Factors That Affect the Folding Ability of Proteins

Aaron R. Dinner,1,2 Victor Abkevich,1 Eugene Shakhnovich,1,2 and Martin Karplus1,2,3*
1Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts
2Committee on Higher Degrees in Biophysics, Harvard University, Cambridge, Massachusetts
3Laboratoire de Chimie Biophysique, Institut le Bel, Université Louis Pasteur, Strasbourg, France

ABSTRACT The folding ability of a heteropolymer model for proteins subject to Monte Carlo dynamics on a simple cubic lattice is shown to be strongly correlated with the stability of the native state. We consider a number of estimates of the stability that can be determined without simulations, including the energy gap between the native state and the structurally dissimilar part of the spectrum (Z score) and, for sequences with fully compact native states, the gap in energy between the native and first excited fully compact states. These estimates are found to be more robust predictors of folding ability than a parameter $s$ that requires simulation for its evaluation: $s = 1 - T_f/T_g$, where $T_f$ is the temperature at which the fluctuation of an order parameter is at its maximum and $T_g$ is the temperature at which the specific heat is at its maximum. We show that the interpretation of $T_g$ as the collapse transition temperature is not correct in general and that the correlation between $s$ and the folding ability arises from the fact that $s$ is related to the energy gap (Z score). Proteins 1999;35:34-40. © 1999 Wiley-Liss, Inc.

Key words: heteropolymer; lattice model; Monte Carlo; protein folding

To function as a protein, a polypeptide chain must possess a unique native state that is stable at a physiological temperature and be able to find that state in a reasonable time (milliseconds to minutes) at that temperature in spite of the fact that the number of possible configurations of the chain is too large for an exhaustive search (Levinthal paradox).1 For understanding the mechanism of folding and for protein design, it is important to determine the properties that distinguish polypeptide sequences that satisfy these requirements from those that do not. One approach is to compare sequence attributes with folding ability. Analytical heteropolymer theory based on a random energy model and supplemented with certain kinetic assumptions suggested that, for a protein to fold, its native state must be stable at a temperature high enough to readily overcome barriers on the rough energy landscape. This criterion for folding was expressed by requiring that the folding transition temperature ($T_f)$, at which point during cooling the ground (native) state becomes the global free energy minimum, be higher than a “glass” transition temperature ($T_g$), at which point the dynamics were assumed to become critically slow.2,3 (but see Gutin et al.4). A related formulation of this criterion, referred to as the energy gap criterion, stipulates that a protein must have a large gap in energy between the native state and the states (folds) that are structurally dissimilar.5 Šali, Shakhnovich, and Karplus (SSK) confirmed these ideas for simple models by demonstrating computationally that a large energy gap promotes fast folding for a 27-mer random heteropolymer model subject to Monte Carlo (MC) dynamics on a simple cubic lattice.6,7 Moreover, SSK showed that, when the ground state is maximally compact (a $3 \times 3 \times 3$ cube), the energy gap criterion can be simplified to a consideration of the difference in energy between the ground state and the first fully compact excited state.6 This follows from the fact that the first fully compact excited state generally differs significantly in structure from the ground state; the simplified criterion is not valid for noncompact states where the first excited state is typically almost identical to the ground state. The results and conclusions of SSK have been used in the design of fast folding sequences8 and are consistent with similar studies which focus on exhaustive enumeration of folding paths for two-dimensional chains9,10 or on the ratio $T_f/T_g$ for the 27-mer.11

Based on results obtained with a set of 15-mer and 27-mer sequences, Klimov and Thirumalai (KT) argued that folding ability is not determined by the energy gap but instead by the parameter $s = 1 - T_f/T_g$, where $T_f$ is the “folding” transition temperature (the maximum in the fluctuation of the order parameter defined in Eq. 3), and $T_g$ is the “collapse” transition temperature (defined as the maximum in the specific heat, $C_v$).12,13 KT found that $s$ correlates positively with the folding time (small $s$ yields fast folding). In the present paper, we show that the maximum in the specific heat ($T_g$) is not the temperature of the collapse transition in general and that, when the collapse transition does coincide with the maximum in

Grant sponsor: National Institutes of Health (E.S.), National Science Foundation; Grant number: GM 52126
*Correspondence to: Martin Karplus, Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138. E-mail: mkarplus@tammy.harvard.edu
Received 8 May 1998; Accepted 12 October 1998
FACTORS THAT AFFECT FOLDING ABILITY

$C_v(T)$ (i.e., two-state folding), $\sigma$ is related to the energy gap. In light of these considerations, we reexamine the results for the sequences studied by KT. We find that other measures of the native state stability and the energy gap, such as the $Z$ score, correlate more strongly with folding ability than does the parameter $\sigma$, and that these correlations are robust with respect to changes in the simulation temperature. Moreover, in contrast to $\sigma$, the measures of the energy gap can be estimated without simulation and thus have predictive value.

