Energy optimization for off-lattice protein folding

Wenqi Huang, Mao Chen,* and Zhipeng Lü
School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China
(Received 15 May 2006; published 9 October 2006)

Two three-dimensional *AB* off-lattice protein models consisting of hydrophobic and hydrophilic monomers are studied in this paper. By incorporating an extra energy contribution into the original energy function, the protein folding is converted from a constraint optimization problem into an unconstrained one which can be solved by the well-known gradient method. From the initial configurations randomly generated by the heuristic strategy proposed in this paper, our algorithm can find better results than those by nPERM for the four Fibonacci sequences. Based on the initial configurations obtained by energy landscape paving (ELP) routine, some of our results for the lowest energies are better than the best values reported in the literature.

DOI: 10.1103/PhysRevE.74.041907 PACS number(s): 87.15.By, 87.10.+e, 05.10.−a

I. INTRODUCTION

Prediction of three-dimensional (3D) protein structures is one of the central problems in computational biology. In despite of the great efforts made by the scientists in past decades, it remains an amazingly difficult computational problem not only because the folded 3D structure of a protein is extremely complicated, but also because a general prediction strategy must not depend on the foreknowledge of any specific structural information about the proteins whose structures are to be predicted [1].

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 [2,3], 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 protein is confined as a self-avoiding path on a regular square or cubic lattice, with attractive interactions only between neighboring nonbonded *H* monomers.

Being the most simplified and most popular model, the *HP* model only considers the interactions between neighboring nonbonded *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 *et al.* [4,5] proposed and studied a more realistic simplified model, namely, the *AB* off-lattice model. With some substantial modifications, a new version of the *AB* model is introduced in Ref. [6].

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. Consequently, various heuristic schemes have been proposed for approaching this problem.

For the 2D *AB* model, neural networks [4], simulated tempering [7], Monte Carlo [8,9], molecular dynamics [10], and biologically motivated methods [11,12] were used to find the native state. An improved pruned enriched Rosenbluth method with importance sampling [13], namely, nPERM [14], was proposed by Hsu *et al.*, which found states with lower energy than previously proposed putative ground states for the four Fibonacci sequences [5]. Without modifying the energy function, Hsu *et al.* extended the 2D *AB* model to 3D version and presented some putative lowest energy states for the four sequences. Although the resulting conformation 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 conformations with single hydrophobic cores. Recently, better results in three dimensions for the four sequences were achieved by means of the annealing contour Monte Carlo (ACMC) algorithm [6], energy landscape paving (ELP) minimizer [15,16], and conformational space annealing (CSA) method [17–20].

However, it is still not clear whether the reported ground-state conformations for these sequences are indeed the global minima in the complicated energy landscape that is characterized by a multitude of local minima separated by high-energy barriers [21,22]. In this paper, we propose a quite different class of heuristic algorithm for predicting the native structure for the two *AB* off-lattice models in three dimensions.

II. MODELS

The *AB* model also uses two types of monomers, now called "*A*" (hydrophobic) and "*B*" (hydrophilic). The distances between successive monomers along the chain are assumed to be fixed ($r_{i,i+1}=1$). In the literature, there are two different *AB* models in three dimensions.

The first model, denoted by model I, is the original *AB* model proposed in Ref. [4]. In this model, 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. The energy function for a N monomer chain is expressed as

*E-mail address: mchen_1@163.com

$$U_{\mathrm{I}} = \frac{1}{4}\sum_{i=2}^{N-1}(1-\cos\theta_i) + 4\sum_{i=1}^{N-2}\sum_{j=i+2}^{N}[r_{ij}^{-12} - C_1(\sigma_i,\sigma_j)r_{ij}^{-6}], \tag{1}$$

where r_{ij} is the distance between monomer i and j. Each σ_i is either A or B, and $C_1(\sigma_i,\sigma_j)$ is +1 for an AA pair, +1/2 for a BB pair and −1/2 for an AB pair, 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.

