Predicting the non-compact conformation of amino acid sequence by particle swarm optimization

Yuzhen Guo
Department of Mathematics
Nanjing University of Aeronautics and Astronautics
Nanjing, China
guoyuzhen@nuaa.edu.cn

Yong Wang
Academy of Mathematics and Systems Science
Chinese Academy of Sciences
Beijing, China
ywang@amss.ac.cn

***Abstract*—Hydrophobic-hydrophilic (HP) model serves as a surrogate for the protein structure prediction problem to fold a chain of amino acids into a 2D square lattice. By the fact that the number of amino acids is equal to the number of lattice points or not, there are two types of folding conformations, i.e., the compact and non-compact conformations. Non-compact conformation tries to fold the amino acids sequence into a relatively larger square lattice, which is more biologically realistic and significant than the compact conformation. Here, we propose a heuristic algorithm to predict the non-compact conformations in 2D HP model. First, the protein structure prediction problem is abstracted to match amino acids to lattice points. The problem is then formulated as an integer programming model and we transform the biological problem into an optimization problem. Classical particle swarm optimization algorithm is extended by the single point adjustment strategy to solve this problem. Compared with existing self-organizing map algorithm, our method is more effective in several benchmark examples.**

Keywords—Protein structure prediction; Non-compact conformation; HP lattice model; Particle swarm optimization.

I. INTRODUCTION

Protein's biological function heavily depends on its spatial structure, so protein structure prediction is important to understand the relationship between structure and function, to design drugs with specific therapeutic properties, and to grow biological polymers with the specific material properties. Therefore, protein structure folding, i.e., determining the 3D structure from the 1D amino acid sequence by minimizing the energy function, remains one of the most challenging problems in computational biology and global optimization [1].

HP model was proposed by Dill and his colleagues in 1995 [2]. This model is perhaps the most simplified but widely used one. It is abstracted from three facets: Firstly, the skeleton structure is only studied in geometry; Secondly, twenty amino acids are classified into two groups as hydrophobic (H) or hydrophilic (P); At last, the major contribution of interaction between two amino acids is due to the interaction between hydrophobic amino acids, which are not adjacent in the sequence but are adjacent in the spatial location. This pair of interaction is denoted as HH and its energy is usually treated as -1. HH interaction tends to form a core in the spatial structure, and hydrophilic amino acids shield the core from surrounding solvents. The energy of natural conformation is assumed to be minimal, so protein structure prediction problem will be translated to maximize the number of HH interactions from the non-covalently interacting lattice neighbors. HP model represents the hard core of protein folding problem and is widely used in practice. For example, chemists evaluate new hypothesis of protein structure prediction by HP model. Also the simplicity of model allows a rigorous analysis of efficiency for a folding algorithm. In fact, this model has become a standard in testing efficiency of folding algorithm [3].

However, HP model is hard to solve in computation. Crescenzi has proved that decision problem for 2D HP model is NP-complete, therefor its optimization problem is an NP-hard problem [4]. How to find an effective heuristic method will be a primary research objective for 2D HP lattice model. In the past ten years, many computational strategies have been proposed to find the minimal free energy conformation in 2D HP model, including genetic algorithm [5], self-organizing mapping method [6], elastic net algorithm [7], Monte Carlo algorithm [8], and EE sampling approach [9], etc. The effectiveness of these methods has been tested with benchmark sequences for 2D HP lattice model. But currently, none of these algorithms seems to dominate the others.