The specific model is a self-avoiding heteropolymer chain on a three-dimensional simple cubic lattice. The energy function is the sum over all contacts (nonbonded spatial nearest-neighbors):

$$E = \sum_{i,j=1}^{N} B_{ij} \Delta(r_{ij} - a)$$  \hspace{1cm} (1)

where $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ is the distance between residues $i$ and $j$ located at $\mathbf{r}_i$ and $\mathbf{r}_j$, respectively, and $a = 1$ is the lattice spacing. The function $\Delta(r_{ij} - a)$ selects the interacting residue pairs; $\Delta(0) = 1$ and 0 otherwise. The $B_{ij}$ give the specific interaction energies between residues $i$ and $j$; a complete set of $B_{ij}$ defines a sequence. The $B_{ij}$ are chosen randomly with gaussian probability distribution:

$$P(B_{ij}) = \frac{1}{\sqrt{2\pi B}} \exp \left[ -\frac{1}{2} \frac{(B_{ij} - B)^2}{B} \right].$$  \hspace{1cm} (2)

The quantity $B_0$ ($B_0 < 0$) is a mean attraction between residues that corresponds to an overall hydrophobic term. The 200 27-mer sequences studied in Šali et al. were generated with $B_0 = -2.0$ because that value guaranteed that most sequences had maximally compact ground states which could be found by examining the enumerated ensemble of 103,346 maximally compact structures. In the present study, we use the same sequences as in Refs. 12 and 13; there are nine 15-mers with $B_0 = -2.0$, thirty-two 15-mers with $B_0 = 0.1$, twenty 27-mers with $B_0 = 2.0$, and fifteen 27-mers with $B_0 = -0.1$. KT included the value $B_0 = 0.1$ because it felt that it produced a more complete set of specific interaction energies between residues; that most sequences had maximally compact ground states differing substantially from $B_0$ in the 58 sequences studied.

$-0.805 \leq B_{ij} \leq 0.690$ for sequences with $B_0 = -0.1$ and $-2.413 \leq B_{ij} \leq -1.708$ for sequences with $B_0 = -2.0$.

We follow KT and use as an order parameter the complement of native pairwise distances:

$$\chi = 1 - \frac{1}{(N-2)(N-1)} \sum_{i,j=1}^{N} \Delta(r_{ij} - r_{ij}^0)$$  \hspace{1cm} (3)

where $r_{ij}^0$ is the distance between $i$ and $j$ in the ground state, and we use as a measure of compactness the total number of contacts ($C_v$). Boltzmann weighted averages of these quantities were calculated exactly for the 15-mer (all 93,250,730 self-avoiding conformations can be enumerated) and by Monte Carlo simulation for the 27-mer (see Ref. 13 for the protocol). Following KT, the “folding” transition temperature ($T_f$) is taken to be the temperature at which the fluctuation in the order parameter $\Delta \chi(T) = \langle \chi^2 \rangle - \langle \chi \rangle^2$ has a maximum, and the “collapse” transition temperature ($T_c$) is taken to be the temperature at which the heat capacity $C_v(T) = \Delta E^2/T^2 = \langle (E^2) - \langle E \rangle^2 \rangle/T^2$ has a maximum (units are chosen to make the Boltzmann constant equal to 1). There is one 15-mer sequence with $B_0 = -0.1$ that exhibited two peaks in $C_v(T)$, three that exhibited shoulders with derivatives close to zero in $C_v(T)$, and one that exhibited a shoulder with a derivative close to zero in $\Delta \chi(T)$; in these cases, for consistency with KT, we calculated the corresponding transition temperature ($T_{f\ell}$ or $T_{c\ell}$) by a weighted average of the two temperatures corresponding to the zero derivative points (see Eq. 10 of Klimov and Thirumalai).

From $T_f$ and $T_{c\ell}$, we can calculate the parameter $\sigma = 1 - T_{c\ell}/T_{f\ell}$, which was argued by KT to determine folding ability. To better understand $T_f$, $T_{c\ell}$, and $\sigma$, we examine the heat capacity ($C_v$), the fluctuation in the order parameter ($\Delta \chi$), the fluctuation in the number of contacts ($\Delta C$), and the average number of contacts ($\langle C \rangle$) as functions of temperature for several representative 15-mer sequences (Fig. 1). The $C_v(T)$ curves make clear the degree to which the chain is collapsed, while the $\Delta C(T)$ curves provide an unambiguous location for the collapse transition. When $B_0 = -2.0$, the collapse is determined by the overall attraction between residues; $C_v(T)$ reaches its maximum ($T = 1.340$ for sequence 91 and $T = 1.605$ for 95) at a much lower temperature than does $\Delta C(T)$ ($T = 3.160$ and $T = 3.320$, respectively). At the maximum in $C_v(T)$, the chain is collapsed: $\langle C(1.340) \rangle = 10.148$ (of 11) for sequence 91 and $\langle C(1.605) \rangle = 9.414$ for sequence 95. The differences in the $C_v(T)$ curves for these two sequences in spite of the striking similarities in the $C(T)$ and $\Delta C(T)$ curves underscore the idea that $T_{c\ell}$ (as KT have defined it) is not the collapse transition temperature when $B_0 = -2.0$. When $B_0 = -0.1$, the collapse transition is driven by specific contacts so that the folding and collapse transitions are closer in $T$. In the case of sequence 62 ($\sigma = 0.006$), the peak in $C_v (T = 0.860)$ is closer to the peak in $\Delta \chi (T = 0.855)$ than to the peak in $\Delta C (T = 0.965)$; $\langle C(0.860) \rangle = 8.269$. In the case of sequence 5 ($\sigma = 0.621$), the peak in $C_v (T = 0.580)$ is closer to the peak in $\Delta C (T = 0.850)$ than to the peak in $\Delta \chi (T = 0.220)$, but there is a substantial
shoulder in $C_v$ at the temperature corresponding to the peak in $\Delta \chi$, which is not uncommon for sequences with larger $\sigma$; $\langle C(0.580) \rangle = 7.452$ (of 10). Although $T_f$ is clearly associated with the folding transition, $T_s$ cannot be interpreted as the temperature of the collapse transition in general.

It is apparent from the above that $\sigma$ must have a different physical meaning than that given in Klimov and Thirumalai. KT argue that, when $\sigma$ is small ($\sigma < 0.1$), "the folding process and the collapse is [sic] synchronous, and this leads to fast folding" (Klimov and Thirumalai, 1982). To better understand $\sigma$ when collapse and folding coincide, we consider a simple two-state model with free energy $F_i = E_i - TS_i$, ($i = 1, 2$). By substitution into $\Delta \chi = \langle \chi^2 \rangle - \langle \chi \rangle^2$, we find

$$\Delta \chi = (\chi_1 - \chi_2)^2 \frac{\exp \left( \frac{F_1 - F_2}{T} \right)}{1 + \exp \left( \frac{F_1 - F_2}{T} \right)^2}.$$  \hspace{1cm} (4)

Similarly,

$$C_v = \left( \frac{E_1 - E_2}{T} \right)^2 \frac{\exp \left( \frac{F_1 - F_2}{T} \right)}{1 + \exp \left( \frac{F_1 - F_2}{T} \right)^2}.$$  \hspace{1cm} (5)

To determine the maxima ($T_i$ and $T_s$, respectively), we take the derivatives with respect to temperature and set them equal to zero; it is important to note that $\chi_1 - \chi_2$ and $E_1 - E_2$ are temperature dependent. We solve the resulting equations by expanding around the true transition point $[T_i = T(F_1 = F_2)]$:

$$\exp \left( \frac{F_1 - F_2}{T} \right) = 1 - (E_1 - E_2) \frac{T - T_i}{T_i^2}.$$  \hspace{1cm} (6)

