Folding of small proteins using a single continuous potential

Seung-Yeon Kim
School of Computational Sciences, Korea Institute for Advanced Study, 207-43 Cheongryangri-dong, Dongdaemun-gu, Seoul 130-722, Korea

Julian Lee
School of Computational Sciences, Korea Institute for Advanced Study, 207-43 Cheongryangri-dong, Dongdaemun-gu, Seoul 130-722, Korea and Department of Bioinformatics and Life Sciences, Soongsil University, Seoul 156-743, Korea

Jooyoung Lee[a]
School of Computational Sciences, Korea Institute for Advanced Study, 207-43 Cheongryangri-dong, Dongdaemun-gu, Seoul 130-722, Korea

Extensive Monte Carlo *folding* simulations for four proteins of various structural classes are carried out, using a *single* continuous potential (united-residue force field). In all cases, collapse occurs at a very early stage, and proteins fold into their nativelike conformations at appropriate temperatures. We also observe that glassy transitions occur at low temperatures. The simulation results demonstrate that the folding mechanism is controlled not only by thermodynamic factors but also by kinetic factors: The way a protein folds into its native structure is also determined by the convergence point of early folding trajectories, which cannot be obtained by the free energy surface. © *2004 American Institute of Physics.* [DOI: 10.1063/1.1689643]

I. INTRODUCTION

Understanding how a protein folds is a long-standing challenge in modern science. Computer simulations[1–14] have been carried out to understand the folding mechanism. However, simulation of protein-folding processes with an atomistic model is a very difficult task. The difficulties come from two sources: (i) Direct folding simulations using an all-atom potential require astronomical amounts of CPU time, and typical simulation times are only about a few nanoseconds. (ii) Atomic pairwise interactions including solvation effects may not be accurate enough. An extensive folding simulation has been carried out for the villin headpiece subdomain (HP-36), where a 1-μs molecular dynamics simulation with an all-atom potential has been performed, producing only candidates for folding intermediates.[1] For this reason, direct folding simulations have been mainly focused on simple models, such as lattice models,[2,3] models where only native interactions are included (Go-type models),[4–7] and a model with discrete energy terms whose parameters are optimized *separately* for each protein.[8] Alternative indirect approaches have also been proposed including unfolding simulations[7,9–11] starting from the folded state of a protein. However, it is not obvious whether the folding is the reverse of the unfolding.[3,7] Moreover, to the best of our knowledge, no one has yet succeeded in folding more than one protein into its native state using a *single* potential, even with simplified models.[8]

In this work, we propose a method to fold several of small proteins using Monte Carlo dynamics. This method uses a *single* continuous potential which includes *all* (native and non-native) interactions, and yet allows us to carry out *folding* simulations starting from non-native conformations. We observe that all proteins fold into their nativelike conformations at appropriate temperatures. We find that the folding mechanism is controlled by both kinetic and thermodynamic factors: The way a protein folds into its native structure is determined not only by the free energy surface, but also by the convergence point of early folding trajectories relative to the native state.

II. METHODS

We study the folding dynamics of proteins using a united-residue[15–17] (UNRES) force field where a protein is represented by a sequence of α-carbon (C^α) atoms linked by virtual bonds with attached united sidechains (SCs) and united peptide groups located in the middle between the consecutive C^α's (Fig. 1). All the virtual bond lengths are fixed: the C^α–C^α distance is taken as 3.8 Å, and C^α–SC distances are given for each amino acid type. The energy of a protein is given by

$$E = U_{dis} + \sum_{i<j} [U^{(4)}_{el\text{-}loc}(i,j) + U_{ss}(i,j)] + \sum_{i \neq j} U_{sp}(i,j)$$

$$+ \sum_{i<j-1} U_{pp}(i,j) + \sum_i [U_b(i) + U_t(i) + U_r(i)].$$

Here, U_{dis} denotes the energy term which forces two cysteine residues to form a disulfide bridge. The four-body interaction term $U^{(4)}_{el\text{-}loc}$ results from the cumulant expansion of the re-