Particle swarm optimization (PSO) was presented by Kennedy and Eberhart in 1995, inspired by the social behaviors of bird flocking [10]. PSO can be used to solve nonlinear problem, non-differentiable problem, and multimodal problem [11]. PSO only requires that the problem is computable and the parameters need to be adjusted are very few. In addition, its principle is simple, and this method can be easily implemented and applied in practice. Currently, PSO has been used to successfully solve travelling salesman problem (TSP) [12]. HP lattice model is similar to TSP and both can be formulated as matching problems. So it is feasible that protein structure folding problem will be attacked by PSO. In this paper, we will define adjustment operators and adjustment sequences to extend PSO for predicting conformations of amino acid sequences.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

II. COMBINATORIAL OPTIMIZATION FOR HP MODEL

Suppose that the number of amino acids in a protein sequence is n and the number of lattice points is m. If $m = n$, the conformation of the sequence in the lattice is defined as compact. If $m > n$, the conformation is defined as non-compact. Searching the optimal compact conformation in 2D lattice was studied by PSO in [13]. In this paper, we will consider to find the optimal non-compact conformation in 2D lattice. We will formally formulate the problem as follows. If the i-th amino acid occupies the vertex j of the lattice, denote $x_{ij} = 1$, else $x_{ij} = 0$. Let $N(\mathrm{j})$ be the set of all adjacent vertexes of the j-th vertex, where $|N(\mathrm{j})| = \{0,1,2,3\}$ is the number of elements in $N(\mathrm{j})$. f is the mapping from hydrophobic and hydrophilic to a binary value. If the i-th amino acid is H, then $f(\mathrm{i}) = 1$, else $f(\mathrm{i}) = 0$. Y_i is the coordinate of the i-th amino acid in lattice. Then the number of HH pairs is calculated by following equation

$$\frac{1}{2}\sum_{j=1}^{m}\left[\sum_{i=1}^{n} f(i)x_{ij}\sum_{s\in N(j)}\sum_{i=1}^{n} f(i)x_{is}\right]. \qquad (1)$$

Then we present minimum free energy model of 2D HP protein folding problem with constraints as follows:

$$\begin{aligned} max \quad & \frac{1}{2}\sum_{j=1}^{m}\left[\sum_{i=1}^{n} f(i)x_{ij}\sum_{s\in N(j)}\sum_{i=1}^{n} f(i)x_{is}\right] \\ \text{s.t.} \quad & \sum_{i=1}^{n} x_{ij} = 0/1, \quad \mathrm{j}=1,2,\cdots \\ & \sum_{j=1}^{m} x_{ij} = 1, \quad \mathrm{i}=1,2,\cdots \\ & |Y_i - Y_{i+1}| = 1, \quad \mathrm{i}=1,2,\cdots \end{aligned} \qquad (2)$$

In this way, we translate the protein structure prediction pro blem to a combinational optimization problem. This problem is known as NP-hard. As the next step, we will design efficient h euristic algorithm to solve it.

III. IMPROVED PARTICLE SWARM OPTIMIZATION

A. Definitions

In order to efficiently solve the optimal problem (2) by particle swarm optimization, we will make some preparations as follows.

- Definition 1. The state vector of the i-th particle is denoted as $X_i = (x_{i1}, x_{i2}, \cdots$, where x_{ij} is the label of lattice point which is visited by the j-th amino acid.
- Definition 2. The operator $\mathrm{T}(k,l)$ shows that the k-th element is placed after the $(l-1)$-th element, then the l-th element and other elements after it will be ranked in the original sequence for particle X_i. This operator is defined as adjustment operator. The new particle is denoted as $X_i^{'}$. For example, particle $\mathrm{X} = (1,2,3,4,5)$, so $\mathrm{X}^{'} = \mathrm{X} + \mathrm{T}(5,2) = (1,5,2,3,4)$, where "+" indicates that the particle X is adjusted by the adjustment operator $\mathrm{T}(5,2)$.
- Definition 3. If the particle is adjusted by a number of adjustment operators, then these ordered operators are defined as adjustment sequence V, in which arbitrary operator is not satisfied with commutative law.
- Definition 4. The basic adjustment sequence includes the least adjustment operators which can reach the adjustment purpose.

B. Improved Particle Swarm Optimization

Step 1. Initialize the particle swarm X_i^0 and the adjustment sequence V_i^0 called as the initial velocity, $i \in \{1,2,\cdots$ where l is the number of particles.

