Determination of the structure of the energy landscape for coarse-grained off-lattice models of folding heteropolymers

Handan Arkin*
Faculty of Engineering, Department of Physics Engineering Tandoğan, Ankara University, Ankara, Turkey
(Received 8 July 2008; revised manuscript received 21 August 2008; published 22 October 2008)

The structure of energy landscapes for a minimalist coarse-grained off-lattice protein model is presented to investigate the folding behaviors of heteropolymers. The obtained energy landscapes serve as a useful tool for visualization of the funnel-like structure of a considered system in the configuration space. Despite the simplicity of the model, the knowledge of the free-energy landscape enables us to show different folding characteristics known from real proteins and synthetic peptides, such as two-state folding and metastability.

DOI: 10.1103/PhysRevE.78.041914 PACS number(s): 87.15.Cc, 05.10.−a, 87.15.A−

I. INTRODUCTION

Prediction of three-dimensional protein structures and the folding mechanism is one of the most challenging objectives in modern interdisciplinary science. Despite great efforts made by scientists, the complexity of proteins as being macromolecules consisting of numerous atoms, the influence of quantum chemical details on long-range interactions make an accurate analysis of the folding process of realistic proteins extremely difficult. The difficulty is especially coming from the folding process which is so slow (microsecond to seconds) that molecular dynamics simulations of the whole folding trajectory are currenty still impossible for realistic protein models. This is due to a complex, rugged shape of the free-energy landscape [1].

The configuration space of proteins presents a complex energy profile consisting of a tremendous number of local minima, barriers, and further topological features. The topography of the energy landscape, especially near the global minimum is of particular importance; this is because the potential energy surface defines the behavior of the system. Because of energy barriers, studies of folding kinetics and thermodynamics are difficult with the commonly used thermodynamic simulation techniques. These techniques are not efficient in sampling a rugged landscape. Therefore, one of the important aspects in this field is studying simple, effective, coarse-grained models [2]. A number of such models were designed, some of which only incorporate two types of aminoacids: Hydrophobic and polar residues, e.g., HP or *AB* models. These models allow the understanding of tertiary heteropolymer folding (for example, hydrophobic-core formation [3–5])—which it is believed that tertiary folding is mainly based on hydrophobic interactions such that atomic details have a minor role.

Another known class of models designated for studies of the folding process are Gō models, where in the simplest lattice formulation, the energy is proportional to the number of native contacts. These models usually require knowledge of the native conformation and serve, e.g., as models for studies of folding pathways [6,7] and native topology [8,9]. There exist also a growing interest in modeling the folding behavior of single-domain proteins with simplified models, while many of them show a simple two-state kinetics [10] without intermediary states that would slow down the folding dynamics ("traps"). It seems, however, that the native-centric models such as those of Gō type require modifications for a qualitatively correct description of sharp folding transitions [11,12]. On the other hand, the kinetics of these minimal models is not biased to a given structure and therefore different types of folding behaviors may be observed.

In this paper, results obtained by the standard hydrophobic-polar off-lattice *AB* model [13] in three dimensions are presented. A big advantage of these more physical models than Gō models is that a different kind of folding types can be studied [14], because the kinetics is not guided towards a certain structure. This give the ability to see glassy folding into metastability as well as two-state folding. The objective of this work is to show that with a simple, minimal model it is possible to see clear indications of different folding characteristics.

II. MODEL

A manifest off-lattice generalization of the hydrophobic-polar (HP) model [15] is the *AB* model [13], where the hydrophobic monomers are labeled by *A* and the polar or hydrophilic ones by *B*. The contact interaction is replaced by a distance-dependent Lennard-Jones type of potential accounting for short-range excluded volume repulsion and long-

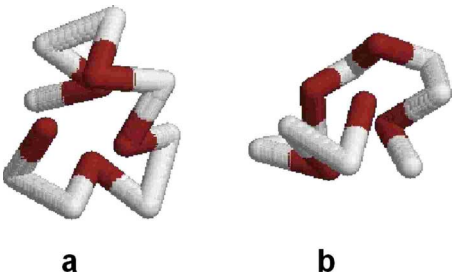

FIG. 1. (Color online) (a) The lowest-energy conformation of the 13-mer sequence (the reference state of the overlap parameter) which is in the nativefold funnel. (b) The second-lowest-energy conformation of 13-mer sequence which is a metastable conformation.

