Dependence of Folding Rates on Protein Length†

Mai Suan Li,‡ D. K. Klimov,§ and D. Thirumalai*,§

Institute of Physics, Polish Academy of Sciences, Al. Lotnikow 32/46, 02−668 Warsaw, Poland, and Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742

Received: March 28, 2002; In Final Form: June 8, 2002

Using three-dimensional Go lattice models with side chains for proteins, we investigate the dependence of folding times on protein length. In agreement with previous theoretical predictions, we find that the folding time τ_F grows as a power law with the chain length, i.e., $\tau_F \sim N^{\lambda}$, where $\lambda \approx 3.6$ for the Go model, in which all native interactions (i.e., between all side chains and backbone atoms) are uniform. If the interactions between side chains are given by pairwise statistical potentials, which introduce heterogeneity in the contact energies, then the power law fits yield large λ values that typically signify a crossover to an underlying activated process. Accordingly, the dependence of τ_F on N is best described using $\tau_F \sim e^{\sqrt{N}}$. The study also shows that the incorporation of side chains considerably slows folding by introducing energetic and topological frustration.

I. Introduction

Protein folding mechanisms depend not only on the architecture of the native state but also on the external conditions (pH, salt concentration, temperature, and molecular environment). Several recent studies have argued that the folding rates (presumably, under the conditions of neutral pH, zero salt concentration, room temperature, and the absence of molecular crowding) are determined solely by the architecture of the native structure.[1] Although the native topology does constrain the ensemble of transition states (the folding nuclei have to be topology preserving[2]), other factors, such as protein size and the native state stability, also play a role in determining folding rates and mechanisms. For instance, a direct correlation between folding rates and stability has been noted by Clarke and coworkers.[3] They showed that for five proteins, all with immunoglobulin-like fold, the folding rates k_F correlate well with the native state stability. On the other hand, there is a poor correlation between k_F and the relative contact order,[1] which quantifies the balance of local vs nonlocal native interactions. Improved correlation may still be expected for the proteins with α-helical or α/β architecture.[4]

Although the importance of native state stability in determining k_F has been demonstrated, limited experimental data have been used to argue that the length of proteins (i.e., the number of amino acids N) should not affect k_F.[1] From the polymer physics perspective, this is somewhat surprising, because the relaxation rates even for ideal polymer chains depend on N. For example, the largest relaxation time in a Rouse chain scales as N^2.[5] Because the size range of single domain proteins is limited (typically less than about 200 residues), the dependence of k_F on N cannot be sharply demonstrated. In proteins, other factors, such as amino acid sequence and the nature of local and nonlocal interactions in the native state, could be more dominant. Nevertheless, the mere fact that proteins are polymers implies that N should play some role in determining the folding rates.[6]

The dependence of k_F on N has been investigated in a number of theoretical studies.[7−13] Several folding scenarios emerge depending on the characteristic folding temperatures, namely, the collapse temperature, T_θ, the folding transition temperature, T_F, and the glass transition temperature, T_g.[7,14] For optimal folding, for which $T_\theta \approx T_F$, Thirumalai has predicted that the folding time τ_F should scale with N as[7]

$$\tau_F \sim N^{\lambda} \tag{1}$$

The dimensionality dependent exponent λ for two-state folders is expected to be between 3.8 and 4.2.[7] Simulation studies using Go lattice models (LMs) without side chains suggest a smaller value of about 3.[11,12] These numerical studies are in broad agreement with the theoretical predictions. The heteropolymer nature of protein-like models could make λ temperature dependent. For two-state optimized folders there is a relatively broad range of temperatures, where τ_F remains relatively insensitive to T.[15] The N dependence of τ_F outside this range may not obey eq 1 or λ may be different.

