Chain Length Scaling of Protein Folding Time

A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich

Department of Chemistry, Harvard University, 12 Oxford Street, Cambridge Massachusetts 02138

(Received 17 April 1996; revised manuscript received 20 June 1996)

Folding of protein-like heteropolymers into unique 3D structures is investigated using Monte Carlo simulations on a cubic lattice. We found that the folding time of chains of length N scales as N^{λ} at the temperature of fastest folding. For chains with random sequences of monomers $\lambda \approx 6$, and for chains with sequences designed to provide a pronounced minimum of energy to their ground state conformation $\lambda \approx 4$. Folding at low temperatures exhibits a simple Arrhenius-like behavior. [S0031-9007(96)02004-2]

PACS numbers: 87.15.By, 05.70.Fh, 64.60.Cn, 82.20.Db

Understanding the dynamics of protein folding into unique conformation is one of the major challenges in molecular biophysics and theory of complex systems. The number of all possible conformations (shapes) of a protein of N amino acids scales as z^N, where z is the number of conformations per one amino acid residue. Thus random search in the conformational space for the native, biologically active shape would require folding time $t \sim z^N$, which is unrealistic for a typical protein with $N \sim 100$ (Levinthal paradox).

Many factors affect folding dynamics. One of them, which is relatively well understood, is the dependence of folding time on temperature. At very high temperature all conformations are equally favorable and folding time should indeed be close to Levinthal's estimate z^N. At very low temperature one should expect Arrhenius-like slowing down caused by energy barrier(s) on a folding pathway. Thus there should be some finite temperature optimal for folding in the sense that the process is fastest at this temperature. Such temperature dependence of the folding time was observed indeed in a number of computer simulations for different models [1–3].

Much less is known about the length dependence of folding time. Lessons from other complex systems, such as polymers [4], spin glasses, and random manifolds [5], suggest that size dependence of dynamics is absolutely crucial for the basic understanding of the properties of free energy landscape on which dynamics occur. Experimental data on the length dependence of folding time are not readily available. The problem here is that folding can involve some kinetic events which are specific to a given protein, such as isomerization of prolines and formation of disulphide bonds. In addition, large proteins typically consist of few domains with independent, to some extent, folding.

Some attempts to estimate the folding time from scaling and other simple arguments [6,7] have been made in the past. They were based mostly on phenomenological considerations and employed a number of *ad hoc* assumptions.

More rigorous, microscopic information can be derived from simulations of protein folding. At present such simulations are possible only for highly simplified (mostly lattice) models, a situation common to studies of many complex systems. However, despite their simplicity, lattice models have been useful in elucidating such crucial, experimentally observed features of protein folding as cooperativity [8–10], nucleation mechanisms [11], and parallel pathways [1,3,12,13]

However, even lattice simulations have been limited to very short chains, in most cases not exceeding 27 monomers [3,13,14]. A crucial step has been made when a sequence design procedure was developed for these models, which allowed us to overcome the 27-mer limitation and fold much longer chains [8]. Most importantly, this development made it possible, after all, to address the outstanding issue of the chain length dependence of the folding rate. This is the subject of the present paper.

A protein chain is modeled as a self-avoiding walk on an infinite cubic lattice [1]. Energy of a given conformation of a chain is given by

$$H = \sum_{i<j} U(a_i, a_j)\Delta_{ij}\,, \tag{1}$$

where $\Delta_{ij} = 1$ if monomers i and j are in contact and $\Delta_{ij} = 0$ otherwise. Monomers i and j are considered to be in contact if they are separated by one lattice bond and $|i - j| \neq 1$. The sequence of amino acids is given by a_i with $i = 1, ..., N$, so that a_i can take 20 different values corresponding to 20 natural amino acids. Finally, $U(a, b)$ is the energy of a contact between amino acids a and b [15]. A number of previous studies suggest [11] that the basic results do not depend on a particular choice of interaction matrix. The dynamics of a chain is modeled by a standard Monte Carlo (MC) procedure for a cubic lattice, in which, at each MC step, a monomer is picked up at random and corner flips and crankshaft moves are considered. This algorithm has been discussed extensively in the literature [16],

First, we studied the folding of chains with random sequences of monomers. Five random sequences were generated for each chain length in the range of 10 to

50. A long Monte Carlo simulation was performed for each of these sequences to identify the conformation with the lowest energy. Figure 1 shows the lowest energy conformation for one of the random chains of 40 monomers. In order to make sure that there are no conformations with lower energies, an additional ten runs, starting from different unfolded conformations, were performed for each sequence. In each of these runs a chain folded to the putative lowest energy conformation, and no conformation with lower energy was encountered.

