Rotation Crossover and K-Site Move Mutation for Evolutionary Protein Folding in 3D FCC HP Model (preliminary version)

Shih-Chieh Su
Department of Computer Science and Information Engineering
National Chung Cheng University
Chiayi County, Taiwan
ssc95p@cs.ccu.edu.tw

Jyh-Jong Tsay
Department of Computer Science and Information Engineering
National Chung Cheng University
Chiayi County, Taiwan
tsay@cs.ccu.edu.tw

Abstract **In this paper we present a new evolutionary algorithm for the protein folding problem. We study the problem in the 3D FCC HP model which has been widely used in previous research. Our focus is to develop evolutionary algorithms (EA) which are robust, easy to operate and can handle various energy functions. We propose lattice rotation for crossover and K-site move for mutation, which form the key components of our evolutionary algorithms. Experiment shows that our algorithms are able to find minimum-energy conformations for many sequences whose optimal conformations are not found in previous EA-based algorithms. Furthermore, our idea can be easily integrated into Monte Carlo and Tabu searches as approaches for local searches.**

Keywords: Protein Structure Prediction; Evolutionary Algorithms; 3D FCC Model; HP Model

I. INTRODUCTION

Proteins are essential biological molecules playing vital roles in nearly all biological processes. Therefore the predication of a protein's tertiary structure based on its primary amino acid sequence has long been the most important and challenging subject in biochemistry, molecular biology and biophysics.

HP model [1] has been used by many researchers, and applied in various lattice [2] algorithms such as 2D square, 2D Triangular, 2D Hexagonal, 3D Cubic, 3D Triangular and 3D FCC.

Due to angle restriction, the structure predicated on 2D Square and 3D cubic lattice models are not relatively close to the real 3D structure. The three dimensional face-centered-cube (3D *FCC*) lattice model [3-5] is one of the improved models that can yield the closest structure to a protein in its native state. This model has thus been extensively studied recently in protein structure predication.

Current major algorithms used in lattice models in protein structure predication include Evolutionary Algorithms [6-11], Constraint Programming [12-15], Monte Carlo [16-17], Growth-based [18] and Branch and Bound [19].

So far Constraint Programming is the best algorithm to find minimum-energy conformation on the HP lattice model. However when the experimental peptide sequence cannot converge, no close answer can be obtained by constraint programming. In addition this approach is not suitable for more complicated energy functions such as any 20 amino acid pairwise interactions energy function. Consequently, Evolutionary Algorithms (EA) so far have been the most robust and widely used approach.

In this paper, we investigate geometric structures of the 3D *FCC* lattice model, and develop Rotation Crossover and K-site Move mutation which are combined with pull move local search to form our evolutionary algorithms. Experiment shows that our algorithms are able to find minimum-energy conformations for many sequences whose optimal conformations are not found in previous EA-based algorithms. Furthermore, our idea can be easily integrated to improve Monte Carlo and Tabu searches.

II. PRELIMINARIES

A. FCC Lattice Model

Raghunathan and Jernigan made an effort in 1997 to find and define a basic unit for the 3D arrangement surrounding one amino acid [20]. In this model, there are 8 cubes with 14 faces and 12 vertices, which form a unique convex polygon containing two regular polygons: a triangle and a square. As a result, every lattice point has 12 neighbors in the 3D FCC lattice model. The 3D FCC lattice can be described by a diagram as in Figure 1 in which each lattice point has 12 neighbors denoted as *1-FR, 2-FL, 3-BR*, *4-BL, 5-FU*, *6-FD*, *7-BU, 8-BD, 9-RU, 10-RD*, *11-LU* and*12-LD,* respectively.