All of the numerical studies mentioned above have been done using LMs in which each residue is represented by a bead confined to the vertices of an appropriate (usually cubic) lattice. Side chain packing effects, which are crucial in the folding process, cannot be considered in this class of LMs. A simple way to include these in the context of LMs is to attach an additional bead to each α-carbon atom in a sequence.[16,17] Thus, an amino acid consists now of two beads, one representing a backbone (BB) and the other a side chain (SC). In this polypeptide model there are $2N$ beads. If an appropriate heterogeneous potential between side chains is included, then the cooperative transition reminiscent of folding can be reproduced.[18] Thermodynamics and kinetics of lattice models with side chains (LMSC) have been recently reported.[17,19,20]

In this paper we examine the effect of rotamer degrees of freedom on the exponent λ (eq 1) using Go-like models.[21] In these highly simplified models, which have received consider-

† Part of the special issue "John C. Tully Festschrift".
* Corresponding author.
‡ Polish Academy of Sciences.
§ University of Maryland.

10.1021/jp025837q CCC: $22.00
Published on Web 06/27/2002

able attention in recent years, only interactions present in the native state are considered. Non-native interactions, which can play an important role in the thermodynamics and kinetics of folding,[19,20,22] are ignored. Nevertheless, several studies have showed that the Go models provide a reasonable caricature of certain aspects of folding.[19,23,24] For this model system, we find that the exponent λ is altered by rotamer degrees of freedom. More importantly, λ depends on the details of interactions and, in this sense, is nonuniversal. As a technical byproduct of our investigation we show that robust results for λ are obtained only for those Monte Carlo move sets, which are effectively ergodic.

II. Methods

Model. In the LMSC, the energy of a conformation is[17]

$$E = \epsilon_{bb} \sum_{i=1,\, j>i+1}^{N} \delta_{r_{ij}^{bb},a} + \epsilon_{bs} \sum_{i,j=1,\, j\neq i}^{N} \delta_{r_{ij}^{bs},a} + \epsilon_{ss} \sum_{i=1,\, j>i}^{N} \delta_{r_{ij}^{ss},a}$$

where ϵ_{bb},ϵ_{bs}, and ϵ_{ss} are BB−BB, BB−SC, and SC−SC contact energies. The terms r_{ij}^{bb}, r_{ij}^{bs}, and r_{ij}^{ss} are the distances between the ith and jth residues for the BB−BB, BB−SC, and SC−SC pairs, respectively. Each lattice site can only be occupied by a single bead (BB or SC) so that the self-avoidance condition is satisfied.

We consider two versions of the Go model. In the model GM1, ϵ_{bb}, ϵ_{bs}, and ϵ_{ss} are chosen to be -1 for native contacts and 0 for non-native ones. In GM2, $\epsilon_{bb} = \epsilon_{bs} = -0.2$, and the values of ϵ_{ss}, which depend on the nature of amino acids, are given by Betancourt-Thirumalai statistical interaction potentials.[25] Thus, GM2 incorporates diversity in the interaction energies that is known to be important in the design of foldable sequences. The fraction of hydrophobic residues in a sequence is approximately 0.5 as in wild-type proteins. By setting all of non-native contact interactions to 3.0 (this value is larger than any of Betancourt-Thirumalai couplings[25]), we ensure that non-native interactions do not contribute to folding. Thus, for all practical purposes both models exhibit Go-like characteristics.

The maximum length of sequences N examined in our work is 40. Investigation of scaling behavior of LMSC beyond this limit is computationally expensive. However, scaling trends may be reasonably established for LMs without side chains, even for $N \leq 40$ (see Figure 2 in ref 11). Therefore, we are fairly confident that considering LMSC with $N \leq 40$ would not hamper our ability to analyze scaling of folding times with N.

Sequences. Protein-like sequences were obtained using the standard Z-score optimization or by minimizing the energy of the native state.[26,27] Z-score optimization is based on Monte Carlo simulations in sequence space aimed at minimizing $Z = (E_0 - E_{ms})/\delta$, where E_0 and E_{ms} are the energy of a sequence in the target conformation and the average energy of misfolded structures, respectively, and δ is the dispersion in the native contact energies. We took $E_{ms} = c\,\langle B\rangle$, where c is average number of nearest neighbor contacts in the manifold of misfolded structures, and $\langle B\rangle$ is the average contact energy for a given sequence. The Monte Carlo simulations in sequence space were done using simulated annealing protocol by generating 20 independent trajectories for each sequence. The optimized sequence is the one with the lowest Z-score. For each N, the target conformations for GM1 and GM2 are identical, so the effect of the "realistic" SC interactions can be directly addressed.

Despite its simplicity, Z-score and energy optimizations[26,27] have been proven to be effective techniques for generating designed foldable sequences with $N \lesssim 100$.[11] Other, more elaborate, technical methods for designing lattice protein-like sequences have been proposed.[28] For the scope of our paper these methods are not relevant, because we are interested only in generating foldable sequences spanning a reasonable range of N.

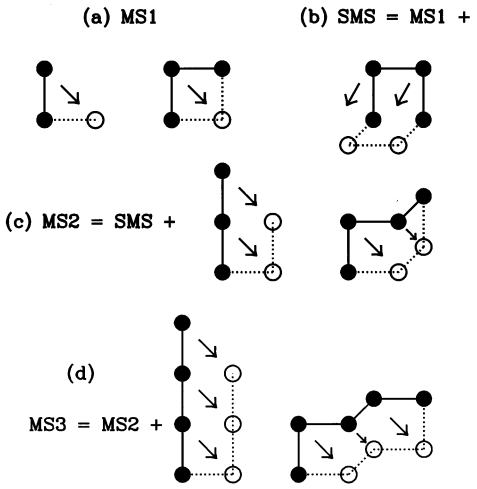

Figure 1. MC move sets examined in this study: (a) MS1 is based on single monomer corner flips (including tail ones); (b) SMS incorporates MS1 and the crankshaft moves; (c) MS2 involves SMS and additional two-monomer moves; (d) MS3 contains MS2 and three-monomer moves.[28]

Move Sets in Monte Carlo Simulations. To assess the efficiency of the Monte Carlo (MC) simulations and to check the robustness of the results, we used four distinct move sets (Figure 1). Move set MS1 involves only single corner (and also tail) monomer moves. In addition to single flips, the standard move set (SMS) also contains the crankshaft motion.[29] Figure 1 shows the moves in the set MS2, which includes SMS and additional two-monomer moves (not crankshaft ones). Move set MS3 is implemented as described in ref 28. The validity of MS3 has been verified for short sequences without side chains by comparing MC results with those obtained by full enumeration of lattice conformations, which is tantamount to performing ensemble average. Thus, for MS3 ergodicity has been established and implementation of detailed balance condition has been also discussed.[28] We have found that due to its flexibility, MS3 is far more efficient than others. The purpose of using different MC move sets is to ensure the robustness of our results. In our study, one MC step consists of N MC moves, i.e., on an average each bead in a sequence is attempted to move once during one MC step.

Computation of Folding Times and Temperatures. For each sequence and temperature we computed the distribution of first passage times τ_{1i}, where τ_{1i} is the number of MC steps needed to reach the native state starting from the unfolded state i. The structure is considered folded if the overlap function $\chi = 0$.[17] The folding time $\tau_F = (1/M)\sum_{i=1}^{M}\tau_{1i}$, where $M = 100$ is the number of individual trajectories. The folding time to reach the native backbone, τ_F^{bb}, has been calculated in a similar way.

In our study, folding times τ_F have been obtained at the temperature of fastest folding T_{min} located by scanning a temperature range for each sequence. If the folding transition temperature T_F is identified with the maximum in the fluctuations of overlap function,[17] then we expect $T_F \approx T_{min}$. This conclusion, which naturally follows from the dual requirement of the stability of protein native state and its kinetic accessibility, is illustrated in Figure 2 for two-state LMSC sequence with $N = 15$. The observation that $T_F \approx T_{min}$ serves as a convenient operational condition to pinpoint the temperature of folding

