Laboratoire d'Informatique Fondamentale, de Recherche Opérationnelle, de Combinatoire et d'Économétrie

Availaible on line at http://www.liforce.usthb.dz

bulletin-liforce@usthb.dz

Bulletin du Laboratoire 01 (2021) 01 - 13

An Improved Simulated Annealing Algorithm for Optimization of Protein Folding Problem

BOUMEDINE Nabil[1], BOUROUBI Sadek[2,*]

[1,2] USTHB, Faculty of Mathematics,
P.B. 32 El-Alia, 16111, Bab Ezzouar, Algiers, Algeria.

nabil.doukou@gmail.com,
sbouroubi@usthb.dz or bouroubis@gmail.com

Résumé : The biological functionality of a protein is determined by its specific three-dimensional native structure. The challenging task of determining the native structure of a protein from its primary sequence is commonly known as the protein folding problem (PFP). The objective of the PFP problem is to predict the three-dimensional structure (conformation) of a given protein based on its primary amino acid sequence. Based on the folding process of the protein, the PFP problem can be formulated as a Combinatorial Optimization Problem with the use of a simplified model to decrease its complexity. Several approaches and heuristics have been proposed to solve PFP in different lattice models, including 2D square lattice, 2D triangular lattice, 3D cubic lattice, and 3D triangular lattice model. In this paper, we present Improved Simulated Annealing (ISA) to solve the problem of PFP in a 2D square lattice model.

Mots clés : Protein folding problem, H-P model; Minimal energy conformation; 2D square lattice; Improved Simulated Annealing algorithm (ISA).

*Corresponding author.

1 Introduction

Proteins are chains of amino acid residues that fold into specific configurations that are called native or tertiary 3D structures. The prediction of proteins based on their amino acid sequences (i.e., their primary structure) is commonly known as the Protein Folding Problem and referred to as PFP. The protein's biological function is essentially dependent on its native 3D structure [1]. Nevertheless, the phenomena of protein misfolding are the cause of many serious diseases such as Alzheimer's disease, Parkinson's disease and mad cow disease prions (diseases caused by a change in the native conformation of proteins) [2], [3], [4], [5]. It is very important to understand the folding process that guides proteins to achieve their native three-dimensional structure to develop treatments for these diseases. Anfinsen's, assumed that the native state of a protein is its minimum free energy conformation [6]. Besides, the folding process is driven basically by the hydrophobic interactions between amino acids which are the key to the development of the native conformation. The functional folding of proteins is mainly encoded by their amino acid sequence. Due to the complexity of the PFP problem, some simplified models have been introduced to reduce the level of complexity of this problem. In 1985, Dill has been proposed as one of the most widely used models in the study of PFP, called the hydrophobic-polar model and referred to as H-P model [7]. The H-P model is based on an important phenomenon in the folding process, the fact that the native state in real proteins is mainly guaranteed by the hydrophobic interactions between the amino acids of the primary structure that form the core of the protein [8]. The quality of folding in the H-P lattice model is expressed by the number of topological contacts between the hydrophobic amino acids (H-H contact) [9]. For every H-H contact, an energy value of -1 is assigned. The free energy is minimal if the number of H-H contacts is maximal. As a result, the PFP problem can be written as an optimization problem such that the goal is to identify the optimal conformation c^* that has the maximum number of H-H contacts in the conformational space as $E(c^*) = min\{E(c)/c \in C\}$, where C is the set of all possible conformations [9], [10]. Even, with a simplified model, it has been proved that the problem of finding the minimum energy is difficult to solve for both two-dimensional and three-dimensional lattices [11].

