Journal of Zhejiang University SCIENCE B
ISSN 1673-1581
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

Heuristic algorithm for off-lattice protein folding problem*

CHEN Mao (陈 矛)†, HUANG Wen-qi (黄文奇)

(*School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China*)

†E-mail: mchen_1@163.com

Received Aug. 10, 2005; revision accepted Nov. 28, 2005

Abstract: Enlightened by the law of interactions among objects in the physical world, we propose a heuristic algorithm for solving the three-dimensional (3D) off-lattice protein folding problem. Based on a physical model, the problem is converted from a nonlinear constraint-satisfied problem to an unconstrained optimization problem which can be solved by the well-known gradient method. To improve the efficiency of our algorithm, a strategy was introduced to generate initial configuration. Computational results showed that this algorithm could find states with lower energy than previously proposed ground states obtained by nPERM algorithm for all chains with length ranging from 13 to 55.

Key words: Protein folding, AB off-lattice model, Gradient method
doi:10.1631/jzus.2006.B0007 **Document code:** A **CLC number:** TP3; Q67

INTRODUCTION

Protein folding problem, or protein structure prediction problem, is one of the central problems in the field of bioinformatics. Studies indicated that proteins' biological functions are determined by their dimensional folding structures (Anfinsen, 1973). Since the structure of a protein is strongly correlated with the sequence of amino acid residues, predicting the native states of a protein from its given sequence by using theoretical computing method is a feasible approach and of great significance for protein engineering (Lau and Dill, 1989).

Since the problem is too difficult to be approached with fully realistic potentials, the theoretical science community has introduced and examined several highly simplified models, one of which is the HP lattice model of Dill (1985) where each amino acid is treated as a point particle on a regular (quadratic or cubic) lattice, and only two types of amino acids−hydrophobic (H) and hydrophilic (P)−are considered. The energy between any two neighboring non-bonded hydrophobic monomers (H-H) is defined as −1, otherwise 0.

Being the most simplified and most popular model, HP model only considers the interactions between neighboring non-bonded H monomers, neglecting the other nonlocal effects caused by P-P, H-P and non-neighbored H-H pairs, which also exert significant statistical influence on the conformation of the monomers in the properly folded state.

To illustrate the influence of nonlocal effects on protein folding, Stillinger (1995) proposed a more realistic simplified model, namely, AB off-lattice model, which also uses only two types of monomers, now called "A" (hydrophobic) and "B" (hydrophilic). The distances between consecutive monomers along the chain are held to be 1, while nonconsecutive monomers interact through a modified Lennard-Jones potential. In addition, there is an energy contribution called bending energy from each bond angle θ_i between successive bonds. Hence, the total energy function U_1 for an n monomers chain is expressed as

$$U_1=\sum_{i=2}^{n-1}V_1(\theta_i)+\sum_{i=1}^{n-2}\sum_{j=i+2}^{n}V_2(r_{ij},\zeta_i,\zeta_j), \qquad (1)$$

* Project supported by the National Basic Research Program (973) of China (No. 2004CB318000) and the National Natural Science Foundation of China (No. 10471051)

where

$$V_1(\theta_i)=(1-\cos\theta_i)/4, \quad (2)$$

$$V_2(r_{ij},\zeta_i,\zeta_j)=4[r_{ij}^{-12}-C(\zeta_i,\zeta_j)r_{ij}^{-6}]. \quad (3)$$

Here r_{ij} is the distance between monomer i and j (with $i<j$). Each ζ_i is either A or B, and $C(\zeta_i, \zeta_j)$ is +1, +1/2 and –1/2 respectively, for AA, BB, and AB pairs, thus producing strong attraction between AA pairs, weak attraction between BB pairs, and weak repulsion between AB pairs, roughly analogous to the situation in real proteins.

Even in this highly simplified model, it is not easy to predict the native state for the protein folding problem. This problem has been recognized to be NP-complete, which means that it is not solvable in polynomial time, even for an optimal algorithm (Crescenzi *et al*., 1998). Consequently, various heuristic schemes have been proposed for approaching this problem.

For its two-dimensional (2D) version, neural networks (Stillinger, 1995), Monte Carlo (Irback *et al*., 1997) and biologically motivated methods (Torcini *et al*., 2001) were used to find the native state. An improved pruned enriched Rosenbluth method with importance sampling, namely, nPERM was proposed by Hsu *et al*.(2003), which found states with lower energy than previously proposed putative ground states for all four Fibonacci sequences with chain lengths≥13. Without modifying the energy function, Hsu *et al*.(2003) extended the 2D AB model to 3D version and presented some putative lowest energy states for the four sequences. Although the resulting configuration corresponding to the lowest energy has a single hydrophobic core for the short sequence with length 13, the longer sequences with length ranging from 21 to 55 do not fold into configurations with single hydrophobic cores. Recently, better results in three dimensions for the four sequences were achieved by means of energy landscape paving (ELP) minimizer (Bachmann *et al*., 2005) and conformational space annealing (CSA) method (Kim *et al*., 2005).

In this paper, we propose a quite different class of heuristic algorithm for predicting the native structure for the 3D AB off-lattice model. The proposed algorithm integrates the well-known gradient method and a novel strategy of generating promising initial configuration in order to find the globally optimal state. Compared with nPERM, the experimental results showed that our algorithm can find lower energy states and that each of the four resulting configurations has single hydrophobic cores.

PROPOSED ALGORITHM

Mathematical formulation

Consider the problem in 3D Euclidean space. Consider an amino acid sequence as a chain of black balls (A) and white balls (B) with radius R=0.5, with the balls being numbered from 1 to n. Denote the coordinates of the center of the ith (i=1, 2, …, n) ball by (x_i, y_i, z_i). At any moment, the entirety of the coordinates of the center of the n balls, x_1, y_1, z_1, …, x_n, y_n, z_n, is called a configuration.

Now, the protein folding problem can be described as the following mathematical model:

$$\min(U_1), \quad (4)$$

subject to

$$\sqrt{(x_i-x_{i+1})^2+(y_i-y_{i+1})^2+(z_i-z_{i+1})^2}=1,$$
$$i=1, 2, \ldots, n-1 \quad (5)$$

$$-\infty<x_i, y_i, z_i<+\infty, \ i=1, 2, \ldots, n-1. \quad (6)$$

In this model, there are 3n continuous deterministic variables and n−1 constraints where n is the number of balls. Constraint Eq.(5) ensures that the distances between the centers of two consecutive balls along the chain are equal to 1. A configuration that satisfies constraint Eq.(5) is termed a legal configuration.

Eqs.(4)~(6) form a specific type of nonlinear constraint-satisfied problem. This is just the mathematical model for our protein folding problem. It is rather difficult to solve this kind of problem directly due to the loss of smoothness in the solution space. Therefore, a scheme is proposed below to convert this problem into an unconstrained optimization problem which is smooth in the solution space.

New mathematical description

Instead of fixing the distances between two successive balls, we imagine the centers of two consecutive balls i and i+1 (i=1, 2, …, n−1) are connected by a fictitious spring with natural length held to be 1. Springs have the tendency to return to their

natural length after being compressed or stretched. So springs can be used to relax the requirement on the solvability of the original constraint-satisfied problem.

Under any configuration, the length of a spring connecting the centers of two consecutive balls along the chain is

$$l_{i,i+1}=\sqrt{(x_i-x_{i+1})^2+(y_i-y_{i+1})^2+(z_i-z_{i+1})^2},\quad i=1, 2, \ldots, n-1. \tag{7}$$

If $l_{i,i+1}>1$, it means that the spring is extended; if $l_{i,i+1}<1$, the spring is compressed. According to Hook's law, the elastic potential energy of a spring is

$$u_{i,i+1}=K_s(l_{i,i+1}-1)^2/2,\ i=1, 2, \ldots, n-1. \tag{8}$$

Here, K_s is the spring coefficient, $K_s>0$. Then the total spring potential energy of the whole configuration is

$$U_s=\sum_{i=1}^{n-1}u_{i,i+1}=\frac{1}{2}K_s\sum_{i=1}^{n-1}(l_{i,i+1}-1)^2. \tag{9}$$

Now, the total potential energy function of the whole configuration consists of three types of contributions: bond angle, Lennard-Jones and spring. The new energy function can be rewritten as:

$$U=K_1U_1+U_s=K_1\sum_{i=2}^{n-1}V_1(\theta_i)+K_1\sum_{i=1}^{n-2}\sum_{j=i+2}^{n}V_2(r_{ij},\zeta_i,\zeta_j)+\frac{1}{2}K_s\sum_{i=1}^{n-1}(l_{i,i+1}-1)^2. \tag{10}$$

Here K_1 is a proportional coefficient, whose use will be discussed later. It can be seen from Eqs.(1)~(3) to Eqs.(7)~(10) that the potential energy U of the whole configuration is a known function of the coordinates $x_1, y_1, z_1, \ldots, x_n, y_n, z_n$ of the centers of all the balls:

$$U=U(x_1, y_1, z_1, \ldots, x_n, y_n, z_n). \tag{11}$$

$U(x_1, y_1, z_1, \ldots, x_n, y_n, z_n)$ is defined on the entire $3n$-dimentianal Euclidean space $(-\infty, +\infty)^{3n}$, smooth, continuous and differentiable everywhere. Based on this new energy function, the protein folding problem is converted to a problem of optimization of the total potential energy $U(x_1, y_1, z_1, \ldots, x_n, y_n, z_n)$. The aim is to find a configuration $(x_1^*, y_1^*, z_1^*, \ldots, x_n^*, y_n^*, z_n^*)$ with minimum energy:

$$U(x_1^*, y_1^*, z_1^*, \ldots, x_n^*, y_n^*, z_n^*)=\min(U). \tag{12}$$

Obviously, this problem is an unconstrained optimization problem, for which, there exists a ready-made algorithm for its solution, the gradient method, or the steepest descent method (Wang *et al*., 2002).

Eq.(8) and Eq.(9) show that the spring potential energy is non-negative. According to Eq.(8), if the coefficient K_s is set to be large enough, a spring with length differing slightly from the natural length 1 can considerably increase the whole energy of the configuration. Accordingly, if a configuration is not a legal one, that is, there are some springs compressed or stretched, the total energy of the configuration will not be very low. Therefore, we can see that the total elastic energy of the springs acts as a penalty function of the degree of departure of a configuration from a legal one, thus ensuring that the resulting configuration is legal.

Gradient method

Randomly define $3n$ real numbers in 3D Euclidean space as the initial configuration $(x_1^{(0)}, y_1^{(0)}, z_1^{(0)}, \ldots, x_n^{(0)}, y_n^{(0)}, z_n^{(0)})$. Calculate grad$U$ at $(x_1^{(0)}, y_1^{(0)}, z_1^{(0)}, \ldots, x_n^{(0)}, y_n^{(0)}, z_n^{(0)})$:

$$\text{grad}U=\left(\frac{\partial U}{\partial x_1}, \frac{\partial U}{\partial y_1}, \frac{\partial U}{\partial z_1}, \ldots, \frac{\partial U}{\partial x_n}, \frac{\partial U}{\partial y_n}, \frac{\partial U}{\partial z_n}\right), \tag{13}$$

where

$$\frac{\partial U}{\partial x_i}=K_1\frac{\partial U_1}{\partial x_i}+\frac{\partial U_s}{\partial x_i},\quad \frac{\partial U}{\partial y_i}=K_1\frac{\partial U_1}{\partial y_i}+\frac{\partial U_s}{\partial y_i},\quad \frac{\partial U}{\partial z_i}=K_1\frac{\partial U_1}{\partial z_i}+\frac{\partial U_s}{\partial z_i},\quad i=1, 2, \ldots, n. \tag{14}$$

