International Journal of Electronics Signals and Systems

Volume 1 | Issue 2 Article 5

September 2011

Prediction of Protein Tertiary Structure using Genetic Algorithm

G. Sindhu
Department of Computer Science and Engineering, Thiagarajar College of Engineering, Madurai, India, sindhug3@gmail.com

S. Sudha
Department of Computer Science and Engineering, Thiagarajar College of Engineering, Madurai, India, ssj@tce.edu

Available online at www.interscience.in

Prediction of Protein Tertiary Structure using Genetic Algorithm

G.Sindhu, S.Sudha
Department of Computer Science and Engineering, Thiagarajar College of Engineering, Madurai, India
Email: sindhug3@gmail.com, ssj@tce.edu

Abstract - Proteins are essential for the biological processes in the human body. They can only perform their functions when they fold into their tertiary structure .Protein structure can be determined experimentally and computationally. Experimental methods are time consuming and high-priced and it is not always feasible to identify the protein structure experimentally. In order to predict the protein structure using computational methods, the problem is formulated as an optimization problem and the goal is to find the lowest free energy conformation. In this paper, Genetic Algorithm (GA) based optimization is used. This algorithm is adapted to search the protein conformational search space to find the lowest free energy conformation. Interestingly, the algorithm was able to find the lowest free energy conformation for a test protein (i.e. Met enkephalin) using ECEPP force fields.

Keywords - Protein Structure prediction problem, ECEPP force field, Genetic Algorithm, SMMP tool.

I. INTRODUCTION

The protein function is related to the protein structure. The protein structure can be described in four levels: primary, secondary, tertiary and quaternary. The primary structure is a sequence of amino acids connected by peptide bonds. Amino acids are the building blocks of the protein. There are 20 amino acid types where each amino acid consists of a main or backbone and side chain. The main chain is the same in all the 20 amino acid type. Differences are in the side chain. Proteins differ from each other by the order or number of amino acids. The secondary structure occurs when the sequence of amino acids are attracted by hydrogen bonds. Tertiary structure is the three dimensional arrangements of the atoms. Quaternary structure consists of more than one amino acid chain [20].

The protein structure prediction problem is regarded as a grand challenge and is one of the great puzzling problems in computational biology. It is how to get the structure of the protein given only its sequence. This problem can be solved experimentally using experimental methods such as NMR and X-ray Crystallography. Experimental methods are the main source of information about protein structure and they can generate more accurate results. However, they are also time consuming where the determination of the structure of a single protein can take months and they are expensive, laborious and need special instruments as well. Moreover and due to some limitations in the experimental methods, it is not always feasible to determine the protein structure experimentally which results in creating a big gap between the number of protein sequences and known protein tertiary structures. In order to bridge this gap, other methods are much needed to determine the protein structure. Scientists from many fields have worked to develop theoretical and computational methods which can help provide cost effective solutions for the protein structure prediction problem. Accordingly, the best existing alternative is using computational methods which can offer cost effective solutions. Computational methods can be traditionally divided into three approaches: Homology Modelling, Threading and Ab initio [11]. In Homology Modelling and Fold Recognition methods, the prediction is performed using the similarities between the target protein sequence and the sequences of already solved proteins structures. So, these methods are limited to predict the structure of proteins which belong to protein families with known structures. On the contrary, Ab initio methods are not limited to protein families with at least one known structure [3]. They are based on the Anfinsen hypothesis which states that the tertiary structure of the protein is the conformation with the lowest free energy. To predict the protein structure using Ab initio method, the problem is formulated as an optimization problem with the aim to find the lowest free energy conformation. In order to perform that, protein conformation must be represented in a proper representation. This representation is ranged from all atoms representation to simplified representation. Then, an energy function is used to calculate the conformation energy and a conformational search algorithm is utilized to search the conformation search space to find the lowest free energy conformation [2].

