Criterion that Determines the Foldability of Proteins

D. K. Klimov and D. Thirumalai

Institute for Physical Science and Technology and Department of Chemistry and Biochemistry, University of Maryland, College Park, Maryland 20742

(Received 4 December 1995)

We show, using lattice models of proteins, how the kinetic accessibility of the native state of proteins is encoded in the primary sequence itself. The folding times for various sequences correlate extremely well with the single parameter intrinsic to the sequence, namely, $\sigma = |T_\theta - T_f|/T_\theta$ where T_θ and T_f are the collapse and folding transition temperatures. Fast folding sequences have small values of σ (typically less than 0.1) whereas σ values for slow folding sequences exceed 0.6. The folding times change by 5 orders of magnitude when σ goes from 0.05 to a value exceeding about 0.6. [S0031-9007(96)00175-5]

PACS numbers: 87.15.By, 61.41.+e, 64.60.Cn, 87.10.+e

The Levinthal argument [1] that the time scale for random search of all possible conformations of a polypeptide chain would be incompatible with biological folding times has served as a major intellectual challenge to understanding the kinetics of folding of proteins. The implication of the Levinthal paradox is that since the underlying energy landscape contains numerous minima separated by barriers of varying heights how can a polypeptide chain, on a relatively fast time scale, reach the unique native conformation [2]. Thus it is natural to wonder if proteins have been designed so that the requirement of kinetic foldability and the associated stability of the native state have somehow been simultaneously satisfied. If this were the case then a logical conclusion is that the requirements for the kinetic accessibility of the native state should be encoded in the primary sequence itself. Furthermore, because protein folding is a self-assembly process then these requirement(s) should be describable in terms of the properties intrinsic to the sequence. This reasoning would further indicate that thermodynamic properties associated with sequences may well dictate the overall kinetic foldability of proteins.

In recent years many aspects of protein folding kinetics have been predicted using very simple representations of proteinlike features [3–13]. In this Letter we use lattice models and Monte Carlo simulations to show that foldability of sequences with a unique native state is entirely determined by two key temperatures that are intrinsic to the sequence. The principal results of this study, which were obtained by systematically investigating about sixty sequences, follow.

(a) The spectra of sequences with a unique ground state show that there are many low energy states that are noncompact. These make significant contributions to thermodynamic properties and are also involved in the folding kinetics.

(b) Folding in this model is essentially determined by two characteristic temperatures, which are intrinsic properties of the sequence. One of them is the collapse transition temperature T_θ, above which the chain is in an extended random coil state. The other is the folding transition temperature T_f, at which the chain undergoes a first order transition to the folded state. In particular, folding times correlate extremely well with the dimensionless parameter $\sigma = |T_\theta - T_f|/T_\theta$. Extremely fast folding sequences have small values of σ (≤ 0.1) while slow folders are characterized by $\sigma > 0.6$.

(c) There appears to be no useful correlation between folding times and the energy gap between the native conformation and the first excited state. There are a large number of sequences, all with roughly the same folding times, but with very different values of the energy gap.

In order to establish the results quoted above we have considered a simple lattice model of a protein. In the lattice model the polypeptide chain is taken to be a chain of N connected beads that are confined to the vertices of a cubic lattice. The self-avoidance criterion is satisfied by the restriction that each lattice site be occupied at most once. Conformation of the chain is described by the coordinates r_i ($i = 0, 1, 2, \ldots, N$) which are the positions on the lattice. A topological contact in this lattice is formed whenever two nonbonded beads i and j ($|i - j| \geq 3$) are nearest neighbors on the cubic lattice, i.e., $r_{ij} \equiv |r_i - r_j| = a$, the lattice spacing. The energy of a given conformation is assumed to be the sum of the energies associated with each topological contact

$$E = \sum_{i<j+3} B_{ij}\delta(r_{ij} - a), \qquad (1)$$

where B_{ij} is the strength of the interaction energy between the beads i and j, and $\delta(0) = 1$ and 0, otherwise. The heterogeneity of interactions in proteins is mimicked by assuming that the interaction energies B_{ij} have a Gaussian distribution [10,11]

$$P(B_{ij}) = \frac{1}{(2\pi B^2)^{1/2}} \exp\left(-\frac{(B_{ij} - B_0)^2}{2B^2}\right), \qquad (2)$$

where B_0 is the mean value and B is the standard deviation. Energies are measured in units of B. We

 0031-9007/96/76(21)/4070(4)$10.00

consider two values of N and two values of B_0. The values of N are 15 and 27 and the values of B_0 are -0.1 and -2.0. The smaller value of N was chosen so that all allowed conformations of the chain could be explicitly enumerated so that the thermodynamic properties can be determined exactly. The value of $B_0 = -0.1$ makes the fraction of hydrophobic residues to be around 0.55, which is close to that found in natural proteins. For sequences with $B_0 = -2.0$ the low energy spectrum is expected to be dominated by compact structures. This allows us to compare the folding kinetics as a function of the size of the conformation space.