Step 2. Based on (1), compute the fitness value of every particle according to the local optimal value P_{ibest} and global optimal value P_{gbest}. If the value is better, update P_{ibest} and P_{gbest}, otherwise the values of P_{ibest} and P_{gbest} are not changed.

Step 3. Update the values of all particles by $\mathrm{X}_i^{k+1} = \mathrm{X}_i^k + \mathrm{V}_i^k$. If the updated particle is not satisfied with the folding constants of amino acid sequence on square lattice, it needs to be adjusted . Otherwise compute $\mathrm{U}_i^{k+1} = \mathrm{X}_i^{k+1} - \mathrm{X}_i^k$.

Step 4. Update the velocity of particle

$$\mathrm{V}_i^{k+1} = \mathrm{U}_i^{k+1} + \mathrm{r}_1(\mathrm{P}_{gbest} - \mathrm{X}_i^{k+1}) + \mathrm{r}_2(\mathrm{P}_{ibest} - \mathrm{X}_i^{k+1}) \qquad (3)$$

in which r_1 and $\mathrm{r}_2 \in (0,1)$ are random numbers, $(\mathrm{P}_{gbest} - \mathrm{X}_i^{k+1})$ and $(\mathrm{P}_{ibest} - \mathrm{X}_i^{k+1})$ are the basic adjustment sequences.

IV. RESULTS

To assess its performance, we applied the improved PSO to four instances for 2D HP model on square lattice. These instances come from standard benchmark dataset. In the following figures, polar will be depicted as "◆", hydrophobic as "●", and the black lines show the covalent bond between the adjacent amino acids.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

A. Sequence 1 HPPHHPPHHPPHHPPHH

This sequence contains seventeen amino acids and is embedded into a 5×5 lattice of R^2. The lattice points in the first row are tabbed as 1-5 from left side. The points of other rows are done sequentially. The 13th point is treated as the origin of the lattice. We can find the optimal conformation with 8 HH pairs by our method. One of the optimal conformations is demonstrated in Fig. 1. The conformation in Fig.1 is non-compact. The result of this sequence indicates the advantage of non-compact conformation, i.e., the optimal conformations of amino acid sequence, whose length is not the product of two integers, can be found by our modified PSO algorithm. This instance was also computed by self-organizing map method with 8 HH pairs in [6]. It indicates that our method is suitable for this small benchmark sequence.

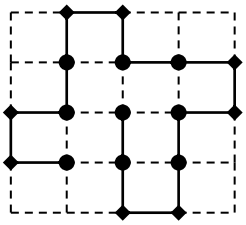

Fig. 1 One of optimal conformations for Sequence 1 in *5×5* lattice

B. Sequence 2. HPHPPHHPHPPHPHHPPHPHs

This benchmark example has twenty amino acids and is embedded into a 5×5 lattice of R^2. The lattice points in the first row are signed as label 1-5 from left side. The points of other rows are ordered sequentially. The 13-th point is treated as the origin. We can find the optimal conformations with 9 HH pairs. One of the optimal conformations is demonstrated in Fig. 2. When this instance is embedded into a 4×5 lattice of R^2, we can find the optimal conformations with 6 HH pairs. One of the optimal conformations is demonstrated in Fig. 3.

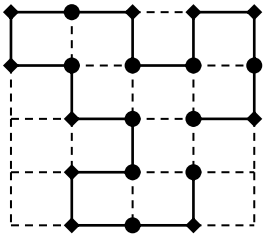

Fig.2 One optimal conformation for Sequence 2 in 5×5 lattice

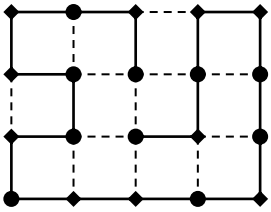

Fig.3 One optimal conformation for Sequence 2 in 4×5 lattice

The conformation in Fig. 2 is non-compact, while the conformation in Fig. 3 is compact. It indicates that our optimal model and improved PSO algorithm can be used to find conformations not only in compact lattice but also in non-compact lattice. For sequence 2. we know that the minimal free energy of non-compact lattice is lower than that of compact lattice. This kind of phenomenon is consistent with the biological insights. Only compact conformation was found by self-originizing map method with 6 HH pairs. It indicates that our method and model can not only find compact conformations but also fold sequence in non-compact lattice**.**

