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 04 (2021) 73 - 85

Protein Structure Prediction in the HP Model Using Scatter Search Algorithm

Nabil BOUMEDINE[1], Sadek BOUROUBI[2]

[1,2] Faculty of Mathematics, Laboratory L'IFORCE,
University of Sciences and Technology Houari Boumediene (USTHB),
B.P. 32 El-Alia, Bab-Ezzouar, 16111 Algiers, Algeria.

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

Résumé : The aim of Protein Structure Prediction PSP is to predict the best conformation of a given amino acid sequence that has the lowest free energy. Because of its complexity, this problem has been widely studied under the HP simplified model with a different type of lattice as a conformational representation. Many algorithms have been proposed for solving the PSP problem using a set of benchmark instances. In this work, we propose an efficient Scatter Search Algorithm (SS) for the PSP problem in a 3D cubic lattice.

Mots clés : PSP problem; H-P model; SS algorithm; LS algorithm; Lowest energy conformation.

1 Introduction

The folding of a protein is guided by a complex process by which a chain of amino acids (primary structure) folds into a specific three-dimensional structure which is called the native structure. The biological function of a protein is determined by its native structures. Based on the thermodynamic hypothesis proposed by Anfinsen's, the folding quality of protein is directly related to its structure energy. In order to play their biological role, the proteins fold into the structure that has the lowest free energy among among all possible configurations using only their amino acid sequences (i.e. its primary sequences)[1]. However, changes in the polypeptide sequence or in the amino acids change the folding process and cause the misfolding of the protein. Misfolding of proteins can cause several diseases and allergies that called conformational diseases, such as Alzheimer's disease [15], Prions and mad cow disease [20], Parkinson's disease [6]. This is the main reason to study the protein folding problem and the process that guides the protein to folds into its native structure, in order to predict it based on the real mechanisms of the protein folding.

A number of experimental approaches were proposed in the literature for predicting the 3D native structure of proteins as the X-Ray Crystallographic Diffraction method [2] and Nuclear Magnetic Resonance (NMR) method [5]. However, these methods are very complex and need a lot of time to achieve their results. Furthermore, serval computational Methods have been proposed and developed by researchers for solving the protein structure folding problem based on simplified models. The most popular one is the Hydrophobic-Polar model (denoted by H-P model) which proposed by Dill et al in [12], this model is also based on the observation that hydrophobic interactions between amino acids have the biggest impact on the folding process and have a key role in the development of the native state of proteins. Although The use of the H-P model is simple and decreases the complexity of the PSP problem, it has been demonstrated that the PSP problem in the H-P model is an NP-hard problem in two-dimensional (2D) [7] and in the three-dimensional (3D) [4].

The Protein Folding Problem (PSP) consists of The folding of a protein is guided by a complex process by which a chain of amino acids (primary structure) folds into a specific three-dimensional structure which is called the native structure[14]. The biological function of a protein is determined by its native structures. Based on the thermodynamic hypothesis proposed by Anfinsen's, the folding quality of protein is directly related to its structure energy. In order to play their biological role, the proteins fold into the structure that has the lowest free energy Since the H-P model has been proposed, the problem of predicting the structure of proteins has become one of the most challenging open problems in computer seance, computational biology, and other fields. A large number of heuristics and meta-heuristics have been proposed for solving the H-P protein folding problem based on different type lattices in 2D and 3D [16]. Particularly, here we inter to the PSP problem in the 3D cubic lattice model. In this work, we propose an efficient scatter search algorithm (SS) for the PSP problem in a 3D cubic lattice model.

The rest of this paper is structured as follows: In section 2, we summarize the related work and a number of state-of-the-art approaches used to solve the PSP problem. In section 3, we present briefly the PSP problem in the H-P model, the energy function,

and the conformational space. In section 4, we give a brief description of our algorithm for solving the PSP problem. Section 5 reports the simulation results, and the analysis thereof, while conclusions and future work are presented in Section 6.