From five random sequences generated for the same chain length, we selected one sequence in which the folding rate was the fastest. Then, for this sequence, we studied folding in a range of temperatures to determine the temperature at which the folding rate was the fastest. Further, at this temperature, we made a more precise estimate of the mean first passage time (MFPT) to the lowest energy conformation by averaging over 50 runs starting from different unfolded conformations. The MFPT for all studied chain lengths is shown in Fig. 2. It is seen that folding time grows with chain length, and the dependence can be well described by a power law that is $t \sim N^{\lambda^{\mathrm{RAN}}}$ with $\lambda^{\mathrm{RAN}} \approx 6$.

Next, we studied the length dependence of the folding rate for sequences designed to have their native conformations as pronounced energy minimum. Such sequences are more likely to exhibit proteinlike behavior [8]. To this end, each of the lowest energy conformations found previously for random sequences was used as a target conformation for sequence design. The design procedure is described in detail elsewhere [17]. We did not study random sequences longer than 50 units, while it is feasible to study folding rate of designed sequences up to 100 units long [8]. To this end we generated random compact conformations to serve as target structures for the sequence design of longer sequences (more than 50 units in length). An additional condition was applied so that variation of energies of native contacts is low enough, in order to eliminate multidomain behavior for longer sequences [17]. In this way five sequences were generated for each studied chain length, and the sequence with the fastest folding was selected. Length dependence of the folding time at optimal temperature for the designed sequences is presented in Fig. 2 for chain lengths ranging from 10 to 100. It can be seen that, for short chains, the folding time of designed sequences is about the same as that of random sequences. For longer chains, though, folding of designed sequences is much faster than folding of random sequences. Overall, length dependence of the folding time for designed sequences is well described by a power law $t \sim N^{\lambda^{\mathrm{MJ}}}$ with $\lambda^{\mathrm{MJ}} \approx 4$.

In order to evaluate how generic this result is we repeated calculations with the other set of parameters, derived by a different procedure [18]. The dependence of the folding rate with this set of parameters (shown in Fig. 2) also fits well by power-law dependence with slightly higher exponent $\lambda^{\mathrm{MS}} \approx 4.5$.

For comparison, we also studied the model introduced by Go and co-workers [19]. In this model, energy of a given conformation is given by

$$H = -\frac{1}{2}\sum_{i<j} \Delta_{ij}^{N}\Delta_{ij}\,, \tag{2}$$

where $\Delta_{ij}^{N} = 1$ if monomers i and j are in contact in the native conformation and $\Delta_{ij}^{N} = 0$ otherwise. In other words, in the Go model all the native interactions are favorable and all non-native interactions do not contribute

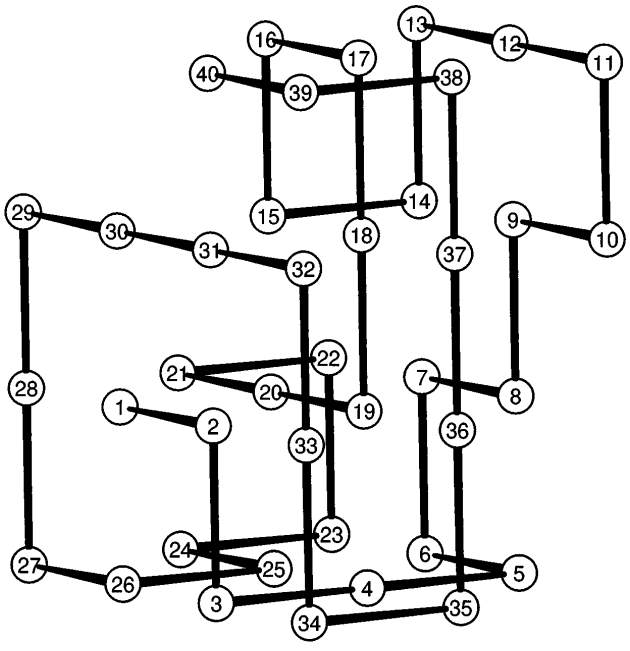

FIG. 1. Lowest energy conformation for a random sequence of 40 amino acid residues.