The other model, denoted by AB model II, was introduced in Ref. [6]. It is a variant of AB model I in that it not only involves bond angle energy and Lennard-Jones potential, but also considers a new energy contribution, the torsional energy. The energy function is given by

$$U_{\mathrm{II}} = \sum_{i=1}^{N-2}\mathbf{u}_i\cdot\mathbf{u}_{i+1} - \frac{1}{2}\sum_{i=1}^{N-3}\mathbf{u}_i\cdot\mathbf{u}_{i+2} + 4\sum_{i=1}^{N-2}\sum_{j=i+2}^{N}C_2(\sigma_i,\sigma_j)(r_{ij}^{-12} - r_{ij}^{-6}), \tag{2}$$

where $\mathbf{u}_i$ is the bond vector between monomers i and $i+1$ with unit length, $C_2(\sigma_i,\sigma_j)$ is +1 for the AA pair, and +1/2 for the BB and AB pair. The first term in Eq. (2) is the part of the energy contribution of bending energy from successive bonds. The second term takes the torsional energy into account.

III. METHODS

A. Problem description

Consider an amino acid sequence as a chain of black balls (A) and white balls (B), numbered from 1 to N. Denote the coordinates of the center of the ith ($i=1,2,\ldots,N$) ball by (x_i,y_i,z_i). At any moment, the positions of the N balls are called a configuration, denoted by $(x_1,y_1,z_1,\ldots,x_N,y_N,z_N)$.

Now, the protein folding problem can be described as follows:

$$\min(U_i), \tag{3}$$

subject to

$$\sqrt{(x_i-x_{i+1})^2+(y_i-y_{i+1})^2+(z_i-z_{i+1})^2} = 1, \quad i=1,2,\ldots,\quad N-1. \tag{4}$$

Note that U_i is U_{I} (for model I) or U_{II} (for model II). Equation (4) ensures that the distances between two consecutive balls along the chain are equal to 1. A configuration that satisfies constraint Eq. (4) is defined as a legal configuration. Equations (3) and (4) form a specific type of nonlinear constraint-satisfaction problem [23,24]. In spite of its simplicity, it is rather difficult to solve this kind of problem directly due to the loss of smoothness in the solution space. Therefore, we propose a scheme to convert this constraint optimization problem into an unconstrained one which is smooth in the solution space.

B. The new mathematical description

Instead of fixing the links between two successive balls, we assume that all the N balls are sequentially linked by springs with natural length to be 1. The spring has the tendency to return to its natural length after being compressed or stretched. We will show how springs can be used to relax the requirement on the solvability of the original constraint optimization problem.

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

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

If $l_{i,i+1}>1$, the spring is stretched. If $l_{i,i+1}<1$, the spring is compressed. According to Hook's law, the elastic potential energy of a spring is

$$u_{i,i+1} = \frac{1}{2}K_s(l_{i,i+1}-1)^2, \quad i=1,2,\ldots,N-1, \tag{6}$$

where $\boldsymbol{K}_s$ is the spring coefficient, $\boldsymbol{K}_s>0$. Then the total spring potential energy of given configuration is

$$U_s = \sum_{i=1}^{N-1}u_{i,i+1} = \frac{1}{2}K_s\sum_{i=1}^{N-1}(l_{i,i+1}-1)^2. \tag{7}$$

Now, apart from the energy contributions from bond angles, Lennard-Jones, and torsion angles (for model II only), the elastic potential energy of springs is also a part of the total potential energy of the configuration. It can be seen from Eqs. (1) and (2) to Eqs. (5)–(7) that the total potential energy U is a known function of the configuration with $3N$ independent variables

$$U = U(x_1,y_1,z_1,\ldots,x_N,y_N,z_N), \tag{8}$$

where $U(x_i,y_1,z_1,\ldots,x_N,y_N,z_N)$ is defined on the entire $3N$-dimensional Euclidean space $(-\infty,+\infty)^{3N}$, smooth, continuous, and differentiable everywhere. The aim is to find a configuration $(x_1^*,y_1^*,z_1^*,\ldots,x_N^*,y_N^*,z_N^*)$ with minimum energy,

$$U(x_1^*,y_1^*,z_1^*,\ldots,x_N^*,y_N^*,z_N^*) = \min(U). \tag{9}$$

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 [25].

Equations (6) and (7) show that the elastic energy is nonnegative. According to Eq. (8), if the coefficient $\boldsymbol{K}_s$ is set to be large enough, a spring with length differing slightly from its natural length can considerably increase the whole energy of the configuration. Therefore, 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.

C. Strategy of generating promising initial configuration

In the course of solution using the gradient method, the initial configuration $(x_1,y_1,z_1,\ldots,x_N,y_N,z_N)$ will be steadily

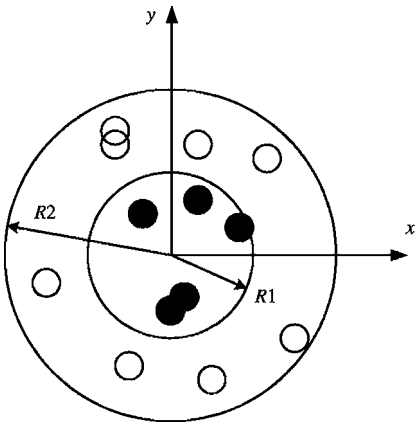

FIG. 1. An initial configuration of 13 balls generated by the strategy of generating promising initial configuration.

evolving. This process will continue until the calculation falls into the trap of local minimum. Then a new initial configuration is randomly generated for another round of calculation. Although it is possible to obtain satisfying results from such calculations in principle, the efficiency is rather low as seen from actual calculations. Thus, promising initial configurations are 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 good heuristic strategy to generate promising initial configurations that simulate 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 $R_2=2R_1$. We set $R_1=\sqrt{N}$ in our algorithm. 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. 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:

$$\sqrt{x_i^2+y_i^2+z_i^2} \leqslant R_1 \tag{10}$$

$$R_1 < \sqrt{x_j^2+y_j^2+z_j^2} \leqslant R_2, \tag{11}$$

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

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.

IV. RESULTS AND DISCUSSIONS

A. Results from the randomly generated initial configurations

Table I shows the lowest energies for all the four Fibonacci sequences for model I obtained by our heuristic algorithm (HA), along with the results by nPERM [14], ELP [16], and CSA [6]. Our results are better than those of the nPERM for all cases, with the energy difference increasing gradually for longer chains. As for sequence with length 13, our result is also slightly better than that of ELP and is equal to that of CSA. Though we cannot reach the energy yielded by ELP and CSA for the other three sequences, our result of sequences $N=21$ and 34 are very close to those by ELP and CSA within less than 0.5 and 1.5%, respectively. For the longest chain with $N=55$; however, our result is higher by about 5%.

It should be pointed out that each of the results is the best one of the solutions iterated from several ($N \leqslant 50$) initial configurations randomly generated by our above-mentioned heuristic strategy. It is likely to find conformations with even lower energy if we try more runs for each sequence, but the computational time is too long to be practical. The runtime of each run is about 38, 75, 347, and 865 s for the four sequences, respectively, on a P4 2.4 GHz PC with 512 MB memory, while the computation time of nPERM was up to 2 days on Linux and UNIX workstation [14]. The runtime of ELP and CSA was not reported in the literature.

B. Results from the initial configurations obtained by ELP

The intensive calculations showed that the initial configuration is crucial for the energy of the resulting conformation. Though the above-mentioned heuristic strategy can generate better initial configurations than those randomly generated in the 3D space, the results are not good enough. In this section, we adopt the lowest-energy conformations obtained by ELP

TABLE I. The four Fibonacci sequences and the lowest energies of model I by our heuristic algorithm (HA) based on the randomly generated initial configurations, in comparison with those by nPERM, CSA, and ELP, respectively.

N	Sequence	nPERM	ELP	CSA	HA[a]
13	*ABBABBABABBAB*	−4.9616	−4.967	−4.9746	−4.9746
21	*BABABBABABBABBABABBAB*	−11.5238	−12.316	−12.3266	−12.2617
34	*ABBABBABABBABBABABBABABBABBA BABBAB*	−21.5678	−25.476	−25.5113	−24.2441
55	*BABABBABABBABBABABBABABBABBA BABBABBABABBABABBABBABABBAB*	−32.8843	−42.428	−42.3418	−40.3177

[a]The lowest energy by HA among the results of 20 runs for sequences $N=13$ and 21, and 50 runs for sequences $N=34$ and 55.

TABLE II. The lowest energies of model I and model II by HA based on the initial configurations that are the resulting minimum-energy conformations by ELP. The results for model I are compared with those by nPERM, ELP, and CSA, respectively. Lowest energies for model II are compared with those by ACMC, ELP, and CSA, respectively.

	Model I				Model II			
N	nPERM	ELP	CSA	HA	ACMC	ELP	CSA	HA
13	−4.9616	−4.967	−4.9746	−4.9746	−26.5066	−26.498	−26.4714	−26.5068
21	−11.5238	−12.316	−12.3266	−12.3266	−51.7575	−52.917	−52.7865	−52.9329
34	−21.5678	−25.476	−25.5113	−25.5113	−94.0431	−97.261	−97.7321	−97.4230
55	−32.8843	−42.428	−42.3418	−42.5181	−154.5050	−172.696	−173.9803	−173.3246

as the initial configurations. From these initial configurations (see http://www.physik.uni-leipzig.de/~bachmann/papers/suppl16/suppl16.html), our algorithm can find new lowest energies for the four Fibonacci sequences and the six sequences proposed in Ref. [16].

Table II shows the lowest energies of model I and model II in three dimensions for the four Fibonacci sequences obtained by our algorithm, along with the results by nPERM, ACMC [6], ELP, and CSA. Obviously, our results for model I are better than those of the nPERM and ELP for all cases, with the energy difference increasing gradually for longer chains. Our results are better by 0.15%, 0.09%, 0.14%, and 0.2% than those by ELP for the four sequences, respectively, and the average improvement is 0.15%. Our results for the three short chains are identical to the results by CSA but slightly better for the longest chain with $N=55$. Results for model II are also better than those by ACMC and ELP for all cases. For model II, our results are better by 0.03%, 0.03%, 0.17%, and 0.36% than those by ELP for the four sequences, respectively, and the average improvement is also 0.15%. Compared with results by CSA, ours are better for the short chains with $N=13$ and 21; however, we cannot reach the energy by CSA for the two long chains.

Since our algorithm found new candidates for global minima with lower energies for most cases (only the results of sequences $N=34$ and 55 of model II by CSA are better than ours), it indicates that the results obtained by ELP, ACMC, and CSA may not correspond to ground-state energies.

Since our calculations are based on deterministic initial configurations, we run our algorithm only once for each case. It should be pointed out that since K_s is rather large ($K_s > 10^6$), the resulting configurations approximately satisfy the constraint Eq. (4). In other words, the distance between any two successive monomers meet the following requirement:

$$|r_{i,i+1} - 1| < \delta, \quad i = 1,2,\ldots,N-1, \tag{12}$$

where δ is named solution quality. δ is 10^{-5} for our results, whereas δ is 2×10^{-4} for the results by ELP.

Figure 2 displays the lowest-energy conformations of model I obtained by our algorithm, where black circles denote hydrophobic monomers (A) and white circles denote hydrophilic monomers (B). It is clear that each conformation has a single hydrophobic core, as observed in real proteins. Figure 3 shows the lowest-energy conformations of model II. Similarly, all sequences fold into conformations with single hydrophobic cores. From Fig. 2 and Fig. 3, it can be seen that the conformations of model II are more compact than

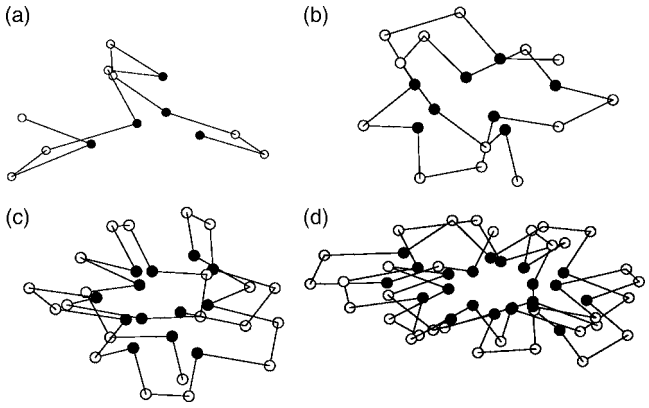

FIG. 2. The lowest-energy conformations of AB model I found with our algorithm for the four Fibonacci sequences: (a) $N=13$, (b) $N=21$, (c) $N=34$, and (d) $N=55$. The black circles represent hydrophobic (A) monomers and the white ones hydrophilic (B) monomers.

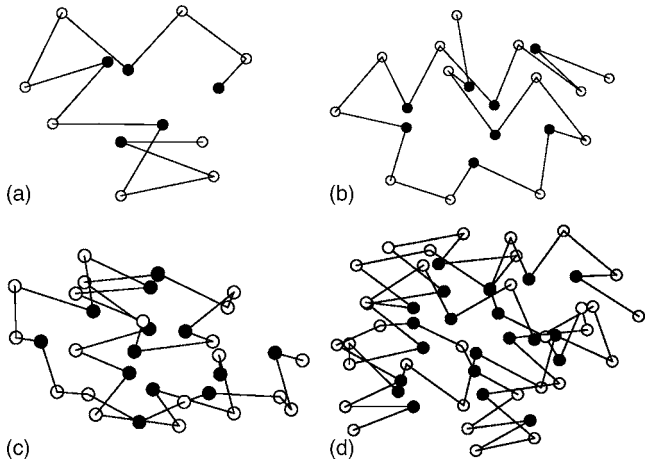

FIG. 3. The lowest-energy conformations of AB model II found with our algorithm for the four Fibonacci sequences: (a) $N=13$, (b) $N=21$, (c) $N=34$, and (d) $N=55$. The black circles represent hydrophobic (A) monomers and the white ones hydrophilic (B) monomers.

TABLE III. The lowest energies of both models obtained by HA for the six sequences introduced in Ref. [16], in comparison with those by ELP. Similarly, the initial configurations for HA are the lowest-energy conformations obtained by ELP.

		Model I		Model II	
No.	Sequence	ELP	HA	ELP	HA
20.1	*BAAAAAABAAAABAABAABB*	−33.810	−33.8429	−58.317	−58.3226
20.2	*BAABAAAABABAABAAAAAB*	−33.926	−33.9445	−58.914	−58.9174
20.3	*AAAABBAAAABAABAAABBA*	−33.578	−33.6093	−59.338	−59.3462
20.4	*AAAABAABABAABBAAABAA*	−34.498	−34.5261	−59.079	−59.0887
20.5	*BAABBAAABBBABABAABAB*	−19.653	−19.6614	−51.566	−51.5682
20.6	*AAABBABBABABBABABABA*	−19.326	−19.3469	−53.417	−53.4217

those of model I for the same sequences. As pointed out in Ref. [17], it is mainly due to the attractive long-range interactions among *AB* monomers.

We also test our algorithm for the six sequences used in the study of thermodynamic properties of heteropolymers introduced in Ref. [16]. Similarly, the initial configurations are also the lowest-energy conformations obtained by ELP. For each sequence, the heuristic algorithm runs for one time, and the CPU time of each run is within 80 s. The sequences and the corresponding ground energies by ELP are listed in Table III, together with the renewed ground energies obtained by our heuristic algorithm.

For each sequence, our result is better than that of ELP for both models. For model I, the average improvement for the six sequences is about 0.076%. (Our result is lower by 0.11% than that of ELP for sequence 20.6, while the result for sequence 20.5 is just better by 0.031%.) For model II, the average improvement for the six sequences is only about 0.0098%. This implies that the results of the six sequences obtained by ELP may only correspond to metastable local minima.