2 Related Work

In recent years, several approaches have been proposed for predicting the native structure of proteins on a 3D cubic lattice using the H-P model. In [17], the authors proposed the first genetic algorithm to solve the PSP problem. In [6], the authors developed an Evolutionary Algorithm hybridized with Backtracking (denoted by EA) which achieved a high-quality solution on a set of data benchmark instances. In [17], a hybrid algorithm has been developed by combining the Genetic and Particle Swarm Optimization algorithms (HGA-PSO). Two versions of the immune algorithm were proposed: aging-AIS [8] and ClonalgI [10]. The Artificial Bee Colony algorithm (MABC) is also used to solve PSP in a 3D cubic simplified model [18]. Furthermore, a number of approximation algorithms have been proposed for PSP in 3D with an approximation ratio of $\frac{3}{8}$ (i.e, 38%) in [15] and improved to $\frac{6}{11}$ (i.e, 54%) in [11].

3 Protein folding problem in H-P model

3.1 The Hydrophobic-Polar Model

In the hydrophobic-polar model (denoted by the H-P model), the amino acids of the primary structure which represent the components of proteins, have been divided into two categories. The hydrophobic class H and the polar class P. Let S the primary structure of a given protein and n its number of amino acids. Based on the hydrophobicity of amino acid we define the components of H-P sequence S' associated with S as follow:

$$\forall 1 \leq i \leq n,\ S'_i = \begin{cases} H, & \text{if the amino acid } i \text{ is of hydrophobic nature,} \\ P, & \text{if the amino acid } i \text{ is of polar nature.} \end{cases} \quad (1)$$

3.2 Solutions Representation

In the H-P model, we represent the structure of a protein by a chain in the lattice, such that each amino acid is presented by a vertex of the lattice. For simplicity, we use the direction of the movement in the lattice to encoding the structure. A feasible structure for a given protein is encoded by a movements sequence S_i of $n-1$ components in the 3D cubic lattice such that $S_i \in \{F, L, R, B, U, D\}^{n-1}$, where the symbols F, L, R, U, B,

D refer the direction Forward, Left, Right, Backward, Up and Down respectively. Each direction generates a neighbor for the current location.

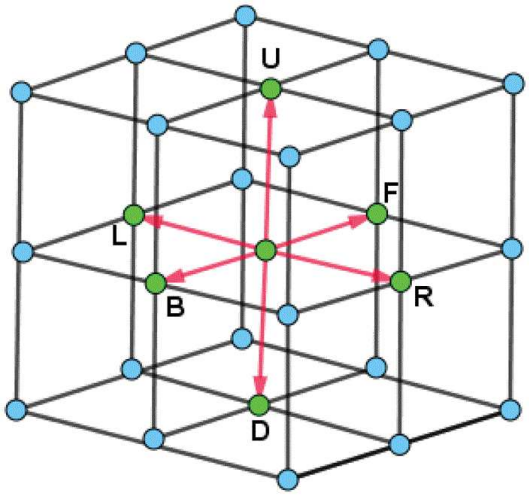

Figure 1: The relative movement in 3D cubic lattice model.

All valid conformations are self-avoiding paths on the 3D cubic lattice which achieve the following constraints:

1. Each amino acid in the protein sequence must be located at a vertex of the lattice.
2. A vertex of the lattice can contain a single amino acid at most.
3. Two amino acids that are located adjacent in the sequence also occupy two adjacent lattice points.