C. Sequence 3. HHHHHPHHHHHHPHHHHPHH

This benchmark sequence has twenty amino acids and is embedded into a 5×5 lattice of R^2. The lattice points in the first row are labeled as 1-5 from left side. The points of other rows are labeled sequentially. The 13-th point is treated as the origin. The minimal free energy of this sequence is -12. One of the optimal conformation is demonstrated in Fig.4. When it is embedded into a 4×5 lattice of R^2, one of the optimal conformations is demonstrated in Fig.5 with 12 HH pairs.

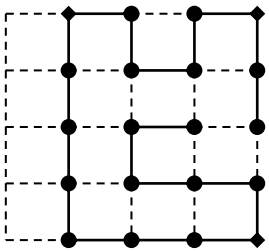

Fig.4 One optimal conformation for Sequence 3 in 5×5 lattice

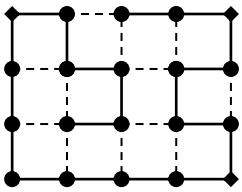

Fig.5 One optimal conformation for Sequence 3 in 4×5 lattice

The computational results show that our method can be used to find the optimal conformation of sequences with length as product of two integers. The minimal free energies of compact conformation and that of non-compact conformation are the same, but their shapes are different. This demonstrates that compact conformation can be treated as a special case of non-compact conformation. It means that our model and method can be applied to fold this kind of amino acid sequences.

D. Sequence 4. PPHPPHHPPPPHHPPPPHHPPPPHH

This benchmark sequence has twenty five amino acids and is embedded into a 5×6 lattice of R^2. The lattice points in the first row are signed as 1-6 from left side. The points of other rows are labeled sequentially. The 15-th point can be treated as the origin. The minimal free energy of this sequence is -8. Two optimal conformations are demonstrated in Fig. 6 and Fig. 7. When it is embedded into a 5×5 lattice of R^2, one of the

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

optimal conformations is demonstrated in Fig. 8 with 8 HH pairs.

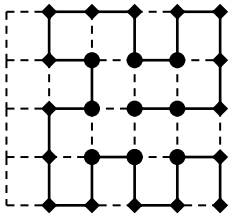

Fig.6 Sequence 4 in 5×6 lattice

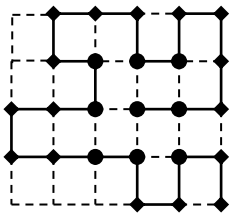

Fig.7 Sequence 4 in 5×6 lattice

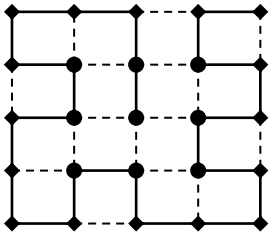

Fig.8 Sequence 4 in 5×5 lattice

The above results indicate that our method can find different structures of sequence with the same minimal free energy. According to Fig. 6 and Fig. 7, we know that the folding conformations can be solved in a bigger lattice and non-compact structure is more flexible and more realistic. Compared with non-compact conformation, predicted compact structure in Fig. 8 requires less computational time but is not very close to the biological nature.