Since the H-P model has been available, various heuristic algorithms have been developed to solve the PFP problem in several different lattice models. In this work, we focus on the resolution of PFP using the H-P model in the two-dimensional square lattice. In this model, Unger and Moult [12], proposed the use of the Genetic Algorithm (GA), which is based on the number of genetic operators (i.e., selection, crossing, and mutation) to explore the search space. The mutation operator is used as a diversification technique and the crossover to create new high-quality solutions by exchanging a part of each other information between pairs of selected solutions. In the proposed GA, the probability of selecting a given solution is related to its fitness (i.e., the roulette selection operator). Therefore, other versions of GA were later proposed in [13], [9],[14], [15]. In [16], the authors applied an Ant Colony Optimization (ACO) algorithm for the PFP problem in the 2D H-P square lattice model, and then applied an extended version in [17]. This algorithm was also successfully applied in the HP model [10]. A Particle Swarm Optimization (PSO)

algorithm has been proposed in [18]. The use of the Immune Algorithm (IA) has been proposed in [19] and [20]. Besides, a hybrid approach has been developed for PFP, such as the GTS algorithm for the 2D H-P model [21], which combines GA with Tabu Search (TS).

In this paper, we propose an Improved Simulated Annealing for the PFP problem in the 2D square model, the proposed algorithm represented by ISA.
The rest of the paper is structured as follows: In section 2, we present the H-P model in the 2D square lattice, and we define the energy function. Next, in section 3, we describe the proposed approach in detail and the obtained results on different benchmark instances are compared with some of the algorithms. Finally, in section 4, we conclude the work and suggest some directions for future studies.

2 The H-P simplified model in the 2D square Lattice.

In the H-P model, the twenty amino acids existing in the literature are divided into two groups H and P. Let s be an amino acid sequence for a selected protein and n be the total number of amino acids in s. H-P model consists of transforming the s sequence into another s' such as:

$$s'_i = \begin{cases} H & \text{if the amino acid } i \text{ is of hydrophobic type,} \\ P & \text{if the amino acid } i \text{ is of polar type.} \end{cases}$$

As we show in Figure1, the single node in the 2D square lattice has four neighbors. To simplify, we encode the direction that generates the neighbors of each node with four numbers from 1 to 4.

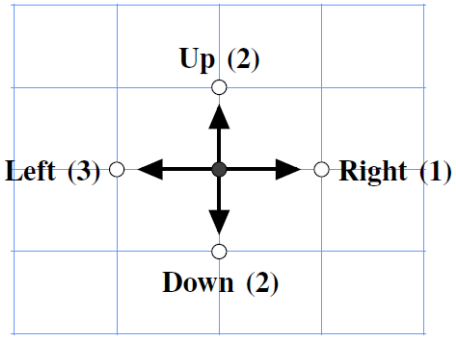

Figure 1: A numerical Encoding neighbors for a given node in the 2D square lattice.

2.1 The constraints of PFP using the H-P model in the lattice.

The native conformation of the sequence of n amino acids of protein s is characterized by the sequence of movements in the network, which are called self-avoidance pathways, that doesn't pass through the same node more than once. A valid configuration can be encoded by a vector of directions (i.e., $n-1$ list of movement directions in the lattice). The challenge is to find a conformation c^* for a given protein s in the lattice that contains the maximum number of H-H topological contacts to minimize the energy value. We define a valid conformation for a given protein sequence as the graph G in the lattice, which achieves the three following constraints:

1. For each amino acid in the protein sequence, it must occupy a node in the lattice.
2. A single node in the network can contain at most one amino acid.
3. Two adjacent amino acids in the sequence also occupy two adjacent lattice nodes.

2.2 The energy function.

Let s be a particular protein sequence in the H-P model, n the number of amino acids in s, and $x_{i,j}$ a binary variable. The energy value of a given valid conformation c for s is calculated by the following expression [9]:

$$E(s) = -\sum_{i=1}^{n-2}\sum_{j=i+2}^{n} x_{ij}y_{ij},$$

$$x_{ij} = \begin{cases} 1; & \text{if } (s_i = H) \text{ and } (s_j = H), \\ 0; & \text{otherwise,} \end{cases}$$

and

$$y_{ij} = \begin{cases} 1; & \text{if the amino acids } i \text{ and } j \text{ form an contact} \\ 0; & \text{otherwise.} \end{cases}$$

For example, the conformation $HPPPHHHHHHPHPPHPPH$ given in Figure2, has 7 H-H contacts, so the value of free energy associated with this conformation is -7 (i.e., $E = -7$).