3.3 Objective Function

Based on the impact of hydrophobic interactions in the development of the active structure of the protein, the energy function in the H-P model adds a value of -1 for each pair of hydrophobic amino acids that occupy two adjacent vertices on the lattice but are not consecutive in the primary sequence (called the H-H topological contact). Let C a feasible conformation for a H-P protein sequence s of n amino acids and (x_i, y_i, z_i) is the position of the amino acid S_i in the 3D lattice, the free energy E(C) of the conformation C is calculated by the following formula:

$$E(C) = - \sum_{1 \leq i \leq j-2 \leq n} Q(i,j) \tag{2}$$

Where

$$Q(i,j) = \begin{cases} 1, & \text{if } \mid x_i - x_j \mid + \mid y_i - y_j \mid + \mid z_i - z_j \mid = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Based on the impact of hydrophobic interactions in the development of the active structure of the protein, the energy function in the H-P model adds a value of -1 for each pair of hydrophobic amino acids that occupy two adjacent vertices on the lattice but are not consecutive in the primary sequence (called the H-H topological contact). Let c a feasible conformation for an H-P protein sequence S of n amino The prediction of protein structure in the H-P model is a combinatorial optimization problem such that the goal is to find the optimal conformation that has the lowest free energy. This is equivalent to finding a feasible conformation c^* in the Conformational Space (CS) that has a maximum number of H-H topological contacts (i.e., $E(c^*) = min\{E(c)/c \in CS\}$).

4 Scatter Search optimization algorithm (SS)

The Scatter Search (SS) algorithm is an evolutionary method introduced by Glover (1977) for combinatorial optimization problems. The initial phase of the SS algorithm consists of creating a set of reference solutions P using a diversification generation method to ensure a level of diversity, then applying a heuristic approach to improve the quality of the solutions and selecting a set of high-quality and diverse solutions to be used as Reference set of Solutions (RS). The reference set RS can be characterized by two distinct subsets RS_{best} and RS_{div}, corresponding to the subsets of the high-quality and diverse solutions, respectively, were $RS = RS_{div} \cup RS_{best}$. In each iteration, SS applies a number of improvement methods on the combined solutions and updates RS according to the objective function. This procedure is iteratively repeated by the SS algorithm until the stopping criteria are satisfied. The SS process can be summarized by the following steps:

A) Generate the initial population according to the following steps:

1 Diversification generation strategy.

1 Improvement method.

2 Reference set update method.

B) Repeat the above SS process until the stopping criteria are met:

1 Method for generating subsets.

2 Combination method for the solutions.

3 Improvement solutions method.

4 Reference Set update method.

5 Our proposed SS for protein folding problem

In this section, we present an effective SS algorithm for the protein folding problem with a detailed description for its each step.

5.1 Initial phase

The first step of the SS algorithm is to generate the initial population P of m solutions using a diversification method. In our algorithm, firstly we generate all the solution in a random way where each solution in the search space is represented by a vector v of $n-1$ component then for each solution, we select two positions i and j such that $1 < i < j < n-1$, and we rotate the subsequence between i and j based on the rotation rules given in [18]. This technique is used to ensure a certain level of diversity. To improve the quality of the obtained solution by the diversification method, we use a local search (LS) as an improvement method. The proposed LS uses the diagonal move as neighbors search technique. Figure 2 presents an example of the diagonal move. For a given solution S, we choose a vertex i such that $1 < i < n-1$, if exists a free position in the lattice, such that this 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. The new solution S' will be accepted and replace S if its energy value is better than of S or a random value u satisfies $u < p$ such that $p \in [0, 1]$, the proposed LS is presented in algorithm 1:

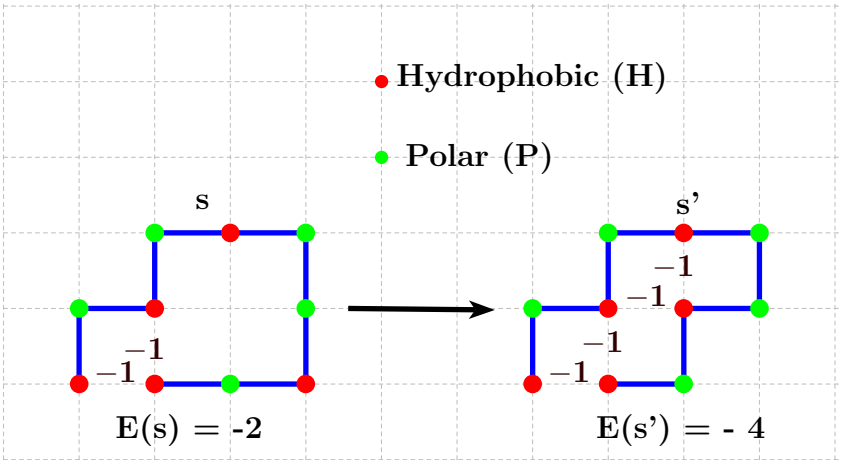

Figure 2: An example of the diagonal move used to define the neighbors in the 2D square lattice.