*handan.olgar@eng.ankara.edu.tr

TABLE I. The three *AB* sequences: 13-mer, 21-mer, and 34-mer and the values of the putative global energy minimums. For 13-mer and 21-mer note that there are two degenerate states (metastable) which are rendered in Fig. 1 and in Fig. 2.

Label	Sequence	Global minimum energy
13-mer	*ABBABBABABBAB*	−26.39, −26.37
21-mer	*BABABBABABBABBABABBAB*	−52.67, −52.54
34-mer	*ABBABBABABBABBABABBABABBABBABABBAB*	−97.32

range interaction, the latter being attractive for *AA* and *BB* pairs and repulsive for *AB* pairs of monomers. An additional interaction accounts for the bending energy of any pair of successive bonds. This model was first applied in two dimensions [13] and generalized to three-dimensional *AB* proteins [16,17], partially with modifications taking implicitly into account additional torsional energy contributions of each bond.

In this study, an effective off-lattice model of *AB* type for heteropolymers with N monomers was employed. The *AB* model, as proposed in Ref. [16] has the energy function

$$E = -\kappa_1 \sum_{k=1}^{N-2} \mathbf{b}_k \cdot \mathbf{b}_{k+1} - \kappa_2 \sum_{k=1}^{N-3} \mathbf{b}_k \cdot \mathbf{b}_{k+2} + 4 \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} C(\sigma_i, \sigma_j) \left(\frac{1}{r_{ij}^{12}} - \frac{1}{r_{ij}^{6}} \right), \quad (1)$$

where $\mathbf{b}_k$ is the bond vector between the monomers k and $k+1$ with length unity. In Ref. [16] different values for the parameter set (κ_1, κ_2) were tested and finally set to $(-1, 0.5)$ as this choice led to distributions for the angles between bond vectors $\mathbf{b}_k$ and $\mathbf{b}_{k+1}$ as well as the torsion angles between the surface vectors $\mathbf{b}_k \times \mathbf{b}_{k+1}$ and $\mathbf{b}_{k+1} \times \mathbf{b}_{k+2}$ that agreed best with distributions obtained for selected functional proteins. Since $\mathbf{b}_k \cdot \mathbf{b}_{k+1} = \cos \vartheta_k$, the choice $\kappa_1 = -1$ makes the coupling between successive bonds "antiferromagnetic" or "antibending." The second term in Eq. (1) takes torsional interactions into account without being an energy associated with the pure torsional barriers in the usual sense. The third term contains now a pure Lennard-Jones potential, where the $1/r_{ij}^6$ long-range interaction is attractive whatever types of monomers interact. The monomer-specific prefactor $C(\sigma_i, \sigma_j)$ only controls the depth of the Lennard-Jones valley,

$$C(\sigma_i, \sigma_j) = \begin{cases} +1, & \sigma_i, \sigma_j = A, \\ +1/2, & \sigma_i, \sigma_j = B \text{ or } \sigma_i \neq \sigma_j. \end{cases} \quad (2)$$

For technical reasons, a cutoff $r_{ij} = 0.5$ was introduced for the Lennard-Jones potentials below which the potential is hard-core repulsive (i.e., the potential is infinite).

Simulations of this model were performed with multicanonical algorithm [18] and the update mechanism is a spherical update which is described in Ref. [5] in detail.

III. RESULTS AND DISCUSSIONS

After calculating the multicanonical weights, 2×10^7 iterations were performed for each sequence in the production run. The sequences and the global minimum energy values of each sequence examined in this study are listed in Table I.

In the literature, many recent papers [19–22] give only the minimum energy values of these sequences, but energy profiles and folding behaviors to a native state are not given. In this study, I show that by employing a simple coarse-grained hydrophobic-polar model and monitoring a simple "order" parameter, it is indeed possible to identify different complex folding characteristics such as two-state folding, metastability, etc., known from real proteins. In order to investigate the folding behavior, I calculate here the overlap parameter (Q), which is as an extension of the torsion-angle based variant [23,24], defined as

$$Q(\mathbf{X}, \mathbf{X}') = \frac{N_t + N_b - d(\mathbf{X}, \mathbf{X}')}{N_t + N_b}, \quad (3)$$

where (with $N_t = N-3$ and $N_b = N-2$ being the numbers of torsional angles Φ_i and bond angles $\Theta_i = \pi - \vartheta_i$, respectively)

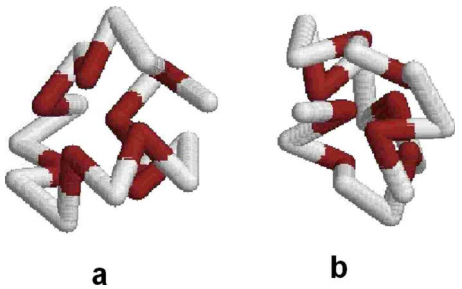

FIG. 2. (Color online) Same as Fig. 1 for the 21-mer sequence. (a) is the lowest-energy conformation of the 21-mer sequence which is used as a reference state in the overlap parameter formula and (b) is the second-lowest-energy conformation which is more probable in the energy landscape.

