RESEARCH ARTICLE OPEN ACCESS

Comparison of the 3D Protein Structure Prediction Algorithms

Fadhl M. Al-Akwaa, Husam Elhetari, Noman Al Naggar and Mahmoud A. Al-Rumaima
Biomedical Eng. Dept., Univ. of Science & Technology, Sana'a, Yemen

Abstract
Determining protein 3D structure is important to known protein functions. Protein structure could be determined experimentally and computationally. Experimental methods are expensive and time consuming whereas computational methods are the alternative solution. From the other hand, computational methods require strong computing power, assumed models and effective algorithms. In this paper we compare the performance of these algorithms. We find that Genetic Algorithm with improved search techniques can represents the best heuristic algorithm to predict 3D protein structure from AB off or HP lattice model.

I. Introduction

Protein is essentially sequence of amino acids which forms a particular 3D structure. Understanding 3D structure is very important to study protein function, understanding critical diseases and in drug design [1, 2].

Protein structure could be determined experimentally and computationally. Experimental methods are expensive and time consuming whereas computational methods are the alternative solution. From the other hand, computational methods require strong computing power, assumed models and effective algorithms.

1.1 Experimental methods:

1.1.1 X-ray crystallography:

X- ray crystallography is a technology that uses X ray to determine atoms positions inside the crystal with using electron density map the atom positions will be obtained[3]. This technique includes place protein under X-ray exposure and analyses electron direction patterns. X- ray crystallography is Excellent for rigid proteins and difficult for flexible proteins. This method is time consuming and expensive so it is not feasible[4].

1.1.2 NMR spectroscopy

In this method protein is purified then placed in strong magnetic field and radio frequency waves. This method is suitable for structure of flexible proteins but is limited to small or medium proteins [4].

1.1.3 Electron microscope

In this method a beam of electrons is used to image the molecule directly. This technique is used to determine large macromolecular complexes structures but researchers can't see each atom[4].

1.2 Computational methods:

1.2.1 Homology

In this method the query protein structure is compared with known structure of one or more proteins from Protein Data Bank to find homologous structure. If sequence similarity between two structure is good then sequence alignment is more accurate[4]. It is used when sequence similarity between two structures is greater than 35% [5]. When the sequence similarity is more than 40% that mean the less gaps be found and about 90 % of protein structure can be modeled with RMSD error about 1-2 Å and that can be equivalent to low resolution X-ray accuracy predictions[6] or medium resolution of NMR accuracy[4].

When sequence similarity between about 30-40% correct alignment becomes difficult and about 80 % of protein structure can be modeled with RMSD error about 3.5 Å that means large errors during modeling. When the sequence similarity is less than 30% the correct alignment becomes much more difficult and it is demanding on computer processing power[6].

1.2.2 Threading or fold recognition

Threading is improved technique that uses both sequence alignment and structural information of secondary or tertiary structure of protein to find the correct fold for the target protein[4, 7]. It is used when sequence similarity between two structures is greater than 25% [5].

This method is the best to predict protein with length about 100 residues but protein structure can be modeled with RMSD error about 2-6 Å. On the other hand, this method is very demanding on computer processing power and there is still a need for target sequences identification[4].

1.2.2 Ab anitio methods

Ab initio methods predict protein 3D structure without using any database information of previously solved protein structures. This method comes from the fact that the protein native structure conformation happens when its molecules attached together with the lowest free energy. So, when we predict the bath of lowest energy based on possible interactions between query protein sequences, we can say this bath is represent the protein real structure[6, 7].

Ab initio methods identify new structure not depend on comparison to known structure so it is alternative in some cases when comparative modeling may not be available[6].

Also, this method is the most promising with regard to providing reliability , accuracy , usability and flexibility in checking the functional divergence of a protein or drug[4]. From the other hand, Ab initio requires strong computing power and effective algorithm and also just applicable for small proteins sequences (<120 residues) [2, 4, 6]

In ab intio modeling, there are two important components for protein structure prediction; first is the structure models design which represent skeleton to predict final confirmation of real protein structure. Second is the optimization technology design to find the lowest free energy conformation between protein sequences to predict the native conformation of protein sequences structure.

In this paper, we compare the performance of some optimization methods when using HP and AB models.

1.3 Structure models design:

Lattice model is very important because it represent skeleton to predict final confirmation of real protein structure. So to predict protein structure the lattice model is essential in modeling of protein fold and it reduce complexity of computer processing by present amino acid with AB or HP sequences into 2D or 3D regular structure as cubic, square, face center cube(FCC) lattice models or any other Bravais Lattice[8].

Predication of optimal conformation of protein 3D structure using lattice model still an NP-complete, so we need optimization algorithms which will be described in below[8].

AB OFF and HP lattice model are the important design models which hydrophobic residues are wrapped inside by hydrophilic residues and hydrophilic residues expose to the surface which contacts with water[9].

