Estimations of the Size of Nucleation Regions in Globular Proteins[†]

Jie Chen,[‡] J. D. Bryngelson,[⊥] and D. Thirumalai*[,‡,§]

Biophysics Program, Institute for Physical Science and Technology, and Department of Chemistry and Biochemistry, University of Maryland, College Park, Maryland 20742, and Physical Sciences Laboratory, Division of Computer Research and Technology, National Institutes of Health, Bethesda, Maryland 20892

Received: July 12, 2008; Revised Manuscript Received: September 3, 2008

Folding of many single-domain proteins has been described using the nucleation−collapse (NC) mechanism. According to NC, folding (formation of secondary structures and tertiary interactions) and chain collapse occur synchronously upon formation of native-like structures involving a critical number of residues. Using simple nucleation theory together with structure-based thermodynamic data, the average size of the most probable nucleus N_R^*, for single-domain proteins, is estimated to be between 15 and 30 residues. We argue that finite-sized fluctuations in this estimate can be large so that nearly half of the residues of a 100 residue protein can be part of the folding nucleus. Inclusion of surface area changes in the folded and unfolded states are important in the determination of N_R^*.

I. Introduction

Several proteins (usually small) reach their native states rapidly (on the time scale of tens of msecs). The folding of many of these proteins can be approximately described thermodynamically and kinetically by a two-state model.[1−3] In this simplified description, the molecule exists in either a folded (F) conformation or in an unfolded (U) conformation. Under folding condition, the fraction of unfolded molecules decays exponentially $P_U(t) \sim e^{-t/\tau_F}$, where τ_F is the folding time. A logical explanation of these findings is that the folded state is separated from the unfolded conformations by a barrier. Theoretical studies[4−7] and some experiments[8,9] further suggest that efficient folding of these proteins is consistent with a nucleation−collapse (or condensation, NC) mechanism. According to the NC mechanism, the rate-limiting step involves a search for one of the folding nuclei, which consists of a critical number of tertiary contacts and possibly native-like secondary structures. Once the critical nucleus is formed, the polypeptide chain becomes compact and with overwhelming probability reaches the native conformation. Because the formation of the folding nucleus and the collapse of the chain are nearly synchronous, this process is referred to as the nucleation−collapse mechanism.[10]

The nature of the folding nuclei,[11,12] the width of the transition state,[11,13] and the heterogeneity of the structures in the transition-state ensemble[14] continue to be topics of great interest for experimentalists[2,15] and theorists alike. Despite the lack of consensus on a variety of issues concerning the NC mechanism, it is clear that the concept of a folding nucleus provides a cogent explanation of how many proteins fold. It is the purpose of this work to use thermodynamic data to estimate the average size of the nucleation region in proteins. We should stress that the estimates are tentative,[4] but the equations are expressed in terms of variables for which precise experimental values can be obtained. Therefore, this exercise provides a convenient framework for analysis of a number of experiments and serves as a supplement to previous theoretical estimates[6,13,16] of the average number of residues that participate in the folding nucleus of single-domain globular proteins.

II. The Essential Idea of a Critical Nucleus: Analogy to the Gas−Liquid Transition

The reasoning that leads to the concept of a critical nucleus is very elegant and simple. To illustrate this reasoning in its clearest form, we shall briefly take a detour from protein folding and consider a gas of small molecules, for example, water. The random thermal fluctuations of this gas will occasionally cause small droplets of the liquid phase to form, but these liquid droplets will disappear quickly because they are thermodynamically unstable. Now, suppose that the temperature of the gas is lowered to a temperature below its boiling point. What happens to a liquid droplet formed by the thermal fluctuations? To be more specific, consider a liquid droplet of radius R and suppose that the difference of the chemical potential between the liquid phase and the gas phase is $-\Delta\mu$ (in the context of protein folding, $\Delta\mu$ is the driving force for structure formation or, at the least, collapse of the chain). Since the temperature of the gas is below its boiling point, the liquid phase is more stable than the gas phase so that $\Delta\mu > 0$, that is, the liquid has a lower chemical potential than the gas. There is also a free energy associated with the surface of the droplet. This free energy is proportional to the droplet's surface area and is responsible for the familiar phenomenon of surface tension in bulk liquids. Putting all of these considerations together leads to a well-known relation between the free energy of a liquid droplet and an equivalent number of gas molecules.[17] Let ΔG represent this free-energy difference, ρ represent the density of the liquid, and γ represent the surface free energy per unit area. Then, adding together the two free-energy terms gives

$$\Delta G(R) = -\left(\frac{4\pi}{3}\right)\rho\Delta\mu R^3 + 4\pi\gamma R^2 \qquad (1)$$

The first term in eq 1 is the bulk term, and the second term is the surface term. Notice that ΔG has a maximum at a critical radius R^*, where

[†] Part of the "Karl Freed Festschrift".