FIG. 3. (Color online) The lowest-energy conformation of the 34-mer sequence which is the native state and is also used in the overlap parameter as a reference state.

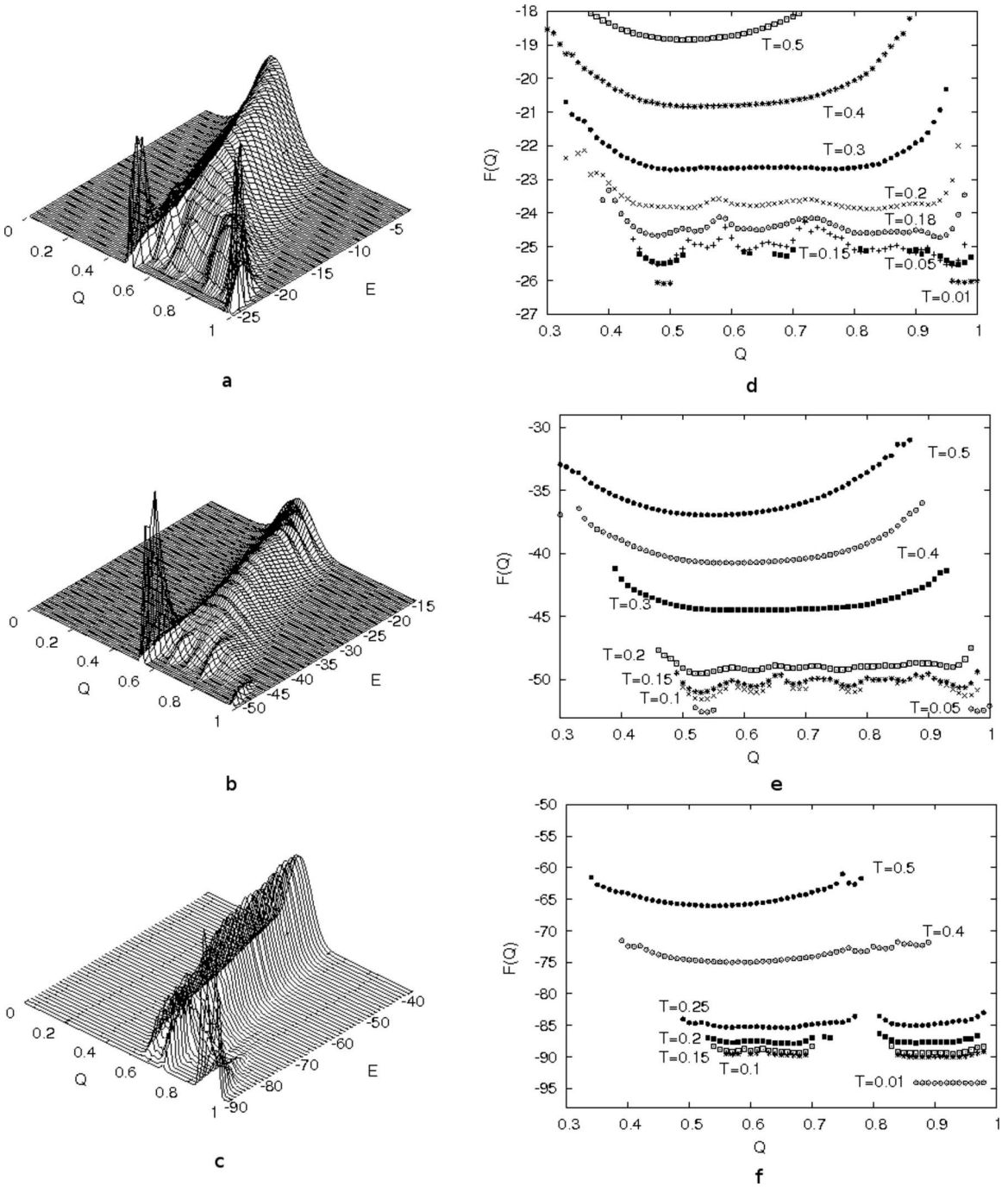

FIG. 4. The energy landscapes obtained by the multicanonical simulation run plotted against energy and the overlap parameter for the (a) 13-mer sequence, (b) 21-mer sequence, and (c) 34-mer sequence and the free energy as a function of overlap parameter for the sequences (d) 13-mer, (e) 21-mer, (f) 34-mer sequences, respectively.