Substitution of Eq. 6 into $d\Delta \chi/dT = 0$ and $dC_v/dT = 0$ and straightforward rearrangement yield

$$T_n - T_i \propto 1/(E_1 - E_2)^2.$$  \hspace{1cm} (7)

Division by $T_n = T_i + O(1/(E_1 - E_2)^2)$ to yield $\sigma = 1 - T_i/T_n$ does not change the dependence of the lowest order term on the gap between the two states. Eq. 7 demonstrates that, due to the temperature dependence of the averages of the energy and of the order parameter, $T_i$ need not coincide with $T_s$ (as defined by KT) when the model exhibits two-state behavior, in contrast to the arguments of KT. Moreover, $T_s$ and $T_f$ deviate from the true transition ($T_0$) in such a way that their difference (or its normalized form, $\sigma$) is inversely proportional to the square of the energy gap. For the thirty-two 15-mers with $B_0 = -0.1, r_{0.1}$, and $r_{0.85}$, while for the fifteen 27-mers with $B_0 = -0.1, r_{0.25}$, the correlations fall below unity because lattice models (particularly the short chains considered here) deviate from two-state behavior even at $B_0 = -0.1$. As $B_0$ increases in magnitude, Eq. 7 breaks down. In a recent publication which estimates $\sigma$ for a number of proteins from experimental data [using a definition for $T_s$ based on the fraction of molecules completely folded rather than the radius of gyration Eq. 16 in Klimov and Thirumalai, KT state that proteins with $\sigma \approx 0.25$ should exhibit two-state folding, while proteins with $\sigma \geq 0.25$ should exhibit three-state folding. This conclusion is inconsistent with the hydrodynamic data presented above; sequence 91 has a very low value of $\sigma$ ($\sigma = 0.071$) but clearly exhibits a three-state folding mechanism (black solid line in Fig. 1). Again, the inconsistency stems from a definition of $T_s$ that monitors the fraction of molecules that are folded rather than the fraction of molecules that are collapsed. This leads KT to identify the melting temperature of about 350 K as the collapse temperature. For the latter, there are no good experimental data, although simulations of protein models, suggest that it is higher than 400 K.

We now compare the correlation between $\sigma$ and folding ability to that obtained with other measures of the native state stability. These include the transition temperatures themselves ($T_n$ and $T_i$) and the $Z$ score of the native state, which more directly measures the degree to which a native state is separated from the majority of states in the spectrum. For all sequences, we estimated...
the Z score by 

\[ Z_B = \frac{1}{2} \left( \frac{E_0 - C_0}{\sigma_B} \right) \]  

where \( E_0 \) is the energy of the native state, \( C_0 \) is the number of contacts in that state, \( \sigma_B \) is the average \( B_{ij} \), and \( \sigma_E \) is the standard deviation of \( B_{ij} \). For the 15-mers, it was possible to compare this estimate to the exact Z score: 

\[ Z_E = \frac{1}{2} \left( \frac{E_0 - \mathbf{E}}{\sigma_E} \right) \]  

where \( \mathbf{E} \) is the unweighted average energy and \( \sigma_E \) is the standard deviation of energies. As expected, the two quantities are correlated (Tables I and II). For the sequences that have fully compact ground states, the Z score is closely related to the energy gap between the ground and first excited compact states (\( \Delta_{\alpha} \)).

Table I and II list the same as Table I for 27-mers with \( B_0 = -0.1 \) All (Above) and Those That Have Fully Compact Ground States (Below).

The logarithm of the MFPT is plotted against the time for finding the ground state (MFPT). The MFPT was calculated from 100 Metropolis Monte Carlo trials with a move set identical to that used previously. Trials began in a random configuration and were allowed to proceed for up to \( 10^9 \) steps. Following KT, the simulation temperature \( T_x \) was chosen to yield a constant value of the order parameter: 

\[ T_x = T(x = 0.8) \]  

Analogous criteria based on a different order parameter were used for 27-mers with \( B_0 = -0.1 \) and 3 × 3 × 3 ground states. Inclusion of the six sequences that have noncompact ground states would be inappropriate, because \( \Delta_{\alpha} \) is unrelated to the Z score in these cases (\( r_{x,y}Z_B = -0.306 \)). We did not include statistics for \( T_f/T_g \) because the thermodynamic definitions for \( T_f/T_g \) appear to be inappropriate for short chains such as those studied here. Kinetic definitions require both the introduction of arbitrary cutoffs and evaluation of the folding time, which we wish to predict. However, as noted above, the properties of the energy landscape which maximize the energy gap (Z score) also maximize \( T_f/T_g \) (given a reasonable definition of \( T_g \)), so that we would expect any quantity that is correlated with the gap to be correlated with \( T_f/T_g \) as well.

