J Biol Phys (2009) 35:245–253
DOI 10.1007/s10867-009-9152-1

ORIGINAL PAPER

Structure optimization of the two-dimensional off-lattice hydrophobic–hydrophilic model

Jingfa Liu · Shengjun Xue · Duanbing Chen · Huantong Geng · Zhaoxia Liu

Received: 2 May 2008 / Accepted: 6 April 2009 /
Published online: 14 May 2009

Abstract A two-dimensional off-lattice protein model with two species of monomers, hydrophobic and hydrophilic, was studied. Low-energy configurations in the model were optimized using the improved energy landscape paving (ELP+) method. In ELP+, the energy landscape paving (ELP) was first applied to search for the low-energy states. After the ELP led to the basins of the local energy minima, the additional degree-of-freedom of bond length was introduced, and the gradient descent method was then used to search for lower energy states near the local minima. Numerical results show that the proposed methods are quite effective for finding the ground states of proteins. A comparison between ELP+ and other methods is made.

Keywords Protein structure prediction · Two-dimensional off-lattice model · Energy landscape paving · Local search

1 Introduction

Predicting the native structure of proteins from their amino acid sequences is one of the most challenging problems in bioinformatics. Despite many decades of research we still lack a detailed understanding of the relation between chemical composition and structure

J. Liu (✉) · S. Xue · H. Geng
Computer and Software Institute, Nanjing University of Information Science and Technology, Nanjing 210044, China
e-mail: ljf720622@163.com

D. Chen
School of Computer Science & Engineering,
University of Electronic Science and Technology of China, Chengdu 610054, China

Z. Liu
Network Information Center, Nanjing University of Information Science and Technology, Nanjing 210044, China

(and consequently function) of proteins. The theoretical analysis of the protein structure prediction problem, or the protein folding problem, faces two major difficulties. 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 detailed protein models are often characterized by a rough energy landscape with a huge number of local minima separated by high-energy barriers.

Since solving such a problem is too difficult for realistic protein models, in recent years the theoretical community has introduced and examined some highly simplified, but still nontrivial, models which make the following assumption: instead of considering all 20 different kinds of amino acids in real proteins, the models comprise only two prototypes of residues: hydrophobic (H) and the hydrophilic (or polar, P) monomers. These models include a large family of HP lattice models [1–3] and HP off-lattice models [4, 5]. Even though these models are highly simplified, to solve the corresponding folding problem remains NP-hard.

In this article, we introduced a so-called AB model by Stillinger et al. [6, 7], where the hydrophobic monomers are now labeled by A and the hydrophilic or polar ones by B. Based on the minimum free-energy theory [8], many authors have developed various techniques to search for the low-energy states of AB proteins in the two-dimensional (2D) case. The methods of optimizing low-energy states include conventional Metropolis-type Monte Carlo (MMC) procedures [7], the memory Tabu search (MTS) method [9], the off-lattice pruned–enriched Rosenbluth method (PERM) and, after subsequent conjugate gradient minimization (PERM+) [10], the annealing contour Monte Carlo (ACMC) method [11], as well as conformational space annealing (CSA) [12] and statistical temperature annealing (STA) methods [13]. For four amino acids chains with lengths $13 \leq n \leq 55$ studied in Refs. [6, 7, 9–13], in this article we introduced another global optimization method, the energy landscape paving (ELP) method [14, 15], to optimize low-energy configurations. After the ELP led to the basins of the local energy minima, the additional degree-of-freedom of bond length was introduced, and the local search based on gradient descent was then used to search lower energy states near the local minima. Numerical results show that the proposed methods are very promising for finding the ground states of proteins.