3 The proposed Simulated Annealing.

In this work, we propose an Improved version of Simulated Annealing (SA) algorithm to solve the protein folding problem in the 2D H-P square lattice model. The proposed

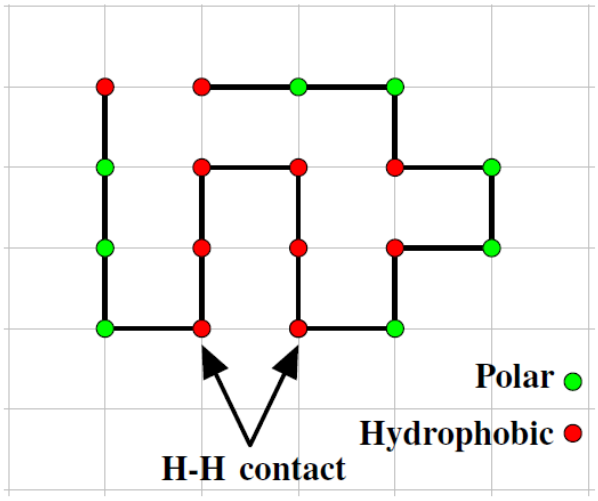

Figure 2: Feasible conformation for the sequence in the 2D square lattice.

algorithm uses an effective strategy to explore the search space by rotating a subsequence in the current solution. The latter improves the capacity of the standard SA for discovering a new region space. The process of the suggested algorithm is described in Algorithm 1 and described below.

3.1 Improved Simulated Annealing (ISA)

SA is one of the most efficient metaheuristic algorithms. SA has proven its efficiency in some methods to avoid local optima. Given a solution s, in order to escape the local optima, SA accepts movement $s' \in N(s)$ that worsens in the neighborhood of s, with a $P(s, s')$ probability inversely dependent on $\Delta(f(s) - f(s'))$ Of course, if the chosen move improves the quality of the solution, the new solution is always chosen (see algorithm 1). To improve the capacity of the SA in the diversification phase, we have incorporated an efficient strategy called rotation phase in the standard SA, this technique is used to explore a new regional space if there is no improvement in the quality of the best-found solution s^* after a fixed number of iterations (k_{max}), it consists in choosing two random positions i and j such that $i, j \in \{2, \ldots, n-1\}$ of the current solution and to rotate the subsequence $s_i, s_{i+1}, \ldots, s_j$. The rotation rule used is given in [22].
To control the frequency of acceptance of worsening movements, we use another parameter T as the temperature parameter, which is initially set to the value T_0. At an iteration i, the temperature is updated with the following formula:

$$T_i = \alpha T_{i-1}, \tag{3.1}$$

where

$$0 < \alpha < 1. \tag{3.2}$$

Algorithm 1 The Psuedo Code of ISA algorithm.

Require: The instance.

Ensure: The best solution s^*.

Begin
$s \leftarrow$ generate an initial solution in random way;
$s^* \longleftarrow$ the best solution s;
$T \leftarrow$ initial temperature;
$T_{min} \leftarrow$ minimal temperature;
$k_{max} \leftarrow$ maximum number of iteration after the use of diversification technique ;
$i \longleftarrow 0$;
While $(T > T_{min})$
$s' \leftarrow$ generate new solution by applying diagonal move;
$\Delta(s, s') = f(s') - f(s)$;
If $(\Delta(s, s') < 0$ or $random\ u \in (0, 1) < e^{\frac{-\Delta(s,s')}{T}})$
$s \longleftarrow s'$
$T \longleftarrow \alpha T$
End If
If $(f(s') < f(s^*))$
$s^* \longleftarrow s'$
Else
$i \leftarrow i + 1$;
End If
If $(i = k_{max})$
$s \leftarrow$ generate new solution by rotating a random
subsequence in s;
$i \leftarrow 0$;
End If
End while
Return the best found solution.
End