Then a new configuration can be calculated following the gradient method:

$$x_i^{(1)}=x_i^{(0)}+\varepsilon(-\partial U/\partial x_i),\quad y_i^{(1)}=y_i^{(0)}+\varepsilon(-\partial U/\partial y_i),\quad z_i^{(1)}=z_i^{(0)}+\varepsilon(-\partial U/\partial z_i),\quad i=1, 2, \ldots, n, \tag{15}$$

where the partial derivatives $\partial U/\partial x_i$, $\partial U/\partial y_i$, $\partial U/\partial z_i$, i=1, 2, …, n are defined at $(x_1^{(0)}, y_1^{(0)}, z_1^{(0)}, \ldots, x_n^{(0)}, y_n^{(0)}, z_n^{(0)})$ in a 3n-dimensional space. ε is step size, which is a small positive real number. We let ε be 10^{-6} in our procedure. Using vector representation,

$$(x_1^{(1)}, y_1^{(1)}, z_1^{(1)}, \ldots, x_n^{(1)}, y_n^{(1)}, z_n^{(1)}) = (x_1^{(0)}, y_1^{(0)}, z_1^{(0)}, \ldots, x_n^{(0)}, y_n^{(0)}, z_n^{(0)}) + \varepsilon(-\mathrm{grad}U). \quad (16)$$

After moving towards the opposite direction of the gradient by $\varepsilon|\mathrm{grad}U|$, configuration $(x_1^{(0)}, y_1^{(0)}, z_1^{(0)}, \ldots, x_n^{(0)}, y_n^{(0)}, z_n^{(0)})$ becomes $(x_1^{(0)}, y_1^{(0)}, z_1^{(0)}, \ldots, x_n^{(0)}, y_n^{(0)}, z_n^{(0)})$. The physical meaning of the negative gradient, $-\mathrm{grad}U$, in the gradient method is the generalized force in the system. $(-\partial U/\partial x_i, -\partial U/\partial y_i, -\partial U/\partial z_i)$ represents the magnitude and direction of the total force exerted on the ith ball. It should be pointed out that the evolution of $(x_1, y_1, z_1, \ldots, x_n, y_n, z_n)$ in the gradient method is a series of movements of the positions of the n balls to a legal configuration with minimum energy.

To adjust the proposition of U_1 in the total energy, we multiply U_1 by a proportional coefficient K_1. At the initial phase of the iteration process, we let K_1 be much larger than K_s so that U_1 dominates the evolution of the configuration to low energy states. As pointed out earlier, to ensure that the resulting configuration is a legal one, the coefficient K_s should be large enough so that a little deformation of the springs away from the natural length will cause considerable increase of the total energy. So we increase K_s and decrease K_1 gradually as the iteration continues, which will increase U_s to drive the configuration to a legal configuration. At the end of the iteration process, the configuration becomes a legal one with low energy.

The calculating procedure is presented as follows:

(1) Randomly give n points (x_1, y_1, z_1), …, (x_n, y_n, z_n) in 3D Euclidean space as the initial configuration $(x_1^{(0)}, y_1^{(0)}, z_1^{(0)}, \ldots, x_n^{(0)}, y_n^{(0)}, z_n^{(0)})$. Let t=0, K_1=4001, K_s=1. Choose a very small positive number, λ, as the criterion for judging grad$U$ to be zero approximately.

(2) Calculate $|\mathrm{grad}U|$ under configuration $(x_1^{(t)}, y_1^{(t)}, z_1^{(t)}, \ldots, x_n^{(t)}, y_n^{(t)}, z_n^{(t)})$. If $|\mathrm{grad}U|<\lambda$, go to Step (6).

(3) $(x_1^{(t+1)}, y_1^{(t+1)}, z_1^{(t+1)}, \ldots, x_n^{(t+1)}, y_n^{(t+1)}, z_n^{(t+1)}) \Leftarrow (x_1^{(t)}, y_1^{(t)}, z_1^{(t)}, \ldots, x_n^{(t)}, y_n^{(t)}, z_n^{(t)}) + \varepsilon(-\mathrm{grad}U)$.

(4) If $K_1>1$, then $K_1 \Leftarrow K_1-0.001$.

(5) $K_s \Leftarrow K_s+10$, $t \Leftarrow t+1$ and turn to Step (2).

(6) Now, the gradient is approximately zero. Calculate the energy of the resulting configuration according to Eq.(1) as the solution and then stop the computation procedure.