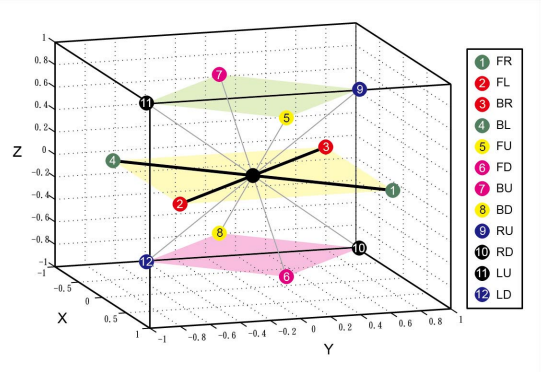

Figure 1. The FCC lattice model. Each lattice point has 12 neighbors.

B. Fitness Function

In the HP model, the free energy E for a protein structure can be calculated by using the fitness function below [10].

(1)

$$E = \sum_{i,j} \Delta r_{ij} \epsilon_{ij},$$

(2)

where the parameter

$$\Delta r_{ij} = \begin{cases} 1 & S_i \text{ and } S_j \text{ are adjacent but not connected amino acids} \\ 0 & \text{others} \end{cases}$$

(3)

Hence, the problem of protein folding is transformed into the problem of finding the conformation which yields minimal free energy, and is defined as follows: given an HP sequence $s = s_1 s_2 ... s_n$, where each s_i is either an H or a P, find an energy-minimizing conformation of s; that is to find $c^* \in C(s)$ such that $E(c^*) = \min\{E(c) \mid c \in C(s)\}$, where $C(s)$ is the set of all valid conformations for s [21].

The energy of a given conformation is defined as the number of topological neighboring (TN) contacts between those Hs, which are not connected in the sequence.

III. Evolutionary Algorithms

The basic components of EA-based protein structure predication include crossover, mutation, local search, and fitness function. In this paper, new methods are proposed for crossover and mutation which are sketched next.

A. Rotation Crossover

The idea of Rotation Crossover was first proposed by Unger and Moult [6] and used on the 2D square lattice model. However, the 3D FCC lattice model is a very tight model composed of triangles and squares. Not all rotations on the matrix are effective as all the points must remain on the matrix after rotation.

It is found from our study that every neighbouring point of the centre point is in the same square and hexagon with any other neighbouring ones. Therefore, we identify two different types of rotation: rotation along a square (type I) or a hexagon (type II). Each type can have different rotation angles. Furthermore, all the rotations can be represented by permutations of the labels of neighouring points. Our rotations are summarized as follows.

Type I: rotations along a square with rotation angle 90°, 180° and 270°.

Type II: rotations along a hexagon with rotation angle 120° and 240°.

During crossover, rotations are performed to increase success rate and to extend the exploration range of local search.

B. K-site Move Mutation

Single or double site move was also used frequently in previous local search methods [16-17][22-23]. After a complete and in-depth study on the methods of moves in local search, we propose a general k site move which is implemented by a systematic Breadth-First-Search (BFS) improved with a lower bounding technique. In principle, all possible moves will be searched and serve as the basis for mutation and local search. However the complexity of the searching space is $1.26K^{0.16}(10.0364)^K$ as calculated in [24], which increases exponentially with K. We propose a *lower bounding method* to avoid illegal conformations as early as possible to reduce the complexity of searching.

C. Local Search of Pull Move

Pull Moves were first proposed by Lesh et al. [25] and used in the local search on the 2D square HP protein folding problem. Böckenhauer et al. [26] later applied pull move in 2D triangular and 3D *FCC* lattice models and proved that this method is reciprocal and complete. Pull Move as described in Böckenhauer et al. is identical with the method proposed by Lesh et al. [25]. Both methods pull the next site to its previous position. However, it is observed that in the *FCC* lattice model pull move can be preserved without moving the next site to its previous position. In the 3D *FCC* lattice model there are four common neighbouring points to be selected (between +60° and -60°). When operating pull move, if there is more than one position on the matrix to be chosen, a point can be chosen randomly for the pull moves.

