A Quasi-physical Algorithm for the Structure Optimization in an Off-lattice Protein Model

Jing-Fa Liu[1,2]* and Wen-Qi Huang[1]

[1] *School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China;* [2] *Department of Mathematics, Hengyang Normal College, Hengyang 421008, China.*

In this paper, we study an off-lattice protein AB model with two species of monomers, hydrophobic and hydrophilic, and present a heuristic quasi-physical algorithm. First, by elaborately simulating the movement of the smooth solids in the physical world, we find low-energy conformations for a given monomer chain. A subsequent off-trap strategy is then proposed to trigger a jump for a stuck situation in order to get out of the local minima. The algorithm has been tested in the three-dimensional AB model for all sequences with lengths of 13–55 monomers. In several cases, we renew the putative ground state energy values. The numerical results show that the proposed methods are very promising for finding the ground states of proteins.

Key words: protein folding, off-lattice model, quasi-physical algorithm, off-trap strategy

Introduction

Predicting the native structure of a protein from its amino acid sequence is one of the most challenging problems in biophysics and bioinformatics. The difficulty of the problem comes from two aspects. One is the determination of the potential energy function. The effective energy function can generally distinguish the native states from non-native states of protein molecules. The other is that the potential energy landscape of the system can be characterized by a multitude of local minima separated by high-energy barriers.

In order to simplify and clarify these two aspects of protein folding phenomena, in resent years the theoretical community has introduced and examined several highly simplified, but still nontrivial models, including a large family of hydrophobic-polar (HP) lattice (*1–5*) and off-lattice models (*6*, *7*). Even though these models are highly simplified, to solve the corresponding folding problem remains to be NP-hard. In recent years, a wide variety of approximate algorithms have been employed to analyze these models, including the sequential importance sampling with pilot-exploration (SISPER; ref. *8*) based on the important sampling, the multi-self-overlap ensemble (MSOE) approach (*9*) based on the Monte Carlo scheme, and the pruned-enriched Rosenbluth method (PERM; ref. *10–13*). All these approaches have been applied to the simplified protein folding problems. However, their efficiency still needs to be improved.

The present paper attempts to find a highly efficient heuristic algorithm that can obtain the low-energy states for a given monomer chain. We use a so-called quasi-physical method (*14–16*), whose working path is to find a natural phenomenon equivalent to the protein folding problem in the physical world. Then we observe the evolution of the motion of matter in it so as to be inspired to obtain a formalistic algorithm for solving the problem.

To see whether the quasi-physical method can be efficient for energy minimization in the protein folding problem, in this paper we introduce a so-called AB model by Stillinger *et al* (*17*, *18*), where the hydrophobic monomers are labeled by A and the hydrophilic or polar ones are labeled by B. This model has been studied in several papers (*17–26*). The methods used to find low-energy states of the AB model include neural networks (*17*), conventional Metropolis type Monte Carlo procedures (*18*), the annealing contour Monte Carlo method (*19*), the simulated tempering (*20*), biologically motivated methods (*21*, *22*), multicanonial methods (*23*, *24*, *26*), and the new PERM with importance sampling (nPERMis; ref. *25*). For its two-dimensional (2D) version, the putative ground states for various AB sequences with various chain lengths have been given in previous

*** Corresponding author.**
E-mail: ljf720622@163.com

studies (*17–19*, *21*, *25*), and for its three-dimensional (3D) version, the putative ground states for four Fibonacci sequences with lengths of 13–55 monomers have also been obtained (*25*, *26*). The present paper studies the 3D version of this model. The numerical results show that the proposed methods are very promising for finding the ground states of proteins.

Algorithm

The AB model

For an n-monomer chain, the distances between the consecutive monomers along the chain are fixed to the unit length, while the non-consecutive monomers interact through a modified Lennard-Jones potential. In addition, there is an energy contribution from each angle $\theta_i(-\pi \leq \theta_i < \pi)$ between consecutive bonds. More precisely, the total energy function (*17–19*, *21*, *25*, *26*) for an n-monomer chain can be written as

$$E = \sum_{i=2}^{n-1} \frac{1}{4}(1 - \cos\theta_i) + \sum_{i=1}^{n-2} \sum_{j=i+2}^{n} 4[r_{ij}^{-12} - C(\zeta_i, \zeta_j) r_{ij}^{-6}] \tag{1}$$

Here r_{ij} is the distance between monomers i and j $(i < j)$. Each ζ_i is either A or B. The first term is the bending energy, favoring the alignment of the three successive monomers $i - 1$, i, and $i + 1$. The second term is the Lennard-Jones potential with a species-dependent coefficient $C(\zeta_i, \zeta_j)$, which is taken to be 1 for an AA pair, 1/2 for a BB pair, and −1/2 for an AB pair, giving strong attraction, weak attraction, and weak repulsion, respectively.