---
[a]Author to whom correspondence should be addressed. Electronic mail: jlee@kias.re.kr
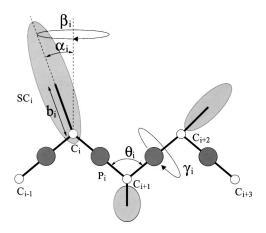
FIG. 1. United-residue representation of a protein. The interaction sites are sidechain ellipsoids of different sizes (SC) and peptide-bond centers (p) indicated by shaded circles, whereas the α-carbon atoms (small open circles) are introduced to define the backbone-local interaction sites and to assist in defining the geometry. The virtual C^α–C^α bonds have a fixed length of 3.8 Å, corresponding to a trans peptide group; the virtual-bond (θ) and dihedral (γ) angles are variable. Each sidechain is attached to the corresponding α-carbon with a different but fixed bond length b_i, variable bond angle α_i formed by SC_i, and the bisector of the angle defined by C_{i-1}^α, C_i^α, and C_{i+1}^α, and with a variable dihedral angle β_i of counterclockwise rotation about the bisector, starting from the right side of the C_{i-1}^α, C_i^α, C_{i+1}^α frame.

FIG. 2. Typical folding trajectories of betanova at $T=40$ and HP-36 at $T=70$ starting from non-native conformations. Only the first 10^8 MCS are shown among 10^9 MCS for clarity. RMSD, R_g, and Q are plotted for every 100 MCS in the early part and for every 2×10^5 steps in the subsequent part. The dotted lines represent the values of R_g of the native states.

stricted free energy of the protein. $U_{ss}(i,j)$ represents the mean free energy of the hydrophobic (hydrophilic) interaction between the sidechains of residues i and j, which is expressed by Lennard-Jones potential, $U_{sp}(i,j)$ corresponds to the excluded-volume interaction between the sidechain of residue i and the peptide group of residue j, and the potential $U_{pp}(i,j)$ accounts for the electrostatic interaction between the peptide groups of residues i and j. The terms $U_b(i)$, $U_t(i)$, and $U_r(i)$ denote the short-range interactions, corresponding to the energies of virtual angle bending, virtual dihedral angle torsions, and sidechain rotamers, respectively.

The parameters of the UNRES force field were *simultaneously* optimized[18,19] for four proteins of betanova[11,20] (20 residues, three-stranded β-sheet), 1fsd (Ref. 21) (28 residues, one β-hairpin, and one α-helix), HP-36 (Refs. 1 and 12) (36 residues, three-helix bundle), and fragment B of staphylococcal protein A (Refs. 5, 7, 9, 10, 13, and 22) (46 residues, three-helix bundle). The low-lying local-energy minima for these proteins were found by the conformational space annealing method.[23,24] The parameters were modified in such a way that the nativelike conformations are energetically more favored than the others. The global minimum-energy conformations found using the optimized force field are of the root-mean-square deviation (RMSD) values 1.5 Å, 1.7 Å, 1.7 Å, and 1.9 Å from the *experimental* structures for betanova, 1fsd, HP-36, and protein A, respectively. For betanova, the "native" structure is shown to be only marginally stable and is in equilibrium with less ordered structures.[25] For this reason, it might be more appropriate to call this structure as the "predominant" one. In this paper, we choose to use the term "native" to refer to "predominant" for betanova, to be consistent with the other three proteins. After the parameter optimization, *one* set of the parameters is obtained for four proteins. The optimized parameters are not overfitted to the
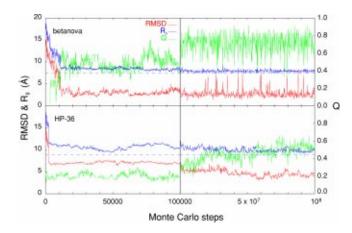
four proteins, but are transferable to other proteins to some extent.[19] In Ref. 19 the procedures to obtain the optimized parameters used in this work are described in detail.