3.2 Diagonal Move Set in the 2D square lattice.

The proposed local move (diagonal move) is used in [23], to determine the neighborhoods of the given conformation, it consists of transferring the current conformation s to another one s'. We choose a vertex i from s, if there exists a free position in the lattice such that the position adjacent to vertex i and its predecessor $i-1$ or successor $i+1$ in the chain, we move the amino acid i to this free position. In the following figure, we present an example of the application of two diagonal moves. The conformation s' is obtained by transferring two amino acid (3 and 6) from the conformation s to their adjacent free positions. We show that the number of H-H contacts in initial conformation is increased by 3 units $(E(s') = -5 < E(s) = -2)$.

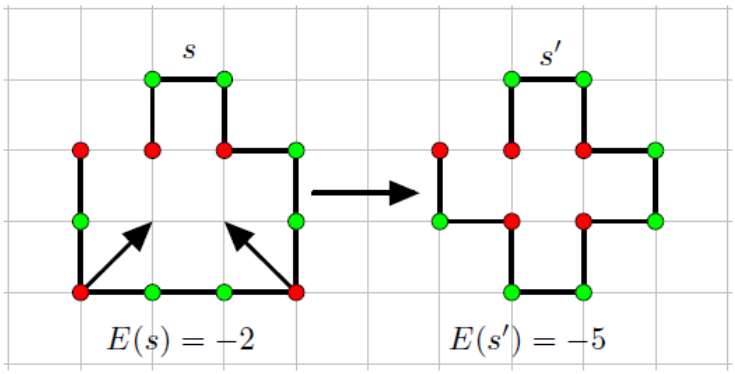

Figure 3: An example of the diagonal move set in 2D square lattice.

4 Experimental Results

The objective of this experimental study is to show the performance of the suggested algorithm (ISA) and to compare the results obtained by the suggested algorithm with a set of other existing algorithms, which are used to solve the PFP problem in a 2D square lattice model. To perform this experiment, we use a computer equipped with an Intel Core i5 processor and 4 GB RAM and MATLAB as the programming language to implement the proposed algorithm. The values of the parameters used for the proposed algorithm are given in Table 1. In order to assess its performance, we applied our improved ISA

Parameters	Values
Initial temperature T_0 ;	10000
α	0.95
k_{max}	10

Table 1: Parameters settings of ISA algorithm.

algorithm to the 8 standard benchmark instances for the 2D H-P Protein Folding Problem shown in Table 2. This table give a set of information for each benchmark sequence (i.e., number, length, the sequence in the H-P model), such that the symbol $(\ldots)^m$ is used to abstracted the m-fold duplicate of subsequence within parenthesis, for example, the sequence $HPHP$ is abstracted as $(HP)^2$. These instances have been widely used in the literature [9], [13], [14], [15], [16]. Experiments on these standard benchmark instances were conducted by performing 20 independent runs for each problem instance. Table 3 resumes the best-obtained results by the suggested algorithm (column ISA), and those taken from some algorithm that used to solve the PFP problem in the 2D square lattice model, Genetic Algorithm (column GA), Ant Colony Optimization algorithm (column ACO), and Monte Carlo algorithm (column MC).

Seq.	L	Protein sequence in the H-P model
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	40	$P^2H(P^2H^2)^2P^5H^{10}P^6(H^2P^2)^2HP^2H^5$
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 2: The used benchmark instances in the H-P model.

Seq.	Length	GA	ACO	MC	ISA
1	20	**-9**	**-9**	-8	**-9**
2	24	**-9**	**-9**	-8	**-9**
3	25	**-8**	**-8**	-7	**-8**
4	36	-12	**-14**	-12	**-14**
5	48	-22	**-23**	-18	**-23**
6	50	-21	**-21**	-19	**-21**
7	60	-34	-34	-31	**-35**
8	64	-35	-32	-31	**-39**

Values in bold indicate the best obtained result.

NA indicate that these data is not available in literature.