The obtained solutions in the first phase will be representing the initial reference set RS for the SS process.

Algorithm 1 Pseudo-Code of the proposed LS algorithm

Input: The initial solution; $p \in [0, 1]$.

Output: The best solution s_{best}.

Begin
Initialization: $x = (x_1, x_2, ..., x_n) \longleftarrow$ an initial solution;
Calculate $f(x)$ using the fitness function;
$iter = 0$;
while the stoping criterion is not met **do**
Create a new population by the following steps:
$k \leftarrow 0,\ P_{new} \leftarrow \emptyset$;
while ($iter \leq$ *Max-iteration*) **do**
$x' \longleftarrow x$;
Choose a random amino acid i among n;
$x' \longleftarrow$ apply the diagonal move rule to the i^{th} amino acid;
if ($f(x') \leq f(x)$ or $u < p$) **then**
$x \longleftarrow x'$;
end if
iter=iter+1;
end while
End

5.2 Scatter search phase

5.2.1 Subset Generation Method

This method aims to select a subset of the reference set RS as a basis for creating combined solutions. In this work we choose tow solutions $\{S_1, S_2\}$ from RS_{best} and one solution S_2 from RS_{div} in a random way in order to used in combined solutions method for the next step.

5.2.2 Solution Combination Method

Two combination method are used in this step. Firstly, we create two offsprings $\{O_1, O_2\}$ (i.e., new solutions) by applying the 1-point crossover operator to S_1 and S_3. Furthermore, we generate new solution O_3 by rotate a sub-sequence of S_2.

5.2.3 Improvement Method

As an improvement method, we use the same local search algorithm that is used in the initial phase in the Scatter Search.

5.2.4 Reference Set Update Method

We update the reference set based on the energy values by replacing the worst solution of RS_{best} with the new best solution among $\{O_1, O_2, O_3\}$ if the latter is better, and we

replace a random solution of RS_{div} by the worst solution among $\{O_1, O_2, O_3\}$.

6 Experimental Results

The experiments are conducted to compare our SS algorithm with exciting state-of-the-art optimization algorithms in the literature that used to solve the PSP problem in a 3D cubic lattice model. We are performing this experimental study using a reference H-P data set. These examples have been widely used in the literature. [3, 9, 21]. The parameters that are used for the SS algorithm are summarized in table 1.

Parameters	Values
Population size	25
RS_{best} size	15
RS_{div} size	10
p value for LS	0.1
Number of generation	50

Table 1: Parameters settings of SS algorithm.

The 3D H-P benchmark instances considered in our study are presented in Table 2, this table contains a set of information for each benchmark instance (identification number, length), such that the symbol $(...)^m$ represent m-fold duplicate of subsequence within parenthesis.

Seq.	Length	Protein sequence in the H-P model
$B1$	20	$(HP)^2PH(HP)^2(PH)^2HP(PH)^2$
$B2$	24	$H^2P^2(HP^2)^6H^2$
$B3$	25	$P^2HP^2(H^2P^4)^3H^2$
$B4$	36	$P(P^2H^2)^2P^5H^5(H^2P^2)^2P^2H(HP^2)^2$
$B5$	46	$P(PH^3)^2P^3(PH^2)^2P^2HPH^4PHP^2H^5(PH)^2HP^2H^2P$
$B6$	48	$P^2H(P^2H^2)^2P^5H^{10}P^6(H^2P^2)^2HP^2H^5$
$B7$	50	$H^2(PH)^3PH^4PH(P^3H)^2P^4(HP^3)^2HPH^4(PH)^3PH^2$
$B8$	60	$P(PH^3)^2H^5P^3H^{10}PHP^3H^{12}P^4H^6PH^2PHP$
$B9$	64	$H^{12}(PH)^2((P^2H^2)^2P^2H)^3(PH)^2H^{11}$

Table 2: The 3D H-P benchmark instances used in the our study.