For these models a sequence is specified by a set of random numbers drawn from the distribution function given in Eq. (2). We generated one-third of all sequences totally randomly. The remaining sequences were optimized [13] by performing Monte Carlo simulations in sequence space using the ground state of randomly selected sequences as target structures. This method of creating sequences, which gave rise to a range of energy gap Δ, generated native states with different topology. For a given sequence we determined the spectrum of the low energy states. For $N = 15$ this was done by enumerating all the allowed conformations which is on the order of 10^8. The spectra for sequences with $N = 27$ had to be determined numerically. This was done using the following procedure. All the sequences for $N = 27$ were optimized by initially running Monte Carlo simulations in sequence space [13]. This yields ground state structures which are (in all likelihood) the lowest energy structures. In order to obtain other low energy states we heated the system to a high enough temperature and slowly cooled it to a very low temperature. During the course of slow cooling we monitored the energy $E(t, T_i)$ as a function of time t and temperature T_i at the ith cooling step. The energies are arranged in increasing order yielding spectrum for a given sequence. In order to ensure that the calculated spectrum is an equilibrium spectrum this procedure was repeated for several (typically 100) initial conditions and in each instance we obtained identical values of the energies. It should be stressed that this method is not very accurate for obtaining high energy states which are expected to be spaced close together.

The low energy spectra for ten sequences with $N = 27$ and $B_0 = -0.1$ are displayed in Fig. 1. For each sequence there are two columns. The left side of each column corresponds to the energy levels obtained from the slow cooling Monte Carlo simulation described above and the right hand column corresponds to the exact spectrum obtained by retaining only the compact structures. The values of the gap Δ are listed below each column. It is clear that for some sequences the lowest energy in the spectrum of compact structures is near the continuum of the true spectrum. This figure shows that the gap calculated using the compact structure alone exceeds the true gap provided the native conformation is itself compact. This is because for $N = 27$ small excitations

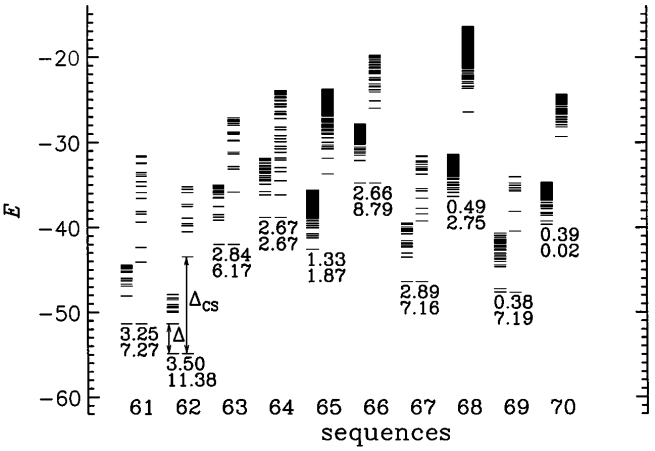

FIG. 1. Lower part of energy spectra for ten optimized sequences ($N = 27$, $B_0 = -0.1$). The left columns of each sequence are the energy levels obtained from slow cooling Monte Carlo simulations. The columns on the right represent the energy spectra of compact structures. Below each pair of spectrum columns the energy gap Δ calculated from Monte Carlo simulations (upper quantity) and Δ_{CS} calculated from the spectra of compact structures (lower quantity) are indicated.

are easily created by flipping one surface bond and this would result in a noncompact structure. Thus, unless the value of $|B_0|$ is artificially high, one expects that several low energy structures would be noncompact, and these would play an important role in the kinetics and thermodynamics of folding.