Since K_s is rather large ($K_s>10^7$) at the end of the calculation, the resulting configuration satisfies Eq.(5) approximately, that is, the length of the springs satisfies the following requirement:

$$|l_{i,i+1}-1|<10^{-6}, \quad i=1, 2, \ldots, n-1. \quad (17)$$

Strategy of generating promising initial configuration

It should be pointed out that the solution of the algorithm above might just be a local (and hopefully also global) minimum. Since gradient method is a deterministic algorithm and the initial configurations are generated randomly, the resulting solutions are very unstable. So we start from a new initial configuration and the above-described computation resumes over again. From many solutions, we choose the best one. Experiments showed that a good result would be obtained from more than one hundred times computation. Thus ensuring that the initial configuration is certainly desirable.

Inspired by the phenomenon that hydrophobic amino acids are lumped together as a compact core surrounded by hydrophilic amino acids in a protein molecule, we put forward a heuristic strategy to generate promising initial configuration that simulates the real protein structure.

We define two spherical spaces with radii R_1 and R_2, respectively, where R_1 and R_2 are positive numbers with $R_2=2R_1$. The two spherical spaces have the same center, which is the origin of the 3D Cartesian coordinate system. For a black ball in initial configuration, its center position can only be generated randomly in a 3D space confined in the spherical space with radius R_1. We set $R_1=\sqrt{n}$ in our algorithm. For a white ball in initial configuration, its center position can only be generated randomly in a 3D space confined in the ball with radius R_2 but excluding the space of ball R_1. In a more formal way, it can be

stated as follows:

$$\sqrt{x_i^2+y_i^2+z_i^2} \le R_1, \tag{18}$$

$$R_1 < \sqrt{x_j^2+y_j^2+z_j^2} \le R_2, \tag{19}$$

where i is black ball and j is white ball, and x, y, z are the coordinates of the center of a randomly generated ball.

Experimental results showed that this strategy could generate relatively better initial configurations. To illustrate this strategy, an initial configuration of 13 balls is shown in Fig.1. For ease of visualization, the illustration is confined to two dimensions.

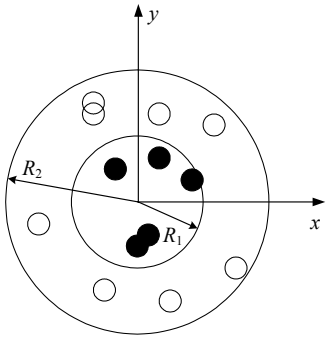

Fig.1 An initial configuration of 13 balls generated by the strategy of generating promising initial configuration

RESULTS

Table 1 shows the lowest energies obtained by our heuristic algorithm, along with the results by nPERM, ELP and CSA. It can be seen that our results are better than those of the nPERM for all the four sequences, with the energy difference increasing gradually for longer chains. For sequence with length 13, our result was also slightly better than that of ELP, and was equal to that of CSA. For other cases, however, we cannot reach the energy yielded by ELP and CSA.

Fig.2 shows the lowest energy configurations obtained by our heuristic algorithm, where black circles indicate hydrophobic monomers (A) and white circles indicate hydrophilic monomers (B). It can be seen that the configuration has single hydrophobic core for all four sequences, which is analogous to the real protein structure.

It should be pointed out that each of the results is the best one of the solutions iterated from several (≤10) randomly generated initial configurations. The runtime for all the four sequences was less than 2 h on a P4 2.4 GHz PC with 512 MB memory, while the computation time of nPERM was up to 2 d on Linux and UNIX workstation. Obviously, HA is much faster than nPERM. Note that the runtime of ELP and CSA was not reported in the literature.

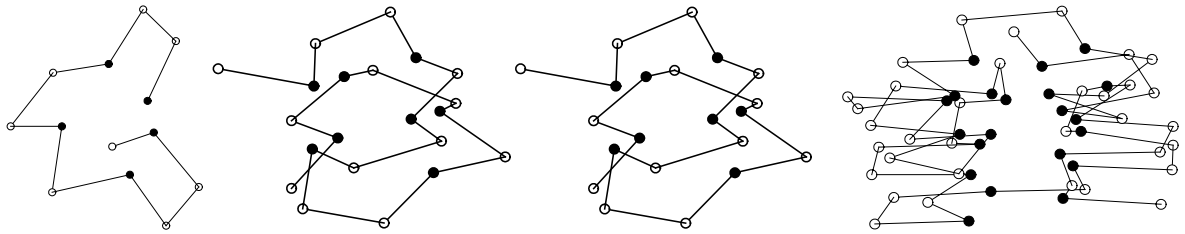

Fig.2 The lowest energy configurations for the four sequences obtained by heuristic algorithm (a) n=13; (b) n=21; (c) n=34; (d) n=55