* To whom correspondence should be addressed. Phone: 301-405-4803. Fax: 301-314-9404. E-mail: thirum@glue.umd.edu.

[‡] Institute for Physical Science and Technology University of Maryland.

[§] Department of Chemistry and Biochemistry, University of Maryland.

[⊥] National Institutes of Health.

10.1021/jp806161k CCC: $40.75
Published on Web 11/11/2008

$$R^* = \frac{2\gamma}{\rho\Delta\mu} \tag{2}$$

If a liquid droplet has a radius less than R^*, then it will have a tendency to shrink and disappear. If the droplet does manage to grow until it has a radius that is larger than R^*, then it will tend to grow.[17] The point of recapping these well-known results is to merely point out that the there is a driving force for formation of a nucleus that depends on the problem, while the surface tension opposes it.

III. Nucleation Theories for Late and Early Transition States

Matheson and Scheraga[16] were the first to estimate the size of nucleation regions in protein folding. By assuming that the major force of assembly of the structure is hydrophobic, they estimated the free-energy cost of forming the nucleus. They concluded that for a number of single-domain globular proteins, the critical nucleus, on an average, consists of about 15−18 residues. Subsequently, theories that view the formation of the folding nucleus of a single protein molecule as a mobile droplet[4,6] have been proposed. These theories are similar in spirit to ones that describe the condensation of the liquid from the vapor phase. Bryngelson and Wolynes[6] (BW) proposed a model in analogy with classical nucleation theory. In a subsequent treatment, Wolynes[13] refined the BW picture by accounting for fluctuations in the "fuzzy interface" between the folding nucleus and the disordered portion of the protein. In the resulting "capilliarity" model, which is similar to that used by Reiss et al.[18] in their treatment of the liquid−solid coexistence in finite-sized rare gas clusters, the free-energy cost of forming a critical nucleus containing N_R residues may be written as

$$\Delta F_{BW}(N_R) \approx -\Delta f(T)N_R + 4\pi\gamma a^2 N_R^{2/3} \tag{3}$$

where the temperature (T)-dependent $\Delta f(T)$ is the gain in free energy (driving force) upon forming a native contact relative to its value in the ensemble of denatured conformations and γ is the average surface tension. The size of the critical nucleus is obtained by maximizing $\Delta F_{BW}(N_R)$ and is given by

$$N^*_{R,BW} \approx \left(\frac{8\pi\gamma a^2}{3\Delta f(T)}\right)^3 \tag{4}$$

Recently, folding of a number of proteins has been analyzed in terms of the capillarity model.[19]

A different form of the free energy of forming a droplet, which is appropriate when the nucleation barrier occurs early in the folding reaction, was introduced by Guo and Thirumalai.[4] By assuming that the principal driving force for structure formation is the hydrophobic interaction, a mean field estimate for the droplet free energy was given by

$$\Delta F_{GT}(N_R) \approx -\frac{\epsilon_H}{2}N_R(N_R - 1) + 4\pi\gamma a^2 N_R^{2/3} \tag{5}$$

In principle, the T-dependent ϵ_H should be taken to be proportional to the second virial coefficient V_2 (<0 when $T < T_\Theta$, the collapse temperature[20]). Given the crudeness of the model (eq 5), such refinements are not needed. In this picture, the driving force for the initial chain compaction is the average attractive hydrophobic interaction, $\epsilon_H < 0$. We have assumed that the most probable nucleus is compact so that the first term in eq 5 represents the free energy of a polypeptide chain that is constrained to be in a sphere with dimensions on the order of the nucleus size. The average critical size of the nucleus according to eq 5 is

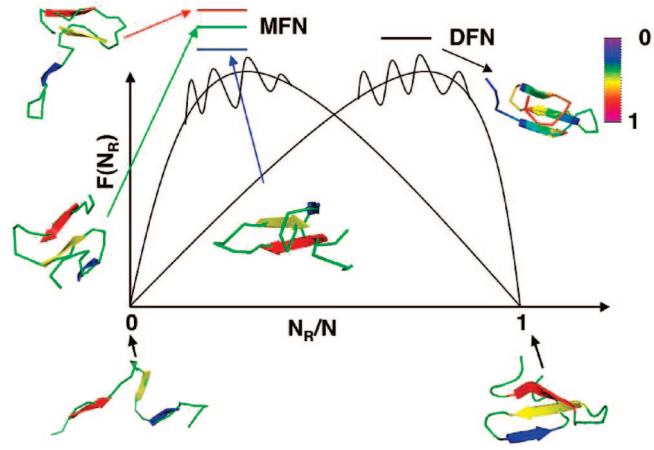

Figure 1. Schematic of the free-energy profile in terms of the putative reaction coordinate N_R normalized by the number of residues N. In this picture, $N_R^\ddagger$ corresponds to a maximum in the free-energy profile. Unfolded states have $N_R/N = 0$, while in the folded state, this ratio is unity. The wiggles in the free-energy profiles are "fine structure" (see ref 13). If the transition state occurs early, then the MFN model is appropriate, whereas for late transition states, the DFN model is accurate. The structures in the transition states (maxima in the free-energy profiles), unfolded state, and the folded state were generated using a coarse-grained model for the WW domain. The structures in the transition state in the MFN show that different parts of the structure are folded in different folding trajectories, which is a characteristic of the multiple folding nuclei model. The scale on the right of the structure of the DFN is meant to show (schematically) that different residues are ordered to a varying extent in the transition state.

$$N^*_{R,GT} \approx \left(\frac{8\pi\gamma a^2}{3\epsilon_H}\right)^{3/4} \tag{6}$$

In order to compute $N_R^\ddagger$, the parameters in the free energy have to be estimated from experimental data (see next section). However, for typical values of γ and ϵ_H, one finds that $N^\ddagger_{R,GT} \approx 15$.[5] In using the first term in eq 5 as the driving force, it is assumed that the range of the hydrophobic force is on the order of the size of the folding nucleus. Clearly, this results in an overestimate of the stabilizing force of the folding nucleus. Similarly, the estimate of the surface tension term is not precise (see the next section). Nevertheless, the naive theory gives reasonable estimate of $N_R^\ddagger$ provided the transition state is reached early.

The theories given above are simplistic for several reasons. (1) They assume that the nuclei are compact. In addition, it is supposed that in the GT picture the nucleus is formed relatively early in the folding process. Such an assumption is not always valid. Within the more flexible capillarity model, movements of the transition states, in the reaction coordinate N_R, can be accounted for by determining the parameters that provides the best fit to experiments. In particular, if full compensation for native interactions is made, resulting in the folding nucleus (see Figure 1) being close to the native conformation, then the nucleus should be considered diffuse. In general, the folding nuclei are not compact, and hence, the effects of surface

contributions in eq 5 have to be refined (see below). (2) More importantly, it is appreciated that the transition region for proteins folding via the NC mechanism is broad[11,13] so that one generically expects a rather large dispersion in the size of the folding nucleus. Thus, it appears that the estimate of the size of the nuclei given in the original formulation in ref 6 and eq 5 may be suitable upper and lower bounds for N_R^*, respectively. (3) The major difficulty in adopting the droplet picture is that surface areas of proteins, as they fold, are better described as fractals. Hence, the changes in the surface areas in the unfolded and folded states have to be estimated more precisely in order to obtain N_R^*. In what follows, we propose such a model in which a fluctuations in the sizes of the unfolded and folded states and calorimetry data are combined to obtain estimates of N_R^*.

IV. Nucleation of the Native Structure in a Protein

Analysis of a number of structures in the Protein Data Bank shows that the radius of gyration, R_G, scales as $R_G \approx 3N^{1/3}$ Å, where N is the number of amino acids in a protein.[21] Therefore, we expect that the volume of a protein is $V \propto M_W \propto R_G^3$, where the M_W is molecular weight in Daltons. More precisely, the typical volume of the folded state is found to scale as $V = 1.27M_W$ Å^3 and have found an accessible surface area of $A_s(F) = 6.3(M_W)^{0.73}$ Å^2, where F stands for the folded state. The deviation of $A_s(F)$ from the expected result $A_s(F) \sim R_G^2 \propto M_W^{2/3}$ shows that even the folded state is not maximally compact in all aspects. In contrast, the accessible surface area of the same protein, when unfolded (U), is approximately $A_s(U) = (1.48M_W + 21)$ Å^2 (this number, which is based on adding the ASA of residues in a Gly-X-Gly tripeptide, is an overestimate because spatial correlation between residues are ignored). Of this accessible surface area, the fraction of $f_{np} \approx 0.57$ of residues is nonpolar[22] in both the unfolded and folded states.[23] The remaining surface area, $f_p = 1 - f_{np} = 0.43$ of the total, is polar.