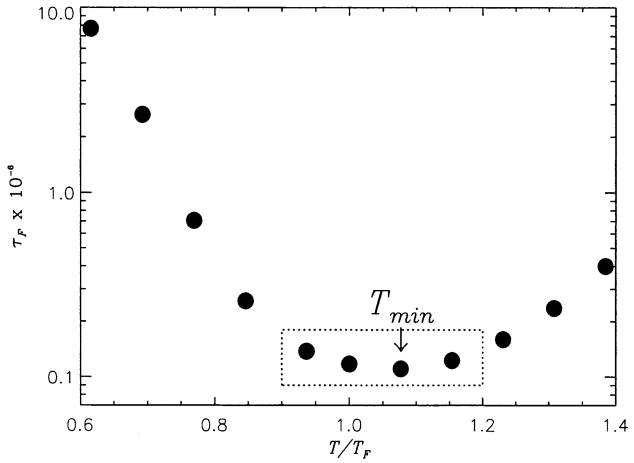

Figure 2. Characteristic U-shape dependence of the folding time τ_F on temperature measured in the units of the folding temperature T_F for the LMSC sequence A ($N = 15$) studied in ref 17. T_F is computed using multiple histogram method as the temperature, at which the fluctuations in the overlap function reach maximum. The temperature at which τ_F is minimum is $T_{min} = 1.07T_F$. An observation that $T_F \approx T_{min}$ may be used for crude estimation of T_F for a large dataset of sequences. An almost flat region in the vicinity of T_F is indicated by a dotted box[15] (see also Figure 3).

simulations without the need of expensive equilibrium simulations. Furthermore, it has been shown[15] that for highly optimized sequences there is a large plateau in the temperature dependence of τ_F. This flat temperature range typically includes T_{min}, T_F, and collapse temperature T_θ. We expect that scaling behavior would be similar for these temperatures.

In a previous study[12] that has examined the N dependence of folding times, T_F has been defined as the temperature at which the probability of occupancy of the microscopic native conformation is 0.5. Such a highly restrictive definition of T_F is not physically meaningful. It is realized that the fluctuations in the overlap function or the equilibrium fraction of native contacts are more appropriate quantities for defining T_F.[17,20,30] The physically relevant definition, which also coincides with the experimental definitions, is based on the notion of the native state as a *collection* of structurally similar conformations belonging to the native basin of attraction (NBA).

III. Results

We have monitored the length dependence of the folding times τ_F, which register folding of the entire native structure, as well as τ_F^{bb}, which is the average first passage time to the folded backbone. The characteristic U-shape for the temperature dependence of τ_F noted for optimized sequences[11] is also found for τ_F^{bb} (Figure 3). Over the range of temperatures, where τ_F remains roughly constant, $\tau_F \approx \tau_F^{bb}$. However, this is not the case at low and high temperatures (see below). The extent of the plateau region in T is larger for GM1 (Figure 3). Narrow shape of the temperature dependence of τ_F for GM2 resembles that computed for the random sequence without side chains (Figure 3 in ref 11). In the rest of the paper we focus on the variations of the folding times at $T_{min} \approx T_F$.

The N-dependence of folding times obtained by four types of MC move sets for GM1 at $T = T_{min}$ is presented in Figure 4. The number of targets used for $N = 9$, 15, 18, 24, 28, 32, and 40 are 100, 50, 50, 20, 17, 15, and 15, respectively. For MS1 the calculations were performed only up to $N = 32$. Ergodic move sets (i.e., SMS, MS2, and MS3), which efficiently sample the conformational space, were used for $N = 40$.