IV. Experimental Results

A. Parameter Settings

Due to the limitation of experimental time, the parameter settings of our experiment are set as listed in Table I. Better results could be obtained if the population size was allowed to set larger.

TABLE I. Parameter Settings

Operations/Parameters	Setting
Crossover rate	0.85 [7]
Mutation rate	0.4 [7]
Size of K-site move	3
Parents selection	Tournament selection
Survival selection	$\mu+\lambda$
Termination	Maximum of generations

B. Benchmarks

We carried out experiments on two benchmarks which were often used in previous research on 2D square and 3D cube lattice models. The protein sequences of the benchmarks are listed in Tables II and III.

TABLE II. Benchmark I. A group of eight benchmarks with 20-64 amino acids.

Seq.	Len.	Protein Sequence
1	20	$(HP)^2PH(HP)^2(PH)^2HP(PH)^2$
2	24	$H^2P^2(HP^2)^6H^2$
3	25	$P^2HP^2(H^2P^4)^3H^2$
4	36	$P(P^2H^2)^2P^5H^5(H^2P^2)^2P^2H(HP^2)^2$
5	48	$P^2H(P^2H^2)^2P^5H^{10}P^6(H^2P^2)^2HP^2H^5$
6	50	$H^2(PH)^3PH^4PH(P^3H)^2P^4(HP^3)^2HPH^4(PH)^3PH^2$
7	60	$P(PH^3)^2H^5P^3H^{10}PHP^3H^{12}P^4H^6PH^2PHP$
8	64	$H^{12}(PH)^2((P^2H^2)^2P^2H)^3(PH)^2H^{11}$

TABLE III. BENCHMARK II. A GROUP OF TEN BENCHMARKS WITH 48 AMINO ACIDS.

Seq.	Len.	Protein Sequence
1	48	$HPH^2P^2H^4PH^3P^2H^2P^2HPH^3PHPH^2P^2H^2P^3HP^8H^2$
2		$H^4PH^2PH^5P^2HP^2H^2P^2HP^6HP^2HP^3HP^2H^2P^2H^3PH$
3		$PHPH^2PH^6P^2HPHP^2HPH^2(PH)^2P^3H(P^2H^2)^2P^2HPHP^2HP$
4		$PHPH^2PH^6P^2HPHP^2HPH^2(PH)^2P^3H(P^2H^2)^2P^2HPHP^2HP$
5		$P^2HP^3HPH^4P^2H^4PH^2PH^3P^2(HP)^2HP^2HP^6H^2PH^2PH$
6		$H^3P^3H^2PH(PH^2)^3PHP^7HPHP^2HP^3HP^2H^6PH$
7		$PHP^4HPH^3PHPH^4PH^2PH^2P^3HPHP^3H^3(P^2H^2)^2P^3H$
8		$PH^2PH^3PH^4P^2H^3P^6HPH^2P^2H^2PHP^3H^2(PH)^2PH^2P^3$
9		$(PH)^2P^4(HP)^2HP^2HPH^6P^2H^3PHP^2HPH^2P^2HPH^3P^4H$
10		$PH^2P^6H^2P^3H^3PHP^2HPH^2(P^2H)^2P^2H^2P^2H^7P^2H^2$

C. *Experimental results for Benchmark I*

Benchmark I is composed of eight peptides of 20-64 amino acids. To enable the comparison with previous studies, the crossover rate and mutation rate were set as described in Hoque et al. [7]. In addition, the best setting of population size and generation is identified for different peptides.

For benchmark I, we compare our approach with MA, TS+PM, HGA and HGA+TR, where MA is a Memetic Algorithm developed from our previous research [27], TS+PM is a Tabu search proposed by Böckenhauer et al. [26] using Bioinformatics Utilities (BIU) library and pull moves strategy, HGA is the hybrid genetic algorithms (HGA) proposed by Hoque et al. [7], and HGA+TR is a combination of HGA and twin removal strategy proposed in [28]. Each algorithm runs 30 times. Table IV gives the experimental result. Notice that the approach proposed in this paper performed better than MA, TS+PM, HGA and HGA+TR. Although TS+PM can also find the best solution, our approach does not use other libraries as in TS+PM which cannot run independently.