To compute N_R^*, we shall assume that these relations between molecular weight, volume, and surface area also hold for a nucleating fragment of native protein structure. We shall express our estimate of the size of the nucleus in terms of the typical number of amino acid residues in the nucleus. The average number of amino acid in a protein fragment with a molecular weight of M_W Dalton is $N = M_W/111$ where 111 is the average molecular weight of an amino acid residue. The free energy per unit area of exposed surface area has been estimated from transfer experiments, from experiments on bilayers, and from calorimetry experiments,[24] from calculations with detailed semiempirical potential functions,[25] and from mutation studies. The results of these studies can be used to construct a structure-based model for protein thermodynamics. Such a model has been constructed by Freire and collaborators.[26,27] Their model has been used to predict successfully many thermal and structural properties of proteins and is particularly convenient for our purposes. In the Appendix, we apply this model to obtain an average free energy of unfolding a protein fragment of N amino acids as a function of the absolute temperature, the change in the accessible polar surface area, A_p, and the change in accessible nonpolar area, A_{np}. At room temperature, 298.15 K, this expression, eq A-17, simplifies to

$$\Delta G_U = 10.1\Delta A_{np} + 31.4\Delta A_p - 2030N \qquad (7)$$

where the accessible surface areas are expressed in terms of square Angstroms and the free energy is in units of calories per mol. A discussion of the generality of eq 7 is given at the end of this section.

In order to use eq 7 for estimating the size of the critical nucleus for folding, we must make two simple changes. First, we are interested in the free energy of folding, ΔG_F, rather than the free energy of unfolding, which is the subject of eq 7. Fortunately, $\Delta G_F = -\Delta G_U$; therefore, this change is simple. Second, we must relate the changes in accessible surface area to the number of amino acids. To derive this relation, we first calculate the changes in the accessible surface area as a function of the molecular weight of the protein fragment. The change in the total accessible surface area upon unfolding, ΔA_T, is given by the difference of the unfolded surface area and the folded surface area

$$\Delta A_T = A_s(U) - A_s(F) = \left[1.48M_W - 6.3M_W^{0.73} + 21\right] \text{Å}^2 \qquad (8)$$

Recalling that the average molecular weight of an amino acid residue is about 111 Daltons, this relation can be expressed as

$$\Delta A_T = [164N - 196N^{0.73} + 21] \text{ Å}^2 \qquad (9)$$

for an average N residue protein fragment. In both the folded and unfolded states, a fraction f_p of this is polar, and f_{np} is nonpolar; therefore

$$\Delta A_{np} = f_{np}\Delta A_T = [93.5N - 112N^{0.73} + 12] \text{ Å}^2 \qquad (10)$$

$$\Delta A_p = f_p\Delta A_T = [70.5N - 84.0N^{0.73} + 9] \text{ Å}^2 \qquad (11)$$

Substituting these relations into eq 7 for the free energy of unfolding yields the desired average free energy for folding an N amino acid fragment of a protein at room temperature

$$\Delta G_F = [-1.13N + 3.77N^{0.73} - 0.4] \text{ cal mol}^{-1} \qquad (12)$$

The size of the critical nucleus for the formation of the native structure is the number of residues that maximize the change in free energy given by eq 12. This can be found by taking the derivative of $\Delta G_F(N)$ with respect to N and setting it equal to zero. Since

$$\frac{d\Delta G_F}{dN} = -1.13 + 2.75N^{-0.27} \qquad (13)$$

is 0 at N_R^*, our ruminations lead to an estimate of

$$N_R^* = (2.75/1.13)^{1/0.27} = 27 \text{ residues} \qquad (14)$$

for the critical nucleus for a single domain protein.