V. CONCLUSION AND DISCUSSION

Because of the difficulty in solving the protein folding problem, there are two ways out: a simplified model and a fast configuration search method. 2D HP lattice model was broadly studied by chemists. So we study the protein structure prediction in 2D equicistant lattice in this paper. We propose a mathematical model, which has the following advantages. Firstly, it is highly simplified; Secondly it leads itself to exact results; Thirdly, it is easy to perform simulation. In this paper, the amino acid sequence is embedded in equidistant lattice to find low-energy non-compact conformations. To our best knowledge, particle swarm optimization algorithm has been directly used to solve TSP. Because of the connection between TSP and protein structure prediction problem, we have extend PSO algorithm for 2D HP model. The results indicate that the improved PSO can find new conformations with minimal free energy. It is faster than other stochastic sampling methods. It can be applied to find compact and non-compact conformations thus can fold the sequence with arbitrary length. An empirical study demonstrates the effectiveness of improved PSO and optimal model for solving 2D HP model.

There are many future directions to pursuit. Since the free energy in the model is given only by the number of nonspecific hydrophobic contact, the positions of polar segments are not directly optimized when searching for optimal structures. This may result in unnatural structures if these segments are too long to be located at the ends of the sequences. A modification is required to try to obtain more nature-like structures for the HP model's scoring system. In addition, we intend to develop and study modified PSO algorithm for other types of protein folding problems such as 3D HP lattice model, 2D triangular lattice model, etc. Overall, we strongly believe that the modified PSO method offer considerable potential for protein structure prediction problem.

ACKNOWLEDGMENT

This work is supported by Basic Scientific Research Foundation of Nanjing University of Aeronautics and Astronautics (Grant No. NN2012012).

REFERENCES

[1] F. Wang, J. Song and Y. Song. Application of BP neural network in protein secondary structure prediction. Computer Technology and Development. Vol. 19, pp. 217-218, 2009.

[2] K.A. Dill, S. Bronmberg, K. Yue. Principles of protein folding a perspective from simple exact models. Protein Science. Vol. 4, pp. 561-602, 1995.

[3] F. Liang and W. H. Wong. Evolutionary Monte Carlo for protein folding simulations. Journal of Chemical Physics. Vol. 7, pp. 3374-3380, 2001.

[4] P. Crescenzi, D. Goldman, C.H. Papadimitriou, A. Piccolboni and M. Yannakakis. On the complexity of protein folding. Journal of Computational Biology. Vol. 3, pp. 423, 1998.

[5] T. Jiang, Q. Cui, G. Shi and S. Ma. Protein folding simulation of the hydrophobic-hydrophilic model by combining tabu search with genetic algorithms. Journal of Chemical Physics. Vol. 8, pp. 4592-4596, 2003.

[6] Y. Wang, Z. Zhan. L. Wu and X. Zhang. An improved self-organizing map algorithm for protein folding and its realization. Journal of System Science and Mathematical Science. Vol. 1, pp. 1-12, 2004.

[7] Y. Guo, E. Feng and Y. Wang. Exploration of two-dimensional HP lattice model by combining local search with elastic net algorithm. Journal of Chemcal Physics. Vol. 15, pp. 4102, 2006.

[8] R. Unger and J. Moult. Genetic algorithm for protein folding simulations. Journal of Molecular Biology. Vol.231, pp.75-81,1993.

[9] S. C. Kou, J. Oh and W. H. Wong. A study of density of states and ground states in hydrophobic-hydrophilic protein folding models by equi-energy sampling. Journal of Chemical Physics. Vol. 124, pp:244903, 2006.

[10] J. Kennedy and R.C.Eberhart. Particle Swarm Optimization. Proceeding of IEEE International Conference on Neural Networks. Piscateway, IEEE. pp. 1942-1948, 1995.

[11] J. Tang. Principle and application of PSO algorithm. Computer Technology and Development. Vol. 20, pp. 215-216, 2010.

[12] C. Wang and J. Zhang. Modified particle swarm optimization algorithm for traveling salesman problem. Journal of North China Electric Power University. Vol. 32, pp. 47-51, 2005.

[13] W. Yan and Y. Guo. Modified particle swarm optimization algorithm for protein structure prediction problem. Computer Technology and Development. Vol. 21, pp. 109-112, 2011.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE