Analysis of Toy Model for Protein Folding Based on Particle Swarm Optimization Algorithm*

Juan Liu[1], Longhui Wang[1], Lianlian He[2], and Feng Shi[2]

[1] School of Computer, Wuhan University, Wuhan, 430079, China
liujuan@whu.edu.cn
[2] School of Mathematics, Wuhan University, Wuhan, 430072,China

Abstract. One of the main problems of computational approaches to protein structure prediction is the computational complexity. Many researches use simplified models to represent protein structure. Toy model is one of the simplification models. Finding the ground state is critical to the toy model of protein. This paper applies Particle Swarm Optimization (PSO) Algorithm to search the ground state of toy model for protein folding, and performs experiments both on artificial data and real protein data to evaluate the PSO-based method. The results show that on one hand, the PSO method is feasible and effective to search for ground state of toy model; on the other hand, toy model just can simulate real protein to some extent, and need further improvements.

1 Introduction

The structure of protein determines its function in molecular. Experimental methods of determining protein structure include X-ray crystallography and NMR-spectroscopy. However some proteins are hard to crystallize, and NMR-spectroscopy method only works on small proteins. Moreover, these two methods are expensive and time-consuming [1]. So predicting protein structure by computational method is very necessary, and it has become one of the most important research topics in modern molecular biology. However, it is very complex to determine the native three-dimensional structure of a protein when only given the sequence of amino acid residues that compose the protein chain [2].

Due to the complexity of the protein-folding problem, scientists have proposed a variety of models such as hydrophobic-polar (HP) model to simplify the problem by abstracting only the "essential physical properties" of real proteins. Generally speaking, there are three representative simplified HP models for protein folding: lattice model [3], triangle lattice model [4], and toy model[5]. In lattice model, the three dimensional space is represented by a lattice, and the 20 amino acids are grouping into two classes: hydrophobic (or non-polar) residues and hydrophilic (or polar) residues, where P represents polar residues, and H represents hydrophobic residues. Residues that are adjacent in the primary sequence must be placed at adjacent points in the lattice. A conformation of a

* This work was supported by the National Natural Science Foundation of China under grant no. 60301009.

L. Wang, K. Chen, and Y.S. Ong (Eds.): ICNC 2005, LNCS 3612, pp. 636–645, 2005.

protein is a self-avoiding walk along the lattice. The protein folding problem is to find a conformation of the protein sequence on the lattice such that the overall energy is minimized, for some reasonable definition of energy [6]. Dill *et.al.* surveyed some works on this model in [7].

Richa Agarwada *et.al.* tested the HP model on a triangle lattice [4]. They examined the choice of a lattice by considering its algorithmic and geometric implications and argued that triangular lattice is a more reasonable choice. Though the structures derived from triangle lattice model are probably still far from biological reality, it's much better than basic lattice model [4].

Stillinger *et al.* had done further improvements and presented the toy model [5]. In this model, amino acid residues are still classified into two kinds: hydrophobic and hydrophilic, but what differences from previous lattice models are that there is only one bond between two consecutive residues, and the angle between the two bonds can change freely. So it is more like the real protein structure than the previous two lattice models [5]. One major advantage of the toy model is that it becomes feasible to determine a complete database of ground state structures for all "polypeptides" up to some modest (but non-trivial) degree of polymerization.

How to find the ground state of toy model of protein? People have tried many methods such as Neural Network [5,8] and Simulated Annealing Algorithm [9]. In this paper, we will try to use PSO algorithm to search the ground state and analyze the experiment results.

PSO is a recently proposed algorithm by J.Kennedy and R. C. Eberhart in 1995 [10], motivated by social behavior of organisms such as bird flocking and fish schooling [11]. In a PSO system, **particles**(individuals) **fly** around in a multidimensional search space. During flight, each particle adjusts its **position**(state) according to its own experience, the experience of a neighboring particle, making use of the best position encountered by itself and its neighbor. Thus, as in modern GAs and memetic algorithms, a PSO system combines local search methods with global search methods, attempting to balance exploration and exploitation [10,12]. In the past several years, PSO has been successfully applied in many research and application areas.