The protein folding problem for the AB model can be formally defined as follows: given a monomer chain $s = \zeta_1 \zeta_2 \zeta_3 \dots \zeta_n$, we try to find an energy-minimizing conformation of s, that is, to find $X^* \in C(s)$ so that $E(X^*) = \min\{E(X) | X \in C(s)\}$, where $C(s)$ is the set of all the valid conformations of s.

The quasi-physical method

Imagine that all n monomers involved in the model are smooth solids with the radius of each being 1/2, which are marked by 1, 2, ..., n and are cast randomly in the 3D Euclidean space, then the potential energy E is a known function of the coordinates of all n monomers with the constraint:

$$r_{i,i+1} = 1 \ (i = 1, 2, \dots, n-1) \tag{2}$$

To remove Constraint (2) and convert the constrained optimization problem into the unconstrained optimization problem, we conceive that there exists a spring with the original length 1 between the centers of the i^{th} and $(i+1)^{\text{th}}$ monomers $(i = 1, 2, \dots, n-1)$. According to the Hooks' law, the spring's elastic potential energy between two adjacent monomers is proportional to the square of the length of the spring transformation. So we can give the elastic potential energy as follows:

$$E' = \sum_{i=1}^{n-1} \frac{1}{2} k_1 (r_{i,i+1} - 1)^2 \tag{3}$$

where k_1 is a physical coefficient characterizing the rigidities of all the springs.

The sum of the potential energy of the whole system is $U = E + E'$. Obviously, the potential energy U is a known function of the coordinates X of all the monomers, and E' is the "penalty" term in the potential energy U. When the physical coefficient k_1 is great enough, the optimal solution of the unconstrained optimization problem of the known function $U(X)$ is also the optimal solution of the constrained optimization problems (1)–(2). Thus, the protein folding problem is converted into an unconstrained optimization problem of $U(X)$.

For this optimization problem, we can employ a ready-made algorithm, the gradient method, or the steepest descent method. By integrating the quasi-physical idea into the gradient method, we gain a quasi-physical algorithm. Assume that $X^{(0)}$ is the initial conformation for iterations. If the conformation $X^{(t)}$ $(t \geq 0)$ is not a local minimum, a new conformation $X^{(t+1)} = X^{(t)} - \lambda_t \nabla U^{(t)}$ is obtained in the anti-gradient direction of the energy function $U(X)$ at $X^{(t)}$, where λ_t is the iterative step length and $-\nabla U^{(t)}$ is the iterative search direction. From the undated conformation $X^{(t+1)}$, we repeat this course of iterations until a global energy minimum conformation X^* is found, or a trap of local minimum X^* $(|\nabla U^{(t)}| < 10^{-6})$ occurs. In the latter case, a completely new round of quasi-physical calculation should be initiated from a new initial conformation. In our simulations of computation, we let $\lambda_0 = 0.5 \times 10^{-6}$, $k_1 \in [1,000, 2,000]$.

In order to speed up the ground state search, when calculating the gradient ∇U in iterations, we modify the Lennard-Jones potential through multiplying it by a physical coefficient k_2= 10,000, meaning the Van Del Waals interaction coefficient. Another trick is to modify the physical coefficient k_1 by multiplying the factor 1.3 and k_2 by 0.7 until $k_2 < 1$ per 50,000 iterative steps in the course of optimization. In the later stage of calculation, or when $\left|\nabla U^{(t)}\right| < 10^{-2}$, we modify k_1 by multiplying the factor 1.1 and decrease the step length λ_t by multiplying a step shrinking factor 0.9 per 50,000 iterative steps. Therefore, in the beginning of calculation, the physical coefficient of all the springs is small so that all monomers can move freely and attain easily low-energy states. Thereafter, along with the execution of calculation, the physical coefficient k_1 increases gradually so as to increase the penalty and make Constraint (2) satisfied gradually, and at last the interactions of springs disappear. Obviously, when the physical coefficient k_1 rises to a big number, for example $\geq 10^{10}$, that is, when the springs turn rigid, Constraint (2) is satisfied naturally, and a global energy minimum conformation is found, or a trap of local minimum occurs.

The off-trap strategy

The calculating experience tells us when all the A-monomers fold into a hydrophobic core, the potential energy of the whole system will turn low. For jumping out of local minima, we can pick out all B-monomers squeezed among A-monomers, and place them in certain spots in 3D space to speed up the lowest-energy state search. The concrete description of the off-trap strategy is as follows: (1) Calculate the center of all A-monomers; (2) Compute the distance from the center to every A-monomer, signing the greatest distance as d; (3) Calculate respectively the distance from every B-monomer to the center. For every B-monomer, as long as the distance $< d$, it is our strategy to pick out the corresponding B-monomer and place it somewhere three times of the distance away from the center, in the vector direction from the center to the B-monomer picked. Keeping all A-monomers and the rest of B-monomers at their current positions, we can obtain a new conformation, where in addition to the changes of bending energy and elastic potential energy, long-range Lennard-Jones interactions of the monomers, with their relative position to each other changed, have to be computed anew after the update.

By integrating the off-trap strategy into the quasi-physical algorithm, we gain a heuristic quasi-physical (HQP) algorithm. The calculation is executed by using the quasi-physical algorithm until a certain minimum conformation is reached. If now the energy E is lower than the target value and the system is satisfied with Constraint (2), the energy will be scored and the computation will terminate, otherwise the calculation point will jump to a new position through the off-trap strategy and the computation will proceed with the quasi-physical algorithm. If the off-trap strategy is repeated up to ten times and the simulation does not find states with lower energy E than the target value, it is our practice to randomly choose another initial conformation for a new round of HQP calculation.

Results and Discussion

We implemented the HQP algorithm in the C++ language on a Pentium IV, 2.0 GHz computer. For the sake of examining the calculation, in this paper we restricted ourselves to the AB model with the Fibonacci sequences in previously studies (*17–19*, *21*, *25*, *26*). The Fibonacci sequences are defined recursively by S_0=A, S_1=B, S_{i+1}=S_{i-1} * S_i. Here "*" is the concatenation operator, for example, the first few sequences are S_2=AB, S_3=BAB, S_4=ABBAB, and so on. They have the lengths given by $n_{i+1} = n_{i-1} + n_i$, that is, given by the Fibonacci numbers. Following the previous studies (*25*, *26*), in this paper we considered the sequences with lengths n=13, 21, 34, and 55, which are listed in Table 1.

In our simulations, the initial conformations were chosen according to the following strategy. Given two concentric spheres with the origin as their centers and r_1=$2n$, r_2=$5n$ as their radii respectively, where n is the length of the chain considered. Cast randomly all A-monomers in small sphere and B-monomers in the region between small and big spheres.

After producing the initial conformation, we can execute the HQP algorithm to compute the position of every ball hereafter at every time. The calculating results showed that along with the increment in the calculation steps, at last all monomers would tend to be stable. In the four conformations that were the solution of the problem, Constraint (2) was satisfied approximately. The error margin was smaller than 10^{-6}, that is: $|r_{i,\,i+1} - 1| < 10^{-6}$ $(i = 1, 2, \ldots, n-1)$.

We employed the HQP algorithm to compute these sequences and produced the final results that are shown in Table 2 in comparison with those of other studies (*25*, *26*). The conformations of putative ground states are shown in Figure 1.

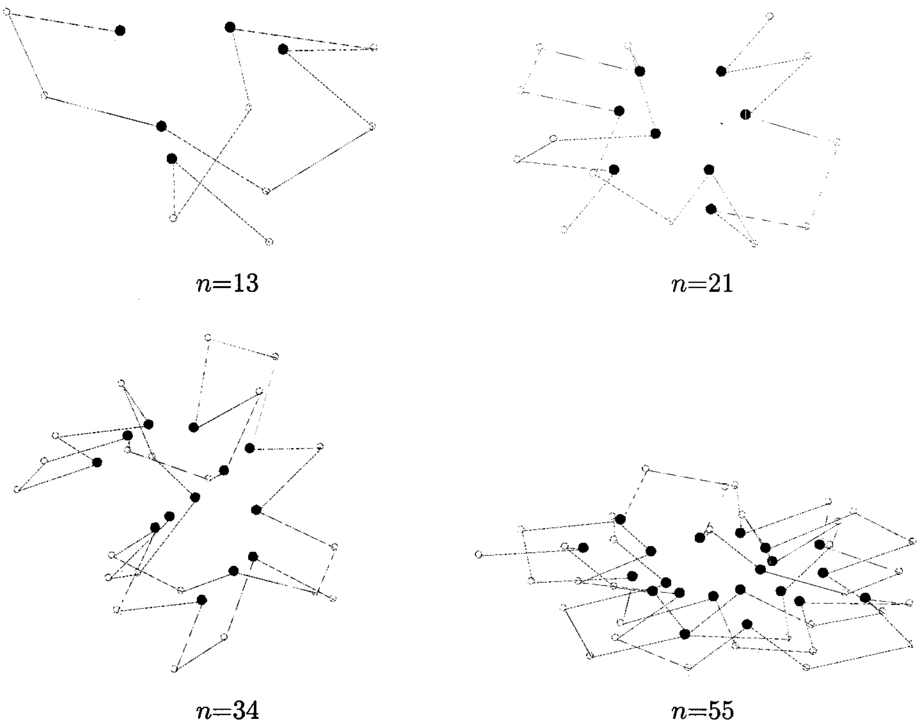
Table 1 The Four Fibonacci Sequences in the AB Model