In this paper, we propose a simple GA for protein tertiary structure prediction. The performance of two real coded crossover operators of GA in protein structure prediction is compared. The target protein is Metenkephalin. The results show that GA has the higher searching capability. In this investigation we utilize the ECEPP energy model as a fitness function; the protein structure is determined by minimizing the energy fitness function.

The rest of the paper is organized as follows. Section 2 deals with the survey of related work. Section 3 highlights the proposed work of this paper. The experiments and results are presented in Section 4.Finally Section 5 concludes the paper stating its future scope.

II. RELATED WORK

Md Tamjidul et al. [1] proposes the impact of twins and the measures for their removal from the population of Genetic Algorithm when applied to effective conformational searching. Twins cause a population to lose diversity, resulting in both the crossover and mutation operation being ineffectual. In this paper the efficient removal of twins from the GA population is achieved with the help of two factors: 1) Chromosome Correlation Factor (CCF) and 2) Correlated Twin Removal (CTR) algorithm. It highlights the need for a chromosome twin removal strategy to maintain consistent performance.

Yunling Liu and Lan Tao [5] considering the deficiency of simple Genetic Algorithms, such as pre-maturity and slow convergence, they propose HPGA/GBX (Hybrid Parallel GA/Guide Blend Crossover) which is an improvement of GA and the algorithm evaluated with three standard test functions. In case of simple Genetic Algorithm, they had been taken the whole population as an input. But in the improved GA, the entire population is randomly divided into M sub-populations, which causes the resultant structure to handle the prematurity and slow convergence problem in a better way. The result shows that HPGA/GBX performs better in terms of searching and finding the minimum energy for small proteins. In this investigation they utilize the ECEPP energy model as a fitness function. The target protein is Met-enkephalin.

R.Day et al. [6] focuses on an energy minimization technique and the use of a multiobjective Genetic Algorithm to solve the Protein Structure Prediction (PSP) problem. They propose a multiobjective fast messy Genetic Algorithm (fmGA) to obtain a solution to this problem. They utilize the CHARMM force field as a energy function. This paper use binary string representation of proteins and it covers the analyses of two proteins: [Met]-enkephalin and Polyalanine. The operators used were cut and splice operator.

Madhusmita et al. [7] uses a real valued Genetic Algorithm, a powerful variant of conventional GA to simulate the PSP problem. The conformations are generated under the constraints of Ramachandran plot along with secondary structure information, which are then screened through a set of knowledge based biophysical filters,viz.persistence length and radius of gyration. This method uses Torsion angles representation. FoldX force field used as a fitness function. They use the Genetic Operators such as Mutate, Variate and crossover. The crossover operator further splited into two types one is 2-point crossover and another one is 1-point crossover. In this work they proposed a fast, efficient GA based approach for PSP.

Pallavi M.Chaudhri et al., [8] just shown that how Genetic Algorithm (GA) is efficiently used for predicting the protein structure. The test protein is crambin protein-a plant seed consisting of 46 amino acids. They used to describe the structure of protein as a list of three dimensional coordinates of each amino acid, or even each atom. Genetic Algorithms proved to be an efficient search tool for structural representations of proteins. It results in highly optimized fitness value.

Jie Song et al. [17] shown that Genetic Algorithm is an efficient approach to find lowest-energy conformation for HP lattice model. They had introduced some new operators to speed up the searching process and give the result with more biology significance. The operators used in addition are symmetric and corner change operators. They suggest that high rates of mating, mutation and relatively high elitism is good for getting an optimized result. The additional operators can speed the evolution and reduce the computation time.