In this paper, we will discuss the application of PSO on toy model for protein folding. The rest part of the paper is organized as following: In section 2, we give a brief description of toy model for protein folding. In section 3, we introduce the basic ideas of PSO. Section 4 includes the experiments and the results. The final section is the conclusion part of this paper.

2 Description of Toy Model

In 1993, Stillinger *et al.* presented the toy model for protein sequence [5]. This model incorporates only two "amino acids", to be denoted by **A** and **B**, in place of the real 20 amino acids. **A** and **B** are linked together by rigid unit-length bonds to form linear un-oriented polymers that reside in two dimensions. As figure 1 illustrates, the configuration of any n-mer is specified by the $n-2$

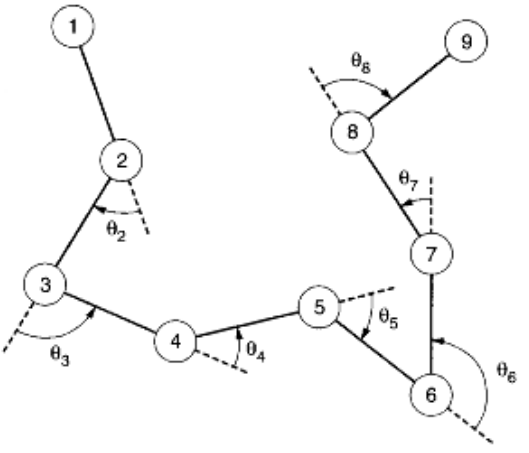


Fig. 1. A schematic diagram of a generic 9-mer, with serially numbered residues, and backbone bend angles

angles of bend $\theta_2 \ldots \theta_{n-1}$ at each of the non-terminal residues. We adhere to the conventions that: $-\pi < \theta_i < \pi$, $\theta_i = 0$ corresponds to linearity of successive bonds, and positive angles indicate counterclockwise rotations.

In the following,we do not consider intermolecular interactions. We suppose that two kinds of interactions compose the intra-molecular potential energy for each molecule: backbone bend potentials (V_1) and non-bonded interactions (V_2). The former is independent of the **A, B** sequence, whereas the later varies with the sequence and receives contribution from each pair of residues that are not directly attached by a backbone bond. Residues along the backbone can be conveniently encoded by a set of binary variables $\xi_1 \ldots \xi_n$, where $\xi_i = 1$ means that the ith residue is **A**; and ξ_i = -1 means that it is **B**. Thus for any n-mer, the intra-molecular potential-energy function Φ can be expressed as formula (1):

$$\Phi = \sum_{i=2}^{n-1} V_1(\theta_i) + \sum_{i=1}^{n-2} \sum_{j=i+2}^{2} V_2(r_{ij}, \xi_i, \xi_j) \tag{1}$$

Where, the distances r_{ij} can be written as functions of the intervening angles (backbone bonds have unit length):

$$r_{ij} = \left\{ \left[1 + \sum_{k=i+1}^{j-1} \cos \left[\sum_{l=i+1}^{k} \theta_l \right] \right]^2 + \left[\sum_{k=i+1}^{j-1} \sin \left[\sum_{l=i+1}^{k} \theta_l \right] \right]^2 \right\}^{1/2} \tag{2}$$

Toy model assigns a simple trigonometric form to V_1:

$$V_1(\theta_i) = \frac{1}{4}(1 - \cos \theta_i) \tag{3}$$

And the non-bonded interactions V_2 have a species dependent Lennard-Jones 12, 6 form:

$$V_2(r_{ij}, \xi_i, \xi_j) = 4[r_{ij}^{-12} - C(\xi_i, \xi_j) r_{ij}^{-6}] \tag{4}$$

Where,

$$C(\xi_i, \xi_j) = \frac{1}{8}(1 + \xi_i + \xi_j + 5\xi_i\xi_j) \tag{5}$$

On account of Equation (4), successive bonds would tend towards linearity ($\theta_i = 0$), if nothing else mattered.