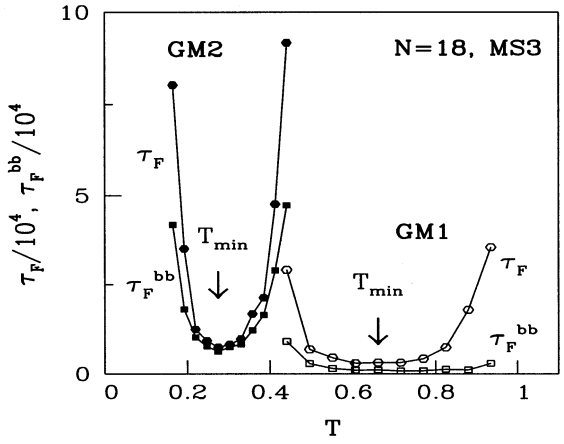

Figure 3. Temperature dependence of τ_F and τ_F^{bb} for $N = 18$ GM1 and GM2 sequences, both of which share the same native conformation. Data are obtained using MS3. The arrows indicate T_{min}, the temperature at which τ_F is minimal. The GM2 sequences have a narrower plateau region than GM1. Therefore, GM2 sequences are not as well optimized as GM1.

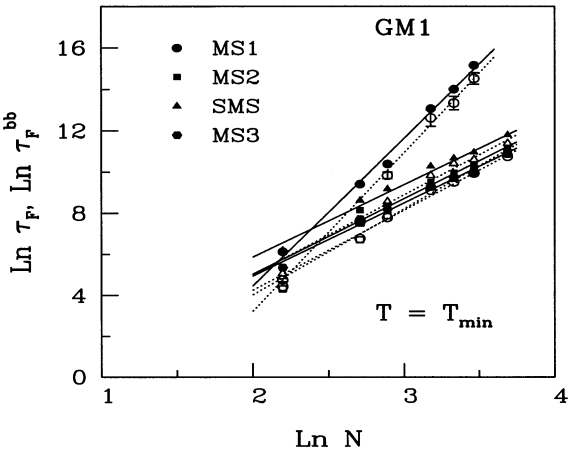

Figure 4. Scaling of τ_F and τ_F^{bb} for GM1 at $T = T_{min}$. The results are shown for all four types of MC move sets. The solid symbols indicate folding times for the entire sequence τ_F, the open ones represent backbone folding times τ_F^{bb}. Straight solid and dotted lines are corresponding power law fits. Scalings of τ_F and τ_F^{bb} are virtually identical, i..e, $\lambda \approx \lambda_{bb}$. The scaling exponents for the MS1 move set differ substantially from the others, which reflects its inherent lack of ergodicity. The results are averaged over 100, 50, 50, 20, 17, 15, and 15 target conformations for $N = 9$, 15, 18, 24, 28, 32, and 40, respectively.

It is well known that MS1 based on single monomer corner and tail flips is not ergodic.[29] Consequently, the folding rates obtained using MS1 are not reliable. For GM1 we computed $\lambda = 3.7 \pm 0.3$, 3.6 ± 0.2, and 3.6 ± 0.2 for MS2, MS3, and SMS, respectively (Figure 4). The power law also holds for folding of the backbone with $\lambda_{bb} \approx \lambda$ at the simulation temperatures T_{min}. Interestingly, exponents λ and λ_{bb} obtained by the three ergodic move sets are almost the same, which establishes the robustness of our results. Because the scaling exponent λ for Go LMSC models is higher than for those without SCs,[11,12] we assume that dense packing of side chains creates additional barriers to folding. We also note that λ for GM1 sequences is in the range proposed by Thirumalai for fast folding sequences.[7]

Side chains alter λ values for GM2 with realistic interactions. It is still possible to fit the folding times obtained for GM2 using a power law with the large exponent $\lambda \approx 6.5 \pm 0.4$, which is similar to that reported for random sequences without side