In Table 4, for each benchmark instance, the best-found result achieved by the suggested algorithm and the mentioned algorithms are reported. We compare our SS algorithm with four algorithms including the Genetic Algorithm (column GA) [21], hybrid Genetic algorithm with the Particle Swarm Optimization (column HGA-PSO) [17], the Immune Algorithm (column IA) [8], Ant Colony Optimization algorithm (column ACO) [20] and Evolutionary Algorithm hybridized with Backtracking (column EA) [6]. The first column contains the best know solution for the corespondent instance (BSK).

Seq.	length	BKS	ACO	GA	HGA-PSO	IA	EA	SS
B1	20	**-11**	-10	**-11**	**-11**	**-11**	**-11**	**-11**
B2	24	**-13**	-8	**-13**	**-13**	**-13**	**-13**	**-13**
B3	25	**-9**	-6	**-9**	**-9**	**-9**	**-9**	**-9**
B4	36	**-18**	-10	**-18**	**-18**	**-18**	**-18**	**-18**
B5	46	**-33**	-21	NA	NA	NA	**-NA**	**-31**
B6	48	**-31**	NA	-25	-29	-29	-25	**-30**
B7	50	**-31**	NA	-23	-26	-23	-23	**-30**
B8	60	**-55**	NA	-37	-49	-41	-39	**-51**
B9	64	**-58**	NA	NA	NA	-42	-39	**-53**

Table 3: The best energy conformations obtained by SS algorithm compared with state-of-the-art algorithms for 9 H-P sequences in 3D cubic lattice model.

For the majority of instances tested our best result is better than those obtained by the state of-the art algorithms. Furthermore, we can also show that the best energy values obtained by the suggested algorithm, achieve a strong improvement for some instances with a significant difference when compared to those obtained by GA and ACO algorithms. The proposed SS algorithm finds the best know energy value for instances B1, B2, B3, B4, and B7, and achieves a near-optimal solution for instances B5, B6, B5 with a little difference even when compared to the best know energy value BKV. As we show in Figure 3, the proposed SS algorithm is more effective in solving the PSP in the term of best results.

Table 4 and Figure 4 show a comparison in terms of metric values (i.e., the best and average results obtained by the proposed SS algorithm and those taken from IA). The aim here is to compare the stability of our algorithm again to IA and EA after all 20 independent runs for each instance. Table 4 shows clearly the superiority of our algorithm than IA and EA algorithm in terms of the best and average energy values. Figure 6 shows the best conformation obtained by our SS algorithm for the instance B6, where the red color is used to present the hydrophobic amino acids and green color for the polar amino acids, this confirmation contains 30 topological contacts of H-H type, as a result of the free energy associate to this confirmation is $E = -30$.

	EA		IA		SS	
Seq.	Best	Ave.	Best	Ave.	Best	Ave.
B1	**-11**	-10.32	**-11**	**-11**	**-11**	**-11**
B2	**-13**	-10.90	**-13**	**-13**	**-13**	**-13**
B3	**-9**	-7.98	**-9**	**-9**	**-9**	**-9**
B4	**-18**	-14.38	**-18**	-16.76	**-18**	**-16.83**
B6	-25	-20.80	-29	-25.16	**-30**	**-26.04**
B7	-23	-20.20	-23	-22.60	**-30**	**-27.12**
B8	-39	-34.18	-41	-39.28	**-51**	**-45.91**
B9	-39	-33.01	-42	-39.08	**-53**	**-47.21**

Table 4: The best and average results obtained by SS algorithm compared with IA and EA algorithms for 20 independent runs.