Toy model is also based on the famous judgement presented by Anfinsen in 1960s: The native structure of protein is the structure with the lowest free energy [13]. This conclusion is the thermodynamics base of using energy minimization method to predict protein structure. For a protein sequence with n residues, we need to search out a group of suitable θ_i ($i = 2, \ldots, n-1$), $\theta_i \in (-\pi, \pi)$, to make the energy function (1) achieve the minimal value in the solution space.

3 Particle Swarm Optimization

PSO algorithm is similar to other genetic algorithms (GA). What makes it different with GAs is that, PSO does not use evolutionary operators to evolve the population, instead, it takes each individual as a particle without weight and volume in the n-dimensional search space, the particle flies at certain speed in the search space. The flying speed of the particle adjusts dynamically according to its flying experience and population's flying experience [14].

3.1 Basic Particle Swarm Optimization Method

Considering the minimal problem, given a particle i, let $X_i = (x_{i1}, x_{i2}, \cdots, x_{in})$ be its current position, $V_i = (v_{i1}, v_{i2}, \cdots, v_{in})$ be its current flying speed, $P_i = (p_{i1}, p_{i2}, \cdots, p_{in})$ be the best position it has experienced. Suppose $f(X)$ is the objective function, obviously, P_i would minimize $f(X)$. P_i is called as the best individual place. Suppose that the particle number in the swarm is s, the best position P_g that all particles in the swarm have experienced is called the global optimal position, so we have $P_g \in \{P_1, P_2, \cdots, P_s\}$, and $f(P_g) = \min_{i \in \{1,2,\cdots,s\}} \{f(P_i)\}$.

With the definition presented as above, basic PSO function can be described as following:

$$\begin{aligned} v_{ij}(t+1) = v_{ij}(t) + c_1 r_{1j}(t)(p_{ij}(t) - x_{ij}(t)) \\ + c_2 r_{2j}(t)(p_{gj}(t) - x_{ij}(t)) \end{aligned} \tag{6}$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1) \tag{7}$$

Where, j indicates the jth dimension of particle, i indicates the ith particle, t indicates the tth generation, c_1, c_2, varying from 0 to 2, are the acceleration speed constants, they determine the relative influence of the social and cognitive components, and are usually both set the same to give each component equal weight as the cognitive and social learning rate. $r_1 \sim U(0,1), r_2 \sim U(0,1)$ are two independent random function [15].

3.2 Canonical Particle Swarm Optimization Method

Due that basic PSO usually failed in some applications, Carlisle and Doziert presented the following typical PSO Algorithm model (Canonical PSO) [16].

$$v_{ij} = \begin{cases} K(v_{ij} + c_1 r_1 (P_{ij} - x_{ij}) + c_2 r_2 (P_{gj} - x_{ij})), & X_{\min} < x_{ij} < X_{\max} \\ 0 & otherwise \end{cases}$$
$$x_{ij} = \begin{cases} x_{ij} + v_{ij}, & X_{\min} < x_{ij} < X_{\max} \\ X_{\max}, & (x_{id} + v_{ij}) > X_{\max} \\ X_{\min}, & X_{\min} < (x_{ij} + v_{ij}) \end{cases} \tag{8}$$

Where K is the constriction factor,

$$K = \frac{2}{\left|2 - C - \sqrt{c_2 - 4C}\right|} \tag{9}$$

In the following experiments, we use this kind of PSO to analyze toy model. We use the classic parameter set [16], in which, $c_1 = 2.8$, $c_2 = 1.3$, $C = c_1 + c_2$, population size $N = 30$. In each generation, we produce new candidate solutions, and calculate the energy function, if the result of the function becomes smaller, we reserve the solution, otherwise we reject the solution. The iteration procedure repeats until the terminal conditions are satisfied. In this article, the procedure will stop when it reaches the maximal iteration steps.

4 Experiments and Results

In this section, we do several experiments to analyze the toy model for protein folding. Canonical PSO described in secion 3.2 is used to search the ground state of the toy model that minimizes Equation(1).

4.1 Experiments on Artificial Sequences

We use some artificial sequence to do two kinds of experiments. First, we use the same sequences as [5] to see whether our method can get the ground state. For these short sequences, the maximal iteration step $L = 30$. From the results illustrated in table 1, we can see that our method can also reach the ground state presented by Stillinger [5].

To explore whether our method can get the correct protein secondary structure elements, we then use two testing sequences "AABABB" and "AAABAA" just like [5] for experiments. The secondary structures on the 2D toy model is shown in figure 2. Figure 3 shows the computational results. Because these two sequences are short, our program got the results in a very short time.

From figure 3 we can see that our method is effective to simulate protein folding as it can correctly give out the secondary structure motif: α-helix and β-sheet.

Table 1. Ground state properties of toy-model polypeptides

Molecular	Φ	Molecular	Φ
AAA	-0.658 21	AAAAA	-2.848 28
AAB	0.032 23	AAAAB	-1.589 44
ABA	-0.658 21	AAABA	-2.444 93
ABB	0.032 23	AAABB	-0.546 88
BAB	-.0.030 27	AABAA	-2.531 70
BBB	-0.030 27	AABAB	-1.347 74
		AABBA	-0.926 62
AAAA	-1.676 33	AABBB	0.040 17
AAAB	-0.585 27	ABAAB	-1.376 47
AABA	-1.450 98	ABABA	-2.220 20
AABB	0.067 20	ABABB	-0.616 80
ABAB	-0.649 38	ABBAB	-0.005 65
ABBA	-0.036 17	ABBBA	-0.398 04
ABBB	0.004 70	ABBBB	-0.065 96
BAAB	0.061 72	BAAAB	-0.521 08
BABB	-0.000 78	BAABB	0.096 21
BBBB	-0.139 74	BABAB	-0.648 03
		BABBB	-0.182 66
		BBABB	-0.240 20
		BBBBB	-0.452 66

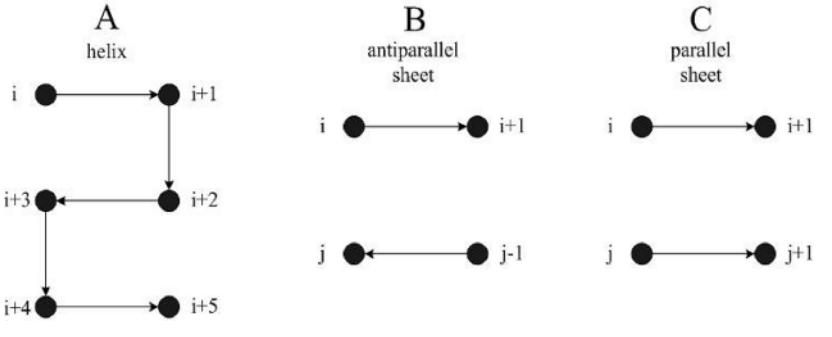
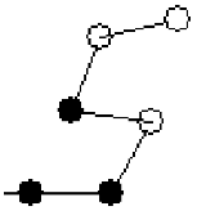


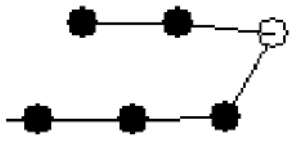
Fig. 2. Secondary structures on the 2D toy model. A: helix, at least two sequential non-covalent contacts between residues [(i, i+3), (i+2,i+5)... (i+2n, i+2n+3)]. B: Antiparallel sheet [(i, j), (i+1, j-1)... (i+n, j+n)]. C: Parallel sheet [(i, j), (i+1, j+1)... (i+n, j+n)]

4.2 Experiments on Real Protein Sequences

Then we test our method on real protein sequences.When sequence becomes long, the determination of the objective function value is extremely time-consuming. So only two real proteins with short sequences are discussed in our paper, i.e., 1agt and 1aho. All information of these two proteins can be downloaded from PDB (http://www.rcsb.org/pdb/).



(a) The lowest-energy conformer of sequence AABABB may be classified as "helical". Φ=-1.335366, x=[-1.045231 -1.951874 1.738942 0.147911]



(b) The lowest-energy conformer for sequence AAABAA is a "β-sheet" motif. (Φ=-3.697501, x= [0.020746 1.040153 1.958890 0.133675]

Fig. 3. Testing sequence results. In fig.3 and the following pictures, the circle indicates hydrophilic residue, and the black dot indicates hydrophobic residue

In the experiments, we use K-D method to distinguish hydrophobic and hydrophilic residues of 20 amino acids in real proteins. Briefly speaking, amino acids I, V, L, P, C, M, A, G are hydrophobic and D, E, F, H, K, N, Q, R, S, T, W, Y are polar [17].

Experiment on 1AGT. First, we experimented on 1agt. The information about its sequence and secondary structure from PDB are as follows:

```
1 GVPINVSCTG SPQCIKPCKD QGMRFGKCMN RKCHCTPK
   EE   B SS STTHHHHHHH HTBSEEEEET TEEEEEE
```

The first line is amino acid sequence, and the second line is its secondary structure. It contains 38 residues, one helical segment and two β-sheet segments.

With the maximal iterate steps L = 5,000, we got ground state shown in figure 4, from which we can see that the final toy model can simulate the real protein to some extent.

Experiment on 1AHO. And then, we discussed on protein 1aho; its protein sequence and secondary structure information are as follows:

```
 1 VKDGYIVDDV NCTYFCGRNA YCNEECTKLK GESGYCQWAS PYGNACYCYK
    B EEEE TT S B   S HH HHHHHHHHTT  SEEEEETTB TTBSEEEEES

51 LPDHVRTKGP GRCH
   B TTS B  S S
```

It contains 64 residues. Residue 19 to 28 is a helix segment in native conformation. With L = 10,000, we got the result shown in figure 5, which also approaches to the real protein structure.

To evaluate the performance of our method, we also compared it with Simulated Annealing (SA) Algorithm implemented in [9] on 1agt and 1aho sequences.

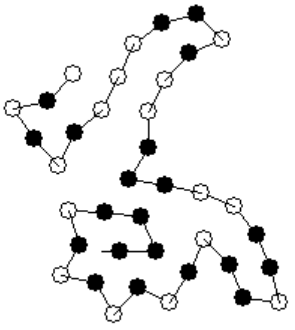


Fig. 4. Φ = -19.616866, x = [1.968670 1.039088 0.068094 1.922932 -0.834257 1.907747 -0.833636 1.912368 -1.340518 1.479550 0.137488 -1.933330 -0.375798 1.044901 1.953578 0.125628 0.280929 0.528956 0.144413 0.067585 -1.937305 0.497480 -0.420421 -0.306854 -0.404344 1.946600 1.041268 0.396669 0.504622 -0.058998 -0.411684 0.426404 -1.939082 -0.130507 -1.945389 0.570014]

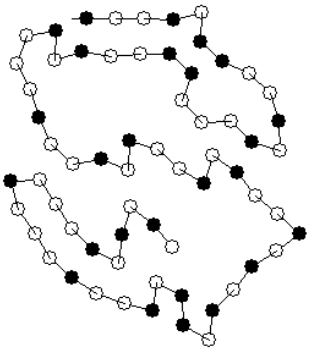


Fig. 5. Φ = -15.181101, x = [-0.010702 -0.060948 0.362086 -1.926352 0.904857 0.301411 -0.299284 -0.573455 -0.201756 -1.900080 -0.531997 0.810784 -0.829126 -1.096663 1.186948 0.746497 0.050294 -0.262349 0.501073 -1.922822 1.787451 1.047013 0.815521 -0.145761 0.093422 0.404816 0.928052 -0.562520 1.924269 -1.820003 -0.455601 0.188326 1.842072 -1.918896 -0.259529 0.200091 -0.056049 -1.756343 -0.071092 0.340538 -0.165433 0.691833 -1.951029 -1.040509 1.052306 1.944196 -1.725629 -0.051463 -0.258637 -0.097700 -0.364711 0.076348 -0.312131 -1.820869 -0.995589 -0.052073 0.215089 0.307311 1.937550 -0.175043 -1.938866 -0.222515]

Both methods were used to search the minimal energy state of toy model for protein folding, and the comparison results are listed in table 2 and table 3.

From table 2 and table 3, we can see that PSO is much faster than SA and it can search better results. This may due that PSO has less parameters than SA, furthermore, since SA often lead to huge computational task, thus it usually can not get the global minimal in reasonable time.