In the UNRES force field there are two backbone angles and two sidechain angles (see Fig. 1) per residue (no sidechains for glycines). The values of these angles are perturbed one at a time, typically about 15°, and the backbone angles are chosen three times more frequently than the sidechain angles. The perturbed conformation is accepted according to the change in the potential energy, following the Metropolis rule. Since only small angle changes are allowed one at a time, the resulting Monte Carlo dynamics can be viewed as equivalent to the real dynamics.[2,8,14] At a fixed temperature, at least ten independent simulations starting from various non-native states of a protein were performed up to 10^9 Monte Carlo steps (MCS) for each run, which we call long-time simulations. Hundreds of long-time simulations were conducted for a protein.

During simulations the values of RMSD from the native structure and the radius of gyration (R_g) were calculated using C^α coordinates. The lowest RMSD values from folding simulations are 0.78 Å, 1.07 Å, 1.58 Å, and 2.07 Å for betanova, 1fsd, HP-36, and protein A, respectively. The fractions of the native contacts (Q and ρ) were also measured during simulations, where Q is calculated from the experimental structure. A native contact is defined to exist when two C^α's separated more than two residues in sequence are placed within 7.0 Å. To define ρ we first characterize the native-state conformations by performing simulations starting from the experimental structures, at the same temperatures where folding simulations were performed. We define ρ as the fraction of native contacts weighted over their contact probabilities in the native state simulations.[7,10,11] Due to the fluctuation of the native conformation, the value of Q is usually lower than that of ρ. The time histories of the typical trajectories from the folding simulations of betanova and HP-36 are shown in Fig. 2. The trajectories for 1fsd and protein A are similar to those in the figure. Distributions of RMSD, Q, ρ, and R_g are also accumulated during the whole

J. Chem. Phys., Vol. 120, No. 17, 1 May 2004
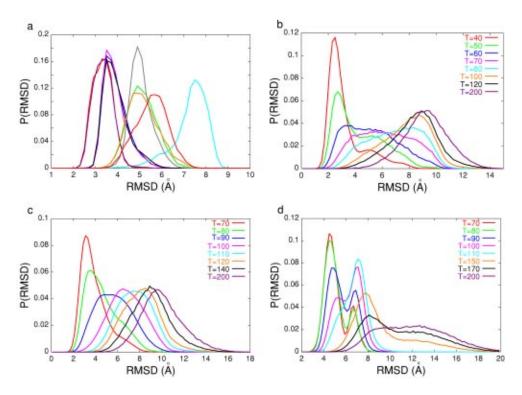
Folding of small proteins   8273

FIG. 3. The probability distributions of RMSD. (a) The distributions of ten independent long-time simulations for betanova, starting from random conformations at $T=30$. The glassy behavior is apparent. (b) RMSD distributions of betanova at various temperatures. As temperature decreases, the value of prominent RMSD dramatically moves from 9.0 Å to 2.5 Å. (c) The distributions for 1fsd. The value of the most probable RMSD drops rapidly from 9.0 Å to 3.1 Å. (d) The distributions for HP-36. The double-peak structure appears for $T \leqslant 100$, representing the cooperative (first-order-like) two-state transition. In (b), (c), and (d) the probability distribution at a temperature for the given protein was obtained by accumulating the results from ten independent long-time simulations.

simulations. The distributions of RMSD are shown in Fig. 3 and those for ρ and R_g are shown in Fig. 4 as contour plots.

To investigate the early folding trajectories, 200 (10^5 MCS) independent simulations were also performed at a fixed temperature for a protein, which we call short-time simulations. The averages of early folding trajectories plotted in Fig. 4 were obtained as follows. We divided the initial 10^5 MCS into 19 intervals (ten 10^3 MCS and subsequent nine 10^4 MCS), and took the average over conformations in each interval. These averages were again averaged over 100 independent simulations starting from random conformations. The same procedure was applied to 100 independent simulations starting from a fully extended conformation.