$$d(\mathbf{X},\mathbf{X}') = \frac{1}{\pi}\left(\sum_{i=1}^{N_t} d_t(\Phi_i,\Phi_i') + \sum_{i=1}^{N_b} d_b(\Theta_i,\Theta_i')\right),$$

$$d_t(\Phi_i,\Phi_i') = \min(|\Phi_i - \Phi_i'|, 2\pi - |\Phi_i - \Phi_i'|),$$

$$d_b(\Theta_i,\Theta_i') = |\Theta_i - \Theta_i'|.$$

Since $-\pi \leqslant \Phi_i \leqslant \pi$ and $0 \leqslant \Theta_i \leqslant \pi$ it follows immediately that $0 \leqslant d_{t,b} \leqslant \pi$. The overlap is unity, if all angles of the conformations $\mathbf{X}$ and $\mathbf{X}'$ coincide, else $0 \leqslant Q < 1$.

The order parameter may be considered as a measure of coincidence of the considered conformation with the reference state.

The global minimum energy conformations for the 13-mer, 21-mer, and 34-mer sequences, which are used in the order parameter as the reference state, are given in Figs. 1(a),

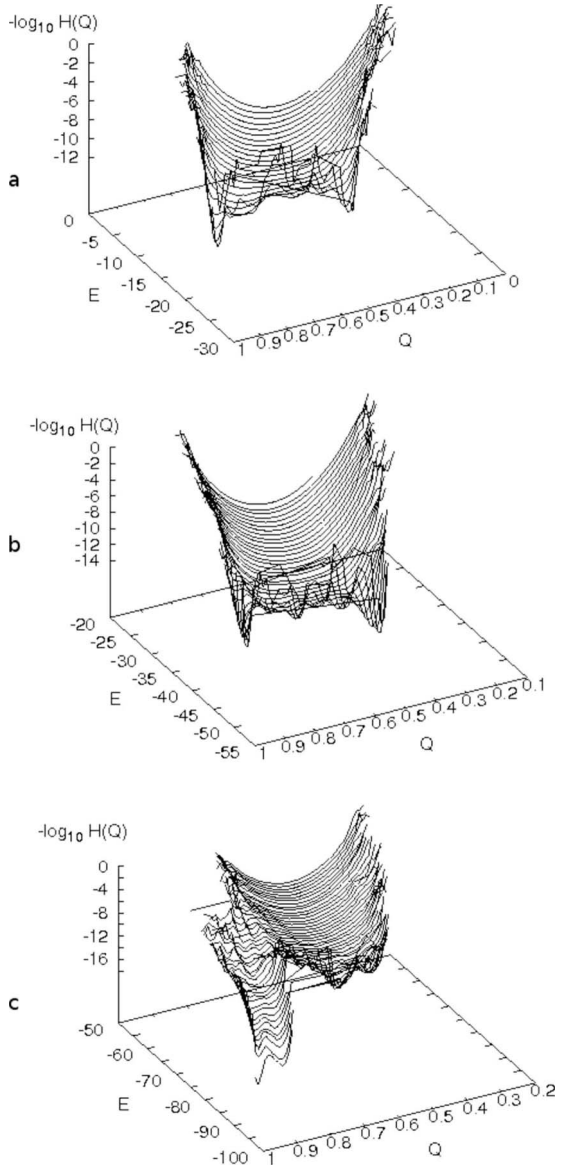

FIG. 5. The minus logarithmic value of Q histogram $[-\log_{10} H(Q)]$ plotted against energy and the overlap parameter for the (a) 13-mer sequence, (b) 21-mer sequence, and (c) 34-mer sequence, respectively.

2(a), and 3(a), respectively. The conformations rendered in Figs. 1(b) and 2(b) has a dejenerate energy values with Figs. 1(a) and 2(a), but have a different geometry. The energy values of the associated conformations are given in Table I. These minimums agree well with previous results [5].

The first part of the Fig. 4 [(a)–(c)] shows the energy histograms (the actual count of states of energy E) obtained by the multicanonical simulation run of 2×10^7 sweeps plotted against the overlap parameter (Q) for the 13-mer, 21-mer, and 34-mer systems. It should be noted that the utilized data is obtained by sampling the conformational space and no minimization procedure is applied.