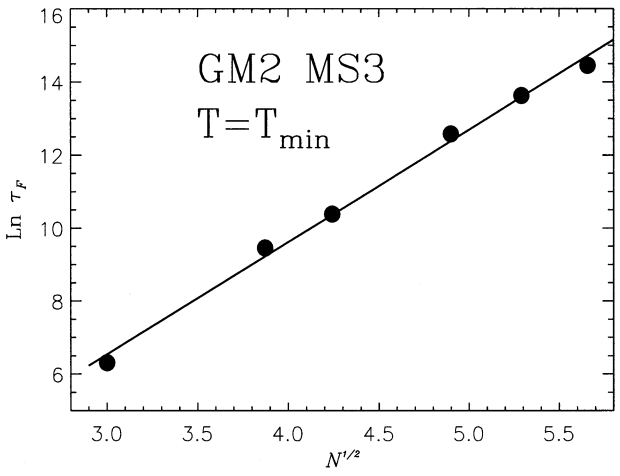

Figure 5. Scaling of τ_F with N for GM2 sequences computed at $T = T_{min}$ using MS3. The solid line represents the exponential fit $e^{\beta\sqrt{N}}$ with $\beta = 3.1 \pm 0.5$. Exponential fit provides a physically sound interpretation for such scaling behavior based on barrier crossing.

chains.[11] Because the same target conformations for both GM1 and GM2 simulations have been used, the large difference in exponents is due to the diversity of interactions. Thus, our study shows that, even though GM2 sequences exhibit two-state folding, heterogeneity in the interactions between side chains (as in GM2) adds roughness to the underlying energy landscape. Furthermore, side chain packing also introduces enhanced topological frustration compared to LMs without side chains. As a consequence scaling behavior of those sequences resembles that of random sequences without side chains.[11] In fact, for random sequences without side chains, Gutin et al had not ruled out the possibility of exponential scaling.[11]

From the physical viewpoint, large λ values are indicative of an activated process with a free energy barrier scaling slower than N.[10] In particular, using the proposal that the activation barrier scales as $\sqrt{N}$,[7] we find that τ_F for GM2 can be fit by $\tau_F \sim e^{\beta\sqrt{N}}$ (see Figure 5). Recent analysis of experimental data also suggests that the barrier height scales as N^α with $\alpha = 0.607 \pm 0.179$,[6] which is consistent with $\sqrt{N}$[7] or $N^{2/3}$[10] scaling.

Although the scaling exponents are the same for all three ergodic move sets, the folding times vary. The dependence of $\tau_F^{SMS}/\tau_F^{MS2}$, and $\tau_F^{SMS}/\tau_F^{MS3}$ on N for GM1 shows that the folding times obtained by the standard move set SMS[20] are about twice as long as those for MS2 and MS3. Because $\tau_F^{SMS}/\tau_F^{MS3} > \tau_F^{SMS}/\tau_F^{MS2}$, we conclude that the MS3 dynamics is the most efficient for folding in LMs. This is not unexpected, because MS3 incorporates flexible choice of multimeric moves, which efficiently sample local conformational space.

Both GM1 and GM2 show that there are no significant differences in the time scales for backbone and side chain folding in the plateau temperature range, i.e., $\tau_F^{bb}/\tau_F \sim O(1)$ (Figure 3). However, outside this temperature range τ_F^{bb} starts to deviate significantly from τ_F. Of particular interest are the temperatures $T < T_F$, where NBA is populated. At low temperatures (T/T_F is relatively small) the backbone ordering occurs considerably faster than does the folding of side chains. The rate determining step is associated with side chain ordering, which might involve transitions over barriers of varying heights. Thus, there may be a relatively narrow temperature window (for example, $0.9 \lesssim T/T_F \lesssim 1.2$ in Figure 2), in which $\tau_F^{bb} \approx \tau_F$.

IV. Conclusions

We have studied the scaling properties of Go lattice sequences with SCs using four different types of Monte Carlo moves. The exponents in the power laws describing the scaling of folding times with sequence length N are sensitive to the ergodicity of the move sets and interaction details.[11] The move set MS3, which is based on flexible selection of multimeric moves, is found to be the most efficient for studying folding in LMs. Strong dependence of folding times for LMSC on sequence length is attributed to side chain packing. The presence of side chains interacting via diverse potentials gives rise to intrinsic roughness in the underlying energy landscape.