V. SUMMARY

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 are mapped onto the final inherent structures by a numerical steepest descent on the potential energy surface. 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 particles to a legal configuration with low potential energy.

Since the gradient method is only a local search algorithm, it is possible to fall into the trap of local minimum. Selecting the best one from many solutions iterated from initial configurations generated by our heuristic strategy may help to find a comparatively good solution, but the computational time is rather long. From much better initial configurations, i.e., lowest-energy conformations obtained by ELP, our method can find conformations of lower energy than the previous results in the literature for most cases.

From our work, it can be seen that the combination of several methods of energy optimization is helpful for protein folding. In our future work, we hope to find some efficient strategy of jumping out of local minimum to develop a more efficient algorithm.

ACKNOWLEDGMENTS

We thank Frank H. Stillinger and Veit Elser for helpful discussions on the understanding of the AB models, and thank Michael Bachmann for giving the coordinates of the minimum-energy conformations from the ELP method. This work is supported by the National Natural Science Foundation of China (Grant No. 10471051) and partially supported by the NKBRPC (Grant No. 2004CB318000).

[1] W. Huang, Z. Lü, and H. Shi, Phys. Rev. E **72**, 016704 (2005).
[2] K. A. Dill, Biochemistry **24**, 1501 (1985).
[3] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S Chan, Protein Sci. **4**, 561 (1995).
[4] F. H. Stillinger, T. Head-Gordon, and C. L. Hirshfeld, Phys. Rev. E **48**, 1469 (1993); T. Head-Gordon and F. H. Stillinger, *ibid.* **48**, 1502 (1993).
[5] F. H. Stillinger and T. Head-Gordon, Phys. Rev. E **52**, 2872 (1995).
[6] F. Liang, J. Chem. Phys. **120**, 6756 (2004).
[7] A. Irbäck, C. Peterson, F. Potthast, and O. Sommelius, J. Chem. Phys. **107**, 273 (1997).
[8] A. Irbäck, C. Peterson, and F. Potthast, Phys. Rev. E **55**, 860 (1997).
[9] A. Irbäck and F. Potthast, J. Chem. Phys. **103**, 10298 (1995).
[10] A. Torcini, R. Livi, and A. Politi, J. Biol. Phys. **27**, 181 (2001).
[11] D. Gorse, Biopolymers **59**, 411 (2001).

[12] D. Gorse, Biopolymers **64**, 146 (2002).
[13] H.-P. Hsu, V. Mehra, W. Nadler, and P. Grassberger, J. Chem. Phys. **118**, 444 (2003); Phys. Rev. E **68**, 021113 (2003).
[14] H.-P. Hsu, V. Mehra, and P. Grassberger, Phys. Rev. E **68**, 037703 (2003).
[15] U. H. E. Hansmann and L. T. Wille, Phys. Rev. Lett. **88**, 068105 (2002).
[16] M. Bachmann, H. Arkin, and W. Janke, Phys. Rev. E **71**, 031906 (2005).
[17] S.-Y. Kim, S. B. Lee, and J. Lee, Phys. Rev. E **72**, 011916 (2005).
[18] J. Lee, H. A. Scheraga, and S. Rackovsky, J. Comput. Phys. **18**, 1222 (1997).
[19] J. Lee, I.-H. Lee, and J. Lee, Phys. Rev. Lett. **91**, 080201 (2003).
[20] J. Lee, S.-Y. Kim, K. Joo, I. Kim, and J. Lee, Proteins **56**, 704 (2004).
[21] P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **89**, 8721 (1992).
[22] J. D. Bryngelson, J. N. Onuchic, N. D Socci, and P. G Wolynes, Proteins: Struct., Funct., Genet. **21**, 167 (1995).
[23] H. Wang, W. Huang, and Q. Zhang, Eur. J. Oper. Res. **141**, 440 (2002).
[24] W. Huang and R. Xu, Sci. China, Ser. E: Technol. Sci. **42**, 441 (2002).
[25] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vettering, *Numerical Recipes* (Cambridge University Press, New York, 1986).