1.3.1 AB off lattice model:

Amino acids can be represented by AB sequences according to strength of the contact between amino acids and water where (A) is hydrophobic amino acids and (B) hydrophilic amino acids[9].

To get minimum energy that predict conformation of real protein 3D structure we use off-lattice model with AB sequences.

In AB off lattice model the AB sequences are placed in three dimensional space as n monomers chain and the energy can be calculated[10].

For n amino acids E of protein sequences is:

$$E = \sum_{i=2}^{n-1} \frac{1}{4}(1-\cos\theta_i) + \sum_{i=1}^{n-2}\sum_{j=i+1}^{n} 4[r_{ij}^{-12} - c(\varepsilon_i,\varepsilon_j)r_{ij}^{-6}]$$

Where θ_i ($0\leq\theta_i\leq\pi$) is the angle between two successive bond vectors.

r_{ij} is the distance between residues i and j with i<j and it depends on both bond and torsional angle.

The constant $c(\varepsilon_i,\varepsilon_j)$ is +1, +1/2, and -1/2 for AA, BB and AB pairs respectively.

There is a strong attraction between AA pairs, a week attraction between BB pairs and a week repulsion between AB pairs.

In off lattice AB model, predicting 3D folding structure problem of n monomer chain is to find suitable n-2 bond angles and n-3 torsional angles which make energy function E minimum. Then we can get the lowest energy conformation of given protein sequence[11].

Considering Interaction between nonadjacent monomers and successive bond make this model more accurate than HP model[11].

AB off lattice still an NP-complete and it needs effective algorithm such as GA, Ant colony to enhance protein 3D structure prediction.

1.3.1 HP lattice model:

HP sequences are another representation of protein amino acid structure depend on relation of water and amino acid. All the amino acids are divided into two groups: hydrophobic H (Gly, Ala, Pro, Val, Leu, Ile, Met, Phe, Tyr, Trp); and hydrophilic or polar P (Ser, Thr, Cys, Asn, Gln, Lys, His, Arg, Asp, Glu)[12].

the hydrophobic amino acids(H) tend to be less exposed to the aqueous solvent than the polar ones(P) ,so a hydrophobic core will be form in spatial structure[13].

when we trace this HP sequence which restricted to self-avoiding paths on 3 dimensional sequence lattice we can get lowest energy that reflect the real folds of protein structure[13].

The free energy of a conformation is the negative number of non-consecutive hydrophobic-hydrophobic contacts[13].

The total energy (E) of a conformation based on the HP model is the sum of the contributions of all pairs of non-consecutive hydrophobic amino acids[6].

The free energy calculation for the HP model, counts only the energy interactions between two non-consecutive amino acid monomers.

$$E = \sum_{i,j:i+1<j} c_{ij} e_{ij}$$

c_{ij} is depend on neighbor of two monomers i and j and it take value of 1 and 0.

e_{ij} is depend on type of amino acids and it take value of -1 and 0.

Minimizing the summation in E equation is equivalent to maximizing the number of non-consecutive H-H contacts. Several other variants of HP-model exist in the literature[12].

1.4 Optimization algorithms:

AB off-lattice model or HP lattice is a typical NP problem for protein structure prediction so we need optimization algorithms such as genetic algorithm ,ant colony optimization , tabu search algorithm , to improve these models efficiency[9].

Predicting Protein 3D structure using optimization algorithms depends on:

-How to make local based search algorithms working with improved technique to escape of trapping in local optima and choose best global heuristic solutions.

- How to make global search algorithms working to find good solutions with adding utilities of some local search algorithm advantages.

When using these improved algorithms with AB off- lattice model or HP lattice model, this will give more accurate result to predict protein 3D structure.

Optimization algorithms are used for finding the lowest free energy conformation between protein sequences to predict the native conformation of protein sequences structure [9].

1.4.1 Genetic algorithm (GA):

Genetic algorithm (GA) is method knows as population based search algorithm to get optimal solution. It chooses set of solutions knows as population and produce a new and better generation by two strategies. First: crossover operation that randomly splits two solution and exchange parts between them. Second: mutation operation change solution at random point.

These two strategies will create new generation that more strong and in protein structure prediction that mean optimal solution near the ideal representation of real protein 3D structure[8]. In general, GA algorithm is simple and effective search algorithm for protein structure prediction but its performance to get optimal solution is decrease if protein sequence length increasing[8].

1.4.2 Tabu Search algorithm (TSA):

A tabu search strategy used techniques such as: Move: set of moves to guarantee fast search among deferent conformations of proteins. Neighborhood: set of solutions called Neighborhood. Tabu list: contains forbidden moves to avoid bad solutions[14]. Get optimal solution at short time than other methods. Tabu search still local base search algorithm and need improvement avoid trapping in global optima.

