A Heuristic Method to Bias Protein's Primary Sequence in Protein Structure Prediction

Nasser Mozayani
Dept. of Computer Science,
Iran University of Science and Technology
Tehran, Iran
mozayani@iust.ac.ir

Hossein Parineh
Dept. of Computer Science
Iran University of Science and Technology
Tehran, Iran
Hossein.Parineh@gmail.com

***Abstract*— Protein Structure Prediction (PSP) is one of the most studied topics in the field of bioinformatics. Regarding the intrinsic hardness of the problem, during last decades several computational methods mainly based on artificial intelligence have been proposed to approach the problem. In this paper we broke the main process of PSP into two steps. The first step is making a bias in the sequence, i.e. providing a very fast yet considerably better energy of conformation compared to the primary sequence with zero energy. The second step, which is studied in the other essay, is feeding this biased sequence to another algorithm to find the best possible conformation. For the first step, we developed a new heuristic method to find a low-energy structure of a protein. The main concept of this method is based on rule extraction from previously determined conformations. We'll call this method Fast-Bias-Algorithm (FBA) mainly because it provides a modified structure with better energy from a primary (linear) structure of a protein in a remarkably short time, comparing to the time needed for the whole process. This method was implemented in Netlogo. We have tested this algorithm on several benchmark sequences ranging from 20 to 50-mers in two dimensional Hydrophobic Hydrophilic lattice models. Comparing with the result of the other algorithms, our method in less than 2% of their time reached up to 62% of the energy of their best conformation.**

Keywords*— Computational Biology ; Protein Structure Prediction; Heuristic method;Fast Bias Algorithm; HP Model ; 2D lattice ; Netlogo

I. INTRODUCTION

Proteins are the building block and functional molecules in human body that play a key role in almost every biological process. Almost all cellular processes on earth are governed or guided by proteins [1]. Progress in handling PSP problem will also improve our understanding of proteins involved in vital processes, including diseases such as cancer [16]. About 39 million non-redundant protein sequences are extracted from DNA sequences and are available at GenBank [14]. Protein 3-D structures which are approximately about 100000, in turn, may be obtained from the Protein Data Bank or PDB [15]. Consequently, there is a huge gap between our capacities to produce protein sequences and to determine 3-D structures of new proteins with yet unknown folds.

In this work we concentrate on *ab-initio* modeling. *Ab-initio* methods are based on the *Anfinsen thermodynamic hypothesis* [2]: the (*native*) conformation adopted by a protein is the most stable one, i.e. the one with minimum free energy. In its native conformation, protein structure is thermodynamically stable, which shows that it complies with Gibbs second law of lowest free energy (second law of thermodynamic) [4].

All-atom computer simulations are typically unpractical, because of the huge amounts of computations that is needed by this process. Even simple abstractions are NP-complete [5]. To overcome the restrictions imposed by computational complexity, several simplified models such as: AB, HP, BLN and Tube model have been proposed.

Algorithm effectiveness is evaluated by the measure of final energy of the predicted structure. Regarding the fact that there are several different models, unfortunately, there is no general agreement on the potential function that should be used with these models, and several different energy functions can be found in literature.

II. BACKGROUND

A. Proteins

Proteins are polymers made up of chains of amino acids. There are only 20 different types of amino acids used in protein structure, so proteins could be represented by a string of characters for computational purposes. The *Primary* or *Linear* structure of a protein is a sequence consisting of amino acids $s_1, \dots, s_i$. The secondary structure is local folding of a sequence which is usually in the form of α-helix, β-sheet or coil. The tertiary structure is then, the composition of these secondary structures which shapes a protein in 3-dimentional space. Quaternary structure is the composition of tertiary structures. The 3-dimensional shape is called as *Functional*, *Native* or *biological fold/ Conformation* [2], [3]. Based on Anfinsen theory [2] Protein in its native conformation is at the most stable state, and has the least free energy. A review of various forces and potentials can be found in [17]. The PSP Problem is about predicting the native conformation of a protein, when its sequence of amino acids is known [18] [19].

978-1-5090-0139-2/15/$31.00 ©2015 IEEE

B. HP Model

This model which was introduced by Lau and Dill in 1989[20] [21], reduces twenty types of amino acids into two main groups, namely H(hydrophobic) and P(hydrophilic).