First, the folding process of sequence 13-mer exhibits two different degenerate states, that is the main channel does not fold in one native state. At high energies (or high temperatures), where the system is in the random coil state, the histogram looks like a Gaussian peak centered around the value of $Q\sim0.3$. When the temperature is lowered, the transition from the random coil to a globular structure is expected. At this temperature ($T=0.18$) the energy landscape deviates from a smooth surface and bifurcates into many channels. Further down in energy (temperature), the two most populated channels finally win: Conformations $Q>0.98$ and conformations around $Q\sim0.5$ [which are rendered in Figs. 1(a) and 1(b), respectively].

Similar folding behavior is also seen in the 21-mer sequence. At high temperatures, the histogram [Fig. 4(b)] looks like a Gaussian peak centered around the value of $Q\sim0.5$. Then in the annealing process (around $T\sim0.2$), several channels are formed and coexist. The two most prominent channels (to which the lowest-energy conformation and the second-lowest-energy conformation belong, that I found in the simulations) eventually lead for $T\sim0.05$ to ensembles of conformations with $Q>0.97$, which are similar to the reference conformation shown in Fig. 2(a), and conformations with $Q\sim0.5$. The second-lowest-energy conformation found in this regime is shown in Fig. 2(b) and is structurally different but energetically degenerate compared with the reference conformation. Distinctly from the 13-mer case, here the conformations with the $Q\sim0.5$ are more probable. This means that, comparing with the second-lowest minima of the 13-mer case, the 21-mer's second-lowest-energy conformation has deeper energy minima, in other words it is surrounded with high-energy barriers so that going out from this minima and searching or finding the lowest-energy state takes a much longer time.

From Fig. 4(c), I identify that folding of sequence 34-mer exhibits typical two-state characteristics. Approaching from high energies (or high temperatures), the conformations in the ensemble have an overlap parameter $Q\sim0.55$, which means that there is no significant similarity with the reference structure (lowest-energy minimum), i.e., the ensemble consist of many unfolded states. Decreasing the temperature or energy a new branch is developed. From there onwards, the developing branch (nativefold conformations) with larger order parameter values ($Q\sim0.9$) wins and more conformations populate that section of the conformational space. In other words, the system folds smoothly towards $Q=1$.

To summarize; the energy landscapes have different branches through which the heteropolymer can go in the folding process. The native-state are located at the bottom rigt corner [$Q=1$ and $E=E_{\min}$; see Figs. 4(a)–4(c)]. The landscape of 13-mer and 21-mer have at the end two metastable states. On the other hand, the landscape of 34-mer shows a two-state folding channel and the folding ends in the native state [bottom right-hand corner of Fig. 4(c)].

The free energy as a function of the overlap parameter at fixed temperature was calculated by the formula [25]

$$F(T,Q) = E - TS(Q) \tag{4}$$

and the entropy in the above formula is

$$S(Q) = \log_{10} H(Q), \tag{5}$$

where $H(Q)$ is the histogram of the overlap parameter Q.

The second part of Fig. 4 gives the free energies as a function of overlap parameter for the three sequences: (d) 13-mer at different temperatures $T=0.5,0.4,0.3,0.2,0.18,0.15,0.05,0.01$. (e) 21-mer for temperatures $T=0.5,0.4,0.3,0.2,0.15,0.1,0.05$. (f) 34-mer for temperatures $T=0.5,0.4,0.25,0.2,0.15,0.1,0.01$, respectively. The folding behaviors can be also seen from the free-energy plots: For the first two sequences (13-mer and 21-mer) there are two branches in the lowest temperature plot. On the other hand, in the 34-mer case, we do not see the other branch in $T=0.01$—there appear only the native-state part in the conformational space.

Finally, all the phenomena can also be shown from the minus logarithmic value of the histogram of Q [$-\log_{10} H(Q)$ which is minus entropy, see Eq. (5)] which are given in Figs. 5(a)–5(c) for all of the sequences. It can be seen that for the 13-mer sequence [Fig. 5(a)], the main folding channel is like a slide. Decreasing the temperature, the channel splits into two parts which one goes to $Q\sim 0.98$ and the other to $Q\sim 0.5$. Similar behavior is also seen in the 21-mer sequence [Fig. 5(b)]. On the other hand, I conclude that folding of sequence 34-mer shows typical two-state characteristics [Fig. 5(c)]. Above the transition temperature, conformations are in random-coil state and close to $T\sim 0.25$ the folding funnel changes towards larger Q values.