The kinetics of folding in these models for both $N = 15$ and 27 was determined using the Monte Carlo method. Briefly, we started from an initial random conformation at a high temperature and quenched the system to a temperature below the folding transition temperature. Many different correlation functions were used to probe the kinetics of approach to the native conformation starting from an ensemble of independent initial high temperature conformations. In this Letter the kinetics are discussed using the structural overlap function $\chi(t)$ which for a given initial trajectory labeled k is given by [9]

$$\chi^k(t) = 1 - \frac{1}{N^2 - 3N + 2} \sum_{i \neq j, j \pm 1} \delta(r_{ij}^k - r_{ij}^N), \quad (3)$$

where r_{ij}^N refers to the coordinates of the lowest energy (native) state and $\delta(x)$ is the Kronecker delta function. The kinetic behavior is inferred by averaging $\chi^k(t)$ over an ensemble of initial conditions. In the simulations we used between 100 and 800 initial trajectories depending on the sequence and the temperature. The collapse transition temperature T_θ is obtained from the peak in the temperature dependence of the specific heat curve while T_f is calculated from the peak in the fluctuation in the structural overlap function [9], $\Delta\chi = \langle\chi^2\rangle - \langle\chi\rangle^2$, where the angular brackets indicate thermodynamic averages. It has been demonstrated that this method

of determining T_θ coincides with the temperature at which the shape of the chain changes from a random coil to a compact structure [14]. The folding times were calculated by a quantitative analysis of the time dependence of approach to the native conformation. For all the sequences examined we find that, after a transient time, $\chi(t)$ is well fit by a sum of exponentials (one, two, or three). The folding time is taken to be the longest time in such fits. Other methods of calculating the folding times, such as the mean first passage time, yield roughly the same trends.

The simulation temperature T_s for each sequence was determined as follows. Since we want to compare folding times for a variety of sequences, it is desirable to subject them to identical folding conditions which in our simulations are determined by the temperature. We chose to run our Monte Carlo simulations at a sequence dependent temperature T_s that is subject to two conditions: (a) T_s must be less than T_f so that the native conformation has the highest occupation probability; (b) the values of $\langle \chi(T_s) \rangle$ must be a constant for all sequences. This allows us to compare all sequences on equal footing.

The major purpose of this Letter is to search for the intrinsic properties of sequences which effectively control the folding times. Since there are only two equilibrium characteristic temperatures, it is likely that the folding times may be determined by these temperature scales. In Fig. 2 we plot the folding times τ_f as a function of σ. The upper panel gives τ_f for $N = 15$ while the lower one is for $N = 27$. The figure shows that the folding time correlates extremely well with the intrinsic thermodynamic parameter σ. The sequences span a range of σ and consequently meaningful conclusions can be drawn. Sequences with small values of σ fold extremely rapidly. The folding times for $N = 15$ and $\sigma \leq 0.1$ rarely exceed 10^4 Monte Carlo steps (MCS) whereas sequences with σ exceeding 0.6 have folding times in excess of 10^7 MCS. For $N = 27$ all the sequences were optimized so that the value of σ does not exceed 0.15. The qualitative trends, namely, sequences with small values of σ fold rapidly while those with larger values of σ have larger folding times, remains unchanged. It is remarkable that the folding time increases dramatically (by several orders of magnitude) when σ changes in the range from 0.05 to 0.6. Thus, it appears that the single parameter σ determines the foldability of the sequence. It is natural to suggest that the kinetic accessibility in real proteins may therefore by encoded in the primary sequence. Physically, if σ is small, then $T_\theta \approx T_f$ and as a result all possible transient structures that are explored by the chain enroute to the native state are of relatively high free energy. Consequently, the usual barriers are very small when σ is small and folding to the native state is very rapid [15].

It has been suggested in the literature that rapid folding of sequences for the model of the sort studied here depends