The folding ability of a sequence was measured by the mean-first passage time for finding the ground state (MFPT). The MFPT was calculated from 100 Metropolis Monte Carlo trials with a move set identical to that used previously. Trials began in a random configuration and were allowed to proceed for up to \( 10^9 \) steps. Following KT, the simulation temperature \( T_x \) was chosen to yield a constant value of the order parameter: 

\[ T_x = T(x = 0.8) \]  

Analogous criteria based on a different order parameter were used for 27-mers with \( B_0 = -0.1 \) and 3 × 3 × 3 ground states. Inclusion of the six sequences that have noncompact ground states would be inappropriate, because \( \Delta_{\alpha} \) is unrelated to the Z score in these cases (\( r_{x,y}Z_B = -0.306 \)). We did not include statistics for \( T_f/T_g \) because the thermodynamic definitions for \( T_f/T_g \) appear to be inappropriate for short chains such as those studied here. Kinetic definitions require both the introduction of arbitrary cutoffs and evaluation of the folding time, which we wish to predict. However, as noted above, the properties of the energy landscape which maximize the energy gap (Z score) also maximize \( T_f/T_g \) (given a reasonable definition of \( T_g \)), so that we would expect any quantity that is correlated with the gap to be correlated with \( T_f/T_g \) as well.

The folding ability of a sequence was measured by the mean-first passage time for finding the ground state (MFPT). The MFPT was calculated from 100 Metropolis Monte Carlo trials with a move set identical to that used previously. Trials began in a random configuration and were allowed to proceed for up to \( 10^9 \) steps. Following KT, the simulation temperature \( T_x \) was chosen to yield a constant value of the order parameter: 

\[ T_x = T(x = 0.8) \]  

Analogous criteria based on a different order parameter were used for 27-mers with \( B_0 = -0.1 \) and 3 × 3 × 3 ground states. Inclusion of the six sequences that have noncompact ground states would be inappropriate, because \( \Delta_{\alpha} \) is unrelated to the Z score in these cases (\( r_{x,y}Z_B = -0.306 \)). We did not include statistics for \( T_f/T_g \) because the thermodynamic definitions for \( T_f/T_g \) appear to be inappropriate for short chains such as those studied here. Kinetic definitions require both the introduction of arbitrary cutoffs and evaluation of the folding time, which we wish to predict. However, as noted above, the properties of the energy landscape which maximize the energy gap (Z score) also maximize \( T_f/T_g \) (given a reasonable definition of \( T_g \)), so that we would expect any quantity that is correlated with the gap to be correlated with \( T_f/T_g \) as well.

The folding ability of a sequence was measured by the mean-first passage time for finding the ground state (MFPT). The MFPT was calculated from 100 Metropolis Monte Carlo trials with a move set identical to that used previously. Trials began in a random configuration and were allowed to proceed for up to \( 10^9 \) steps. Following KT, the simulation temperature \( T_x \) was chosen to yield a constant value of the order parameter: 

\[ T_x = T(x = 0.8) \]  

Analogous criteria based on a different order parameter were used for 27-mers with \( B_0 = -0.1 \) and 3 × 3 × 3 ground states. Inclusion of the six sequences that have noncompact ground states would be inappropriate, because \( \Delta_{\alpha} \) is unrelated to the Z score in these cases (\( r_{x,y}Z_B = -0.306 \)). We did not include statistics for \( T_f/T_g \) because the thermodynamic definitions for \( T_f/T_g \) appear to be inappropriate for short chains such as those studied here. Kinetic definitions require both the introduction of arbitrary cutoffs and evaluation of the folding time, which we wish to predict. However, as noted above, the properties of the energy landscape which maximize the energy gap (Z score) also maximize \( T_f/T_g \) (given a reasonable definition of \( T_g \)), so that we would expect any quantity that is correlated with the gap to be correlated with \( T_f/T_g \) as well.

