A HYBRID GENETIC ALGORITHM FOR THE PROTEIN STRUCTURE PREDICTION PROBLEM

Nabil Boumedine [1], *Sadek Bouroubi* [2]

University of Sciences and Technology Houari Boumediene (USTHB),
Faculty of Mathematics, LIFORCE Laboratory.

ABSTRACT

The biological role of a protein is defined through the specific structure of this protein, which is known as the native structure. The challenge of protein structure prediction (PSP) is the identification of the native structure of proteins from their amino acid sequences. Although a number of experimental methods have been successfully applied to solve some small instances or those with specific properties of the protein structure prediction problem, they require huge computational time and resources. Given the limitations of existing experimental methods, computational methods have become the key to solving this problem. Several optimization algorithms have been successfully applied to solve the PSP in simplified models. In this paper, we propose an efficient hybrid genetic algorithm to predict the three-dimensional conformation of the protein in the hydrophobic-polar model.

keywords : protein structure prediction, minimal energy conformation, genetic algorithm, hill-climbing algorithm.

1. INTRODUCTION

Proteins are defined as biological macromolecules present in all living cells. They consist of one or more polypeptide chains. Each of these chains consists of a sequence of amino acid residues linked together by peptide bonds. Proteins fold in the space to adopt a unique native three-dimensional structure that determines their biological properties and functions [1]. Determining the structure of a protein is an essential part of understanding its biological roles. In addition, predicting the native structure of proteins can help solve several diseases that result from misfolding of protein structures including Alzheimer's disease, Piron's disease, and Parkinson's disease [2, 3]. In 1950, Christian Anfinsen was the first to propose the theory of protein folding, finding that the native conformation of a protein is determined by a set of interactions that occur at the atomic level and thus by the chemical properties of its specific amino acid sequence [4]. The protein structure prediction (PSP) problem is an NP-hard problem that aims to identify the structure of the protein from its amino acid sequence that has the minimum free energy to ensure the stability of the protein in its native state [5]. The experimental approaches such as X-ray crystallography [6], and nuclear magnetic resonance (NMR) spectroscopy [7], are the main techniques for determining the structure of proteins. However, these methods are expensive, time-consuming and are generally limited in their application to small proteins or to those with specific properties. As a result, computational approaches have become fundamental tools for protein structure determination and thus play a major role in developing our knowledge of these essential elements of life. Given the huge number of possible conformations even for small proteins, simplified models have been developed to reduce the complexity of the PSP problem. One such reduced representation that has been widely used in the resolution of the (PSP) is the so-called polar hydrophobic (H-P) model [8]. This model is based on the fact that the native structure of a protein is the conformation that minimizes the overall free energy.

This energy is essentially determined by the number of hydrophobic contacts that exist in this structure. Maximizing this type of contact leads to the minimum free energy that stabilizes the protein in its native (i.e. optimal) structure [9]. Thus, based on this assumption, the PSP problem can be considered as a combinatorial optimization problem where the goal is to finding the structure that maximizes the number of topological contacts between hydrophobic amino acids. Several heuristics and metaheuristics approaches have been applied to address the PSP problem in different types of lattices under the H-P model [10, 11]. In this paper, we propose an efficient hybrid approach to solve the PSP problem under the H-P model using the 3D triangular lattice for conformation representation. The proposed approach combines the genetic algorithm with an efficient hill-Climbing algorithm.

2. THE H-P SIMPLIFIED MODEL IN A 3D CUBIC LATTICE

In the HP model, the twenty amino acids are represented as string H or P [8]. Let s the protein sequence, and n is the number of amino acids existing in s. The HP model consists in converting the sequence s to another one s' such that :

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

The single node in a 3D cubic lattice contains exactly six neighbors. To simplify, we encode the directions that generates the neighbors of each node with the numbers from 1 to 6. A feasible conformation of a given protein sequence s of n amino acids is defined by the sequence of movements in the lattice, called self-avoidance paths, which do not pass through the same node more than once. Each conformation can be represented by $n-1$ movement directions in the lattice. The objective of the protein folding problem is to predict the native conformation, which maximizes the number of H-H topological contacts thus reducing the energy value. For example, the conformation given in 1, has 7 H-H contacts (i.e. E=-7).

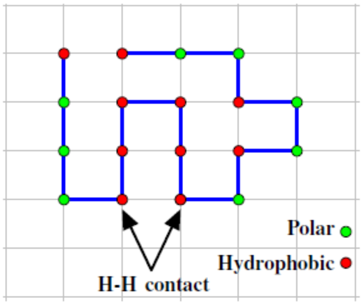

FIGURE 1 – Representation of conformation for the sequence of 18 amino acids in the 2D square lattice.

2.1. The energy function.

let s is a specific protein sequence in the H-P model, n is the number of amino acids in s, and $x_{i,j}$ binary variable. The energy value of a feasible conformation is calculated formula [12, 13] :

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

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

and

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

3. A HYBRID GENETIC ALGORITHM FOR THE PSP PROBLEM

In this section, we present an efficient hybrid method to solve the PSP problem. The proposed algorithm is called GA-HC in reference to the combined methods : the genetic algorithm and (GA) Hill-Climbing (HC) algorithm. The main motivation for this hybridization is to exploit the complementarity of these two different optimization strategies. The main motivation behind this choice of hybridization is the possibility of exploiting a very good complementarity between these two different optimization strategies.This complementarity is justified by the fact that the GA algorithm has a high capacity in diversification and in discovering new regions in the search space whereas the HC algorithm is very efficient in intensification phase. As a result, the suggested hybridization can represent a good balance between the exploration, and the exploitation of search space.

The proposed GA-HC algorithm starts with an initial population P_0 of m random solutions and iteratively improves their quality. At the beginning of each generation i, all solutions presented in the current population are evaluated by the objective function. We generate a new population P_i as follows : we use the tournament selection operator to select two solutions s_1, s_2 that will participate in the reproduction phase (i.e., crossover and mutation operators). Two new solutions will be generated by applying the crossover operation with a single crossover point. In order to improve the quality of these solutions, we introduce each of them as an initial solution in the proposed hill Climbing algorithm. If the quality of the obtained solutions s_1^*, s_2^* is better than the old ones, we replace s_1, s_2 by s_1^* and, s_2^* for the next generation. Otherwise we apply the mutation operator for both parents and introduce it into the new population.

3.1. Hill-Climbing

The Hill-Climbing (HC) method is used to find a local optimum from an initial solution and to improve it progressively at each iteration. Consider an optimization problem P, such that the goal is to find the best solution that minimizes the objective f in the set of feasible solutions X. The local search strategy starts with a random solution $s_0 \in X$ and at each iteration i, a new solution s_{i+1} is selected in the neighborhood $V(x_i)$ of the current solution. The efficiency of the HC algorithm depends strongly on the way the neighbors are defined and the strategy of moving from a given solution to another neighboring solution. In the proposed HC , the neighbors of a given solution are defined by a set of diagonal moves in the lattice. It consists in choosing a random selected amino acid i and changing its position p in the lattice by a another one p' which diagonally adjacent to p. If the resulting solution is better than the old one, it will be the current solution for the next iteration.

For example, Figure 2 shows a feasible conformation for the protein sequence given in the H-P model by PHPHPHHHHPPPH which was obtained by applying the diagonal move to the 4th amino acid in the conformation s, the new solution s' has 6 H-H contacts, which is better than the initial solution which contains 4 H-H contacts.

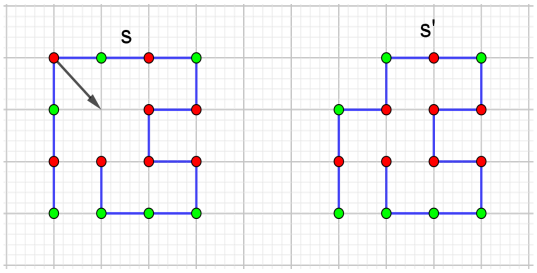

FIGURE 2 – An example of diagonal movement in the 2D square lattice.

4. EXPERIMENTAL RESULTS

The aim of this experimental study is to demonstrate the effectiveness of the proposed algorithm (GA-HC), as well as to compare the obtained result by the proposed algorithm with a set of existing algorithms, that have been applied to solve the PSP problem in the 2D triangular lattice model. To achieve this experiment, we use a computer with an Intel Core i5 processor and 4 GB of RAM and we use MATLAB as a programming language to implement the suggested algorithm. The parameters we used for the proposed algorithm are summarized in the Table 1.