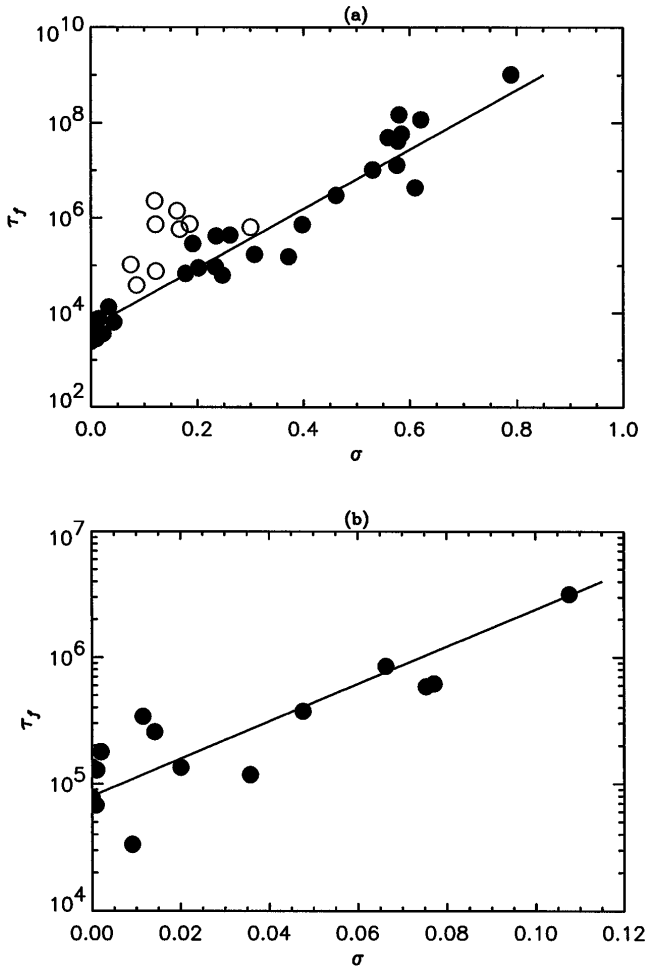

FIG. 2. Dependence of the folding time τ_f on $\sigma = |T_\theta - T_f|/T_\theta$ for a number of sequences. The upper panel (a) corresponds to $N = 15$. The full circles correspond to $B_0 = -0.1$ while the empty circles are for $B_0 = -2.0$. The lower panel (b) is given for $N = 27$, $B_0 = -0.1$. The solid lines are a guide to the eye. The data appear to be consistent with an exponential although other forms cannot be ruled out [15].

exclusively on the energy gap Δ between the native state and the first excited state [11]. We assume that any intrinsic parameter that controls folding times should be dimensionless, and consequently we express Δ in units of the simulation temperature of each sequence. In Fig. 3 we present the folding times τ_f as a function of Δ/T_s. The upper panel of Fig. 3 is for $N = 15$ while the lower panel is for $N = 27$. These figures show in a dramatic fashion the irrelevance of Δ/T_s in determining folding times. It is obvious from Fig. 3 that it is possible to generate arbitrarily a large number of sequences with varying values of Δ all of which will have approximately the same folding times. Thus, it is clear that energy gap alone cannot determine folding times in these models [16].

There are a number of implications of this study for real proteins. (1) Whenever σ is less than about 0.1