III. RESULTS

The simulation of betanova at $T=40$ (arbitrary units) starting from an unfolded conformation demonstrates that rapid collapse occurs in about 10^4 MCS; the value of R_g decreases, whereas the value of Q remains below 0.3 (Fig. 2). During the next 10^8 MCS, more compact states appear, and the value of Q moves up as high as 1.0. There are two values of RMSD that are populated [see also Fig. 3(b)]: one corresponds to less ordered structures (RMSD~5 Å) and the other to nativelike structures (RMSD~2.5 Å). This demonstrates that, although the nativelike structure is the most stable one, thermal fluctuation can temporarily kick the protein out of the native structure.

We observe that collapse occurs at a very early stage ($\sim 10^4$ MCS) for all four proteins, but the details of each folding process appear to depend on the secondary structure contents. We now analyze the details of the folding behavior of each protein. For betanova at low temperatures ($T \leqslant 30$), the probability distributions of various quantities such as RMSD depend on initial conformations, showing its glassy behavior [Fig. 3(a)]. At higher temperatures ($T \geqslant 40$) this nonergodic glassy behavior disappears. It should be noted that nativelike structures are more easily found from the simulation at $T=40$ [Fig. 3(b)] than from the best of ten runs at $T=30$. When temperature decreases from $T=80$ to 60, the location of the RMSD peak dramatically moves from 8 Å to 3 Å. This demonstrates the cooperative folding characteristics of betanova. For $40 \leqslant T \leqslant 60$ betanova folds into its nativelike structure. The initial folding trajectories and the distribution of (ρ, R_g) at $T=40$ are shown in Fig. 4(a) for betanova. Regardless of its initial conformation (either random or fully extended), the average pathways to the folded conformation initially converge to $(\rho, R_g) \sim (0.35, 8.5$ Å), and then they move horizontally to the native structure. This is consistent with the recent folding scenario for proteins with β structure.[7] The populated states[11] around $\rho \sim 0.4$–0.5 and $R_g \sim 11$–12 Å are not from initial folding trajectories, but from the fluctuation of nativelike structures. This kinetic information is difficult to be captured by free energy calculations alone.[11] Figure 5 shows snapshots of a typical folding trajectory for betanova at $T=50$.
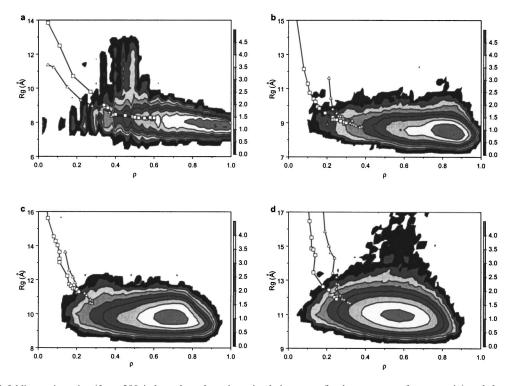
FIG. 4. The initial folding trajectories (from 200 independent short-time simulations at a fixed temperature for a protein) and the contour plots of the population distributions (from 10 independent long-time simulations at a fixed temperature for a protein) as a function of ρ and R_g at appropriate temperatures. The triangles represent the averages of 100 early folding trajectories starting from random conformations. The squares are from 100 early folding trajectories starting from a fully extended conformation. The color scale of the contour plots indicates the exponent x of the population 10^{x+5} at given values of ρ and R_g. (a) Betanova at $T=40$. (b) 1fsd at $T=70$. (c) HP-36 at $T=70$. (d) Protein A at $T=80$.
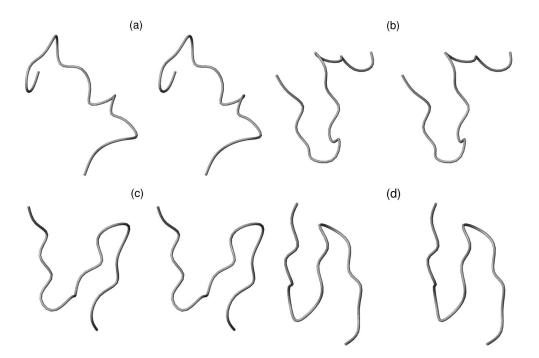