1.4.3 Improved Tabu Search algorithm (ITSA):

Tabu Search algorithm (TSA) with memory system is strong to gets local optimal solution for local search. To improve TS which is poor algorithm for global search problem, we can add aspiration criteria which make algorithm more effective to get global optimal solution [11].

TS algorithm with this improvement method has very good performance and predict 3D structure prediction of proteins effectively. Improved TS algorithm has the good rank for lowest energy conformation but it is used for small proteins only[11].

1.4.4 Genetic tabu search algorithm (GTSA):

Genetic algorithm (GA) is heuristic search algorithm taken from evolutionary of genetic selection which take individuals by fitness function .Individuals that have high fitness values will have high opportunities to generate successors, and GA used for get optimal solution for protein structure prediction.

In spite of GA is strong in global search, it still poor in local search but with using tabu search (TS) algorithm which effective in local search algorithm and has low global search capability the GTSA algorithm will be stronger. In general, GATS can be effective search algorithm for protein structure prediction that overcomes the poor local search capability of GA and low global capability of TS. GATS algorithm prove it has effective capability of protein structure prediction than other algorithms. GATS is used with short protein sequence and need strong computing processing[10].

1.4.5 Particle Swarm Optimizer (PSO):

Particle Swarm Optimizer is an iterative optimization tool. But sill poor in local search optimization and to avoid tapping in local optima Levy flight distribution is used and the algorithm will be more effective to solve protein structure prediction and the new algorithm called LPSO[9].

1.4.6 Improved Particle Swarm Optimization (LPSO):

Particle Swarm Optimizer PSO is an iterative optimization tool. But sill poor in local search optimization and to avoid tapping in local optima Levy flight distribution is used and the algorithm will be more effective to solve PSP and the new algorithm called LPSO[9]. LPSO has it simplicity, convenience, fast convergence and fewer parameters to solve protein structure prediction but still need strong power of processing and more computation cost[9].

1.4.7 Hybrid of Genetic Algorithm and Particle Swarm Optimization (GA-PSO):

GA produces new good solution but still weak in local search whereas PSO algorithm is good in local search. So, GA-PSO attempts to treat this weakness during mutation process [2]. In addition, GA-PSO Performs better than use GA only but it wok only on small proteins sequences and need power computing processing.

II. Results and Discussion

Table 1 and 2 show the energy performance for different searching algorithm applied on AB and HP lattice model respectively. In table 2 energy is decreased when protein length above 60 amino acids, which indicate the low efficiency of these algorithms at large protein structure. Also, when comparing table 1 and 2 we could see the outperformance of ITSA algorithm over TSA and GA-PSO over GA. Also we could see the good performance of the GA either on AB and HP lattice model.

Table 1: Performance comparison of different optimization algorithms applied in AB lattice model for different protein length.

AB lattice model		PSO	LPSO	ITSA	GATS
Protein length	5	-0.2182	-1.0627	--	--
	8	-1.2856	-2.0038	--	--
	13	-2.8218	-4.6159	-6.5687	-6.9539
	21	-4.1515	-6.6465	-13.4151	-14.7974
	34	-4.2235	-7.3375	-27.9903	-27.9897
	55	-8.0209	-13.0487	-41.5098	-42.4746
Average Energy		-3.4	-5.7	-22	-23

Table 2: Performance comparison of different optimization algorithms applied in HP lattice model for different protein length.

HP lattice model		GA	TSA	GA-PSO
Protein length	20	-29	-9	-11
	24	-28	-9	-13
	25	-25	-8	-9
	36	-50	-14	-18
	48	-65	-23	-29
	50	-59	-21	-26
	60	-114	-34	-49
	64	-98	-42	--
Average Energy		-58.5	-20	-22

In addition, Fig 1 and 2 confirm table 1 and 2 results respectively.

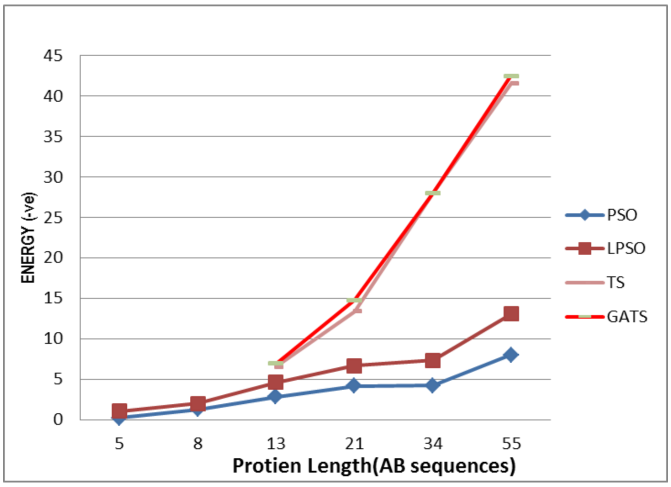

Fig 1: The energy of different optimization algorithms applied on AB protein sequence.