At the end it should be noted that these sequences are studied recently within a different context [5,19–21], and therefore were not designed for this study. Nonetheless, we found complex folding behaviors, which are, in fact, qualitatively comparable to known characteristics of real proteins.

IV. CONCLUSION

In this work, different length of heteropolymer sequences of the off-lattice hydrophobic-polar model proteins are simulated to investigate the folding behaviors of the heteropolymers by exploiting the information withdrawn from the energy landscape topographies. The obtained pictures serve as a useful tool for visualization of the funnel-like structure of considered systems in the configuration space. I would like to point out that the energy landscapes, obtained from simulation data for the whole available energy range are presented here, which is an important achievement for the visualization of the evolution of folding process. As a result, with a minimal heteropolymer model, it is possible to see clear indications for the different folding characteristics, such as two-state folding, metastability, etc., known from bioproteins and synthetic peptides.

ACKNOWLEDGMENTS

I acknowledge the support from Scientific and Technological Research Council of Turkey under Contract No. 104T150 and Turkish Academy of Sciences under the programme to Reward Successful Young Scientists. I would like to thank Professor Tarik Celik for reading this article.

[1] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).
[2] J. N. Onuchic and P. G. Wolynes, Curr. Opin. Struct. Biol. **14**, 70 (2004).
[3] M. Bachmann and W. Janke, Phys. Rev. Lett. **91**, 208105 (2003).
[4] M. Bachmann and W. Janke, J. Chem. Phys. **120**, 6779 (2004); Comput. Phys. Commun. **169**, 111 (2005).
[5] M. Bachmann, H. Arkın, and W. Janke, Phys. Rev. E **71**, 031906 (2005).
[6] V. S. Pande, A. Yu. Grosberg, C. Joerg, and T. Tanaka, Phys. Rev. Lett. **76**, 3987 (1996).
[7] T. R. Weikl and K. A. Dill, J. Mol. Biol. **332**, 953 (2003).
[8] C. Clementi, A. Maritan, and J. R. Banavar, Phys. Rev. Lett. **81**, 3287 (1998).
[9] N. Koga and S. Takada, J. Mol. Biol. **313**, 171 (2001).
[10] H. S. Chan, S. Shimizu, and H. Kaya, Methods Enzymol. **380**, 350 (2004).
[11] H. Kaya and H. S. Chan, J. Mol. Biol. **326**, 911 (2003).
[12] H. Kaya and H. S. Chan, Proteins **52**, 510 (2003).
[13] F. H. Stillinger, T. Head-Gordon, and C. L. Hirshfeld, Phys. Rev. E **48**, 1469 (1993); F. H. Stillinger and T. Head-Gordon, *ibid.* **52**, 2872 (1995).
[14] S. Schnabel, M. Bachmann, and W. Janke, Phys. Rev. Lett. **98**, 048103 (2007).
[15] K. A. Dill, Biochemistry **24**, 1501 (1985); K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[16] A. Irbäck, C. Peterson, F. Potthast, and O. Sommelius, J. Chem. Phys. **107**, 273 (1997).
[17] A. Irbäck, C. Peterson, and F. Potthast, Phys. Rev. E **55**, 860 (1997).
[18] B. A. Berg and T. Çelik, Phys. Rev. Lett. **69**, 2292 (1992); B. A. Berg, Fields Inst. Commun. **28**, 1 (2000).
[19] J. G. Kim, J. E. Straub, and T. Keyes, Phys. Rev. Lett. **97**, 050601 (2006).
[20] V. Elser and I. Rankenburg, Phys. Rev. E **73**, 026702 (2006).
[21] S.-Y. Kim, S. B. Lee, and J. Lee, Phys. Rev. E **72**, 011916 (2005).
[22] H.-P. Hsu, V. Mehra, and P. Grassberger, Phys. Rev. E **68**, 037703 (2003).
[23] U. H. E. Hansmann, M. Masuya, and Y. Okamoto, Proc. Natl. Acad. Sci. U.S.A. **94**, 10652 (1997).
[24] B. A. Berg, H. Noguchi, and Y. Okamoto, Phys. Rev. E **68**, 036126 (2003).
[25] J. N. Onuchic, P. G. Wolynes, Z. Luthey-Schulten, and N. D. Socci, Proc. Natl. Acad. Sci. U.S.A. **92**, 3626 (1995); S. S. Plotkin, J. Wang, and P. G. Wolynes, J. Chem. Phys. **106**, 2932 (1997).