FIG. 5. Snapshots of a typical folding trajectory for betanova at $T=50$. In the figures the left side is the N-terminal part. (a) Initial random conformation: RMSD$=10.0$ Å, $Q=0.08$, and $R_g=12.2$ Å. (b) Conformation at 10^4 MCS: RMSD$=7.2$ Å, $Q=0.21$, and $R_g=10.7$ Å. This conformation corresponds to an initially collapsed state where the β-hairpin structure is formed only in the N-terminal residues. (c) Conformation at 2×10^4 MCS: RMSD$=5.3$ Å, $Q=0.33$, and $R_g=8.8$ Å. This conformation has the same topology as the native state, but it is less ordered. In particular, it corresponds to the convergence point of the early folding trajectories in (Q,R_g) plane. (d) Conformation at 7×10^5 MCS: RMSD$=1.4$ Å, $Q=0.75$, and $R_g=7.7$ Å. The conformation is almost identical to the native state, and is one of the most populated conformations in Fig. 4(a).

(a)   (b)

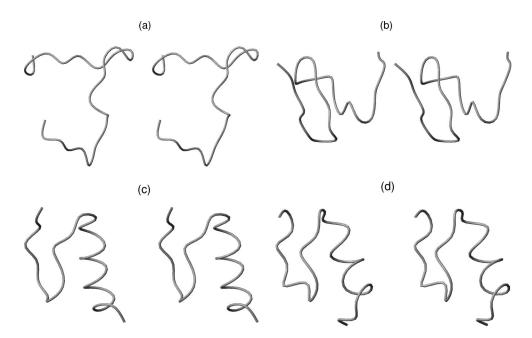(c)   (d)

FIG. 6. Snapshots of a typical folding trajectory for 1fsd at $T=70$. In (a) the upper left side is the N-terminal part. In (b), (c), and (d) the left side is the N-terminal part. (a) Initial random conformation: RMSD$=9.8$ Å, $Q=0.03$, and $R_g=14.4$ Å. (b) Conformation at 4×10^4 MCS: RMSD$=5.9$ Å, $Q=0.18$, and $R_g=9.5$ Å. This conformation corresponds to the convergence point of the early folding trajectories in (Q,R_g) plane. The β-hairpin is formed but the α-helix is not yet initiated. (c) Conformation at 10^6 MCS: RMSD$=3.7$ Å, $Q=0.50$, and $R_g=8.4$ Å. This conformation has the α-helix formed and its topology is that of the native state. It is one of the most populated conformations in Fig. 4(b). (d) Conformation at 2×10^6 MCS: RMSD$=1.9$ Å, $Q=0.68$, and $R_g=8.0$ Å. This conformation is almost identical to the native state.

For 1fsd the distributions of various quantities for ten independent runs (10^9 MCS each) show glassy behavior for $T\leqslant50$. Again, the nonergodic glassy behavior disappears at higher temperatures ($T\geqslant70$). Figure 3(c) shows the RMSD distributions at various temperatures. The strength of the co-operativity for 1fsd is not as strong as that of betanova. Again after initial collapse to $(\rho,R_g)\sim(0.3,9$ Å$)$, the average trajectories move horizontally [Fig. 4(b)], although less

prominently so compared to the case of betanova. Figure 6 shows snapshots of a typical folding trajectory for 1fsd at $T=70$.

For HP-36, the distributions of various quantities for ten independent runs (10^9 MCS each) again show glassy behavior at $T=60$. At higher temperatures we have two RMSD peaks [Fig. 3(d)]. At $T=90$ the peak near the native structure begins to dominate, and as temperature decreases it becomes

(a)   (b)

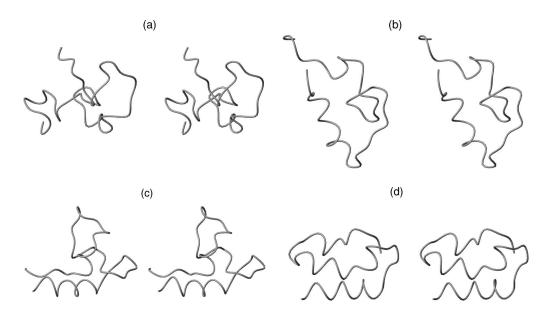(c)   (d)