The prediction problem has been proven to be NP-complete, implying that a polynomial time algorithm is not feasible either. Statistical approaches to the PSP problem include Contact Interaction and Chain Growth. Both these techniques are characterized by exhibiting lower accuracy as the sequence length increases and also by being non-reversible in their move-steps while searching for optimum conformation. Alternative PSP strategies include Artificial Neural Networks (ANN), Support Vector Machines (SVM) and Bayesian Networks (BN), while Hidden Markov Models (HMMs) which are based on Bayesian learning, have also been used to convert multiple sequence alignment into position-specific scoring matrices (PSSM), which are subsequently applied to predict protein structures. These approaches are often dependent on the training set and thus mostly applicable to the homology modelling and threading-based approaches rather than ab initio PSP problems. In particular, if the training sets are unrelated to the test sets, then information relating to a particular motif does not assist in a different motif. For deterministic approaches to the PSP problem, approximation algorithms provide an insight, though they are not particularly useful in identifying minimum energy conformations, and while linear programming (LP) methods have been used for protein threading, they have not been applied in abinitio applications, with the recent LP focus being confined to approximating the upper bound of the fitness value based on sequence patterns only. Therefore, non-deterministic search techniques have dominated attempts to solve the PSP problem, of which there are a plethora including Monte Carlo (MC) simulation, Evolutionary MC (EMC) Simulated Annealing (SA), Tabu Search with Genetic Algorithms (GTB), Ant Colony Optimization, Immune Algorithm (IA) based on Artificial Immune System (AIS), Conformational Space Annealing (CSA), and so on. Due to their simplicity and search effectiveness, Genetic Algorithms are very attractive especially for the crossover operation which can build new conformation by exchanging sub-conformations [1].

In this paper, Genetic Algorithm with Discrete Crossover (DC) and Mid-point Crossover (MC) operators for the test protein Met-Enkephalin has been proposed. Torsion angle representation model is used for protein representation. ECEPP force field is used as a fitness function.

III. PROPOSED SCHEME

This section is devoted to describe how the Genetic Algorithm was adapted to solve the protein conformational search problem in order to find the lowest free energy conformation.

D. Protein Conformation Representation

Each amino acid consists of two parts: the main chain and the side chain (Figure 1) [2]. The main chain torsion angles are: φ, ψ and ω. The side chain torsion angles are χn. As the overall structure of proteins can be described by their backbone and side chain torsion angles, the tertiary structure of a protein can be obtained by rotating the torsion angles around the rotating bonds. So, the protein conformation is represented as a sequence of the torsion angles. This representation is a common protein conformation representation and it is widely used in protein conformational search algorithms.

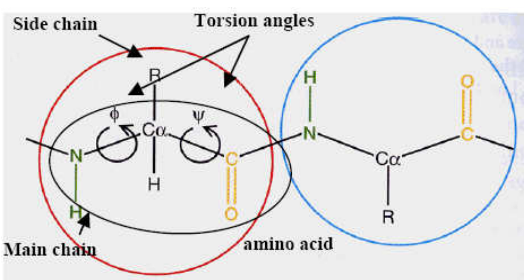

Figure 1. Amino Acid

In the torsion angles representation, each conformation is represented as an array of real values. These values are the values of the amino acid torsion angles. The length of the array represents the number of torsion angles of the protein. Generating conformations is done by changing the values of the torsion angles randomly.

E. Energy Function

The protein energy function is the objective function and the torsion angles are the variables. The conformation energy is calculated using ECEPP force fields which it is implemented as a part of the SMMP (Simple Molecular Mechanics for Proteins)

F. The Algorithm

In a GA, a population of chromosomes, representing a series of candidate solutions (called individuals) to an optimization problem, generally evolves toward better solutions. The evolution usually starts from a population of randomly generated individuals. In each generation, the fitness of every individual is evaluated, the best individuals are selected (elitism), and the rest of the new population is formed by the recombination of pairs of individuals, submitted to random mutations. The new population is then used in the next generation of the algorithm. Commonly, as employed in this problem, the algorithm ends when a maximum number of generations is reached.

