# An efficient hybrid of hill-climbing and genetic algorithm for 2D triangular protein structure prediction

Shih-Chieh Su[1], Cheng-Jian Lin[2], Chuan-Kang Ting[1]
[1] Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan
[2] Department of Computer Science and Information Engineering, National Chin-Yi University of Technology, Taiwan
[1]{ssc95p, ckting}@cs.ccu.edu.tw
[2]cjlin@ncut.edu.tw

*Abstract.* **Proteins play fundamental and crucial roles in nearly all biological processes, such as, enzymatic catalysis, signaling transduction, embryonic development, DNA and RNA synthesis and others. It has been a long-standing goal in molecular biology to predict the tertiary structure of a protein from its primary amino acid sequence. From visual comparison, it was found that a 2D triangular lattice model could give a better structure modeling and prediction for proteins with short primary amino acid sequences. In this paper it is proposed that elite-based reproduction strategy (ERS) genetic algorithm (GA) and a hybrid of hill-climbing and genetic algorithm (HHGA) for protein structure prediction on the 2D triangular lattice. It is hoped that other researchers can note the importance of this model. The simulation results of the experiments show that the elite-based reproduction strategy (ERS) genetic algorithm (GA) and hybrid hill-climbing genetic algorithm (HHGA) can successfully be applied to protein structure prediction problems.**

***Keywords:component; genetic algorithm; protein structure prediction; triangula model; hill-climbing***

## I. INTRODUCTION

Since the HP lattice model developed by Lau and Dill [1] and used by others to derive similar approximation and heuristic search algorithms for a variety of lattice models has been proven useful to explore the relationship between the primary amino acid sequence and its native folding structure, particularly in protein folding problems (PFP) and protein structure prediction (PSP). The main purpose of the HP lattice model is to understand the physicochemical principle of protein folding during the modeling process of searching for the lowest free-energy conformation of a protein.

Despite a range of modeling accuracy, both high- and low-resolution models can contribute to an understanding of the protein structure obtained from different types of experiments, such as NMR and crystallography, and can have various applications in protein modification, protein-ligand and protein-protein interactions as discussed in Sali and Kuriyan [2] and summarized in Table 1.

TABLE I. THE RELATIONSHIP BETWEEN MODELING ACCURACY AND THE RELATED APPLICATION

| Accuracy | Application |
|---|---|
| <30% | Refining NMR structures<br>Finding binding/active sites by 3D motif searching<br>Annotating function by fold assignment |
| 30%-60% | Molecular replacement in crystallography<br>Supporting site-directed mutagenesis |
| >60% | Comparable to medium-resolution NMR, low-resolution crystallography<br>Docking of small ligands, proteins |

To improve the modeling accuracy, many lattice models have been developed and proposed. In the present study, 4 popular lattice models were compared only in terms of visual comparison. The models discussed are, namely, 2D square and triangular lattice models, 3D cubic lattice models. The protein structures obtained from the four modeling types were compared with reported 'real' biological protein structures. It was found that the 2D triangular lattice model could give a better structure modeling and prediction for proteins with short primary amino acid sequences as shown in Figure 1.

Figure 1 shows visual comparisons for PDB id: 1A0Ma. Figure 1(a) shows real protein structure; [1](b) and [1](c) are 2D square and triangular lattice model simulation results. Black-filled dots indicate Hydrophobic/non-polar amino acids and white dots denote hydrophilic/polar amino acid. [1](d) 3D cubic lattice model simulation result from CPSP-tools [5]. In (d), green balls indicate hydrophobic amino acids while the gray balls indicate the hydrophilic amino acids. Hart and Istrail [3] first gave a 1/4(25%) approximation for the problem of the 2D square lattice and a 3/8(38%)–approximation for the problem of the 3D cubic lattice. Agarwala et al. [4] gave a 6/11(54%) approximation for the problem, consistent with our experimental results.