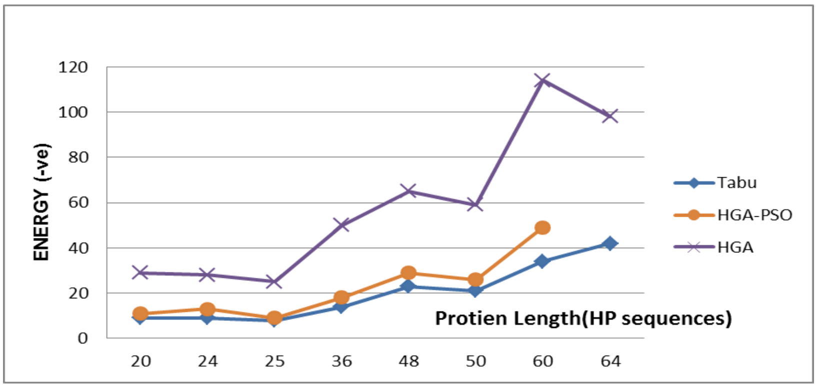

Fig 2: The energy of different optimization algorithms applied on HP protein sequence.

III. Conclusions

Protein 3D structures have many applications in drug design and understanding disease. Protein skeleton lattice models and optimization algorithms selection are two challenges in protein structure prediction problem. In this paper we show the outperformance of the GA either on HP and AB lattice sequence. In the next paper we will compare algorithms power computing requirements to solve protein 3D structure of long protein sequences.

References

[1] M. Rashid, *et al.*, "A New Genetic Algorithm for Simplified Protein Structure Prediction," in *AI 2012: Advances in Artificial Intelligence*. vol. 7691, M. Thielscher and D. Zhang, Eds., ed: Springer Berlin Heidelberg, 2012, pp. 107-119.

[2] C.-J. Lin and S.-C. Su, "Protein 3D HP Model Folding Simulation Using a Hybrid of Genetic Algorithm and Particle Swarm Optimization," *International Journal of Fuzzy Systems*, vol. 13, 2011.

[3] Y. Zhang, "Combining Empirical and Experimental Data in Protein Structure Determination," MSC, ELECTRICAL AND COMPUTER ENGINEERING, University of Massachusetts, 2009.

[4] A. A. Chida, "Protein Tertiary Model Assessment Using Granular Machine Learning Techniques," PhD, Computer Science, GeorgiaState University, 2012.

[5] C. N.ROKDE and M. KSHIRSAGAR, "A COMPARATIVE STUDY OF PROTEIN TERTIARY STRUCTURE PREDICTION METHODS," *International Journal of Computer Science and Informatics*, vol. 3, pp. 2231–5292, 2013.

[6] M. MIHĂŞAN, "BASIC PROTEIN STRUCTURE PREDICTION FOR THE BIOLOGIST: A REVIEW " *Arch. Biol. Sci.*, vol. 62, pp. 857-871, 2010

[7] Hongyu Zhang, "Protein Tertiary Structures:Prediction from Amino Acid Sequences," *ENCYCLOPEDIA OF LIFE SCIENCES*, 2002.

[8] MdTamjidulHoque, *et al.*, "Protein Folding Prediction In 3D FCC HP Lattice Model

Using Genetic Algorithm," presented at the 2007 IEEE Congress o n Evolu tionary Computatio n (CEC 2007).

[9] X. Chen, *et al.*, "An Improved Particle Swarm Optimization for Protein Folding Prediction," *I.J. Information Engineering and Electronic Business*, vol. 1, 2011.

[10] X. Zhang, *et al.*, "3D Protein structure prediction with genetic tabu search algorithm," *BMC Syst Biol*, vol. 4, 2010.

[11] Z. Xiaolong and C. Wen, "Protein 3D Structure Prediction by Improved Tabu Search in Off-Lattice AB Model," in *Bioinformatics and Biomedical Engineering, 2008. ICBBE 2008. The 2nd International Conference on*, 2008, pp. 184-187.

[12] S. Shatabda, *et al.*, "The road not taken: retreat and diverge in local search for simplified protein structure prediction," *BMC Bioinformatics*, vol. 14, p. S19, 2013.

[13] S. Fidanova and I. Lirkov, "Ant colony system approach for protein folding," in *Computer Science and Information Technology, 2008. IMCSIT 2008. International Multiconference on*, 2008, pp. 887-891.

[14] JacekBlazewicz, *et al.*, "A tabo search strategy for finding low energy structure of protein in H model," *Computitional Methods in Science and Technology*, vol. 10, pp. 7-19, 2004.