Table 3: The best Energy value obtained by ISA compared with other algorithms for 9 H-P sequences given in table 2.

Table 3 and Figure 4 show clearly the superiority of the proposed algorithm in terms of the quality solutions when compared to the other mentioned approaches. We show that the best conformation produced by the proposed algorithm are better than those obtained by the other algorithm for the most tested instances. Furthermore, we can also observe that the best results obtained by the suggested algorithm, achieve a strong improvement for instances 7 and 8 with a significant difference when compared to GA, MC, and ACO

algorithms. We notice that all compared approaches can easily achieve the best-known solution for instances ranging from 1 to 3. However, only our proposed method can find the best know solution for sequences 7 and 8. Consequently, we can say that the suggested algorithm ISA is at least comparable to the other algorithms.
To assess the effect of the rotation phase that we used in our suggested algorithm, we implemented the standard SA algorithm (without rotation phase) and we compared it with our improved version ISA that uses the rotation phase as a diversification strategy, in this comparison we use the best and average results as metric values for all fifteen runs of each algorithm. The best and average energy levels achieved by the standard SA and ISA are reported in Table 4.

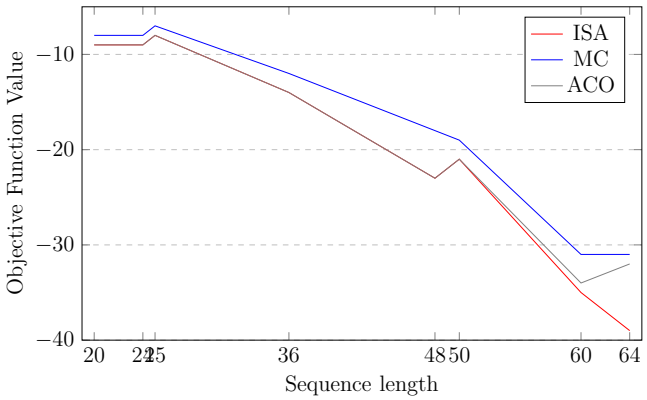

Figure 4: Illustration of the comparison results concerning the lowest energy values obtained using ISA against ACO and MC algorithms in 2D square lattice.

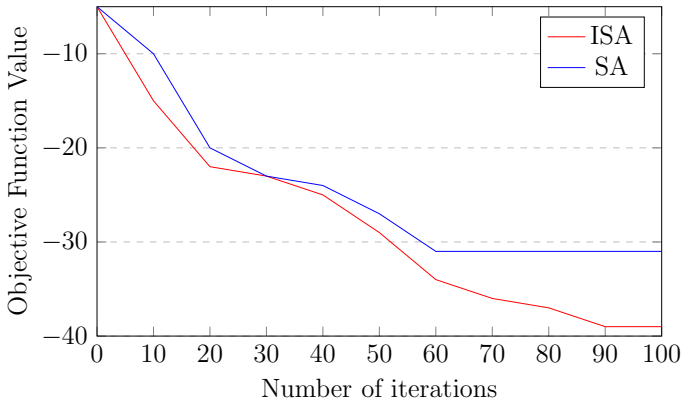

Figure 5: Comparison of the convergence between the standard SA algorithm and the improved version ISA.

From the best and average energy levels shown in bold-face in Table 4, it is evident that, for all the benchmark instances used in this study, ISA algorithm significantly outperforms better than the standard version of the standard SA algorithm in terms of both the best results and stabilities.
In Figure 5, we present a comparative analysis of the convergence between the Improved

SA algorithm and the original version of SA. The two algorithms were tested for sequence 8. To perform this comparison, we used the same initial solution and each algorithm ran for 1000 iterations. We notice that the standard SA algorithm is stagnated in the local minimum (i.e., premature convergence), on the other hand, the improved version of SA has a strong capacity to explore new regions of the search space.