D. *Experimental result for Benchmark II*

Benchmark II is composed of ten peptides of 48 amino acids studied in [29] by Dotu et al. who combined many Tabu-search approaches and numerous benchmarks in FCC model.

Table V gives a comparison of our method with LS, LS-2N-G, LNS-MULT and LNS-3D proposed in [29], where LS is Tabu Search (i.e. local search) with randomized initialization, LS-G is Tabu Search with the new initialization, LS-2N is Two Neighborhoods Tabu Search with randomized initialization, LS-2N-G is Two Neighborhoods Tabu Search with the new initialization, LNS-MULT is Multiple Sequence Reoptimized LNS and LNS-3D is 3D Structure Reoptimized LNS.

TABLE IV. **COMPARISON OF OUR APPROACH WITH MA, TS+PM, HGA AND HGA+TR0.8.** THE FORMAT OF COLUMN ENTRIES IS 'AVERAGE / MINIMUM'. NUMBERS IN BOLD INDICATE THE LOWEST ENERGY. IN ALL CASES, NATIVE ENERGY IS COMPUTED USING CPSP-TOOLS, DESCRIBED IN [13-15]. THE SEQUENCE 1 TO 5 IF THE POPULATION SIZE WAS SET AS 10 AND THE GENERATION NUMBER AS 30. FOR SEQUENCE 6 TO 7 THE BEST SETTING IS 10 FOR POPULATION SIZE AND 100 FOR GENERATION NUMBER.

#	Len.	Native E.	Our Method	MA [27]	HGA [7]	TS+PM [26]	HGA+TR [7][28]
1	20	**-23**	-22.3/**-23**	-22.5/**-23**	-/-29	**-/-23**	-/-29
2	24	**-23**	-22.1/**-23**	-22.6/-23	-/-28	**-/-23**	-/-28
3	25	**-17**	**-17/-17**	**-17/-17**	-/-25	**-/-17**	-/-25
4	36	**-38**	-36.6/**-38**	-36.7/**-38**	-/-50	**-/-38**	-/-51
5	48	**-74**	-70.7/**-74**	-68.5/-72	-/-65	**-/-74**	-/-69
6	50	**-73**	**-66.6/-73**	-62.7/-69	-/-59	**-/-**	-/-59
7	60	**-130**	-124.8/**-130**	-115.9/-122	-/-114	**-/-130**	-/-117
8	64	**-132**	-126.4/**-132**	-107/-115	-/-98	**-/-132**	-/-103

TABLE V. COMPARISON OF OUR APPROACH WITH DOTU ET AL. [29]. THE FORMAT OF COLUMN ENTRIES IS 'AVERAGE / MINIMUM'. NUMBERS IN BOLD INDICATE THE LOWEST ENERGY. IN ALL CASES, NATIVE ENERGY IS COMPUTED USING CPSP-TOOLS, DESCRIBED IN [13-15]. THE POPULATION SIZE IS 30 AND THE GENERATION NUMBER IS 150.