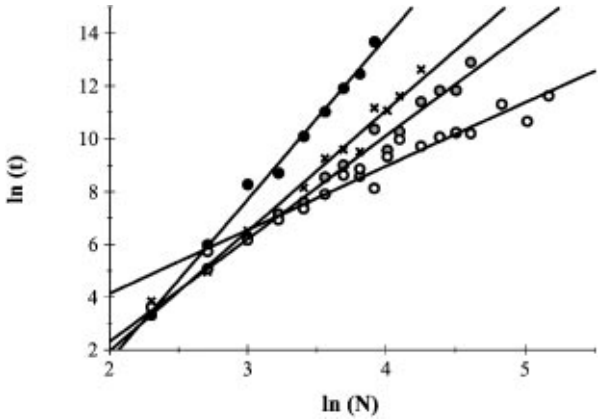

FIG. 2. Dependence of the folding time t on the chain length N for random sequences (black circles), for sequences designed using potentials from [18] (+), for sequences designed with potentials from [15] (gray circles), and for the Go model (white circles). Folding time was estimated by averaging over 50 folding runs for each of the sequences.

to the energy. In some sense, the Go model corresponds to ideally designed sequences.

For the Go model the same conformations as for designed sequences were used as native. Length dependence of the folding time at optimal temperature for the Go model is presented in Fig. 2 for chain lengths from 10 to 175. Again, the dependence is fitted very well by a power law $t \sim N^{\lambda^{\mathrm{Go}}}$ with $\lambda^{\mathrm{Go}} \approx 2.7$.

Figure 3 shows the inverse temperature dependence of the folding time for the sequence designed to fold to the native conformation shown in Fig. 1. It has a clear minimum at some optimal temperature. All the scaling dependencies for the folding rate discussed above were obtained at the conditions corresponding, for each sequence, to such an optimal temperature. What happens at lower temperatures? It is clear from Fig. 3 that at low temperatures the logarithm of folding time depends linearly on inverse temperature exhibiting an Arrhenius-like behavior $t \sim \exp(E_b/T)$. This suggests that at low temperature folding is an activated process which entails overcoming of an energetic barrier E_b. Similar dependencies were obtained for all studied chains, and the corresponding energy barriers E_b were determined for each sequence from the slopes of Arrhenius-like branches of these dependencies. The analysis shows that such defined energy barrier scales linearly with energy of the native conformation with coefficient $\phi \approx 0.18$, suggesting that at low temperature the chain length scaling of folding time may become exponential [20]. However, we cannot exclude that the apparent energy barrier which determines folding rates at low temperature, and even its chain length scaling, may be due to short-scale effects, such as the braking of incorrectly formed contacts. These may depend crucially on the detail of the model and simulation technique involved. Thus we cannot exclude that folding rates determined at low temperature may be very sensitive to details of the model, and therefore the conclusion about exponential chain length scaling of the folding rate at low temperature should be taken with caution.

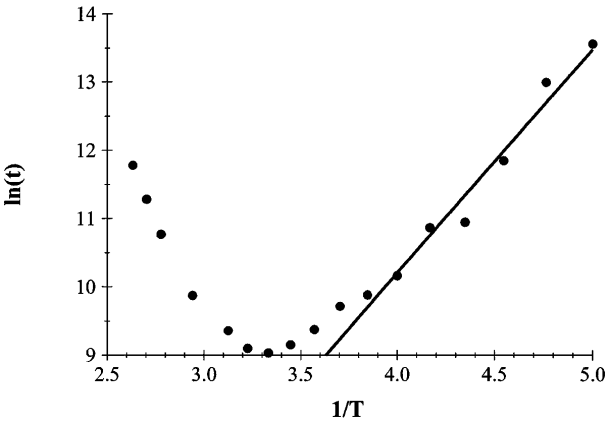

FIG. 3. Dependence of the folding time t on the temperature T for a random sequence of 40 residues with the ground state shown in Fig. 1. Straight line approximates the low-temperature part of the dependence. The slope of the line is the energy barrier E_b.