n	Sequence ("B$_2$" for BB)
13	$\mathrm{AB_2AB_2ABAB_2AB}$
21	$\mathrm{BABAB_2ABAB_2AB_2ABAB_2AB}$
34	$\mathrm{AB_2AB_2ABAB_2AB_2ABAB_2ABAB_2AB_2ABAB_2AB}$
55	$\mathrm{BABAB_2ABAB_2AB_2ABAB_2ABAB_2AB_2ABAB_2AB_2ABAB_2ABAB_2AB_2ABAB_2AB}$

Table 2 Comparison of HQP and Other Algarithms on the Estimation of the Global Energy Minima for the Four Fibonacci Sequences in the AB Model*

n	PERM	PERM+	MUCA	ELP	HQP (CPU time$^{\#}$)
13	−3.9730	−4.9616	−4.967	−4.967	−4.9729 (5,674 s)
21	−7.6857	−11.5238	−12.296	−12.316	−12.2554 (8,924 s)
34	−12.8601	−21.5678	−25.321	−25.476	−24.8083 (24,265 s)
55	−20.1070	−32.8843	−41.502	−42.428	−42.5199 (39,124 s)

*The values are compared with the results quoted in Hsu *et al* (*25*) employing the PERM and the subsequent conjugate gradient (PERM+) minimization, and with the lowest energies listed in Bachmann *et al* (*26*) obtained with the multicanonical (MUCA) Monte Carlo method and the energy landscape paving (ELP) minimization. $^{\#}$CPU time means the time needed in a certain running to get the listed energies on the Pentium IV, 2.0 GHz computer.



n=13

n=21

n=34

n=55

Fig. 1 Stereographic views of putative ground states of the four Fibonacci sequences listed in Table 1. Full dots and empty circles indicate hydrophobic and hydrophilic monomers, respectively.

The experiments showed that the HQP algorithm considerably outperformed the PERM algorithm and the subsequent conjugate gradient (PERM+) algorithm in Hsu *et al* (*25*) for the Fibonacci sequences of Table 1 in the AB model. This was particularly pronounced for the longest chain considered. In addition, for the monomer chains with n = 21, 34, and 55, we found putative ground states different from those given in Hsu *et al* (*25*), which stated that the chains with n= 21 and 34 folded into conformations with single hydrophobic cores (except for a single A-monomer that kept out in both cases), and the chain with n= 55 formed two clearly disjointed main hydrophobic groups. From the conformations (Figure 1) produced by the HQP simulation, we easily see that each of the four Fibonacci sequences has a single hydrophobic core. Indeed, with this fact we are able to refute the claims for putative ground states in Hsu *et al* (*25*), and agree well with what comes out in Bachmann *et al* (*26*).

We also compared our results with the minimum energies listed in Bachmann *et al* (*26*), where the so-called multicanonical (MUCA) Monte Carlo method and the energy landscape paving (ELP) minimization were applied to these Fibonacci sequences. Table 2 shows that HQP runs lower energies for the sequences with 13 and 55 monomers, while the results for those with 21 and 34 monomers are also comparable. Moreover, the CPU time used by the HQP algorithm should be less than or comparable to those used by the PERM+ and MUCA methods. Hsu *et al* (*25*) did not give the exact CPU time of their runs. They just mentioned that their results were obtained on their Linux or Unix workstation with up to 2 days of CPU time, while Bachmann *et al* (*26*) even did not mention the CPU time of their runs.