In this model the primary structure of a protein *P* could be defined as a sequence of *n* amino acids:

$$s = s_1, ..., s_n \text{ where } s_i\{H, P\}, \vee i \ , 1 \le i \le n \qquad (1\text{-}1)$$

A conformation of the protein *P* is a function *C* that maps the protein sequence *S* to the points of a 2D/3D dimensional Cartesian lattice: C: [1…n] $\rightarrow Z^d$

$$s = s_1, \dots, s_n \ \rightarrow \ \{(x_1, y_1),(x_2, y_2),\dots,(x_n, y_n)\} \qquad (1\text{-}2)$$

Where (x_i, y_i) represents the position in the 2D/3D lattice of the amino acid s_i such that:

1. $\forall 1 \le i, j \le n$, with $i \ne j \Rightarrow c(i) \ne c(j)$ (*self-avoiding*). (1-3)
2. $\forall 1 \le i, j \le n$, with $|i - j| = 1 \Rightarrow |x_i - x_j| + |y_i - y_j| = 1$ (*connected neighbors*). (1-4)

C. Enrgy in HP Model

Given a sequence $s = s_1, \dots, s_n$ the energy of a conformation *c* of *s* is defined as follows:

$$E(C) = \sum_{1 \le j+1 < j \le n} \varepsilon_{i,j} \ . \delta(s_i, s_j) \qquad (1\text{-}5)$$

Where $\varepsilon_{i,j}$ is equal to -1 whenever both s_i and s_j are *H* amino acids, 0 otherwise, and $\partial(s_i, s_j)$ is 1 if s_i and s_j are topological neighbors, 0 otherwise [18] [20] [22].

III. Related Works

Protein structure can be determined experimentally by the use of X-ray crystallography [22] and the less popular NMR imaging [23]. These processes are very expensive and time consuming; additionally direct observation of a protein's structure does not give any information about the process of folding itself. Another approach is comparative modeling in which the known structure of one protein is used to predict the structure of another regarding the similarity of amino acid sequences [24]. More information about this method could be found here [25]. Regarding the inherent difficulty of the problem we can cite the Levinthal's paradox [28] which states that for a 100-length chain there will be at least 3^{100} possible conformations (considering three degrees of freedom), characterizes it as an intractable problem [29].

Shmygelska et al (2005) [30] modified a technique called *An Colony Optimization* (ACO) under lattice HP-Model. An advantage of this method to its prior ones was its high range of movements and consequently better achieved conformations. In 2006, Chu et al. [31] proposed *MACO* which works on the same base and yet earned better results comparing to previous versions of ACO. In 1993, stillinger et al. [7] published an essay about using Artificial Neural Network (ANN) in AB-model for the PSP problem. In 1992 Unger and Moult [8] published their impressive result of research about using Genetic Algorithm (GA) combined with Monte Carlo method. Another try to approach the PSP problem was made by Ghosh and Paray [32] in 2008 by the use of some machine learning techniques such as K-NN, Fuzzy K-NN, and Minimum Distance. There were also other tries for implementing heuristic algorithms such as Mann and Backofen in 2014[33] by using Constraint Based Programming (CBP). Also, Hart and Istrail in 1996 [34] suggested a method in which the algorithm was guarantied to find a solution with at least 3/8 of best energy, in a linear time. Some other methods were proposed by the use of taboo search algorithm [10] [11] which claimed to find the best free energy in long sequences. One of the approaches that recently have got more attention is Multi Agent Systems (MAS). Sample works could be found in [12] by Bortulossi et al. in 2007, and Gonzalez Perez et al. in 2009[13].

IV. Fast Bias Algorithm

Before introducing a new heuristic method, which will be called Fast Bias Algorithm (FBA), it is needed to introduce some basic concepts.

A. H and P Moving Tendency

In HP model, all of the amino acids are divided into two categories: H(Hydrophobic), P(Hydrophilic). Regarding the fact that environment is filled with water molecules, it is expected that H-type is forced into center of the structure while *P*-type is forced into outer surface. This is shown in fig. 4-1.

