi. Phy. 115: 3374-3380. DOI: 10.1063/1.1387478.
15. Mcguffin, L.J. and D.T. Jones, 2003. Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics, 19: 874-881. http://bioinformatics.oxfordjournals.org/cgi/content/abstract/19/7/874.

16. Judy, M.V. and K.S. Ravichandran. 2007. 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. DOI: 10.1109/CEC.2007.4425099.
17. Jiang, N., W.X. Wu and I. Mitchell, 2005. 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, Springer-Verlag Berlin Heidelberg, London UK., pp: 462.469. http://books.google.com.pk/books?id=XRqwMciw4TwC.
18. Alione, N., 2006. Parallel evolution strategy on grids for the protein threading problem. J. Parallel Distributed Computing,, 66: 1489-1502. DOI: 10.1016/j.jpdc.2006.08.005
19. Nguyen, D.H., Yoshihara, I. Yamamori, K. and Yasunaga, M. 2002. Aligning multiple protein sequences by parallel hybrid genetic algorithm. Genome Inform., 13: 123-132. https://www.jsbi.org/modules/journal1/index.php/GIW02/GIW02F013.pdf.
20. Raval, A., Z. Ghahramani and D.L. Wild, 2002. A Bayesian network model for protein fold and remote homologue recognition. Bioinformatics, 18: 788-801. http://bioinformatics.oxfordjournals.org/cgi/content/abstract/18/6/788.
21. Schulze-Kremer, S. and U. Tiedemann, 1994. Parameterising genetic algorithms for protein folding simulation. Presented at Colloquium on Molecular Bioinformatics, Februery Institute of Electrical Engineers, IEE Press, London, UK.
22. Taylor, W.R. and I. Jonassen, 2004. A structural pattern-based method for protein fold recognition. Proteins Struct. Funct. Bioinformat., 56: 222-234. DOI: 10.1002/prot.20073.
23. Thomas, S. and N.M. Amato, 2004. Parallel protein folding with STAPL. Proceedings of the 18th International Parallel and Distributed Processing Symposium. Apr. 26-30, IEEE Computer Society, Washington, DC., USA., pp: 189. DOI: 10.1109/IPDPS.2004.1303204.
24. Unger, R., 2004. The genetic algorithm approach to protein structure prediction. Struct. Bond., 110: 153-175. http://faculty.biu.ac.il/~unger/pub/ga_bonding.pdf.
25. Unger, R. and J. Moult, 1993. Genetic algorithms for protein folding simulations. J. Mol. Biol., 231: 75-81. http://www.ncbi.nlm.nih.gov/pubmed/8496967.
26. Wiese, K.C and A. Hendriks, 2006. 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, July 16-21, IEEE Computer Society, Washington, DC., USA., pp: 2323-2330. DOI: 10.1109/CEC.2006.1688595.
27. Xu, J., 2005. Fold recognition by predicted alignment accuracy. IEEE/ACM Trans. Comput. Biol. Bioinform., 2: 157-165. http://www2.computer.org/portal/web/csdl/doi/10.1109/TCBB.2005.24.
28. Yadgari, J., A. Amir and R. Uunger, 2001. Genetic threading. Constraints, 6: 271-292. DOI: 10.1023/A:1011489723652.
29. Yanev, N. and R. Andonoy, 2003. 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. http://portal.acm.org/citation.cfm?id=838237.838402.