GA is a technique of function optimization derived from the principles of evolutionary theory. The Genetic Algorithm is a heuristic method that operates on pieces of information like nature does on genes in the course of evolution. It has good global search characteristics. Three operators are invented to modify individuals: Selection, Mutation and Crossover. The decision about the application of an operator is made during run time and can be controlled by various parameters [5]. The basic outline of a Genetic Algorithm is as follows:

1) Initialize a population of individuals. This can be done either randomly or with domain specific background knowledge to start the search with promising seed individuals.
2) Evaluate all individuals of the initial population.
3) Generate new individuals. Operations to produce new individuals are: Selection, Mutation and Crossover.
4) Go back to step 2 until either a desired fitness value was reached or until a predefined number of iterations was performed (Termination Criteria).

Additionally two real coded crossovers Discrete Crossover (DC) and Mid-point Crossover (MC) are used along with boundary mutation. It produces an optimal solution.

IV. EXPERIMENTS AND RESULTS

The algorithm is implemented using Java in Linux environment. The SMMP package is used for ECEPP energy calculation. The algorithm is applied to find the lowest free energy conformation of Met-enkephalin, i.e. a small protein which is extensively used to test the conformational search methods. It consists of 5 amino acids with 24 torsion angles. Two types of real-coded crossovers are performed. The performances of the two crossovers are compared.

The number of population is set to 120 and the number of iterations is set to 500. The mutation rate is set to 0.01 and the crossover rate is set to 0.8.

TABLE II
PERFORMANCE OF CROSSOVERS

S.No	GA Operators		
	Crossover	**Mutation**	**Result(kcal/mole)**
1	Discrete Crossover (DC)	Boundary Mutation	-12.429
2	Mid-Point Crossover(MC)	Boundary Mutation	-9.3437

The results in table 1 describes that, the two real-coded crossovers produce the conformation which has low energy. It is observed that the success rate of GA with DC and MC is better than GA with simple crossover operators.

V. CONCLUSION AND FUTURE WORK

This paper used Genetic Algorithm with MC and HC/IC crossovers to search the protein conformational search space to find the lowest free energy conformation. The results indicated that the algorithm is able to find the lowest free energy conformation of -12.429 kcal/mol using ECEPP force field. Better results are gained using Discrete Crossover with boundary mutation.

Further work is needed to compare the performance of the algorithm on larger proteins and also to improve the performance of the algorithm by parallelizing and comparing the performance of the algorithm with other existing algorithms for protein conformational search.

REFERENCES

1. Md Tamjidul Hoque, Madhu Chetty, Andrew Lewis, and Abdul Sattar, *"Twin Removal in Genetic Algorithms for Protein Structure Prediction using Low-Resolution Model"*, IEEE/ACM Transactions on Computational Biology and Bioinformatics, TCBB-2008-06-0102.R2, 2009.
2. Heshm Awadh A.Bahamish, Rosni Abdullah and Rosalina Abdul Salam, *"Protein Tertiary Structure Prediction Using Artificial Bee Colony Algorithm"*, IEEE Third Asia International Conference on Modeling and Simulation, 2009.
3. WANG Cai-Yun, ZHU Hao-Dong and CAI Le-Cai, *"A new prediction protein structure method based on Genetic Algorithm and Coarse-grained protein model"*, IEEE 2nd International Conference on Biomedical Engineering and Informatics, 2009. BMEI '09.
4. Pawel Widera, Jonathan M.Garibaldi and Natalio Krasnogors, *"Evolutionary design of energy functions for Protein structure prediction"*. IEEE International Conference-2009.
5. Yunling Liu, Lan Tao, *"Protein Structure Prediction based on An Improved Genetic Algorithm"*, The 2nd IEEE International Conference on Bioinformatics and Biomedical Engineering, Shanghai, 2008, p: 577-580.
6. R.Day, J.Zydallis and G.Lamont, "Solving the Protein Structure Prediction Problem through a Multi-objective Genetic Algorithm", IEEE International Conference-2008.
7. Madhusmita, Harijinder Singh and Abhijitis Mitra, "Real valued Genetic Algorithm based approach for Protein Structure Prediction-Role of Biophysical Filters for Reduction of Conformational Search Space", IEEE International Conference-2008.
8. Mrs.Pallavi M.Chaudhri, Mr.Prasad P.Thute, *"Application of Genetic Algorithms in Structural Representation of Proteins"*, IEEE First International Conference on Emerging Trends in Engineering and Technology-2008.
9. Steffen Schulze-Kremer, *"Genetic Algorithms for Protein Tertiary Structure Prediction"*, Springer-Verlag London, UK, Pages: 262 - 279, ISBN: 3-540-56602-3, 1993.
10. Wen Yuan Liu,Shui Xing Wang,Bao Wen Wang,Jia Xin Yu,*"Protein Secondary Structure Prediction Using SVM with Bayesian Method"*, IEEE 2nd International Conference on Bioinformatics and Biomedical Engineering, 2008. ICBBE 2008.