The folding ability of a sequence was measured by the mean-first passage time for finding the ground state (MFPT). The MFPT was calculated from 100 Metropolis Monte Carlo trials with a move set identical to that used previously. Trials began in a random configuration and were allowed to proceed for up to \( 10^9 \) steps. Following KT, the simulation temperature \( T_x \) was chosen to yield a constant value of the order parameter: 

\[ T_x = T(x = 0.8) \]  

Analogous criteria based on a different order parameter were used for 27-mers with \( B_0 = -0.1 \) and 3 × 3 × 3 ground states. Inclusion of the six sequences that have noncompact ground states would be inappropriate, because \( \Delta_{\alpha} \) is unrelated to the Z score in these cases (\( r_{x,y}Z_B = -0.306 \)). We did not include statistics for \( T_f/T_g \) because the thermodynamic definitions for \( T_f/T_g \) appear to be inappropriate for short chains such as those studied here. Kinetic definitions require both the introduction of arbitrary cutoffs and evaluation of the folding time, which we wish to predict. However, as noted above, the properties of the energy landscape which maximize the energy gap (Z score) also maximize \( T_f/T_g \) (given a reasonable definition of \( T_g \)), so that we would expect any quantity that is correlated with the gap to be correlated with \( T_f/T_g \) as well.

The folding ability of a sequence was measured by the mean-first passage time for finding the ground state (MFPT). The MFPT was calculated from 100 Metropolis Monte Carlo trials with a move set identical to that used previously. Trials began in a random configuration and were allowed to proceed for up to \( 10^9 \) steps. Following KT, the simulation temperature \( T_x \) was chosen to yield a constant value of the order parameter: 

\[ T_x = T(x = 0.8) \]  

Analogous criteria based on a different order parameter were used for 27-mers with \( B_0 = -0.1 \) and 3 × 3 × 3 ground states. Inclusion of the six sequences that have noncompact ground states would be inappropriate, because \( \Delta_{\alpha} \) is unrelated to the Z score in these cases (\( r_{x,y}Z_B = -0.306 \)). We did not include statistics for \( T_f/T_g \) because the thermodynamic definitions for \( T_f/T_g \) appear to be inappropriate for short chains such as those studied here. Kinetic definitions require both the introduction of arbitrary cutoffs and evaluation of the folding time, which we wish to predict. However, as noted above, the properties of the energy landscape which maximize the energy gap (Z score) also maximize \( T_f/T_g \) (given a reasonable definition of \( T_g \)), so that we would expect any quantity that is correlated with the gap to be correlated with \( T_f/T_g \) as well.
a 27-mer, which switched from a two-state folding mechanism to a three-state one as the magnitude of $B_0$ increased. Although there is a high correlation between $s$ and MFPT for sequences with $B_0 = 0.1$, the correlation is low for those with $B_0 = 2.0$. In contrast, $Z_B$ and $T_s$ both do well and are comparable for both values of $B_0$. For the 27-mers, the most sensitive measure of folding ability is clearly $T_s$; these results are in agreement with Sali et al., where it was found that a temperature at which the order parameter had a particular value $(T_x)$ yielded the best predictivity of folding ability. For the sequences with compact ground states, $D_{cs}$ correlates with folding ability almost as strongly as does $Z_B$ and much more strongly than does $s$ (Table II).

The correlation between $\Delta_{cs}$ and folding ability is obscured in the studies by KT (Fig. 22 in Klimov and Thirumalai) because, in contrast to the study of SSK, they include sequences with native states that are not fully compact. They compare folding ability to the gap between the native state and the first excited state from the complete conformational ensemble $(\Delta)$ (Fig. 20 in Klimov and Thirumalai). Typically, the first non-compact excited...

Fig. 2. 15-mer dependence of the mean-first passage time on $s$, $T_s$, and $Z_B$ for sequences with $B_0 = -0.1$ (●), those with $B_0 = -2.0$ (○), and those that were excluded from the statistics (+) (see Table I). For the sequences excluded from the statistics, first passage times of $10^9$ were substituted for the trials that failed to find the native state.