FIG. 7. Snapshots of a typical folding trajectory for protein A at $T=85$. In the figures the lower side is the C-terminal part. (a) Initial random conformation: RMSD$=14.3$ Å, $Q=0.21$, and $R_g=16.9$ Å. (b) Conformation at 1.3×10^5 MCS: RMSD$=7.2$ Å, $Q=0.28$, and $R_g=11.3$ Å. This conformation corresponds to the convergence point of the early folding trajectories in (Q,R_g) plane. The helix III is already formed. (c) Conformation at 1.013×10^7 MCS: RMSD $=10.0$ Å, $Q=0.43$, and $R_g=10.7$ Å. The RMSD value is larger than that of the conformation (b). It should be noted that the helix III is stably formed. (d) Conformation at 5.213×10^7 MCS: RMSD$=2.6$ Å, $Q=0.56$, and $R_g=9.9$ Å. The conformation has the same topology as the native state.
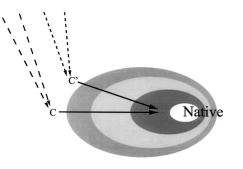
FIG. 8. A schematic of the folding trajectories. The colored contour represents the free energy surface, which is an equilibrium property. Even for proteins with identical free energy landscape, the early folding trajectories (dashed lines) may converge into different points (C or C′). The solid lines represent the later part of the folding trajectories dictated by the free energy landscape. The location (C or C′) of the convergence point of the early folding trajectories for a protein is determined by its kinetic properties. This information can be obtained only by direct *folding* simulations.

stronger. This double-peak feature demonstrates the cooperative two-state transition. The conformations centered at the higher value of RMSD come from a variety of collapsed states. The conformations from the other peak are native like. When we examine them, the helix I (residues 4–8) is stably formed,[12] while the others are fluctuating. The initial folding trajectories [Fig. 4(c)] converge to $(\rho,R_g)\sim(0.2,11\,\text{Å})$. These collapsed structures fold into nativelike structures $(\rho,R_g)\sim(0.7,10\,\text{Å})$. Compared to the case of betanova and 1fsd, the average folding trajectories are more diagonal.

The overall folding characteristics of protein A are similar to those of HP-36. The initial folding trajectories [Fig. 4(d)] converge to $(\rho,R_g)\sim(0.25,12\,\text{Å})$. These collapsed states fold into nativelike structures in a diagonal fashion similar to the case of HP-36. When we examine the native-like conformations with $3\,\text{Å}\leqslant\text{RMSD}\leqslant4\,\text{Å}$, the helix III (residues 42–55) is most stably formed. This is in agreement with recent investigations.[9,13,22] We also observe that a first-order-like collapse transition[5] [from $(Q,R_g)\sim(0.15,18\,\text{Å})$ to $(0.15,12\,\text{Å})$] occurs near $T=120$. Figure 7 shows snapshots of a typical folding trajectory for protein A at $T=85$.

IV. DISCUSSION

By using a single continuous potential, we could observe folding processes of four small proteins in realistic settings. In all cases, rapid collapse is followed by subsequent folding process that takes place in a longer time scale. We also observe that glassy transitions occur at low temperatures. The folding mechanism suggested in this study is as follows. There are two aspects of folding dynamics: (i) nonequilibrium kinetic properties and (ii) equilibrium thermodynamic properties (Fig. 8). The nonequilibrium kinetic properties, relevant to the early folding trajectories (fast process), can be examined only by direct folding simulations. The free energy surface, an equilibrium thermodynamic property, dictates the way an initially collapsed state completes its folding (slow process). The way a protein folds into its native structure—

i.e., either horizontally or diagonally in the (ρ,R_g) plane—is determined by the position (for example, C or C′ in Fig. 8) of (ρ,R_g) where early folding trajectories converge, relative to the native state. It appears that slow folding process of α-proteins occurs in a diagonal fashion compared to that of proteins containing β-strands.[7]

In conclusion, we successfully carried out direct folding simulations of more than one protein using a *single* continuous potential. In this study, we simultaneously optimized[18,19] the parameters of the UNRES force field for four small proteins and used this force field for the study of the folding mechanism of these proteins. The results provide new insights into the folding mechanism.