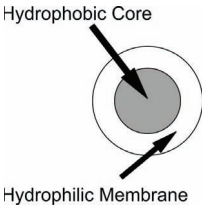

Fig. 4-1: H-P placement in space

This theory is supported by comparing conformations obtained from other methods .As displayed in fig. 4-2, in all situations *H* has a tendency to stay in the center of the conformation while *P* is likely to be placed in outer surface. It could also be conclude that there is a tendency for H-type residues to get as close as possible together, even if they can't join the core, the same as shown in fig. 4-3. After adding some more residues the formation will change as shown in fig. 4-3. As shown in fig. 4-3, free energy for the primary sequence is equal to zero, while in the final conformation the energy has reached to -2.

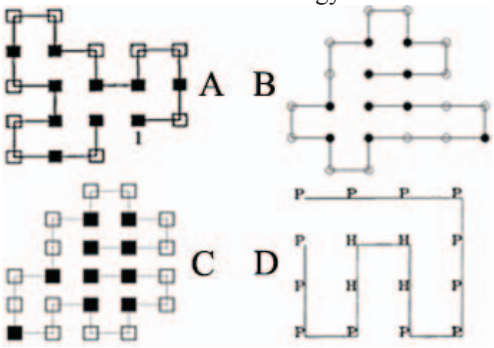

Fig. 4-2: The tendency of *H* and *P* for placement. These pictures are taken from A)[34] B)[11] C)[36] D)[20]

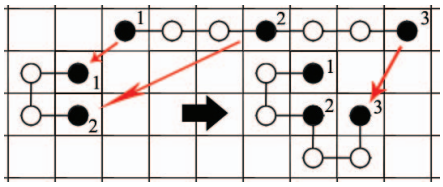

Fig. 4-3: The tendency of H-type to get close

B. *Sectioning Primary Sequence*

The result from previous section, for new *H* placement, implies that input sequence ought to be divided into several substrings. The logic behind this sectioning process is:

1. Each subunit starts with *P* and ends to *H*
2. If the sequence starts with H, then first subunit starts with *H* and ends to *H*
3. Consecutive *H* residues with be allocated to previous subunit
4. Each subunit should include at least one H, and one P

Some sample divisions are shown in fig. 4-4.

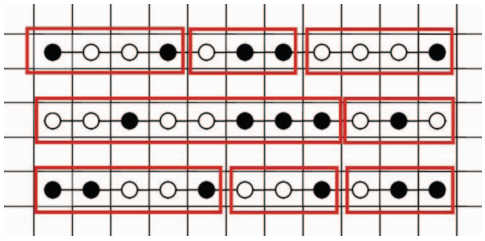

Fig.4-4: sample divisions

C. *Connection Potential (CP)*

From all of the points in lattice board that could be used for placing last *H* ,only some of them are actually useful for energy improvement. These points are the neighborhoods of all Hs in the structure. We call this points "Connection Potential" or *CP*, because they're potantially a H-H connection which promotes energy of the structure. Look at picture 4-5 , each *H* has some graysacled neighbors which represent a *CP*.

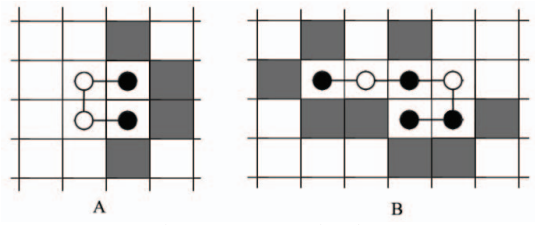

Fig.4-5: *CP*s around each H

D. *H-Center*

Regarding section A, there is a tendency in the structure about forming a hydrophobic core. In addition to *CP*s, it is important that all *H* stay as close as possible. The most important thing for the last *H* is to be placed on a *CP* which is closest to the location that other *H* residues are located. So a new coordination *H-Center(x,y)* is defined to represent this point. If S is a sequence with length n, then:

$$S = s_1, s_2, s_3, \dots, s_n \quad Where \quad s_i \ \{H, P\}, \forall i \ , 1 \le i \le n \quad (1\text{-}6)$$

Then coordination of that point is:

$$H_{Center(x)} = \frac{\sum_{i=0, s_i=\{H\}}^{n} x_i}{\sum_{i=0, s_i=\{H\}}^{n} 1} \quad (1\text{-}7)$$

$$H_{Center(y)} = \frac{\sum_{i=0, s_i=\{H\}}^{n} y_i}{\sum_{i=0, s_i=\{H\}}^{n} 1} \quad (1\text{-}8)$$

This point (*H_Center*) is used for navigating new *H* into the coordination where previous *H*s are located.