Fig. 3. Same as Figure 2 for 27-mers.
### FACTORS THAT AFFECT FOLDING ABILITY

| TABLE III. Pearson Linear Correlation Coefficients Between Measures of Stability and the Logarithm of the Mean First Passage Time for 15-mers Simulated With a Single Temperature |
|-----------------|-----------------|-----------------|-----------------|
| \( T \) | 0.8 | 1.0 | 1.2 |
| \( B_0 \) | -2.0 | -2.0 | -2.0 |
| \( T_s \) | 0.294 | -0.838 | -0.287 | -0.882 | -0.516 | -0.889 |
| \( T_r \) | 0.388 | -0.868 | -0.437 | -0.924 | 0.662 | -0.932 |
| \( T_f \) | 0.427 | -0.852 | -0.552 | -0.895 | -0.727 | -0.911 |
| \( \sigma \) | 0.083 | 0.698 | 0.201 | 0.719 | 0.155 | 0.733 |
| \( Z_B \) | 0.254 | 0.841 | 0.255 | 0.859 | 0.255 | 0.851 |
| \( Z_s \) | 0.619 | 0.870 | 0.638 | 0.873 | 0.673 | 0.868 |

state is a tail flip, and thus that gap need not be correlated with either \( \Delta_0 \) or the Z score. Moreover, it should be noted that, even if the correct gap (\( \Delta_0 \)) is used with the appropriate sequences (those with fully compact native states), it is incorrect to "normalize" by the simulation temperature (\( T_s \)) as KT have done in Fig. 3 in Klimov and Thirumalai\(^{12,13}\) and Fig. 21 in Klimov and Thirumalai\(^{13}\). Because \( \Delta_0 \) and \( T_s \) are closely related, they both increase as the magnitude of the Z score increases. Consequently, their ratio, \( \Delta_0/T_s \), is not expected to correlate with stability or folding ability. The calculations presented up to this point are consistent with those of KT\(^{12,13}\); it is the interpretation that differs and leads us to the additional calculations and analyses that follow.

In the study by KT and the present one, the simulation temperature differs for each sequence. This provides a physically meaningful approach since it is appropriate to compare the folding of sequences under conditions of corresponding stability; a protein sequence must not only be able to find its native state but it must be dominantly populated at equilibrium.\(^6\) It has been suggested that the correlation between folding ability and the energy gap found in Ref. 6 and, by inference, here derives from the variation in simulation temperature.\(^22\) To address this concern, we performed simulations for the 15-mers at three sets of constant temperatures \( T = 0.8, 1.0, \) and \( 1.2 \) (Table III). Although the correlations are somewhat reduced, they remain highly significant. In particular, the correlation between the logarithm of the MFPT and the exact Z score \( (Z_f) \), the most direct measure of the energy gap, is essentially unchanged. The correlation with \( Z_B \) is strong for sequences with \( B_0 = -0.1 \) but weak for sequences with \( B_0 = -2.0 \). The difference stems from the fact that \( Z_B \), by definition \( [Z_B = (E_0 - C_0|B|)/kT_0] \), assumes the average state to be maximally compact \( (C = C_0) \), but non-compact states become important at the relatively high temperatures considered \( (C \) is small). For sequences with \( B_0 = -0.1 \), \( Z_B \) correlates well with the folding time because the term \( C_0|B| \) is made small by the factor \( B_0 \). For sequences with \( B_0 = -2.0 \), \( Z_B \) correlates poorly with the folding time because \( C_0|B| \) remains large in magnitude.

Thus, for the same 58 sequences as were studied in Klimov and Thirumalai,\(^{12,13}\) it is evident that, although \( \sigma \) is well correlated with the folding ability for sequences with small \( B_0 \) \( (B_0 = -0.1) \), it is not for sequences with large \( B_0 \) \( (B_0 = -2.0) \). The difference between the two sets of sequences stems from a difference in the dependence of \( \sigma \) on the Z score (and hence \( T_s \)), which correlates well with folding ability for both values of \( B_0 \) and different simulation temperatures; the relationship between \( \sigma \) and the energy gap \( [\text{Eq. 7}] \) breaks down for sequences with \( B_0 = -2.0 \) because they deviate strongly from two-state behavior due to the large separation in temperature of the collapse and folding transitions. Moreover, \( Z_B \), unlike \( T_s \) and \( \sigma \), can be calculated without either explicit enumeration of all the conformations or Monte Carlo simulation. Consequently, \( Z_B \) is not only more practical for protein design but it is a true predictor of folding ability.