As was pointed out earlier, the direct comparison between our main result and experiments on real proteins is a challenging task for experimentalists. This will require a careful choice of proteins to study as model systems and conditions at which their folding is fastest. We hope that the presented results will stimulate such experiments.

The major finding of this work is that the chain length dependence of the folding rate is relatively weak (much weaker than exponential, suggested by Levinthal's argument and some phenomenological estimates [3,6]). This fact is of fundamental importance since it explains, from the folding perspective, the wide range of lengths of existing proteins (roughly 50–1000 amino acids). Exponential dependence of folding time on protein length would make folding of longer chains prohibitively slow. Another important application of the present result is that it can provide a crucial and delicate test of the existing and future kinetic theories. Phenomenological analysis based on the random energy model (REM) predicts exponential dependence of folding rate on system size at all temperatures [3,6]. This is in contrast with simulation results and clearly points out the inapplicability of the REM (and perhaps mean-field models in general) to tackle kinetic issues.

This can be understood from the general perspective that folding transition is cooperative (first-order-like). Such transitions follow the nucleation mechanism at which the transition state is highly inhomogeneous, consisting of islands of the "new" phase in the sea of "old" phase [21]. Such an inhomogeneous distribution (representing the least-activation path) explains the weak size dependence of the rate of first-order transitions. It cannot be described by any global homogeneous order parameter. The number of native-like contacts Q was used as a kinetic reaction coordinate in [13,22]. Such analysis does not take into account the inhomogeneity of the transition state and thus predicts exponential size dependence of protein folding rate. The inapplicability of the REM-based and other mean-field approaches (e.g., using Q as kinetic reaction coordinate) for longer chains has been noted before [3,13,22]. However, it is not clear what is the maximal chain length up to which dynamics can be satisfactorily described by a mean-field approach. We do not see a pronounced change of dynamic behavior from shortest to longest chains (see Fig. 2). Further, recent analysis (Pande, Grosberg, and Shakhnovich, unpublished results) indicated that Q is not a good reaction coordinate for folding kinetics of chains as short as 18 monomers. These results suggest that the REM-based or other mean-field analysis may be inapplicable to study dynamics of even very short chains.

In this work we studied sequences with different degrees of design, from "ideally" designed (Go model) to random (no design at all). The quantitative measure of the

degree of design is relative energy $z = (E_{nat} - E_{av})/\sigma$, where E_{nat} is energy of the native conformation, E_{av} is average energy of a misfolded conformation, and σ is standard variance of interaction strengths. Interestingly, the exponent of power-law dependence increased monotonically with z, i.e., $\lambda^{Go} < \lambda^{MJ} < \lambda^{MS} < \lambda^{RAN}$ and $z^{Go} < z^{MJ} < z^{MS} < z^{RAN}$ (we note that z is negative). This observation may provide a valuable hint for future theory of folding dynamics.

A possible physical explanation of the power-law rate dependence for designed sequences is close to the arguments presented recently by Thirumalai [7]. Since folding is a cooperative (first-order-like) process [8,9], the transition state is reached when nucleus of the folded conformation is formed [11]. The power-law length scaling of the folding rate implies that the nucleus does not grow with the chain length. In this case the length dependence of the folding rate appears to be due to the entropic cost of loop closure around the folding nucleus. The free energy of loop closure depends logarithmically on its length [4], and this factor, combined with the power-law length dependence of polymer relaxation time [4], apparently translates into the overall power-law dependence of the folding rate.