#	Native E.	Our Method	LS	LS-2N-G	LNS-MULT	LNS-3D
1	**-69**	-67.37/**-69**	-57.50/65	-64.61/-68	-66.77/**-69**	-67.68/**-69**
2	**-69**	-66.97/**-69**	-56.59/-64	-62.51/-68	-66.60/**-69**	-66.73/**-69**
3	**-72**	-68.80/**-72**	-56.69/-66	-62.51/-67	-68.02/**-72**	-68.06/-71
4	**-71**	-68.10/**-71**	-58.08/-65	-63.10/-68	-67.31/**-71**	-67.61/**-71**
5	**-70**	-67.77/**-70**	-57.01/-64	-63.79/-68	-66.98/**-70**	-67.04/**-70**
6	**-70**	-66.93/**-70**	-56.52/-63	-64.91/-68	-67.49/**-70**	-67.43/**-70**
7	**-70**	-67.57/**-70**	-58.15/-63	-63.75/-67	-66.55/**-70**	-66.68/-69
8	**-69**	-66.37/**-69**	-55.31/-63	-62.56/-66	-65.80/**-69**	-65.81/**-69**
9	**-71**	-69.10/**-71**	-58.91/-67	-64.40/-69	-67.95/**-71**	-67.92/**-71**
10	**-68**	-66.47/**-68**	-57.47/-64	-63.61/-67	-65.76/**-68**	-65.67/**-68**

The experiment shows that both our approach and LNS-MULT can find conformations with native energy. Note that our algorithm does not use other libraries as in LNS-MULT and can operate independently. However, the running time of Tabu search is faster than our approach. Both methods have their merits and disadvantages.

V. CONCLUSIONS

In this paper, we investigate the geometric structure of the 3D FCC lattice model, and develop Rotation Crossover and K-site Move Mutation to improve EA-based algorithms for protein folding in the 3D FCC lattice model. Combined with Pull Move local search, our algorithm is the first EA-based approach to find minimum-energy conformation for all sequences in two well-known benchmarks. Furthermore, our algorithm is purely EA-based, does not rely on any library, can be modified to work with any fitness function, and can be easily integrated with Monte Carlo and Tabu search. In the future, we will continue to experiment and improve the

search capability of our algorithm for more data sets, especially for long sequences, as well as for more tedious fitness functions such as 20 amino acid pairwise interaction energy functions. In addition, we will continue to combine more information in the search process to find structures which are closer to real structures.

REFERENCES

[1] K. F. Lau and K. A. Dill, "Lattice statistical mechanics model of the conformation and sequence space of proteins," Macromolecules, 1989, pp. 3986-3997.

[2] A. Dayem Ullah, L. Kapsokalivas, M. Mann, and K. Steinhöfel, "Protein Folding Simulation by Two-Stage Optimization," In Proc. of ISICA'09, Wuhan, China, 2009, pages 138-145.

[3] W. E. Hart and S. Istrail, "Lattice and Off-Lattice Side Chain Models of Protein Folding: Linear Time Structure Prediction Better than 86% of Optimal," Journal of Computational Biology, 1997, pp.241–259.

[4] G. Raghunathan, R. L. Jernigan, "Ideal architecture of residue packing and its observation in protein structures," Protein Science, pp. 2072 – 2083.

[5] B. H. Park, M. Levitt, "The complexity and accuracy of discrete state models of protein structure," Journal of Molecular Biology , 1995, pp. 493–507

[6] R. Unger, and J. Moult, "Genetic algorithms for protein folding simulations," Journal of Molecular Biology, 1993, pp. 75–81.

[7] M. T. Hoque, M. Chetty, and A. Sattar, "Protein Folding Prediction in 3D FCC HP Lattice Model Using Genetic Algorithm," in Bioinformatics special session, IEEE Congress on Evolutionary Computation (CEC), Singapore, 2007, pp. 4138–4145.

[8] M. T. Hoque , M. Chetty, A. Lewis, and A. Sattar, "DFS Based Partial Pathways in GA for Protein Structure Prediction," PRIB 2008, LNBI 5265, 2008, pp. 41–53.

[9] M. T. Hoque, M. Chetty, A. Lewis, A. Sattar, V. M. Avery, "DFS-generated pathways in GA crossover for protein structure prediction," Neurocomputing, Volume 73 Issue13-15, August, 2010.

[10] C. Huang, X. Yang, and Z. He, "Protein folding simulations of 2D HP model by the genetic algorithm based on optimal secondary structures," Computational Biology and Chemistry, 2010, pp. 137–142.