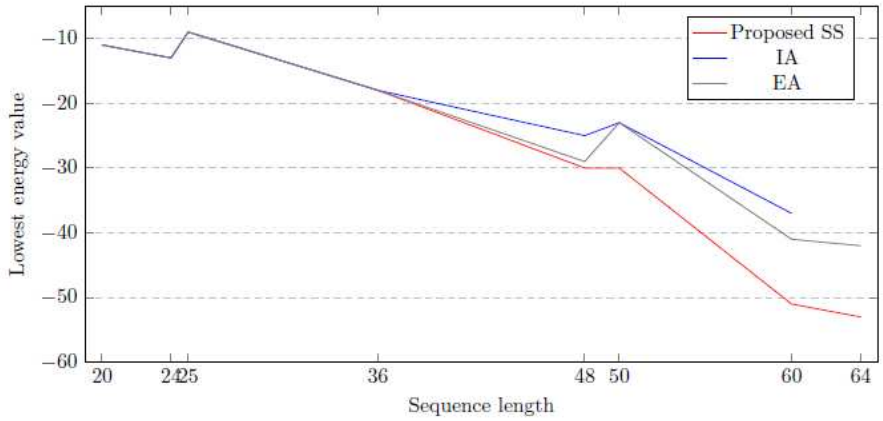

Figure 3: The best results achieved by IA, EA algorithms, and our SS algorithm.

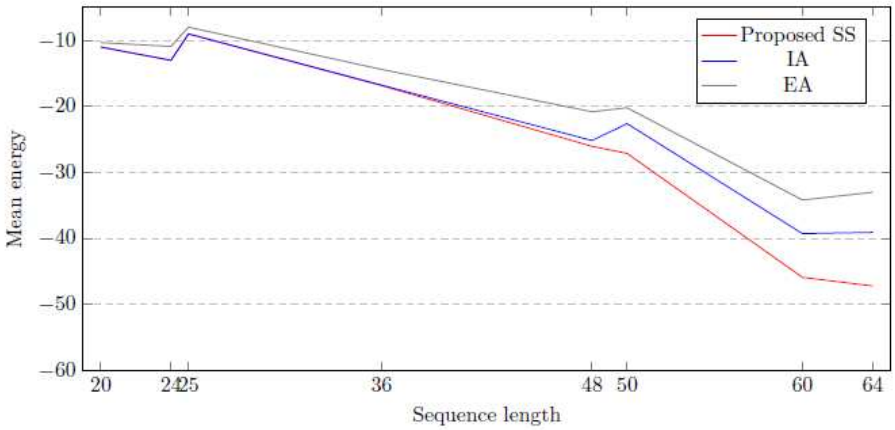

Figure 4: The average results achieved by IA, EA algorithms, and our SS algorithm.