However, due to the many benchmarks associated with the square lattice model, the large amount of data accumulated over the years and the availability of comparison with different strategies and modeling methods, many researchers have favored and focused research on the square lattice model. As a result, little has been done in the direction of the 2D triangular lattice model. In this paper, it

978-1-4244-8302-0/10/$26.00 ©2010 IEEE

is proposed that elite-based reproduction strategy (ERS) genetic algorithm (GA) and a hybrid of hill-climbing and genetic algorithms (HHGA) may be used for protein structure prediction on the 2D triangular lattice. It is hoped that other researchers can note the importance of this model.

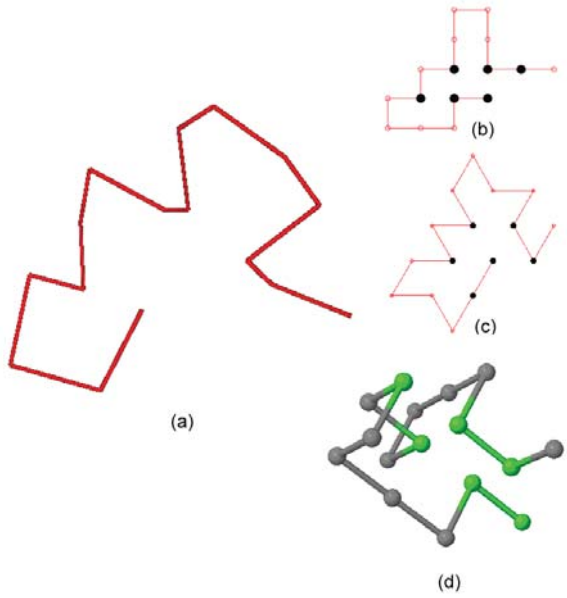


Figure 1. Visual comparison for PDB id: 1A0Ma. (a) is a real protein structure; (b) and (c) are 2D square and triangular lattice model simulation result[s]. Black-filled dots indicate Hydrophobic/non-polar amino acids and white dots denote hydrophilic/polar amino acid. (d) is 3D square lattice model simulation result from CPSP-tools [5]. In (d), green balls indicate hydrophobic amino acids while the gray balls indicate the hydrophilic amino acids. use left alignment.

The remainder of this paper is structured as follows: section 2 gives the preliminaries and the formal definition of the protein structure prediction problem in the HP 2D triangular lattice model. The methodology used in the study is explained in section 3 The comparison results are presented and discussed in section 4 followed by the conclusion in section 5.

## II. PRELIMINARIES

Proteins play fundamental and crucial roles in nearly all biological processes, such as, enzymatic catalysis, signaling transduction, embryonic development, DNA and RNA synthesis and others. It has been a long-standing goal of molecular biology to predict the tertiary structure of a protein from its primary amino acid sequence [6-7].

This paper emphasizes research on *ab initio* modeling. The 2D HP triangular lattice model discussed here is one such simplified method and is thought to be the best two-dimensional model in protein structure prediction at present.

### *A. HP lattice model*

The HP lattice model [1] is the most frequently used simplified model, which is based on the observation that the hydrophobic interaction between the amino acid residues is the driving force for the protein folding and for the development of native state in proteins [8]. In this model, each amino acid is classified based on its hydrophobicity as an H (hydrophobic or non-polar) or a P (hydrophilic or polar). The HP lattice model allows HP protein sequences to be configured as self-avoiding walks (SAW) on the lattice path favoring an energy free state due to HH interaction. The energy of a given conformation is defined as the number of topological neighboring (TN) contacts between those Hs, which are not adjacent in the sequence. Figure 2 shows an example for the 2D triangular lattice model.

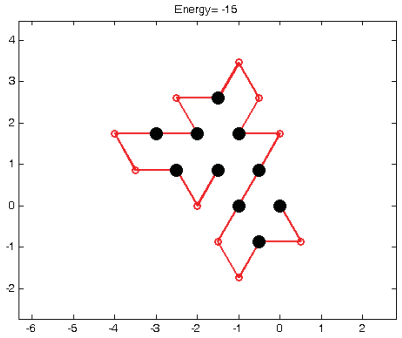


Figure 2. **An optimal conformation for the sequence "$(HP)^2PH(HP)^2(PH)^2H\ P(PH)^2$" in a 2D triangular lattice model.** The black filled dots denote the hydrophobic amino acid and the white open squares denote the hydrophilic amino acids. The H-H contacts (free energy) in the conformation, which are assigned the energy value of -1. The free energy is defined as a minimum value; the maximum number of H-H contact is given in the case of two-dimensional models, Figure 2 illustrates a protein structure with 15 H-H contacts (energy= -15).use left alignment.