The difference between random and designed sequences is crucial. First of all we see that the designed sequences fold faster at their respective optimal folding temperatures and the difference in folding rates between random and designed sequences becomes more pronounced as chains become longer (due to different exponents in the scaling laws describing length dependencies).

Real proteins must not only reach their native state fast but also be stable in their native conformations. Designed sequences are stable in the native state in the whole range of studied lengths at the temperature of their fastest folding, while random sequences get more and more unstable as the chain length is increased. (Data not shown.) This is consistent with earlier hypotheses [19,23] that protein sequences must have been selected to satisfy "maximal consistency" or "minimal frustrations" requirements. The incarnation of these requirements in our model is via sequence design.

Finally, we note that while power-law scaling fits our data best, the range of chain lengths, which we studied, spanned only about an order of magnitude. This length range is most limited for random sequences; for them we cannot rule out a weak stretched exponential, rather than power length scaling [7]. We believe that microscopic analytical theory of protein folding dynamics can give satisfactory answer to the important question of length dependence of protein folding time. It is our hope that presented results will stimulate the development of such theory.

This work was supported by the David and Lucille Packard Foundation and NIH. Interesting discussions with D. Thirumalai and A. Grosberg are gratefully acknowledged.

[1] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, J. Chem. Phys. **101**, 5062 (1994).
[2] H. S. Chan and K. A. Dill, J. Chem. Phys. **100**, 9238 (1994).
[3] J. D. Bryngelson *et al.*, Proteins **21**, 167 (1995).
[4] P. G. DeGennes *Scaling Concepts in Polymer Physics* (Cornell Univ. Press, Ithaca, New York, 1979).
[5] K. Binder and A. Young, Rev. Mod. Phys. **58**, 801 (1986); V. Vinokur, M. C. Marchetti, and L. W. Chen, Phys. Rev. Lett. **77**, 1845 (1996).
[6] J. D. Bryngelson and P. G. Wolynes, J. Phys. Chem. **93**, 6902 (1989); E. Shakhnovich and A. M. Gutin, Europhys. Lett. **9**, 569 (1989); J. Saven, J. Wang, and P. Wolynes, J. Chem. Phys. **101**, 11 037 (1994).
[7] D. Thirumalai, J. Phys. I (France) **5**, 1457 (1995).
[8] E. I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).
[9] M.-H. Hao and H. Scheraga, J. Phys. Chem. **98**, 4940 (1994).
[10] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, J. Mol. Biol. **252**, 460 (1995).
[11] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, Biochemistry **33**, 10 026 (1994); E. I. Shakhnovich, V. I. Abkevich, and O. B. Ptitsyn, Nature (London) **379**, 96 (1996).
[12] S. E. Radford and C. M. Dobson, Philos. Trans. R. Soc. London B **348**, 17 (1995).
[13] A. Sali, E. Shakhnovich, and M. Karplus, Nature (London) **369**, 248 (1994).
[14] E. Shakhnovich *et al.*, Phys. Rev. Lett. **67**, 1665 (1991); N. Socci and J. Onuchic, J. Chem. Phys. **101**, 1519 (1994).
[15] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985).
[16] P. H. Verdier, J. Chem. Phys. **59**, 6119 (1973); H. J. Hilhorst and J. M. Deutch, J. Chem. Phys. **63**, 5153 (1975); A. Sali, E. I. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).
[17] E. I. Shakhnovich and A. M. Gutin, Proc. Natl. Acad. Sci. U.S.A. **90**, 7195 (1993); V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, Folding & Design **1**, 221 (1996).
[18] L. Mirny and E. Shakhnovich, J. Mol. Biol. (to be published).
[19] H. Taketomi, Y. Ueda, and N. Go, Int. J. Peptide Protein Res. **7**, 445 (1975).
[20] D. Takada and P. Wolynes (to be published).
[21] E. M. Lifshitz and L. P. Pitaevskii, *Physical Kinetics* (Pergamon, London, 1981).
[22] P. Wolynes, J. Onuchic, and D. Thirumalai, Science **267**, 1619 (1995); N. Socci, J. Onuchic, and P. Wolynes, J. Chem. Phys. **104**, 5860 (1996).
[23] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987).