We have tested the proposed algorithm on 9 benchmark datasets that have been widely used in the literature to solve the PSP problem in a 2D traingular lattice, these instances are listed in Table 2 which was introduced in [14, 15, 16]. This table contains a set of information related

Parameters	Values
Population size	80
Crossover probability	0.9
Number of generation	40

TABLE 1 – Parameters settings of GA-HC algorithm.

Seq.	Length	Sequence	E*
s_1	20	$(HP)^2PH(HP)^2(PH)^2HP(PH)^2$	-11
s_2	24	$H^2P^2(HP^2)^6H^2$	-13
s_3	25	$P^2HP^2(H^2P^4)^3H^2$	-9
s_4	36	$P(P^2H^2)^2P^5H^5(H^2P^2)^2P^2H(HP^2)^2$	-18
s_5	46	$P^2H^3PH^3P^3HPH^2PH^2P^2HPH^4PHP^2H^5PHPH^2P^2H^2P$	-35
s_6	48	$P^2H(P^2H^2)^2P^5H^{10}P^6(H^2P^2)^2HP^2H^5$	-31
s_7	50	$H^2(PH)^3PH^4PH(P^3H)^2P^4(HP^3)^2HPH^4(PH)^3PH^2$	-34
s_8	60	$P(PH^3)^2H^5P^3H^{10}PHP^3H^{12}P^4H^6PH^2PHP$	-55
s_9	64	$H^{12}(PH)^2((P^2H^2)^2P^2H)^3(PH)^2H^{11}$	-59

TABLE 2 – 3D HP benchmark instances for PSP problem.

to this set of benchmarks (e.g., the length, the sequence in the H-P model). Whereas the symbol $(...)^m$ is used to represent m duplication of the subsequence in brackets. For example, the sequence $HPHP$ is equivalent to $(HP)^2$.

Table 3 illustrates the results obtained by the proposed algorithm (GA-HC), and the best results obtained by some interesting algorithms from the literature that have been used to solve the PSP problem in a 3D cubic lattice model, including Genetic Algorithm (GA) [16], Ant Colony Optimization algorithm (ACO)[17], a Hybrid algorithm that combines Genetic Algorithm and Particle Swarm Optimization algorithm (HGA-PSO) [18], the Best Improvement Local Search (BILS) algorithm, Immune Algorithm (AIS) [19], Evolutionary Algorithm (EA column) [15] and, Scatter Search algorithm (SS)[20]. The results given in Table 3 are reported after 20 independent runs for each instance considered.

Seq.	Length	BKV	ACO	SS	HGA-PSO	AIS	EA	GA	BILS	GA-HC
$3d1$	20	**-11**	-10	**-11**	**-11**	**-11**	**-11**	**-11**	-10	**-11**
$3d2$	24	**-13**	-8	**-13**	**-13**	**-13**	**-13**	**-13**	-9	**-13**
$3d3$	25	**-9**	-6	**-9**	**-9**	**-9**	**-9**	**-9**	-7	**-9**
$3d4$	36	**-18**	-10	**-18**	**-18**	**-18**	**-18**	**-18**	-12	**-18**
$3d5$	46	**-35**	-21	-31	NA	NA	-30	-32	-22	-31
$3d6$	48	**-31**	NA	-30	-29	-29	-29	-31	-19	-30
$3d7$	50	**-34**	NA	-30	-26	-23	-25	-30	-18	**-31**
$3d8$	60	**-55**	NA	-51	-49	-41	-48	-50	-36	**-53**
$3d9$	64	**-59**	NA	-53	NA	-42	-45	-52	-34	**-54**

Bold values are the best energy values.

NA is referring to unavailable data in the literature.

TABLE 3 – The best results were obtained by GA-HC in comparison with state-of-the-art approaches for a 3D cubic lattice model.