### *B. Calculating the free energy*

The free energy for the protein can be calculated by using the following formulae [9]:

$$\epsilon_{ij} = \begin{cases} -1.0 & \text{the pair of H and H residues} \\ 0.0 & \text{others} \end{cases} \quad (1)$$

$$E = \sum_{i,j} \Delta r_{ij} \epsilon_{ij}, \quad (2)$$

where the parameter

$$\Delta r_{ij} = \begin{cases} 1 & S_i \text{and} S_j \text{are adjacent but not connected amino acids} \\ 0 & \text{others} \end{cases} \quad (3)$$

Hence, the problem of the optimization of protein folding into the calculation of the minimal free energy of the protein folding conformation can be transformed.

As a result, the following problem can be formally defined: given an HP sequence $s = s_1 s_2 ... s_n$, find an energy-minimizing conformation of $s$; that is: find $c^* \in C(s)$ such that $E(c^*) = \min\{E(c) \mid c \in C\}$, where $C(s)$ is the set of all valid conformations for $s$. [10]

### C. Triangular lattice model

A significant drawback of the cubic lattice [4] is that if two residues are at any even distance from one another in the primary sequence then they cannot be in topological contact with one another when the protein is embedded in this lattice. For example, on the square lattice, two amino acids in contact in any folding must be an odd distance away in the protein sequence [4]. Joel et al [11] introduced a 2D triangular lattice model. Consequently, the 2D triangular lattice can be described by a diagram as in Figure 3.

In the two-dimensional triangular lattice, each lattice point has six neighbours. Since each residue has two covalent neighbours except the first and the last residues, a residue at a lattice point may be in topological contact with at most four other residues. Thus, each residue may be involved in at most 4 H-H contact [11].

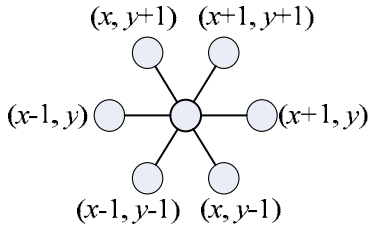


Figure 3. **Neighbors of vertex (*x*, *y*)**

The unit vectors shown in Figure 3 are logically defined. Real units require normalization by $\sqrt{2}$ and are (1,0),(-1,0), $(-1/2, \frac{\sqrt{3}}{2}), (1/2, -\frac{\sqrt{3}}{2}), (1/2, \frac{\sqrt{3}}{2}), (-1/2, -\frac{\sqrt{3}}{2})$.

After the unit vectors are obtained in the triangular lattice, it is much easier to model protein conformation on a two-dimensional triangular lattice and not exhibit the parity problem [4]. However, the lattice model of protein conformation as a self avoiding walk is NP-complete [12]. As a result, there are usually numerous likely conformations even for a protein with a short amino acid sequence. To solve this problem, many heuristic search algorithms [13-18] have been developed for a variety of lattice models. The algorithm developed from the present study is discussed in the next section.

## III. METHODS

In this paper, we propose an elite-based reproduction strategy (ERS), a genetic algorithm (GA) and a hybrid of hill-climbing and genetic algorithms (HHGA) for protein structure prediction on the 2D triangular lattice. Figures 4 and 5 show the flowchart of the proposed ERS-GA and HHGA. The proposed HHGA is mainly a combination of evolutionary algorithms (EA) with local search operators that work within the EA loop to enhance exploitation capability. The details are illustrated as follows:

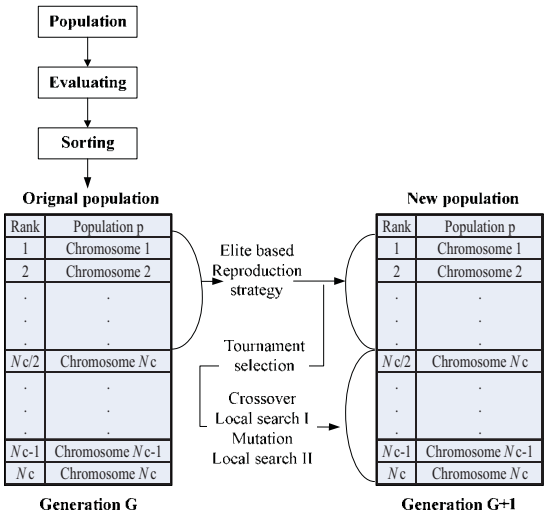


Figure 4. **Flowchart of the proposed ERS.**

### A. Initialization step

If the input amino acid sequence is of length n, then each individual in the population is a string of length n - 1 over the symbols = { *L*; *R*; *LU*; *LD* ; *RU*; *RD* }, and that denotes a valid conformation in the 2D triangular lattice [11,14]. The symbols *L*; *R*; *LU*; *LD*; *RU*; and *RD* are used to denote the fold directions: *L* is for left, *R* is for right, *LU* is for left-up, *LD* is for left-down, *RU* is for right-up and *RD* is right-down in the encoding scheme, respectively. An initial population is generated randomly and initializes an n - 1 dimensional space within a fixed range. Population size was set at 100.

- **Evaluating.** The evaluating step evaluates each chromosome in a population. The goal of the HIGA method is to minimize the fitness value. The lower a fitness value, the better the fitness. The fitness function is used by equation (1-3).

- **Sorting.** After the evaluating step, the next step is to sort the chromosomes in the population. After the whole population is sorted, the chromosomes are sorted in each group in the top half of population. This sorting can help to perform the reproduction

step because the best chromosome in each group can be retained. After sorting the chromosomes in the population, the algorithm goes to the next step.

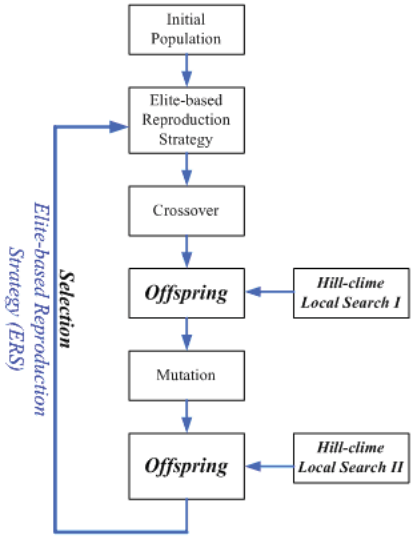


Figure 5. **Flowchart of the proposed HHGA.**

### B. *Elite-Based Reproduction Strategy (ERS)*

Reproduction is a process in which individual strings are copied according to their fitness value. In this study, the tournament selection method was used for this reproduction process. In particular, a height elite-based strategy (elite-based rate 50%) is used and the best performing individuals in the top half of the population [19] advance to the next generation. The other half is generated to perform crossover and mutation operations on individuals in the top half of the parent generation.

### C. *Crossover step*

The operation was performed for a selected pairs with a crossover rate that was set to 0.8 in this study. The first step is to select the individuals from the population for the crossover. Tournament selection is used to select the top-half of the best performing individuals [19]. The individuals are crossed and separated using a two-point crossover operation.

### D. *Local search I*

Local search I chooses the neighboring residues in a way similar to mutation operation, which introduces mutations into a specific gene on the corresponding chromosome. The likely direction can be chosen from {*L; R; LU; LD; RU; RD*}. If the minimum fitness value of a mutation introduced chromosome is better than that of the non-mutation introduced chromosome, the neighbour residues will be accepted to replace the current one on the chromosome. Otherwise, this chromosome will be rejected.

### E. *Mutation step*

When the mutation points are selected, the mutation rate is 0.1. For the HP protein structure prediction problem, every amino-acid residue owns each folding direction that is unique to the site.

### F. *Local search II*

After mutation, Local search II chooses the neighbored residues in a way similar to that used for the crossover operation. New individuals are created by changing the site's direction between point *i* and point *j*.