	SA		ISA	
Seq.	Best	Average	Best	Average
1	**-9**	**-9**	**-9**	**-9**
2	**-9**	-8.61	**-9**	**-9**
3	-7	-6.83	**-8**	**-8**
4	**-14**	-11.92	**-14**	**-13.05**
5	-21	-18.39	**-23**	**-21.22**
6	-19	-17.42	**-21**	**-19.27**
7	-32	-29.31	**-35**	**-32.33**
8	-34	-30.01	**-39**	**-33.47**

Table 4: Comparison of the best solutions and stabilities of ISA with standard SA for 9 H-P sequences given in table 2.

5 Conclusion

In this paper, we presented the initial results of Protein Structure Prediction in the 2D square lattice model by using an improved Simulated Annealing algorithm. Concerning H-P instances and the minimum free energy obtained, our methods outperforms the published results on six out of ten benchmark problems for a 2D square lattice. These results encourage us to use this algorithm for solving the same problems in other types of lattices such as 3D lattices and 3D triangular lattices.

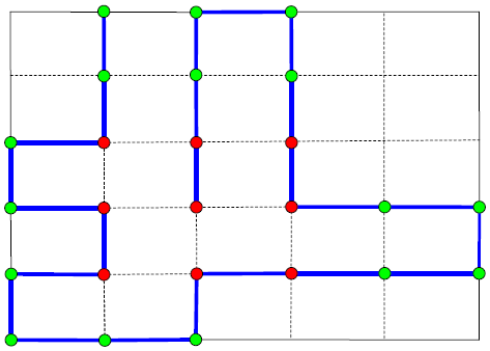

Figure 6: The best conformation obtained by ISA for the instance 3 with 8 H-H contacts (e.g., $E = -8$).

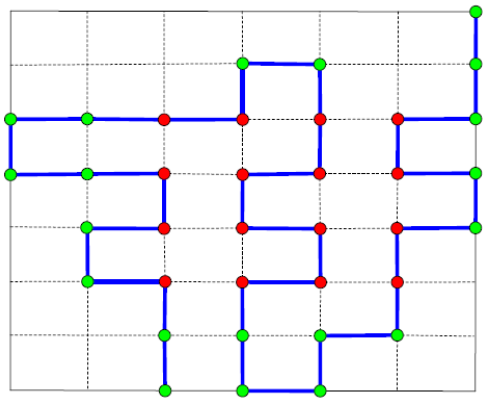

Figure 7: The best conformation obtained by ISA for the instance 4 with 14 H-H contacts (e.g., $E = -14$).

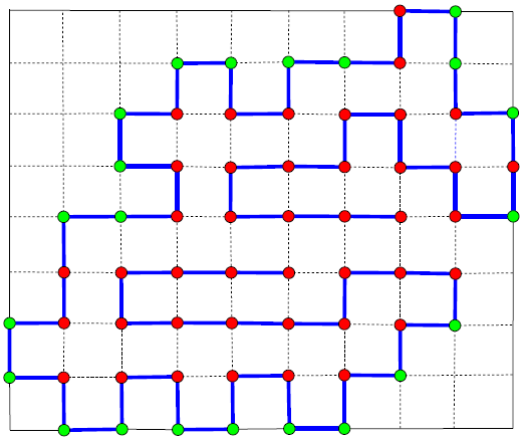

Figure 8: The best conformation obtained by ISA for the instance 6 with 39 H-H contacts (e.g., $E = -39$).

References

[1] Paul J. Hagerman and Ignacio Tinoco Jr. From sequence to structure to function. Current opinion in structural biology, 6(3): 277, 1996.

[2] John A. Hardy and Gerald A. Higgins. Alzheimer's disease: the amyloid cascade hypothesis. Science, 256(5054): 184-186, 1992.

[3] A. B. Singleton, M. Farrer, J. Johnson, A. Singleton, S. Hague, J. Kachergus, M. Hulihan, T. Peuralinna, A. Dutra, and R. Nussbaum. Synuclein locus triplication causes Parkinson's disease. Science, 302(5646): 841-841, 2003.

[4] Brian K. Nunnally and Ira S. Krull. Prions and mad cow disease. CRC Press, 2003.