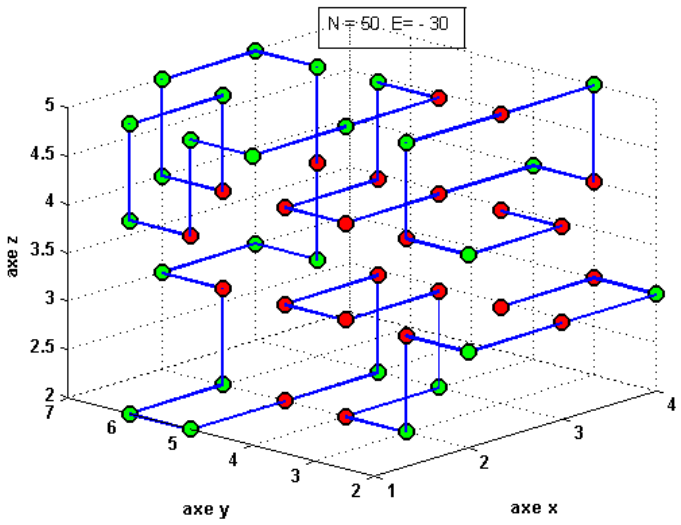

Figure 5: The best conformation obtained by SS algorithm for the instance B6

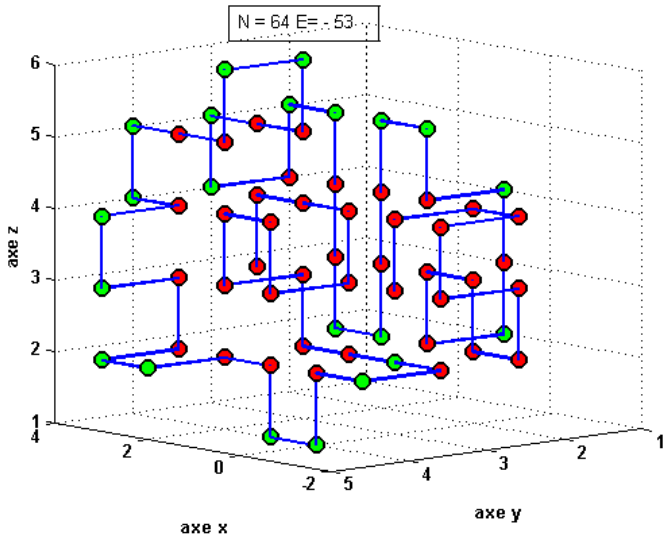

Figure 6: The best conformation obtained by SS algorithm for the instance B9

7 Conclusion

Since the H-P model exists, the protein folding problem has been widely studied and many heuristic and metaheuristics have been used to solve it. In this paper, we proposed an efficient scatter search algorithm for solving the PSP problem in a 3D cubic lattice model. The simulated results on a set of data benchmarks demonstrate the performance of the SS algorithm. In comparison, the results archived by SS are better and more stable than

those obtained by state-of-the-art algorithms in terms of the lowest and average energy. These results are very encouraging to use the same algorithm in other lattices such as 2D and 3D triangular lattice models.

References

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

[2] Eric T. Baldwin, Irene T. Weber, Robert St Charles, Jian-Cheng Xuan, Ettore Appella, Masaki Yamada, Kouji Matsushima, B. F. Edwards, G. Marius Clore, and Angela M. Gronenborn. Crystal structure of interleukin 8: symbiosis of NMR and crystallography. Proceedings of the National Academy of Sciences, 88(2):502-506, 1991.

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

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

[5] William Lawrence Bragg, David C. Phillips, and Henry Lipson. development of X-ray analysis. G. Bell, 1975.

[6] Carlos Cotta. Protein structure prediction using evolutionary algorithms hybridized with backtracking. In International Work-Conference on Artificial Neural Networks, pages 321-328. Springer, 2003.

[7] Pierluigi Crescenzi, Deborah Goldman, Christos Papadimitriou, Antonio Piccolboni, and Mihalis Yannakakis. On the complexity of protein folding. Journal of computational biology, 5(3):423-465, 1998.

[8] 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.

[9] 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.

[10] Carolina P. De Almeida, Richard A. Gonalves, and Myriam R. Delgado. A hybrid immune-based system for the protein folding problem. In European Conference on Evolutionary Computation in Combinatorial Optimization, pages 13-24. Springer, 2007.

[11] S. Decatur and Serafim 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.

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

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

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

[15] William E. Hart and Sorin C. Istrail. Fast protein folding in the hydrophobic hydrophilic model within three-eighths of optimal. Journal of computational biology, 3(1):53-96, 1996.

[16] 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.

[17] Cheng-Jian Lin and Shih-Chieh Su. Protein 3D HP Model Folding Simulation Using a Hybrid of Genetic Algorithm and Particle Swarm Optimization. International Journal of Fuzzy Systems, 13(2), 2011.

[18] 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.

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

[20] N. Thilagavathi and T. Amudha. ACO-metaheuristic for 3D-HP protein folding optimization. ARPN Journal of Engineering and Applied Sciences, 10(11):4948-4953, 2015.

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