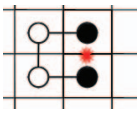

Fig.4-6: *H_Center* for this structure

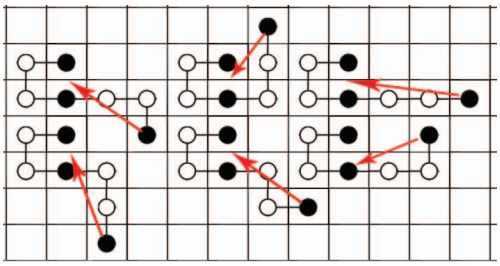

Fig.4-7: Heading last H into *H_Center*

E. *P quantity: Even or odd?*

An important item that consents new *H* to be placed in a candidate *CP* is to determine if both number of *P*s (*P*s between last *H* and new *H*) and candidate *CP* are even or odd. Fig. 4-8 shows how evenness and oddness of *P*s, affects accessible places for new *H*.

F. *Placing Ps*

After placing new *H* to its new location, all the *P*s between last *H* and new *H* must be relocated so that the link length between these *H* and *P* is equal to unit value (1 in models, and 3.8 angstrom in real world). There is a simple pseudo code for it that which states:

I. Set i(=index) the index of last *H* + 1
II. Find vacant neighbor of i which is closer to new *H*
III. Place *P* with index of i in that location
IV. Increase i
V. Go to II until you reach to new *H*

V. ORGANIZING AND USING EXTRACTED RULES

After rule extraction in section V, we need to exploit them to establish a framework. There are several items that should be built, considered or updated at any iteration of the algorithm. These items help the algorithm to manipulate the process of protein folding correctly. In this section these items will be introduced.

A. *HP_List*

In our model the world is a 2D lattice. As a result we may allocate a pair of (x,y) to each patch as shown in fig. 5-1. Consequently, each residue could be represented by four items

consisting: index number, residue type (H/P), X coordination, Y coordination. This is displayed in table 5-1.

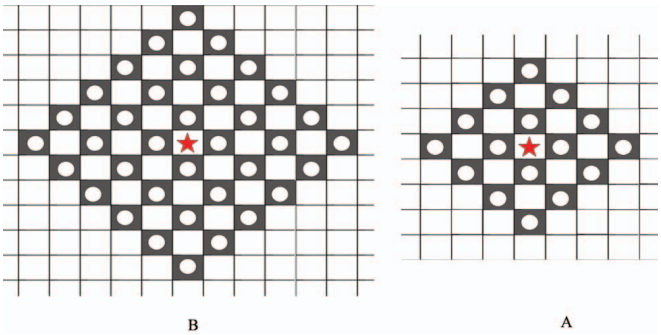

Fig.4-8: Accessible places for new H, for even/odd number of Ps

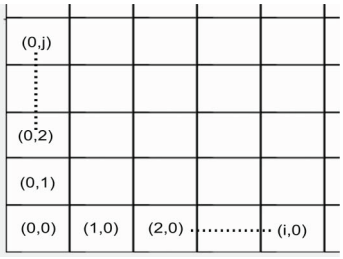

Fig. 5-1: coordination of each location

Table 5-1: All important data about residues in *HP_List*

Index	Type H=1 / P=0	X-pos	Y-pos
0	T_0	X_0	Y_0
1	T_1	X_1	Y_1
…	…	…	…
I	T_i	X_i	Y_i

B. *CP_List*

Another important table is called *CP_List* which stores every *CP* location and determines if they are useable or not. It is shown in table 5-2.