[5] Juha Laurn, David A. Gimbel, Haakon B. Nygaard, John W. Gilbert, and Stephen M. Strittmatter. Cellular prion protein mediates impairment of synaptic plasticity by amyloid- oligomers. Nature, 457(7233): 1128-1132, 2009.

[6] Christian B. Anfinsen. Principles that govern the folding of protein chains. Science, 181(4096): 223-230, 1973.

[7] Ken A. Dill. Theory for the folding and stability of globular proteins. Biochemistry, 24(6): 1501-1509, 1985.

[8] Kit Fun Lau and Ken A. Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. Macromolecules, 22(10): 3986-3997, 1989.

[9] Cheng-Jian Lin and Ming-Hua Hsieh. An efficient hybrid Taguchigenetic algorithm for protein folding simulation. Expert systems with applications, 36(10): 12446-12453, 2009.

[10] Alena Shmygelska and Holger H. Hoos. An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem. BMC bioinformatics, 6(1): 30, 2005.

[11] Bonnie Berger and Tom Leighton. Protein folding in the hydrophobichydrophilic (HP) model is NP-complete. Journal of Computational Biology, 5(1): 27-40, 1998.

[12] Ron Unger and John Moult. Genetic algorithms for protein folding simulations. Journal of molecular biology, 231(1): 75-81, 1993.

[13] Thomas Dandekar and Patrick Argos. Folding the main chain of small proteins with the genetic algorithm. Journal of Molecular Biology, 236(3): 844-861, 1994.

[14] Natalio Krasnogor, William E. Hart, Jim Smith, and David A. Pelta. Protein structure prediction with evolutionary algorithms. In Proceedings of the 1st Annual Conference on Genetic and Evolutionary Computation-Volume 2, pages 1596-1601. Morgan Kaufmann Publishers Inc., 1999.

[15] Rainer Knig and Thomas Dandekar. Improving genetic algorithms for protein folding simulations by systematic crossover. BioSystems, 50(1): 17-25, 1999.

[16] Alena Shmygelska, Rosalia Aguirre-Hernandez, and Holger H. Hoos. An ant colony optimization algorithm for the 2D HP protein folding problem. In International Workshop on Ant Algorithms, pages 40-52. Springer, 2002.

[17] Alena Shmygelska and Holger H. Hoos. An improved ant colony optimisation algorithm for the 2D HP protein folding problem. In Conference of the Canadian Society for Computational Studies of Intelligence, pages 400-417. Springer, 2003.

[18] Andrei Butu and Henri Luchian. Protein structure prediction in lattice models with particle swarm optimization. In International Conference on Swarm Intelligence, pages 512-519. Springer, 2010.

[19] Vincenzo Cutello, Giuseppe Nicosia, Mario Pavone, and Jonathan Timmis. An immune algorithm for protein structure prediction on lattice models. IEEE transactions on evolutionary computation, 11(1): 101-117, 2007.

[20] Vincenzo Cutello, Giuseppe Morelli, Giuseppe Nicosia, and Mario Pavone. Immune algorithms with aging operators for the string folding problem and the protein folding problem. In European Conference on Evolutionary Computation in Combinatorial Optimization, pages 80-90. Springer, 2005.

[21] Tianzi Jiang, Qinghua Cui, Guihua Shi, and Songde Ma. Protein folding simulations of the hydrophobichydrophilic model by combining tabu search with genetic algorithms. The Journal of chemical physics, 119(8): 4592-4596, 2003.

[22] Cheng-Jian Lin and Shih-Chieh Su. Using an efficient artificial bee colony algorithm for protein structure prediction on lattice models. International journal of innovative computing, information and control, 8(3): 2049-2064, 2012.

[23] Hans-Joachim Bckenhauer, Abu Zafer M. Dayem Ullah, Leonidas Kapsokalivas, and Kathleen Steinhfel. A local move set for protein folding in triangular lattice models. In International Workshop on Algorithms in Bioinformatics, pages 369-381. Springer, 2008.