Table 1 Test sequences and the lowest energies obtained by heuristic algorithm (HA), in comparison with those by nPERM, ELP, and CSA, respectively

n	Sequence	nPERM	ELP	CSA	HA
13	ABBABBABABBAB	−4.9616	−4.967	−4.9746	−4.9746
21	BABABBABABBABBABABBAB	−11.5238	−12.316	−12.3266	−12.0617
34	ABBABBABABBABBABABBABABBABBABABBAB	−21.5678	−25.476	−25.5113	−23.0441
55	BABABBABABBABBABABBABABBABBABABBAB BABABBABABBABBABABBAB	−32.8843	−42.428	−42.3418	−38.1977

CONCLUSION

The objective of the protein folding problem is to find inherent structures for a given set of attracting particles (amino acid monomers) that initially are widely dispersed. The elastic potential energy of spring is introduced into the energy function of the configuration to convert the protein folding problem to an unconstrained optimization problem solvable by the steepest descent method. Random initial configurations of the n particles were mapped onto the final inherent structure configurations by a numerical steepest descent on the potential energy surface. You can watch particles move according to the steepest descent algorithm from an initial diffuse random array towards a more compact array with lower potential energy.

Since gradient method is only a local search algorithm, it is possible for the gradient method to fall into the trap of local minimum. Selecting the best one from many solutions iterated from a promising initial configuration in a confined space may help to find a comparably good solution, but that will cost much computation time. In our future work, we hope to find some efficient strategy of jumping out of local minimum to develop more efficient algorithm.

References

Anfinsen, C., 1973. Principles that govern the folding of protein chains. *Science*, **181**:223-230.

Bachmann, M., Arkın, H., Janke, W., 2005. Multicanonical study of coarse-grained off-lattice models for folding heteropolymers. *Phys. Rev. E*, **71**:031906. [doi:10.1103/PhysRevE.71.031906]

Crescenzi, P., Goldman, D., Papadimitriou, C., Piccolboni, A., Yannakakis, M., 1998. On the complexity of protein folding. *Journal of Computational Biology*, **5**(3):409-422.

Dill, K.A., 1985. Theory for the folding and stability of globular proteins. *Biochemistry*, **24**:1501-1509. [doi:10.1021/bi00327a032]

Hsu, H.P., Mehra, V., Grassberger, P., 2003. Structure optimization in an off-lattice protein model. *Phys. Rev. E*, **68**:037703. [doi:10.1103/PhysRevE.68.037703]

Irback, A., Peterson, C., Potthast, F., 1997. Identification of amino acid sequences with good folding properties in an off-lattice model. *Phys. Rev. E*, **55**:860-867. [doi:10.1103/PhysRevE.55.860]

Kim, S.Y., Lee, S.B., Lee, J., 2005. Structure optimization by conformational space annealing in an off-lattice protein model. *Phys. Rev. E*, **72**:011916. [doi:10.1103/PhysRevE.72.011916]

Lau, K.F., Dill, K.A., 1989. A lattice statistical mechanics model of the conformational and sequence space of proteins. *Macromolecules*, **22**:3986-3997. [doi:10.1021/ma00200a030]

Stillinger, F.H., 1995. Collective aspects of protein folding illustrated by a toy model. *Phys. Rev.*, **52**:2872-2877.

Torcini, A., Livi, R., Politi, A., 2001. A dynamical approach to protein folding. *J. Biol. Phys.*, **27**:181-186. [doi:10.1023/A:1013104123892]

Wang, H.Q., Huang, W.Q., Zhang, Q., Xu, D.M., 2002. An improved algorithm for the packing of unequal circles within a larger containing circle. *European Journal of Operational Research*, **141**:440-453. [doi:10.1016/S0377-2217(01)00241-7]