Furthermore, we will apply the methods proposed in this paper to all-atom models with realistic potential by combining it with the simulated annealing or the genetic methods, and design various kinds of higher performance algorithms for the protein folding problem.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (No. 10471051), the National Key Basic Research Program (973) of China (No. 2004CB318000), and the Foundation of Hunan Educational Office (No. 05C649).

References

1. Dill, K.A. 1985. Theory for the folding and stability of globular proteins. *Biochemistry* 24: 1501-1509.
2. Lan, K.F. and Dill, K.A. 1989. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* 22: 3986-3997.
3. Shakhnovich, E.I. 1994. Proteins with selected sequences fold into unique native conformation. *Phys. Rev. Lett.* 72: 3907-3911.
4. Dill, K.A., *et al.* 1995. Principles of protein folding—a perspective from simple exact models. *Protein Sci.* 4: 561-602.
5. Camacho, C.J. and Thirumalai, D. 1993. Kinetics and thermodynamics of folding in model proteins. *Proc. Natl. Acad. Sci. USA* 90: 6369-6372.
6. Honeycutt, J.D. and Thirumalai, D. 1992. The nature of folded states of globular proteins. *Biopolymers* 32: 695-709.
7. Fukugita, M., *et al.* 1993. Kinematics and thermodynamics of a folding heteropolymer. *Proc. Natl. Acad. Sci. USA* 90: 6365-6368.
8. Zhang, J.L. and Liu, J.S. 2002. A new sequential importance sampling method and its application to the two-dimensional hydrophobic-hydrophilic model. *J. Chem. Phys.* 117: 3492-3498.
9. Chikenji, G., *et al.* 1999. Multi-self-overlap ensemble for protein folding: ground state search and thermodynamics. *Phys. Rev. Lett.* 83: 1886-1889.
10. Hsu, H.P., *et al.* 2003. Growth algorithms for lattice heteropolymers at low temperatures. *J. Chem. Phys.* 118: 444-452.
11. Frauenkron, H., *et al.* 1998. A new Monte Carlo algorithm for protein folding. *Phys. Rev. Lett.* 80: 3149-3152.
12. Grassberger, P. 1997. The pruned-enriched Rosenbluth method: simulations of Theta polymers of chain length up to 1000000. *Phys. Rev. E* 56: 3682-3693.
13. Huang, W.Q. and Lü, Z.P. 2004. Personification algorithm for protein folding problem: improvements in PERM. *Chin. Sci. Bull.* 49: 2092-2096.
14. Huang, W.Q. and Kang, Y. 2002. A heuristic quasi-physical strategy for solving disks packing problem. *Simul. Model. Pract. Theory* 10: 195-207.
15. Wang, H.Q., *et al.* 2002. An improved algorithm for the packing of unequal circles within a larger containing circle. *Eur. J. Oper. Res.* 141: 440-453.
16. Huang, W.Q. and Jin, R.C. 1999. Quasiphysical and quasisociological algorithm Solar for solving SAT problem. *Sci. China E* 42: 485-493.
17. Stillinger, F.H., *et al.* 1993. Toy model for protein folding. *Phys. Rev. E* 48: 1469-1477.

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

19. Liang, F. 2004. Annealing contour Monte Carlo algorithm for structure optimization in an off-lattice protein model. *J. Chem. Phys.* 120: 6756-6763.

20. Irbäck, A., *et al.* 1997. Local interactions and protein folding: a three-dimensional off-lattice approach. *J. Chem. Phys.* 107: 273-282.

21. Gorse, D. 2002. Application of a chaperone-based refolding method to two- and three-dimensional off-lattice protein models. *Biopolymers* 64: 146-160.

22. Gorse, D. 2001. Global minimization of an off-lattice potential energy function using a chaperone-based refolding method. *Biopolymers* 59: 411-426.

23. Irbäck, A., *et al.* 1997. Identification of amino acid sequences with good folding properties in an off-lattice model. *Phys. Rev. E.* 55: 860-867.

24. Irbäck, A. and Potthast, F. 1995. Studies of an off-lattice model for protein folding: sequence dependence and improved sampling at finite temperature. *J. Chem. Phys.* 103: 10298-10305.

25. Hsu, H.P., *et al.* 2003. Structure optimization in an off-lattice protein model. *Phys. Rev. E* 68: 037703.

26. Bachmann, M., *et al.* 2005. Multicanonical study of coarse-grained off-lattice models for folding heteropolymers. *Phys. Rev. E* 71: 031906.