Table 3 clearly shows the performance and superiority of the proposed GA-HC algorithm with regard to the quality of the solutions when it is compared to the other five methods mentioned. We can show that the best conformations produced by the proposed algorithm are better than those generated by the other algorithms for the most tested instances. Moreover, we can also observe that the best conformation obtained by the proposed algorithm, archives a strong improvement for instances s_7, s_8, and s_9 with significant deference, when compared to AIS, HGA-PSO, BILS, and EA algorithms. We can see also that the SS, and GA algorithms and suggested algorithm method can easily obtain the best know solution when the length of the instance is smaller than 48. However, our proposed method perform better than the SS, and GA for the long sequences (i.e., s_7, s_8, and s_9), with some exception, for example the best result obtained by the GA algorithm is better than ours for the sequences s_5 and , s_6 withe a little difference. Consequently, we can say that the suggested algorithm GA-TS are at least comparable to the other algorithms.

5. CONCLUSION

The protein structure prediction problem is an NP-hard problem, which consists of predicting the structure of proteins based on their base sequence. In this paper, a successful method was proposed to solve this problem in 3D cubic lattice, the proposed hybrid method combines the genetic algorithm with the Hill-Climbing search method. According to the results obtained, we can say that this approach can be produced good quality solutions compared to other existing approaches, these results encourage us to use the proposed approach to solve other combinatorial optimization problems.

6. REFERENCES

[1] Paul J. Hagerman and Ignacio Tinoco Jr. From sequence to structure to function. *Current opinion in structural biology*, 6(3) :277–280, 1996.

[2] John A. Hardy and Gerald A. Higgins. Alzheimer's disease : the amyloid cascade hypothesis. *Science*, 256(5054) :184–186, 1992. Publisher : American Association for the Advancement of Science.

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

[4] Christian B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096) :223–230, 1973. Publisher : JSTOR.

[5] Bonnie Berger and Tom Leighton. Protein folding in the hydrophobic-hydrophilic (HP) is NP-complete. In *Proceedings of the second annual international conference on Computational molecular biology*, pages 30–39, 1998.

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

[7] 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. Publisher : National Acad Sciences.

[8] Ken A. Dill. Theory for the folding and stability of globular proteins. *Biochemistry*, 24(6) :1501–1509, 1985. Publisher : ACS Publications.

[9] 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. Publisher : ACS Publications.

[10] Nabil Boumedine and Sadek Bouroubi. An Improved Simulated Annealing Algorithm for Optimization of Protein Folding Problem. In *2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH)*, pages 246–251. IEEE, 2021.

[11] Nabil Boumedine and Sadek Bouroubi. Protein folding simulations in the hydrophobic-polar model using a hybrid cuckoo search algorithm. *arXiv preprint arXiv :2105.13226*, 2021.

[12] Cheng-Jian Lin and Ming-Hua Hsieh. An efficient hybrid Taguchi-genetic algorithm for protein folding simulation. *Expert systems with applications*, 36(10) :12446–12453, 2009. Publisher : Elsevier.

[13] Nabil Boumedine and Sadek Bouroubi. A new hybrid genetic algorithm for protein structure prediction on the 2Dtriangular lattice. *Turkish Journal of Electrical Engineering & Computer Sciences*, 29(2) :499–513, 2021. Publisher : The Scientific and Technological Research Council of Turkey.

[14] Tianzi Jiang, Qinghua Cui, Guihua Shi, and Songde Ma. Protein folding simulations of the hydrophobic–hydrophilic model by combining tabu search with genetic algorithms. *The Journal of chemical physics*, 119(8) :4592–4596, 2003. Publisher : American Institute of Physics.

[15] Mario Garza-Fabre, Gregorio Toscano-Pulido, and Eduardo Rodriguez-Tello. Multi-objectivization, fitness landscape transformation and search performance : A case of study on the hp model for protein structure prediction. *European Journal of Operational Research*, 243(2) :405–422, 2015. Publisher : Elsevier.

[16] Mario Garza-Fabre, Eduardo Rodriguez-Tello, and Gregorio Toscano-Pulido. Comparative analysis of different evaluation functions for protein structure prediction under the HP model. *Journal of Computer Science and Technology*, 28(5) :868–889, 2013. Publisher : Springer.

[17] 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.

[18] 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.

[19] 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.

[20] Boumedine Nabil and Bouroubi Sadek. Protein structure prediction in the HP model using scatter search algorithm. In *2020 4th International Symposium on Informatics and its Applications (ISIA)*, pages 1–5. IEEE, 2020.