Acknowledgment. It is a pleasure to dedicate this paper to John Tully on the occasion of his 60th birthday. Fruitful discussions with M. Betancourt and R. Dima are gratefully acknowledged. This work was supported by KBN (Grant No. 2P03B-146-18) and the grant from the National Science Foundation (CHE02-09340).

References and Notes

(1) Plaxco, K. W.; Simons, K. T.; Baker, D. *J. Mol. Biol.* **1998**, *277*, 985.

(2) Guo, Z.; Thirumalai, D. *Biopolymers* **1995**, *36*, 83.

(3) Clarke, J.; Cota, E.; Fowler, S. B.; Hamill, S. J. *Struct. Fold. Des.* **1999**, *7*, 1145.

(4) Lindberg, M. O.; Tangrot, J.; Otzen, D. E.; Dolgikh, D. A.; Finkelstein, A. V.; Oliveberg, M. *J. Mol. Biol.* **2001**, *314*, 891.

(5) de Gennes, P. G. *Scaling Concepts in Polymer Physics*; Cornell University Press: New York, 1985.

(6) Koga, N.; Takada, S. *J. Mol. Biol.* **2001**, *313*, 171.

(7) Thirumalai, D. *J. Phys. I (France)* **1995**, *5*, 1457.

(8) Takada, S.; Wolynes, P. G. *J. Chem. Phys.* **1997**, *107*, 9585.

(9) Finkelstein, A. V.; Badredtinov, A. Y. *Fold. Des.* **1997**, *2*, 115.

(10) Wolynes, P. G. *Proc. Natl. Acad. Sci., U.S.A.* **1997**, *94*, 6170.

(11) Gutin, A. M.; Abkevich, V. I.; Shakhnovich E. I. *Phys. Rev. Lett.* **1996**, *77*, 5433.

(12) Cieplak, M.; Hoang, X. H.; Li, M. S. *Phys. Rev. Lett.* **1999**, *83*, 1684.

(13) Faisca, P. F. N.; Ball, R. C., cond-mat/0110128.

(14) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. *Proteins Struct. Funct. Genet.* **1995**, *21*, 167.

(15) Abkevich, V.; Mirny, L.; Shakhnovich, E. In *Monte Carlo Approach to Biopolymers and Proteins Folding* Grassberger, P., Barkema, G. T., Nadler, W., Eds.; World Scientific: Singapore, 1998; pp 1−18.

(16) Bromberg, S.; Dill, K. A. *Protein Sci.* **1994**, *3*, 997.

(17) Klimov, D. K.; Thirumalai, D. *Folding Des.* **1998**, *3*, 127.

(18) Klimov, D. K.; Thirumalai, D. *J. Comput. Chem.* **2002**, *23*, 161.

(19) Klimov, D. K.; Thirumalai, D. *Proteins: Struct., Funct. Genet.* **2001**, *43*, 465.

(20) Li, L.; Mirny, L.; Shakhnovich, E. I. *Nature Struct. Biol.* **2000**, *7*, 336.

(21) Go, N. *Annu. Rev. Biophys. Bioeng.* **1983**, *12*, 183.

(22) Li, M. S.; Cieplak, M. *Eur. Phys. J. B* **2000**, *14*, 787.

(23) Clementi, C.; Nymeyer, H.; Onuchic, J. N. *J. Mol. Biol.* **2000**, *298*, 937.

(24) Klimov, D. K.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 2544.

(25) Betancourt, M. R.; Thirumalai, D. *Protein Sci.* **1999**, *8*, 361.

(26) Shakhnovich, E. I.; Gutin, A. M. *Protein Eng.* **1993**, *6*, 793.

(27) Shakhnovich, E. I. *Phys. Rev. Lett.* **1994**, *72*, 3907.

(28) Betancourt, M. R.; Thirumalai, D. *J. Phys. Chem. B* **2002**, *106*, 599.

(29) Hilhorst, H. J.; Deutch J. M. *J. Chem. Phys.* **1975**, *63*, 5153.

(30) Shea, J.-E.; Onuchic, J. N.; Brooks, C. L., III. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 12512.