Table 2. Comparison PSO with SA: ground state

	PSO	SA
1AGT	-19.6168 66	-17.3628 15
1AHO	-15.1911 01	-14.9612 73

Table 3. Comparison PSO with SA: searching time

	PSO	SA
1AGT	8,376 s	12, 065 s
1AHO	10,149 s	15, 832 s

From the results shown in figure 4 and figure 5, we can also see that, although the toy model can simulate the real protein to some extent, the results are still some different from the real proteins. That is to say, toy model needs further improvements.

5 Conclusions

Toy model is a great improvement of simplification models of protein folding. Because comparing with lattice model, the angle of its bond can turn freely and thus it is more like real protein structure. In this paper, we applied PSO on toy model for protein folding and got good results. Our experiment results show that PSO has strong ability to search extremum in consecutive space. At present, Our method only considered two kinds of residues and only two kinds of interaction energy. Maybe we can improve the model by considering more interaction energy and more properties of amino acid residues, not just only the polar and non-polar characters. However, we should note that not all properties are mattered with the structure of protein, for unnecessary conditions will make the question too complicated. We will address this direction in the future.

References

1. Park, B.H. and Levitt, M.: The Complexity and Accuracy of Discrete State Models of Protein Structure, J.Mol.Biol. **249** (1995) 493-507
2. Kolinski, A. and Skolnick, J.: (2004) Reduced Models of Protein and Their applications, Polymer **45** (2004) 511-524
3. Dill, K. A.: (1985), Theory for the folding and stability of globular proteins, Biochemistry **24** (1985) 1501-1512
4. Afarwala, R., Batzoglou, S., et al.: Local Rules for Protein Folding on a Triangular Lattice and Generalized Hydrophobicity in the HP Model, Proceedings of the first annual international conference on Computational molecular biology (1997) 1-2
5. Stillinger, F. H., Gordon, T. H., and Hirshfeld, C. L.: Toy Model for Protein Folding, Physical Review E **48** (1993) 1469-1477
6. Morrissey, M.P., Ahmed, Z. and Shakhnovich, E. I.: The Role of Cotranslation in Protein Folding: a Lattice Model Study, Polymer **45** (2004) 557-571
7. Dill, K. A., Bromberg, S., and Yue, K., et al.:(1995) A perspective from simple exact models, Prot. Sci. **4** (1995) 561-602
8. Caspi, S. and Jacob, E.B.: Conformation Changes and Folding of Protein Mediated by Davydov's Soliton, Physics Letters A **272**, (2000) 124-129

9. Wang, L, Zhou, H., et al.: Perspective Roles of Short- and Long-Range Interactions in Protein Folding, Wuhan University Journal of Natural Sciences **9** (2004) 182-187
10. Kennedy, J. and Eberhart, R.C.: Particle Swarm Optimization, In Proceedings of the IEEE Int. Conf. Neural Networks (1995) 1942-1948
11. Shi, Y. and Eberhart, R.C.: A Modified Particle Swarm Optimizer. In: Proceedings of the IEEE International Conference on Evolutionary Computation. (1998) 69-73
12. Eberhart, R.C., Kennedy, J.: Swarm Intelligence, Morgan Kaufmanns (2001)
13. Anfisen, C.B.: Developmental Biology Supplement 2, Academic Press Inc., USA (1968)
14. Parsopoulos, K.E., Vrahatis, M.N.: Particle Swarm Optimization method in Multiobjective Problems, In: Proceedings ACM Symposium on Applied Computing (2002) 603-607
15. Shi, Y., Eberhart, R.C.: Empirical Study of Particle Swarm Optimization, In: Proceedings of the 1999 Congress on Evolutionary Computation, (1999) 1945-1950
16. Carlisle, A., Dozier, G.: An Off-the-Shelf PSO, In :Proceedings of the Workshop on Particle Swarm Optimization (2001) 1-6
17. Thorton, J., Taylor, W. R.: Structure Prediction, In:Findlay J.B.C, Geisow M.J. (eds.): Protein Aequencing, Oxford: IRL Press (1989) 147-190