An important benefit of this table is that at the time of selecting a *CP*, it shows which is accessible or not. A *CP* could be inaccessible by three reasons:

- *CP* is out of reach
- Using that *CP* makes an dead-end
- Inequality in being even/odd for *CP* and number of *P*s

Table 5-1: All CPs in structure and their accessibility

Index	Possibility: Yes=1/No=0	X-pos	Y-pos
0	Y/N	X_0	Y_0
1	Y/N	X_1	Y_1
…	…	…	…
I	Y/N	X_i	Y_i

C. *Algorithm Description*

The first step in algorithm is creating HP_List and CP_List based on the number of residues in input sequence. Then the algorithm searches for the first H. Here the act of sectioning commences. If the next residue is H, the best available CP will be selected for this new H and it will be moved to that place, and if it's P, the algorithm looks at the next residue until it reaches to next H, at the same time, it counts the number of Ps in order to 1) find out if the number of Ps is even or odd and determine available places to go 2)find out the farthest range that the last H may go. After reaching next H, it will be placed on best available CP and then Ps will be placed between Last H and new H. In each step HP_List and CP_List is updated in order to access the actual data of the structure. This will continue until the algorithm reaches to the end of the sequence.

D. *Algorithm*

The algorithm of the whole process is presented in fig. 5-2.

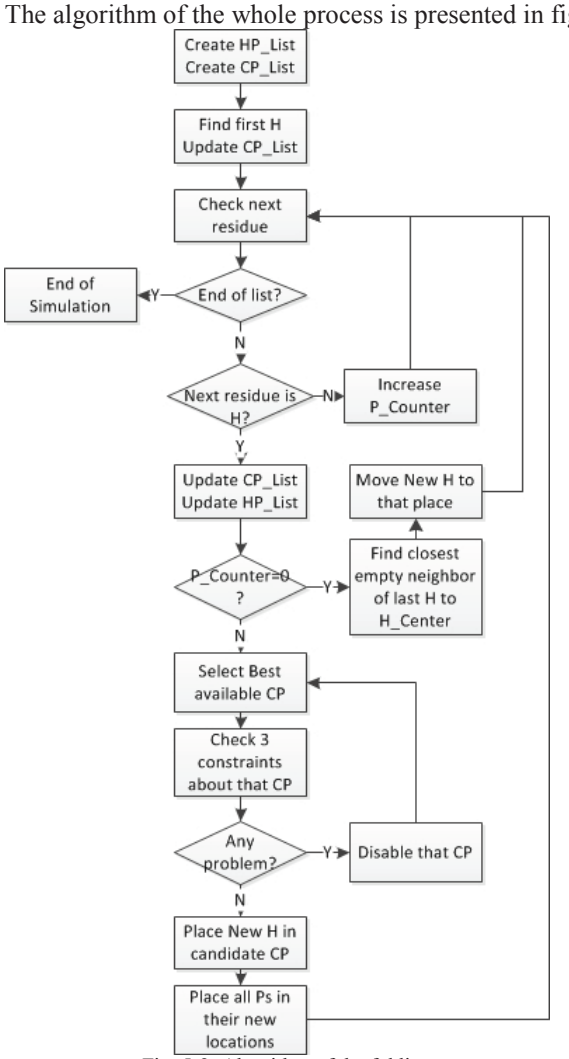

Fig. 5-2: Algorithm of the folding process

VI. EXPERIMENTAL RESUTLTS

In this section the results of implementing FBA algorithm in NETLOGO software will be discussed. In order to evaluate FBA, we will test it with four benchmark sequences that are presented in table 6-1. In table 6-2, the energy of structure for each sequence after implementing FBA is mentioned and compared to the energies acquired by other algorithms.

In table 6-3, the time spent to reach these structures for each sequence is displayed and is compared with other algorithms. It should be considered that FBA were run on a core i3 (2.4GHz with 3G RAM) laptop, while rest of them were run on different computers. For instance, ACO were run on a 1GHz Pentium III processor with 1G RAM.

Table 6-1: Benchmark sequences

Seq. No.	Seq. length	Sequence
1	20	HPHPPHHPHPPHPHHPPHPH
2	25	PPHPPHHPPPPHHPPPPHHPPPPHH
3	36	PPPHHPPHHPPPPPHHHHHHHPPHHPPPPHHPPHPP
4	50	HHPHPHPHPHHHHPHPPPHPPPHPPPPHPPPHPPPHPHHHHPHPHPHPHH

Table 6-2: Energy comparison

Seq. No.	Seq. length	E_{bias} [f]	E_{GA} [a]	E_{MMC} [b]	E_{CG} [c]	E_{ACO} [d]	E_{PFGA} [e]
1	20	-5	-9	-9	-9	-9	-9
2	25	-5	-8	-8	-8	-8	-8
3	36	-5	-14	-13	-14	-14	-14
4	50	-9	-21	-21	-21	-21	-21

a. The result is mentioned from unger and moult [8] with 200 structures for 300 generations.
b. The result is mentioned from unger and moult [8]
c. The result is mentioned from Dill and Beutler [37] by using chain growth algorithm
d. The result is mentioned from Shmygelska [30]
e. The result is mentioned from Thang [38]
f. The shape of each conformation is shown in fig. 6-2

The main idea for using FBA is decreasing the size of conformation space and giving bias to the primary sequence so that in a very short time it is expected to have a remarkable modification in structure energy. By looking at table 6-3, which is taken from Unger and Moult work [8], and based on table 6-2 and 6-1 it is obvious that by using FBA in less than 30ms more than 83 million structures were eliminated. As a result in next step the algorithm will have to search in a remarkably smaller conformation space to find the best conformation .This result illustrates that one of the assigned tasks for FBA, about conformation space decrement, is done.

Based on these tables, using FBA shows several advantages such as:

- Removing a large number of worse energy conformation, in a very short time (miliseconds)
- Reaching a better energy level which ranges between 42.8% - 62.5% of the best possible known conformation, in that short time.

Table 6-3: Time comparison for different algorithms

Seq. No.	Seq. length	T_{bias}	T_{GA} [a]	T_{CG} [a]	T_{ACO} [b]	T_{newACO} [b]
1	20	0.029	5.16	2.16	23.90	3.33
2	25	0.033	6	6.6	35.32	10.62
3	36	0.054	54.6	132	4746.12	11.81
4	50	0.110	3180	18600	3000.28	4952.92

a. From Dill and Beutler [37]
b. From Shmygelska [30]

Table 6-4: Time comparison for different algorithms

Energy Level Distribution	
Energy Level	Possible Structures
0	36,098,079
1	31,656,934
2	12,473,446
3	2,943,974
4	517,984
5	77,080
6	10,364
7	1,194
8	96
9	4

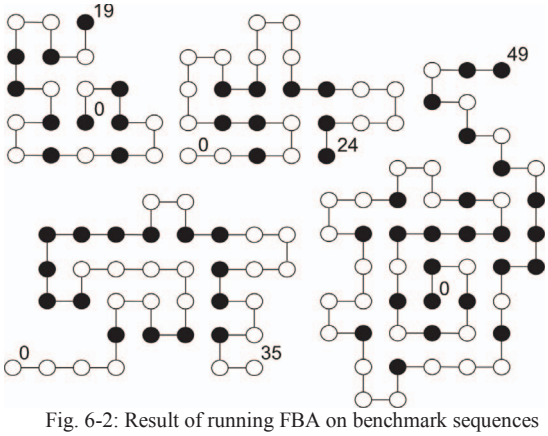

Fig. 6-2: Result of running FBA on benchmark sequences

Table 6-5: Energy time comparison for FBA and other algorithm for 20-mer sequence

Sequence length 20				
%	T_{bias}/E_{bias}	T_{GA} *a	T_{CG} *b	T_{ACO} *c
Time	0.029	%0.56	1.34%	%0.87
Energy	−5	%55.55	%55.55	%55.55

Table 6-6: Energy time comparison for FBA and other algorithm for 25-mer sequence

Sequence length 25				
%	T_{bias}/E_{bias}	T_{GA} *a	T_{CG} *b	T_{ACO} *c
Time	0.033	%0.55	%0.50	%0.31
Energy	−5	%62.5	%62.5	%62.5

Table 6-7: Energy time comparison for FBA and other algorithm for 50-mer sequence

Sequence length 50				
%	T_{bias}/E_{bias}	T_{GA} *a	T_{CG} *b	T_{ACO} *c
Time	0.11	<< %1	<< %1	<< %1
Energy	-9	%42.8	%42.8	%42.8

VII. COMPARISON

Regarding the fact that this method i.e. dividing the whole procces of PSP into two sections and biasing primary sequence in the first section was implemented for the first time, there couldn't be direct comparison with similar works, yet in section VI the result of our work was compared with similar ones.

VIII. COLCLUSION

In this work we suggested to break the whole process of PSP into two steps. In the first step which is in the scope of this essay, by using HP square lattice model, we introduced a novel heuristic algorithm that makes the conformation space as small as possible and biases the primary sequence (with zero energy) to a better energy level, as a result in the next step (which is out of scope of this essay) another algorithm will have to search through a smaller conformation space and find the best conformation in a remarkably shorter time.

To assess the quality of FBA, it was tested with benchmark sequences. The result shows that using FBA would make a big decrease in the size of conformation space, so that the search process for the best conformation would be held in smaller conformation space and thus much faster. This will help next step of the algorithm which finds the best conformation to run in much shorter time comparing other algorithms.

REFERENCES

[1] D.Petrey, T.S.Chen, L.Deng, J.I.Garzon, H.Hwang, G.Lasso, H.Lee, A.Silkov, B.Hoing "Template-based prediction of protein function" *Current Opinion in Structural Biology*, V(32), , PP: 33–38, June 2015.

[2] Anfinsen, C. B.. "Principles that govern the folding of protein chains". *Science*, 181(96),pp. 223-230, 1973.

[3] M.X.Suresh, M.Gromiha and M.Suwa "Development of a Machine Learning Method to Predict Membrane Protein-Ligand Binding Residues Using Basic Sequence Information " *Advances in Bioinformatics* , Jan.2015

[4] C. B. Anfinsen, Haber, E., Sela, M and White, F. H, "The kinetics of formation of native ribonuclease," *Proceedings of the National Academy of Sciences*, , 47(9), pp. 1309–1314,1961.

[5] S.Shatabda, M.A.H.Newton, M.Rashid, D.N.Phum "How Good Are Simplified Models for Protein Structure Prediction?" *Advances in Bioinformatics,* 2014.

[6] A. Shmygelska and Holger H. Hoos, "An Improved Ant Colony Optimisation Algorithm for the 2D HP Protein Folding Problem," *Springer: Lecture Notes in Computer Science Volume ,* vol. 2671, pp. 400-417, 2003.

[7] F. H. Stillinger, Teresa Head-Gordon, and Catherine L. Hirshfeld, "Toy model for protein folding," *PHYSICAL REVIEW E,* vol 48, 1993.

[8] R. unger and John Moult, "Genetic Algorithm for Protein Floding Simulations," *Molecular Biology,* vol. 231, pp. 75-81, 1992.

[9] T.Kalaichelvi and P.Rangarajan, "Criterion based Two Dimensional Protein Folding Using Extended GA," *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT),* vol. 2(5), 2012.

[10] X. Lin, Xiaolong Zhang and Fengli zho, "Protein structure prediction with local adjust tabu search algorithm," in *International Conference on Intelligent Computing ICIC*, Nanning,China, 2013.

[11] N. Lesh, M. Mitzenmacher and S. Whitesides, "A complete and effective move set for simplified protein folding," in *Proceedings of the seventh annual international conference on Research in computational molecular biology*, New York, 2003.

[12] L. Bortolussi and Dovier , Agostino, "Agent-based Protein Folding Simulation," *Multiagent and Grid Systems - Multi-agent systems for medicine,computational biology, and bioinformatics,* vol. 3, pp. 183-197, 2005.

[13] A. Orro, Massimiliano Saba, Eloisa Vargiu and Gianmaria Mancosu, "Using a Personalized, Adaptive and Cooperative MultiAgent System to Predict Protein Secondary Structure," in *First International Workshop on Multi-Agent Systems for Medicine, Computational Biology, and Bioinformatics*, Utrecht (Netherlands), 2005.

[14] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. Genbank. *Nucleic Acids Research*, 41(D1):D36-D42, 2013.