To generate neighboring folding directions, the gene from point *i* to *j* will be rotated by *60°*, *120°*, *180°*, *240°* and *300°* and the fitness value will be calculated individually from these five directions. If the new folding direction is superior to the original direction then one of the two motions is selected by competition. If the new folding direction is not better than the original direction, the original direction will not be changed.

### G. *Updating the local best and the global best*

In this step, the local best and the global best are up-dated. If the fitness value of a particle is higher than that of the local best, then the local best will be replaced with the particle; and if the local best is better than the current global best, than the global best is replaced]with the local best in the swarm.

### H. *Termination condition*

The algorithm is run for a maximum of 500 iterations. The best member of the population is then returned.

## IV. EXPERIMENTAL RESULTS

Sequences 1 through 8 used in this study were described in Jiang et al. [20]. These sequences have been used as the benchmark for the 2D square HP model as shown in Table II. However, the optimal energy of these benchmark sequences in the 2D triangular HP model was unknown; nevertheless, comparisons with previous studies provided a means of demonstrating the effectiveness of the method described here.

TABLE II. **THE BENCHMARKS FOR THE 2D TRIANGULAR LATTICE HP MODEL**

| 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 | $P^2H(P^2H^2)^2P^5H^{10}P^6(H^2P^3)^2HP^2H^5$ |
| 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. **COMPARISON OF THE PROPOSED APPROACH WITH THE SIMPLE GA (SGA), HYBRID GA (HGA).** FIGURES IN BOLD INDICATE THE LOWEST ENERGY.

| Seq. | Length | SGA[14] | HGA[14] | ERS-GA |
|---|---|---|---|---|
| 1 | 20 | -11 | -15 | **-15** |
| 2 | 24 | -10 | -13 | **-14** |
| 3 | 25 | -10 | -10 | **-11** |
| 4 | 36 | -16 | -19 | **-22** |
| 5 | 48 | -26 | -32 | **-34** |
| 6 | 50 | -21 | -23 | **-32** |
| 7 | 60 | -40 | -46 | **-62** |
| 8 | 64 | -33 | -46 | **-51** |

The experiments were conducted in two steps. First, elite-based reproduction strategy (ERS) genetic algorithm (GA) was used to predict the protein structure to evaluate the efficacy of this method. The results obtained were compared with prior work and are summarized in Tables III and IV. As indicated in Table III, the study yielded exceptionally good results for all 2D cases.

Next, the hill-climbing local search in the ERS-GA approach was added for the sake of improving the algorithm performance. This hybrid of hill-climbing and genetic algorithm (HHGA) approach can effectively enhance the performance.

Previously, Backofen and Will [21] utilized advanced techniques such as constraint programming. This is the first method that is able to calculate all optimal side-chain structures of a given sequence, while proving their optimality [5]. In the study performed by Böckenhauer et al. [15], the authors used a similar library and extended the library by implementing the 2D triangular lattice and the pull move set for triangular lattice models, which enabled them to obtain satisfactory results. From a comparison with past studies as shown in Table IV, HHGA is a similarly good approach in protein structure prediction.

TABLE IV. **COMPARISON OF HHGA APPROACH WITH THE TABU SEARCH (TS).** FIGURES IN BOLD INDICATE THE LOWEST ENERGY.

| Seq. | Length | TS[15] | HHGA | Conformation (HHGA) |
|---|---|---|---|---|
| 1 | 20 | **-15** | **-15** | Fig. 6 (a) |
| 2 | 24 | **-17** | **-17** | Fig. 6 (b) |
| 3 | 25 | **-12** | **-12** | Fig. 6 (c) |
| 4 | 36 | **-24** | -23 | Fig. 6 (d) |
| 5 | 48 | **-40** | **-40** | Fig. 6 (e) |
| 6 | 50 | - | **-34** | Fig. 6 (f) |
| 7 | 60 | **-70** | -66 | Fig. 6 (g) |
| 8 | 64 | -50 | **-54** | Fig. 6 (h) |