2 Two-dimensional AB model

For an n-mer chain living in two dimensions, the distances between consecutive monomers along the chain are fixed to unit length, while nonconsecutive monomers interact through a modified Lennard–Jones potential. There is an angle θ_i between bond vectors $\boldsymbol{u}_i$ and $\boldsymbol{u}_{i+1}$, where θ_i $(1 \leq i \leq n-2)$ denotes the bond angle at monomer $i+1$ and satisfies $-\pi \leq \theta_i < \pi$. When the rotating direction of the angle is counterclockwise, its value is positive. The shape of the n-mer is either specified by the $n-2$ bond angles θ_1,..., θ_{n-2} or by the $n-1$ bond vectors $\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_{n-1}$. The definitions of the bond angle θ_i and the bond vector $\boldsymbol{u}_i$ are shown in Fig. 1. The energy function [6, 7, 9–13, 16, 17] consists of two types of contributions: bond angle and Lennard–Jones. It can be written as

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

Here, r_{ij} is the distance between monomers 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 (or BA) pairs, giving

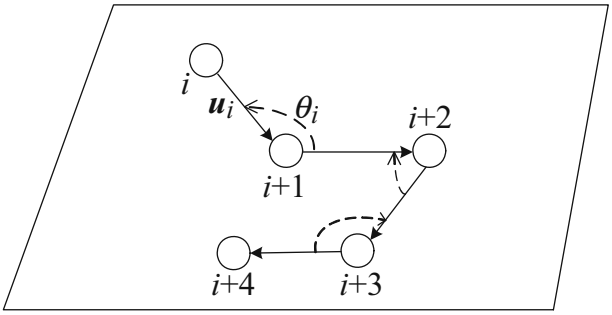

Fig. 1 Definitions of bond vector $\boldsymbol{u}_i$ and bond angle θ_i

strong attraction between AA pairs, weak attraction between BB pairs, and weak repulsion between A and B.

Obviously, the AB model with no explicit representation of side chains or hydrogen bonds is able to provide only a coarse-grained approximation to the complexities of real proteins. However, its off-lattice constructions and the uses of both Lennard–Jones and bond-angle contributions still enable it to capture some of the basic features of protein structures.

The protein structure prediction problem for the AB model can be formally defined as follows: given a monomer chain $s = \zeta_1\zeta_2\zeta_3...\zeta_n$, we try to find an energy-minimizing configuration of s in the 2D plane, i.e., find $X^* \in G(s)$ so that $E(X^*) = \min\{E(X) \mid X \in G(\mathrm{s})\}$, where $G(s)$ is the set of all the valid configurations of s.

3 Theory and methods

3.1 Energy landscape paving method

In order to locate low-energy states of the sequences under consideration, we used the so-called energy landscape paving (ELP) minimizer [14, 15], which combines ideas from energy landscape deformation [18] and Tabu search [19]. The ELP minimization is a Monte Carlo (MC) optimization method, but with a modified energy expression designed to steer the search away from regions that have already been explored. This means that if a state X with energy $E(X)$ is hit, the energy E is increased by a "penalty" and replaced by energy $\widetilde{E} = E + f(H(q,t))$. In this expression, the "penalty" term $f(H(q,t)$ is a function of the histogram $H(q,t)$ in a prechosen "order parameter" q. The histogram is updated at each MC step, thus becoming a function of "time" t. The statistical weight for a state X is defined as

$$\omega\left(\widetilde{E}(X)\right) = \exp\left(-\widetilde{E}(X)/k_B T\right),$$

where $k_\mathrm{B}T$ is the thermal energy at the (low) temperature T, and k_B is the Boltzmann constant. In the ELP simulations of the 2D AB model, we used the potential energy itself as an order parameter and chose $\widetilde{E} = E + H(E,t)$ as the replacement of the energy E. The temperature was set to $T = 5\ K$.

In a regular low-temperature simulation, the probability to escape a local minimum depends on the height of the surrounding energy barriers. Within ELP, the sampling weight of a local minimum state decreases with the time the system stays in that minimum, and consequently the probability to escape the minimum increases until the local minimum is

no longer favored. The system will then explore higher energies until it falls into a new local minimum. In practice, the accumulated histogram function $H(E, t)$ from all previously visited energies at the MC step t helps the simulation escape local entrapment and surpass high-energy barriers more easily.

3.2 Configuration update mechanism

In ELP, each MC step must update the current configuration. We used alternately the following gradient descent update mechanism and a rotary update mechanism to update configuration.

3.2.1 Gradient descent update mechanism

Consider the current configuration $X = (X_1, X_2,....,X_n)$, where $X_i = (x_i, y_i)$ denotes the position vector of the ith amino acid ($i = 1, 2, ..., n$). Since the length of the bonds is fixed ($|\boldsymbol{u}_i| = 1$, $i = 1, 2, ..., n-1$), the ith amino acid lies on a circle with radius of unit length around the $(i-1)$th amino acid. For updating the configuration X, we produced randomly a positive integer j ($2 \le j \le n$, where n denotes chain length) and changed temporarily the position of the amino acid j to that of the amino acid j* in the anti-gradient direction of the energy function $E(X)$ at the amino acid j, that is, $X_j* = X_j - \varepsilon \nabla E_j$, where X_j and X_j* denote the position vectors of the amino acids j and j*, respectively, $\varepsilon = 0.2$ is the iterative step length factor, and $-\nabla E_j = \left(-\frac{\partial E}{\partial x_j}, -\frac{\partial E}{\partial y_j}\right)$ is the iterative search direction. Link up the two points $j-1$ and j*, and the linked line intersects the circle at a point j'. Move amino acid j to the position of j' and amino acids $j+1$, $j+2$, ..., n are wholly translated as a rigid body with fixed trend of the partial chain (see Fig. 2). We thus obtained a new configuration N. Obviously, all the bond vectors in the configuration N are still unit length.

Produce a randomly positive integer k ($2 \le k \le n$). Based on the configuration N, we move the position of the amino acid k to new position k' and amino acids $k+1$, $k+2$, ..., n are wholly translated as a rigid body according to the above-mentioned similar method we obtain a new configuration X'. Thus, based on the configuration X, we obtain the new configuration X' after applying two gradient methods.

In gradient descent update mechanism, the search has a tendency to fall into traps of local minima. ELP must spend a great deal of time to jump out of the traps. To enhance the efficiency of ELP and be able to sample also very narrow and deep valleys of the energy landscape, we executed alternately the gradient descent update mechanism and the

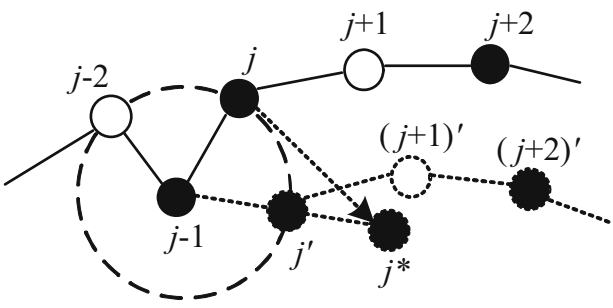

Fig. 2 Gradient descent update mechanism

following rotary update mechanism. After the simulation executed the gradient descent update mechanism 5,000 times, the rotary update mechanism was implemented 50 times.

3.2.2 Rotary update mechanism

The procedure of the rotary update mechanism is displayed in Fig. 3. To update the configuration X, we selected randomly a positive integer i ($2 \leq i \leq n$). Since the ith amino acid lies on the circumference with radius unity around the $(i-1)$th amino acid, we selected randomly point i' on circumference with maximum opening angle $2\theta_{\max}$ (the dark area in Fig. 3). In our simulations of the AB model, we used a very small opening angle, $\cos(\theta_{\max}) = 0.9$, in order to be able to sample very narrow and deep energy basins. The ith amino acid was moved to the point i' and the $(i+1)$th, $(i+2)$th, ..., nth amino acids were wholly translated as a rigid body with fixed trend of the partial chain. Thus an updated configuration X' was obtained.

3.3 Local search based on gradient descent

In ELP, the motivation for introducing the penalty term $H(E, t)$ is to escape local minima. But, there is a technical flaw: the collection of the histogram function $H(E, t)$ is performed by dividing the energy space into finite-size bins. Consider an attempted move in ELP which yields a new lower energy minimum that has never been visited before and happens to fall into the same bin containing other energies previously visited in the earlier steps. Undesirably, due to the penalty term $H(E, t)$, the likelihood of accepting this new energy minimum becomes small. As a result, the ELP minimization may miss this new lower energy state near local minima. For this, we proposed the following local search method.

We introduced the additional degrees-of-freedom of bond lengths so that all monomers can freely move near the local minima derived by the ELP minimization. Suppose that there exists a spring with the original length 1 between the centers of the ith and $(i+1)$th monomers ($i = 1, 2, ..., n-1$). According to Hooke's law, the elastic potential energy between two adjacent monomers is proportional to the square of the length of the spring transformation. So the elastic potential energy of the system can be indicated as follows:

$$E' = \sum_{i=1}^{n-1} k\left(r_{i,\,i+1} - 1\right)^2,$$

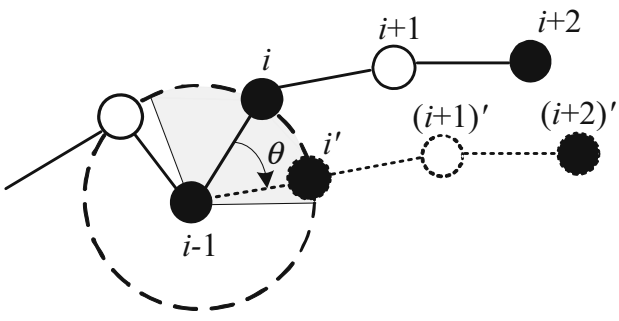

Fig. 3 Rotary update mechanism

where k is a physical coefficient characterizing the rigidities of all the springs. Thus, the sum of the potential energy is $U = E + E'$. Obviously, E' is the "penalty" term in the potential energy U. When the physical coefficient k is great enough, $r_{i,i+1} \to 1$ ($i = 1,2,...,n-1$) and the optimal solution of the unconstrained optimization problem min $U(X)$ is also the optimal solution of the optimization problem min $E(X)$ with the constraint

$$r_{i,i+1} = 1, i = 1,\ 2,\ ...,\ n-1. \tag{1}$$

For the unconstrained optimization problem min $U(X)$, we employed the gradient descent method to find low-energy configurations for a given monomer chain. In the beginning of the computation, we let $k \in [1, 10]$. Subsequently, we modified the physical coefficient k through multiplying by the factor 1.3 per 100,000 iterative steps. During the beginning of the calculation, the physical coefficient of all the springs was small, such that all monomers could move freely and easily attain low-energy states. Subsequently, as the calculation was carried out, the physical coefficient k increased gradually so as to increase the "penalty" and satisfy constraint (1) gradually, and finally the interaction of the springs disappeared. Obviously, when the physical coefficient rose to a large number (e.g., $\geq 10^{10}$), that is, when the springs turned rigid, constraint (1) was satisfied approximately, and a global energy minimum configuration was found, or a new trap of local minimum occurred, which had lower energy than the local minima derived by the ELP minimization.

4 Simulations

To search low-energy configurations for a given monomer chain, the ELP minimization was first executed. In our ELP simulations, the initial configuration was obtained through selecting the lowest-energy one from 10 randomly produced configurations. After the ELP was run up to 8×10^7 MC sweeps, the lowest-energy configuration was picked out from all energy states previously visited. By introducing the additional degree-of-freedom of bond lengths, the local search based on gradient descent was then used to search lower energy states near the lowest-energy state found by the ELP minimization. The gradient descent minimization was repeated up to five times. During the beginning of each round of gradient minimization, the physical coefficient was restored to initial value so that all monomers could freely move and easily attain low-energy states. After five rounds of gradient minimization, if the simulations did not find states with lower energy than the target value, we chose randomly another initial configuration for a new round of ELP simulation and subsequent local search based on gradient descent.

5 Results

In this article, we restricted ourselves to the 2D off-lattice AB model with Fibonacci sequences. The Fibonacci sequences are defined recursively by $S_0 = \mathrm{A}$, $S_1 = \mathrm{B}$, $S_{i+1} = S_{i-1} * S_i$. Here "*" is the concatenation operator. The first few sequences are $S_2 = \mathrm{AB}$, $S_3 = \mathrm{BAB}$, $S_4 = \mathrm{ABBAB}$, and so forth. They have the lengths given by $n_{i+1} = n_{i-1} + n_i$, i.e., given by the Fibonacci numbers. Following Refs. [6, 7, 9–13], in this article we considered the sequences with lengths $n = 13$, 21, 34 and 55 listed in Table 1.

For each sequence, the ELP minimization was run 50 times independently on a PIV 2.0 GHz computer. After each round of ELP simulation, the additional degree-of-freedom

Table 1 Test sequences and the lowest energies obtained by ELP+ in the 2D AB model, in comparison with those by MMC [7], PERM+ [10], ACMC [11], CSA [12], and STA [13], respectively

n	Sequence	$E_{\min}^{\mathrm{MMC}}$	$E_{\min}^{\mathrm{PERM+}}$	$E_{\min}^{\mathrm{ACMC}}$	$E_{\min}^{\mathrm{CSA}}$	$E_{\min}^{\mathrm{STA}}$	$E_{\min}^{\mathrm{ELP+}}$
13	ABBABBABABBAB	−3.2235	−3.2939	−3.2941	−3.2941	−3.2941	−3.2941
21	BABABBABABBABBABABBAB	−5.2881	−6.1976	−6.1979	−6.1980	−6.1980	−6.1980
34	ABBABBABABBABBABABB ABABBABBABABBAB	−8.9749	−10.7001	−10.8060	−10.8060	−10.8060	−10.7453
55	BABABBABABBABBABABB ABABBABBABABBABBAB ABBABABBABBABABBAB	−14.4089	−18.5154	−18.7407	−18.9110	−18.9202	−18.9301

of bond length was introduced, and the local search based on gradient descent was then used to search lower energy states near the lowest-energy state found by the ELP minimization. The calculated results showed that, along with the increment in the calculation steps in the AB model, all monomers ultimately tended to be stable and constraint (1) was satisfied

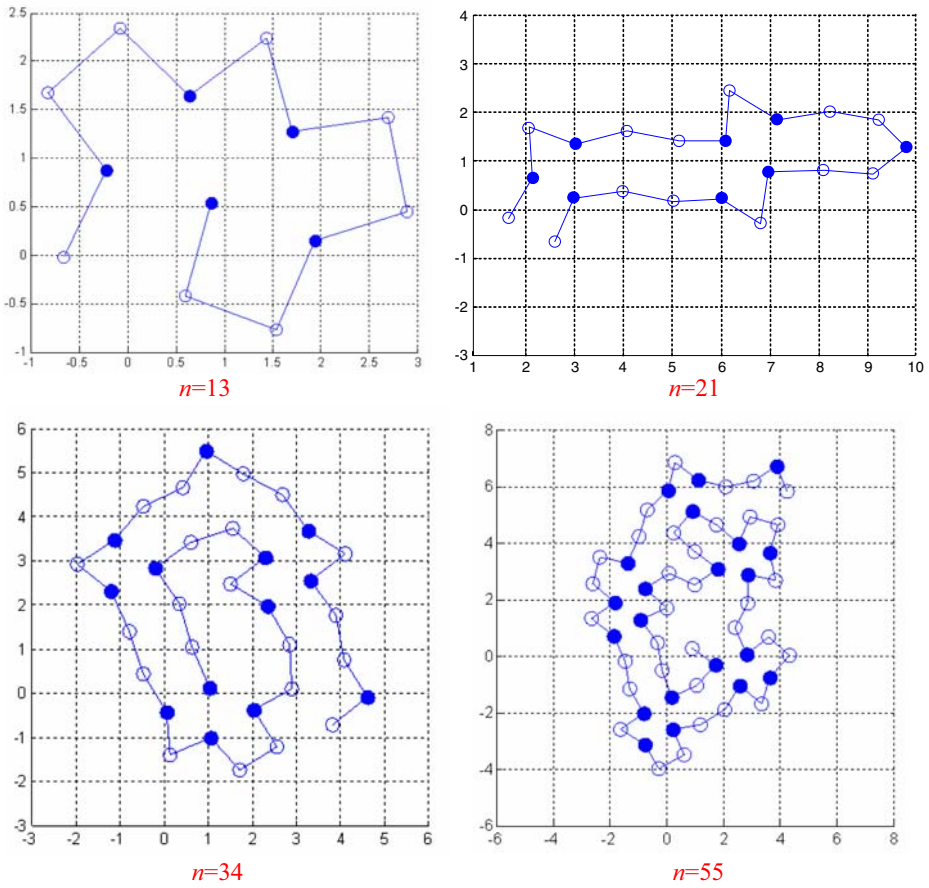

Fig. 4 The lowest-energy configurations for the four sequences listed in Table 1, found by applying the ELP+ algorithm to the 2D AB model

approximately. The error margin was smaller than 10^{-10}, i.e.,$\left|r_{i,i+1} - 1\right| < 10^{-10}, i = 1, 2, ..., n - 1$.

The lowest energies found by the ELP minimization and subsequent local searches based on gradient descent (ELP+) in 50 runs are listed in Table 1. We compare these results with minimum energies by conventional Metropolis-type Monte Carlo (MMC) procedures [7], the off-lattice pruned–enriched Rosenbluth method (PERM) and subsequent conjugate gradient minimization (PERM+) [10], the annealing contour Monte Carlo (ACMC) method [11], as well as conformational space annealing (CSA) [12] and statistical temperature annealing (STA) [13]. Table 1 shows that the results obtained by ELP+ are better than those of MMC and PERM+ for all four sequences. It can also be seen from Table 1 that ELP+ found slightly lower energy states for the 55-mer sequence than ACMC, CSA, and STA, while the putative ground-state energies found by ELP+ for 13-mer and 21-mer sequences are identical to those by ACMC, CSA, and STA. But ACMC, CSA and STA found lower energy states than ELP+ for the 34-mer sequence.

Figure 4 shows the lowest-energy configurations obtained by ELP+ for the four Fibonacci sequences listed in Table 1, where solid dots and empty circles indicate hydrophobic and hydrophilic monomers, respectively. The configurations for 13-mer and 21-mer sequences are identical to those by PERM+, ACMC, CSA, and STA, as can be expected from their almost identical ground-state energies in Table 1. This implies that all five methods find ground-state configurations reasonably well for these short chains. For 34-mer and 55-mer sequences, on the other hand, the lowest-energy configurations by ELP+ are different from those by PERM+, ACMC, CSA, and STA, as the energy difference between them is obvious.

6 Summary and anticipation

In this article, we introduced a global optimization method, energy landscape paving (ELP). We used ELP and subsequent local search (ELP+) to optimize low-energy configurations of a 2D off-lattice AB protein model. The numerical results showed that the proposed methods are very promising for finding the ground states of proteins. Of course, the chosen model still has some factors that are not very realistic. But this does not affect the fact that the AB model is suitable for benchmarking for the protein structure prediction problem.

Although ELP+ has powerful global optimization performance, it is still easy to get into traps of local minima as a result of random sampling. In future work, we hope to improve the sampling method in ELP, and apply it to all-atom models with realistic potentials to design various kinds of more high performance algorithms for the protein structure prediction problem.

Acknowledgements This work was supported by Foundation of Nanjing University of Information Science and Technology and Qing Lan Project.

References

1. Shakhnovich, E.I.: Proteins with selected sequences fold into unique native conformation. Phys. Rev. Lett. **72**(24), 3907–3911 (1994). doi:10.1103/PhysRevLett.72.3907
2. Camacho, C.J., Thirumalai, D.: Kinetics and thermodynamics of folding in model proteins. Proc. Natl. Acad. Sci. U. S. A. **90**(13), 6369–6372 (1993). doi:10.1073/pnas.90.13.6369

3. Dill, K.A., Bromberg, S., Yue, K., Fiebig, K.M., Yee, D.P., Thomas, P.D., Chan, H.S.: Principles of protein folding: a perspective from simple exact models. Protein Sci. **4**(4), 561–602 (1995)
4. Honeycutt, J.D., Thirumalai, D.: The nature of folded states of globular proteins. Biopolymers **32**(6), 695–709 (1992). doi:10.1002/bip.360320610
5. Fukugita, M., Lancaster, D., Mitchard, M.G.: Kinematics and thermodynamics of a folding heteropolymer. Proc. Natl. Acad. Sci. U.S.A. **90**(13), 6365–6368 (1993). doi:10.1073/pnas.90.13.6365
6. Stillinger, F.H., Head-Gordon, T., Hirshfeld, C.L.: Toy model for protein folding. Phys. Rev. E **48**(2), 1469–1477 (1993). doi:10.1103/PhysRevE.48.1469
7. Stillinger, F.H., Head-Gordon, T.: Collective aspects of protein folding illustrated by a toy model. Phys. Rev. E **52**(32), 2872–2877 (1995). doi:10.1103/PhysRevE.52.2872
8. Anfinsen, C.B.: Principles that govern the folding of protein chains. Science **181**(96), 223–230 (1973). doi:10.1126/science.181.4096.223
9. Yue, X.H., Tang, H.W., Guo, C.H.: A Tabu search and its application in 2D HP off-lattice model. Comput. Appl. Chem. **22**(12). 1101–1105 (2005)
10. Hsu, H.P., Mehra, V., Grassberger, P.: Structure optimization in an off-lattice protein model. Phys. Rev. E **68**(3), 037703 (2003). doi:10.1103/PhysRevE.68.037703
11. Liang, F.: Annealing contour Monte Carlo algorithm for structure optimization in an off-lattice protein model. J. Chem. Phys. **120**(14), 6756–6763 (2004). doi:10.1063/1.1665529
12. Kim, S., Lee, S.B., Lee, J.: Structure optimization by conformational space annealing in an off-lattice protein model. Phys. Rev. E **72**(1), 011916 (2005). doi:10.1103/PhysRevE.72.011916
13. Kim, J., Straub, J.E.: Structure optimization and folding mechanism of off-lattice protein models using statistical temperature molecular dynamics simulation: statistical temperature annealing. Phys. Rev. E **76**(1), 011913 (2007). doi:10.1103/PhysRevE.76.011913
14. Hansmann, U.H.E., Wille, L.T.: Global optimization by energy landscape paving. Phys. Rev. Lett. **88**(6), 068105 (2002). doi:10.1103/PhysRevLett.88.068105
15. Schug, A., Wenzel, W., Hansmann, U.H.E.: Energy landscape paving simulations of the trp-cage protein. J. Chem. Phys. **122**(19), 194711 (2005). doi:10.1063/1.1899149
16. Liu, J.F., Huang, W.Q.: Studies of finding low energy configurations in off-lattice protein models. J. Theor. Comput. Chem. **5**(3), 587–594 (2006). doi:10.1142/S0219633606002453
17. Liu, J.F., Huang, W.Q.: Quasi-physical algorithm of an off-lattice model for protein folding problem. J. Comput. Sci. Technol. **22**(4), 574–597 (2007). doi:10.1007/s11390-007-9067-x
18. Besold, G., Risbo, J., Mouritsen, O.G.: Efficient Monte Carlo sampling by direct flattening of free energy barriers. Comput. Mater. Sci. **15**(3), 311–340 (1999). doi:10.1016/S0927-0256(99)00023-3
19. Cvijovic, D., Klinowski, J.: Taboo search: an approach to the multiple minima problem. Science **267**(3), 664–666 (1995). doi:10.1126/science.267.5198.664