[1] Y. Duan and P. A. Kollman, Science **282**, 740 (1998).
[2] J. Skolnick and A. Kolinski, Science **250**, 1121 (1990); A. Sali, E. Shakhnovich, and M. Karplus, Nature (London) **369**, 248 (1994); H. S. Chan and K. A. Dill, Proteins **30**, 2 (1998); D. K. Klimov and D. Thirumalai, *ibid.* **43**, 465 (2001); J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).
[3] A. R. Dinner and M. Karplus, J. Mol. Biol. **292**, 403 (1999).
[4] N. Go, Annu. Rev. Biophys. Bioeng. **12**, 183 (1983).
[5] Y. Zhou and M. Karplus, Nature (London) **401**, 400 (1999).
[6] O. V. Galzitskaya and A. V. Finkelstein, Proc. Natl. Acad. Sci. U.S.A. **96**, 11299 (1999); E. Alm and D. Baker, *ibid.* **96**, 11305 (1999); V. Munoz and W. A. Eaton, *ibid.* **96**, 11311 (1999).
[7] J.-L. Shea and C. L. Brooks, Annu. Rev. Phys. Chem. **52**, 499 (2001).
[8] E. Kussel, J. Shimada, and E. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **99**, 5343 (2002).
[9] D. O. V. Alonso and V. Daggett, Proc. Natl. Acad. Sci. U.S.A. **97**, 133 (2000).
[10] E. M. Boczko and C. L. Brooks, Science **269**, 393 (1995).
[11] B. D. Bursulaya and C. L. Brooks, J. Am. Chem. Soc. **121**, 9947 (1999).
[12] A. Fernandez, M.-Y. Shen, A. Colubri, T. R. Sosnick, R. S. Berry, and K. Freed, Biochemistry **42**, 664 (2003).
[13] A. Ghosh, R. Elber, and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **99**, 10394 (2002).
[14] A. Kolinski and J. Skolnick, *Lattice Models of Protein Folding, Dynamics, and Thermodynamics* (Chapman & Hall, New York, 1996).
[15] A. Liwo, S. Oldziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga, J. Comput. Chem. **18**, 849 (1997).
[16] A. Liwo, M. R. Pincus, R. J. Wawak, S. Rackovsky, S. Oldziej, and H. A. Scheraga, J. Comput. Chem. **18**, 874 (1997).
[17] A. Liwo, R. Kazmierkiewicz, C. Czaplewski, M. Groth, S. Oldziej, R. J. Wawak, S. Rackovsky, M. R. Pincus, and H. A. Scheraga, J. Comput. Chem. **19**, 259 (1998).
[18] J. Lee, K. Park, and J. Lee, J. Phys. Chem. B **106**, 11647 (2002).
[19] J. Lee, S.-Y. Kim, and J. Lee, J. Phys. Chem. B (to be published).
[20] T. Kortemme, M. Ramirez-Alvardo, and L. Serrano, Science **281**, 253 (1998).
[21] B. I. Dahiyat and S. L. Mayo, Science **278**, 82 (1997).
[22] Y. Bai, A. Karimi, H. J. Dyson, and P. E. Wright, Protein Sci. **6**, 1449 (1997).
[23] J. Lee, H. A. Scheraga, and S. Rackovsky, J. Comput. Chem. **18**, 1222 (1997).
[24] S.-Y. Kim, S. J. Lee, and J. Lee, J. Chem. Phys. **119**, 10274 (2003).
[25] M. L. de la Paz, E. Lacroix, M. Ramirez-Alvardo, and L. Serrano, J. Mol. Biol. **312**, 229 (2001).