[11] S-C. S, C-J Lin and C-K. Ting, "An effective hybrid of hill climbing and genetic algorithm for 2D triangular protein structure prediction," Proteome Science ,2011, vol. 9, S19, doi:10.1186/1477-5956-9-S1-S19

[12] A. Dal Palu, A. Dovier, and F. Fogolari, "Constraint Logic Programming approach to protein structure prediction. BMC. Bioinformatics, 5:186, November 2004.

[13] R. Backofen and S. Will, "A constraint-based approach to fast and exact structure prediction in three-dimensional protein models", Journal of Constraints, 2006, pp. 5-30.

[14] M. Mann, S. Will, and R. Backofen, "CPSP-tools - Exact and Complete Algorithms for High-throughput 3D Lattice Protein Studies," BMC Bioinformatics, 9, 230, 2008.

[15] M. Mann, C. Smith, M. Rabbath, M. Edwards, S. Will, and R. Backofen, "CPSP-web-tools: a server for 3D lattice protein studies," Bioinformatics, 2009, pp. 676-677,.

[16] C. Thachuk, A. Shmygelska and H. H. Hoos, "A replica exchange Monte Carlo algorithm for protein folding in the HP model," BMC Bioinformatics, 2007, 8:342

[17] P. H. Verdier, W. H. Stockmayer, "Monte Carlo Calculations on the Dynamics of Polymers in Dilute Solution," The Journal of Chemical Physics , 1962, 36:227-235.

[18] H.P. Hsu , V.Mehra, W. Nadler, P. Grassberger, "Growth-based optimization algorithm for lattice heteropolymers," Physical review, E, Statistical, nonlinear, and soft matter physics , 2003, Vol. 68(2):021113.

[19] S.Y Hsieh, D.W Lai, "A New Branch and Bound method for the Protein Folding Problem in the HP Model," BIOINFORMATICS ,2008.

[20] G. Raghunathan, R. L. Jernigan, "Ideal architecture of residue packing and its observation in protein structures," Protein Science, pp. 2072 – 2083.

[21] A. Shmygelska and H. H. Hoos, "An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem," BMC Bioinformatics, 2005, pp. 30

[22] 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 ,1995, pp. 561-602.

[23] M. T. Gurler, C. C. Crabb, D. M. Dahlin, J. Kovac, "Effect of bead movement rules on the relaxation of cubic lattice models of polymer chains," Macromolecules, 1983, 16(3):398-403.

[24] P. Schuster, P. F. Stadler, "Discrete Models of Biopolymers," In Handbook of Computional Chemistry Edited by: Crabbe MJC, Drew M, Konopka A. Marcel Dekker, New York; 2001.

[25] N. Lesh, M. Mitzenmacher, S. Whitesides, "A complete and effective move set for simplified protein folding", In RECOMB '03: Proceedings of the seventh annual international conference on Research in computational molecular biology New York, NY, USA.

[26] H-J. Böckenhauer, A. D. Ullah, L. Kapsokalivas, and K. Steinhöfel, "A Local Move Set for Protein Folding in Triangular Lattice Models," Algorithms in Bioinformatics, LNCS, 2008, pp. 369-381.

[27] J-J Tsay,S-C Su, "Ab initio protein structure prediction based on memetic algorithm and 3D FCC lattice model," Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference, 2011, pp.315-318.

[28] M. T. Hoque, M. Chetty, A. Lewis, and A. Sattar, "Twin-Removal in Genetic Algorithms for Protein Structure Prediction using Low Resolution Model," IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011, pp. 234 – 245.

[29] I. Dotu, M. Cebrian, P. V. Hentenryck and P. Clote, "On lattice protein structure prediction revisited," IEEE/ACM Trans Comput Biol Bioinform, 2011, vol. 8, pp. 1620-32.