[15] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E.Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235-242, 2000.

[16] Y. Duan and P. A. Kollman. Computational protein folding: from lattice to all-atom. IBM Systems Journal, 40:297-309, 2001. 0018-8670.

[17] A. Neumaier. Molecular modeling of proteins and mathematical prediction of protein structure. *SIAM Review*, 39:407–460, 1997.

[18] R.Backofen. " A constraint optimization approach using a new lower bound. *Constraints"*, 6(2–3) 2001, pp. 223–255.

[19] J. Skolnick and A. Kolinski. "Reduced models of proteins and their applications". *Polymer*, 45:511–524, 2004.

[20] K. Dill and K. Lau. "A lattice statistical mechanics model of the conformational sequence spaces of proteins." *Macromolecules*, 22 (1989), pp. 3986–3997.

[21] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan. "Principles of protein folding - a perspective from simple exact models." *Protein Science*, 4(4), 1995, pp. 561–602.

[22] J. Drenth, *Principles of Protein X-Ray Crystallography,* (Springer, Berlin, 2006)

[23] K. Wüthrich," NMR Studies of Structure and Function of Biological Macromolecules (Nobel Lecture)" *Angew. Chem. Int. Ed. Engl.* **42**, 3340, 2003.

[24] A. Lesk, C. Chothia, " The relation between the divergence of sequence and structure in proteins"*Philos. Trans. R. Soc. London B* **317**, 345, 1986.

[25] H.Hendi, W.Khalifa, A.B.Salem , "A Study of Intelligent Techniques for Protein Secondary Structure Prediction", *International Journal of Information Models and Analyses",* Volume 4,Number 1, 2015.

[26] C. Floudas et al.," Advances in protein structure prediction and de novo proteindesign: A review" *Chem. Eng. Sci.* **61**, 966 ,2006.

[27] C. Levinthal. "Are there pathways for protein folding?" *Journal of Medical Physics*, 65(1):44-45, 1968.

[28] M.Azizi, M.S.Adib, "PROTEIN STRUCTURE PREDICTION BY MEANS OF SEQUENTIAL PATTERN MINING", International Journal of Artificial Intelligence & ApplicationsVol. 6, No. 4, July 2015.

[29] A. Shmygelska and Holger H. Hoos, "An Improved Ant Colony Optimisation Algorithm for the 2D HP Protein Folding Problem, "*Springer: Lecture Notes in Computer Science Volume,*vol. 2671, pp. 400-417, 2003.

[30] D. Chu, M. Till and A. Zomaya, "Parallel Ant Colony Optimization for 3D Protein Structure Prediction using the HP Lattice Model," *Parallel Evolutionry Computations , Springer,* vol. 22, pp. pp 177-198 , 2006.

[31] A. Ghosh and Bijnan Parai, "Protein secondary structure prediction using distance based classifiers," *International Journal of Approximate Reasoning,* vol. 47, p. 37–44, 2008.

[32] M. Mann and R. Backofen , "Exact methods for lattice protein models," *Bio-Algorithms and Med-Systems,* vol. 10(4), pp. 213-225, 2014.

[33] H. We and SC.Istrail, "Fast protein folding in the hydrophobic-hydrophilic model within three-eighths of optimal," *Spring, J Comput Biol.,* vol. 3(1), pp. 53-9, 1996.

[34] K. A. Dill, "Theory for the folding and stability of globular proteins.," *Biochemistry,,* vol. 1501–1509, p. 24(6), 1985

[35] C. Huang, Xiangbo Yang and Zhihong He, "Protein folding simulations of 2D HP model by the genetic algorithm based on optimal secondary structures,"*Computational Biology and Chemistry*,vol.34,pp.137-142, 2010.

[36] K. A. DILL and THOMAS C. BEUTLER, "A fast conformational search strategy for finding low Energy structures of model proteins," *Prorein Science,* vol. 5, pp. 2037-2043, 1996

[37] T. N. Bui and Gnanasekaran Sundarraj, "An efficient genetic algorithm for predicting protein tertiary structures in the 2D HP model," in *GECCO '05 Proceedings of the 7th annual conference on Genetic and evolutionary computation* , New york, 2005.