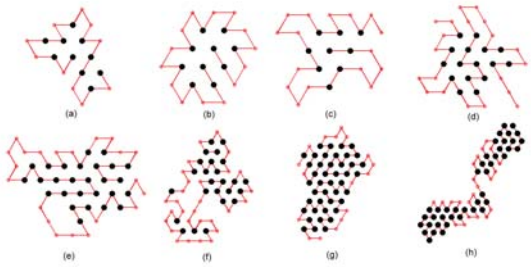

Figure 6. **(a) to (h) Results of the structure of 8 protein sequences**

## V. CONCLUSION

In the *ab initio* technique, the lattice model is one of the most frequently used methods in protein structure prediction. From visual comparison, it was found that the 2D triangular lattice model could yield better structure modeling sequences and prediction for proteins with short primary amino acid sequences. More examples are shown in Figure 6. Meanwhile, it was we realized that the 2D triangular lattice model is rarely used by other researchers in protein structure prediction.

This paper has highlighted this interesting issue and provides a short introduction to the working method for 2D triangular lattice models. Finally, the paper presents are elite-based reproduction strategy (ERS) genetic algorithm (GA) and a hybrid of hill-climbing and genetic algorithms (HHGA) for protein structure prediction on the 2D triangular lattice. The simulation results of the experiments show that ERS-GA and HHGA could successfully be applied to the problem of protein structure prediction It is hoped that this initial research can increase the awareness and interest from other researchers in the 2D triangular lattice model.

## 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. Šali and J. Kuriyan, "Challenges at the frontiers of structural biology", Trends in Genetics, 1999, pp. 20-24

[3] W. E. Hart and S. Istrail, "Fast protein folding in the Hydrophobic-Hydrophilic model within three-eighths of optimal (extended abstract)", Proceedings of 27th Annual ACM Symposium on Theory of Computation (STOC95), 1995, pp. 157–168.

[4] S. Decatur and S. Batzoglou, "Protein folding in the Hydrophobic-Polar model on the 3D triangular lattice", In 6th Annual MIT Laboratory for Computer Science Student Workshop on Computing Technologies, 1996.

[5] 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. 1–2.

[6] A. E. Mirsky, and L. Pauling, "On the structure of native, denatured and coagulated proteins", Proc. Natl. Acad. Sci. USA, 1936, pp. 439-447.

[7] C.A. Orengo, and A.E. Todd , "From protein structure to function", Curr. Opin. Struct. Biol., 1999, pp.374-382.

[8] Y-Z. Guoa, E-M. Fenga, and Y. Wangb, "Optimal HP configurations of proteins by combining local search with elastic net algorithm", Journal of Biochemical and Biophysical Methods, 2007, pp. 335-340.

[9] 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.

[10] 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

[11] G. Joel, M. Martin, and J. Minghui, "RNA folding on the 3D triangular lattice", BMC Bioinformatics, 2009, doi:10.1186/1471-2105-10-369.

[12] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, and M. Yannakakis, "On the complexity of protein folding", J. Comp. Biol., 1998, pp. 409–422.

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

[14] M.T. Hoque, M. Chetty, and L.S. Dooley, "A hybrid genetic algorithm for 2D FCC hydrophobic–hydrophilic lattice model to predict protein folding", in: Proceedings of the 19th ACS Australian Joint Conference on Artificial Intelligence, LNAI, Springer, 2006, pp. 867–876.

[15] 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.

[16] A.A. Albrechta, A. Skaliotisb, and K. Steinhöfelb, "Stochastic protein folding simulation in the three-dimensional HP-model", Computational Biology and Chemistry, 2008, pp. 248-255.

[17] 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.

[18] X. Zhao, "Advances on protein folding simulations based on the lattice HP models with natural computing", Applied Soft Computing, 2008, pp. 1029-1040

[19] C. J. Lin, and Y. C. Hsu, "Reinforcement hybrid evolutionary learning for recurrent wavelet-based neuro-fuzzy systems", IEEE Transactions on Fuzzy Systems, 2007, pp. 729–745.

[20] T. Jiang, Q. Cui, G. Shi, and S. Ma, "Protein folding simulations for the hydrophobic-hydrophilic model by combining tabu search with genetic algorithms", Journal of Chemical Physics, 2003, pp.4592-4596.

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