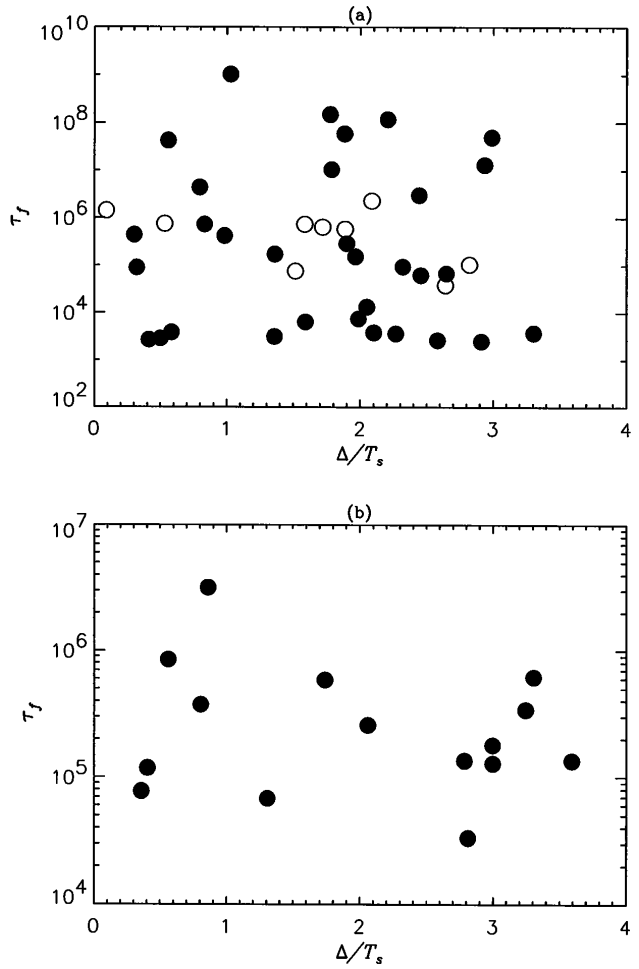

FIG. 3. Plots of the folding time τ_f as a function of the energy gap Δ divided by the simulation temperature T_s. (a) The solid circles correspond to $B_0 = -0.1$ whereas the open circles are for $B_0 = -2.0$. The value of N is 15. (b) This plot is for $N = 27$, $B_0 = -0.1$.

(the precise value depends on the length of the sequence) the folding kinetics appears to be an all-or-none process. In these cases we find that the native conformation is reached by a specific nucleation collapse mechanism, i.e., the folding process and the initial collapse of the polypeptide chain are almost synchronous [15]. Such sequences are highly optimized and in general we expect this behavior to be seen in relatively small proteins. The fast folders can be made to reach the native conformation at relatively high temperatures. (2) Moderate folders (see Fig. 2) have σ values in the range between 0.1 and 0.6. Thus for $\sigma = 0.4$ the folding transition temperature is roughly 36 °C assuming that $T_\theta \approx 60$ °C. Since moderate folders can be created in a more random manner, it is tempting to speculate that longer proteins in nature are not fully optimized. It is likely that these are the proteins which obey the minimal frustration principle [4]. These observations suggest that the simultaneous requirement of kinetic accessibility and the associated thermodynamic stability can be achieved by both fast and moderate folding proteins. (3) From a *de novo* design point of view it appears that one should synthesize proteins so that the σ values lie below about 0.6. For a given sequence both T_θ and T_f are measurable so that by a process of iteration one could arrive at the desired sequence. It would be desirable to understand theoretically the precise factors that determine σ.

In summary, using the random heteropolymer models we have shown that the kinetic accessibility of the native states in proteins may be encoded in the primary sequence. It is remarkable that the single parameter that characterizes the folding kinetics is determined solely by the collapse transition temperature and the folding transition temperature. If the protein sequence and the associated effective potential of mean force are given, these characteristic temperatures can be calculated. Alternatively experiments can be used to obtain T_θ and T_f. Because these temperatures can be altered (by mutations) one can use this encodability aspect to control folding in proteins and presumably other biomolecules.

This work was supported by the Air Force Office of Scientific Research (Grant No. F496209410106) and the National Science Foundation. We are grateful to J. N. Onuchic and P. G. Wolynes for interesting discussions.

[1] C. Levinthal, in *Mossbauer Spectroscopy in Biological Systems*, edited by P. Debrunner, J. C. M. Tsibris, and E. Münck (University of Illinois Press, Urbana, 1968).
[2] C. B. Anfinsen, Science **181**, 223 (1973).
[3] J. Skolnick and A. Kolinski, Science **250**, 1121 (1990).
[4] J. D. Bryngelson and P. G. Wolynes, J. Phys. Chem. **93**, 6902 (1989).
[5] P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **89**, 8721 (1992).
[6] T. R. Garel and H. Orland, Europhys. Lett. **6**, 307 (1988).
[7] R. Zwanzig, Proc. Natl. Acad. Sci. U.S.A. **92**, 9801 (1995).
[8] H. Chan and K. A. Dill, J. Chem. Phys. **99**, 2116 (1993); **100**, 9238 (1994).
[9] C. J. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **90**, 6369 (1993).
[10] E. Shakhnovich G. Farztdinov, A. M. Gutin, and M. Karplus, Phys. Rev. Lett. **67**, 1665 (1991).
[11] A. Sali, E. Shakhnovich, and M. Karplus, Nature (London) **369**, 248 (1994).
[12] N. D. Socci and J. N. Onuchic, J. Chem. Phys. **101**, 1519 (1994); **103**, 4732 (1995).
[13] E. Shakhnovich and A. M. Gutin, Protein Eng. **6**, 793 (1993); E. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).
[14] J. D. Honeycutt and D. Thirumalai, Biopolymers **32**, 695 (1992).
[15] D. Thirumalai, J. Phys. I (France) **5**, 1457 (1995).
[16] H. S. Chan, Nature (London) **373**, 664 (1995).