11. Christine Kehyayan, Nashat Mansour, Hassan Khachfe,*"Evolutionary Algorithm for Protein Structure Prediction"*, IEEE International Conference on Advanced Computer theory and Engineering-2008.

12. T.W.de Lima, P.H.R.Gabriel, A.C.B.Delbern, R.A.Faccioli, I.N.da Silva, *"Evolutionary Algorithm to ab initio Protein Secondary Structure Prediction with Hydrophobic Interactions"*, IEEE International Conference-2007.

13. Michela Taufer, Chahm An, Andreas Kersten, Charles L.Brooks III, *"Predictor@Home: A "Protein Structure Prediction Supercomputer" Based on Global Computing"*, IEEE Transactions on Parallel and Distributed Systems-2006.

14. Yun-Ling Liu, Lan Tao, *"An Improved Parallel Simulated Annealing Algorithm used for Protein Structure Prediction"*, IEEE Fifth International Conference on Machine Learning and Cybernetics-2006.

15. Heshm Awadh A.Bahamish, Rosni Abdullah, Rosalina Abdul Salam, *"Protein conformational search Using honey Bee Colony optimization"*, IEEE Regional Conference on Mathematics, Statistics and Application-2006.

16. Rajkumar Bondugula, Dong Xu, Yi Shang, *"A fast algorithm for Low-Resolution Protein Structure Prediction"*, IEEE International Conference-2006.

17. Jie Song,Jiaxing Cheng,TingTing zheng and Junjun mao, *"A Novel Genetic Algorithm for HP Model Protein Folding"*, IEEE Sixth International Conference on Parallel and Distributed Computing, Applications and Technologies-2005.

18. Wanjun Gu, Tong Zhou, Jianmin Ma Xiao Sunand and Zuhong Lu, *"Folding Type Specific Secondary Structure Propensities of Synonymous Codons"*, IEEE Transactions on NanoBioScience-2003.

19. Richard O.Day, Gray B.Lamont and Ruth Pachter, *"Protein Structure Prediction by Applying an Evolutionary Algorithm"*, IEEE International Conference on Parallel and Distributed Processing-2003.

20. Satya Nanda Vel Arjunan, Safaai Deris and Rosli Md Illias, *"Literature Survey of Protein Secondary Structure Prediction"*, IEEE Journal Technology-2001, 34(C) 2001:63-72.

21. Y. Duan and P. A. Kollman , *"Computational protein folding: From lattice to all-atom,"* IBM System Journal, vol .40, no. 2, 2001

22. T. Pedersen and J. Moult, "Protein folding simulations with Genetic Algorithms and a detailed molecular description," J. Mol. Biol., vol.269, pp.240-259, 1997.

23. Vinicius Tragante do Ó1, Renato Tinós1,*"Diversity Control in Genetic Algorithms for Protein Structure Prediction"*. J R Soc Interface. 2006 February 22; 3(6): 139–151.

❑❑❑