The correlation of the folding time with the energy gap can be understood in terms of its effect on the energy surface. For random 27-mer sequences, SSK found that folding proceeds by a fast collapse to a semicompetent random globule, followed by a slow, non-directed search through the \( (10^2) \) semicompetent structures for one of the \( (10^3) \) transition states that lead rapidly to the native conformation.\(^7\) A large energy gap results in a native-like transition state that is stable at a temperature high enough for the folding polypeptide chain to overcome barriers between random semi-compact states. As the energy gap increases to the levels obtainable in designed sequences, the model exhibits Hammond behavior\(^23\) in that the fraction of native contacts required in the transition state from which the chain folds rapidly to the native state decreases; random sequences with relatively small gaps must form 80% of the native contacts,\(^7\) while designed sequences with large gaps need form only 20%.\(^24,25\) This decrease in the number of native contacts in the transition state as the energy gap increases is consistent with the behavior of homogeneous systems that coalesce into droplets. Strengthening the interparticle interactions accelerates the process of droplet formation by lowering the energy of the transition state and decreasing size.\(^26\) Thus, the stability of the native state (which is determined by the energy spectrum and is generally related to the energy gap) correlates with folding time because it directly affects the transition state of the folding reaction.

**NOTE ADDED IN PROOF**

The present discussion of Klimov and Thirumalai\(^{12,13,15}\) is also applicable to Klimov DK, Thirumalai D. Linking rates of folding in lattice models of proteins with underlying thermodynamic characteristics. J. Chem Phys 1998;109: 4119–4125.

**ACKNOWLEDGMENTS**

Wethank Yaoqi Zhou for helpful discussions and K.T. for making their sequence data available to us. ARD is a Howard Hughes Medical Institute Predoctoral Fellow. This work was supported in part by a grant from the National Science Foundation (MK).

**REFERENCES**

1. Levinthal C. In: Debrunner P, Tsibris JCM, Munck E, editors. Mass spectrometry in biological systems, proceedings of a meeting held at Allerton House, Monticello, IL, Urbana: University of Illinois Press, 1969;22–24.
2. Bryngelson JD, Wolynes PG. Spin glasses and the statistical mechanics of protein folding. Proc Natl Acad Sci USA 1987;84:7524–7528.

3. Bryngelson JD, Wolynes PG. Intermediates and barrier crossing in a random energy model (with applications to protein folding). J Phys Chem 1989;93:6902–6915.

4. Gutin A., Sali A, Abkevich V, Karplus M, Shakhnovich EI. Temperature dependence of the folding rate in a simple model: search for a “glass”transition. J Chem Phys 1998;108:6466–6483.

5. Shakhnovich EI, Gutin AM. Formation of unique structure in polypeptide chains: theoretical investigation with the aid of a replica approach. Biophys Chem 1989;34:187–199.

6. Sali A, Shakhnovich E, Karplus M. Kinetics of protein folding: a lattice model study of the requirements for folding to the native state. J Mol Biol 1994;235:1614–1636.

7. Sali A, Shakhnovich E, Karplus M. How does a protein fold? Nature 1994;369:248–251.

8. Shakhnovich EI, Gutin AM. Engineering of stable and fast-folding sequences of model proteins. Proc Natl Acad Sci USA 1993;90:7195–7199.

9. Chan HS, Dill KA. Transition states and folding dynamics of proteins and heteropolymers. J Chem Phys 1994;100:9238–9257.

10. Chan HS, Dill KA. Protein folding in the landscape perspective: chevron plots and non-Arrhenius kinetics. Proteins 1996;30:2–33.

11. Sali A, Abkevich V, Shakhnovich EI. Specific nucleus as the transition state for protein folding: evidence from the lattice model. Biochemistry 1994;33:10026–10036.

12. Hammond GS. A correlation of reaction rates. J Am Chem Soc 1955;77:334–338.

13. Dinner AR, Sali A, Karplus M. The folding mechanism of larger model proteins: role of native structure. Proc Natl Acad Sci USA 1996;93:8356–8361.

14. Oxtoby DW. Nucleation of crystals from the melt. Adv Chem Phys 1988;70:263–296.