A few comments concerning the estimate of N_R^* are in order. (1) The reliability of the range of values for N_R^* depends on a number of experimentally determined parameters (see Appendix). In order to evaluate ΔG_F, we have relied on the estimates of coefficients that relate heat capacity and enthalpy changes in terms of structure-based solvent-accessible area changes upon folding.[26] While a structure-based thermodynam-

ics scale, derived by Freire and co-workers, has been used to accurately predict ΔG_F for several proteins, the extent of variations in these parameters for a broad class of proteins has not been fully established. Therefore, our estimates of N_R^* are tentative. (2) From the results in the Appendix and the arguments leading to eq 12, it is clear that the free-energy change can be written in a more general form as

$$\Delta G_F = C_0 - C_1 N + C_2 N^{2\nu_U} - C_3 N^{2\nu_F} \qquad (15)$$

In eq 15, $C_2 \propto \gamma_U$ and $C_3 \propto \gamma_F$, where γ_U and γ_F are, respectively, surface tension in the U and F states.We recover eq 12 with $2\nu_U \approx 1$, $2\nu_F \approx 0.73$, $(C_2 - C_1) = -1.13$, and $C_0 = 0.4$. If $2\nu_U \approx 1$ (unfolded proteins behave approximately as Gaussian chains) and if $2\nu_F \approx 2/3$ (native states are maximally compact), then we find $N_R^* \approx (2/3)[C_3/(C_2 - C_1)]^3$. For the coefficients in eq 12, N_R^* changes to about 15, which, given the crudeness of the model, is not that different from 27. If we set $2\nu_U \approx 1.2$ and $2\nu_F \approx 2/3$, then the expression for N_R^* has to be solved numerically assuming that C_0, C_1, C_2, and C_3 are known from experiments. (3) In order to obtain eq 12, we have assumed that $A_s(f) \sim M_W^{0.73}$ instead of $A_s(f) \sim M_W^{2/3}$. The deviation is due to the mesoscopic nature of proteins that gives them a fractal character even when folded. Indeed, detailed analysis of proteins in the PDB have been used to show[28] that ν_F satisfies the bound $[(1 + b/ln\upsilon)/3] \leq \nu_F \leq 1/3$, where b (>0) is a constant. Thus, our use of $A_s(F) \sim M_W^{0.73}$ is appropriate and reflects the noncompact aspect of surface-ordered residues. (4) Even the general form of eq 15 is only approximate. It is well-known that there is a distribution ($P(R_G)$) in the radius of gyration R_G in both the u and f states. This implies that fluctuation effects, which are especially important in finite-sized systems, can play an important role.[29,30] As a result, the surface area contributions (third and fourth terms in eq 15) should involve integrals over $P(R_G)$. As small small-angle X-ray scattering measurements of $P(R_G)$ for a broad class of proteins become readily available, they can be used to account for fluctuations in γ_U and γ_F in eq 15.

V. Discussion

The present analysis shows that, in general, the size of the most probable size of the folding nucleus is about 15−30 residues. Experiments (see, for example, refs 31 and 32) show that typically folding nuclei are small. It is likely that the relatively small value depends on the nature of the driving force for structure formation. If the NC mechanism is driven by formation of the early transition state, with (N_R^*/N) playing the role of the Tanford β-like parameter, then we expect the multiple folding nuclei (MFN) model to hold (see Figure 1). According to MFN,[11] there are several small-sized nuclei whose formation consolidates the rest of the structure. The heterogeneity of the structures in the MFN picture, which could also contain interactions that are absent in the native state,[33] reflects the broad distribution of free energy-barriers[13,34] separating the folded and unfolded states. If the transition state occurs close to the native state, then N_R^* can be large. In this scenario, it is appropriate to view folding in terms of a diffuse nucleus model (DFN) according to which many, if not all, residues are ordered in the transition region to some extent. Even a single protein can fold by either of the mechanisms (MFN or DFN), depending on the external conditions.[11,13]

The exponents in eq 15 that characterize the surface area changes can exceed values expected for compact objects. This is a consequence of the fractal nature of proteins especially when unfolded.[28] In this picture, the precise boundary between the ordered nucleus and the rest of the structure is not well-defined and leads to the possibility of creating MFN with very little free-energy cost. As a result, one can even imagine a scenario when the driving force nearly cancels the opposing force due to surface tension effects. This implies that the free-energy cost of creating a droplet with N_R residues is $F(N_R) \approx -\alpha_1 N_R^{\delta} + \gamma N_R^{\theta}$ with $\theta = 2\nu_U$. If $\nu_U \approx \delta$, then we see that there is effectively no barrier to folding. In such a scenario, folding occurs in a "downhill" manner, as has been observed in recent experiments.[35,36] This scenario is most likely for small proteins under Θ-solvent conditions.

Due to finite-size fluctuations, there can be large dispersion in N_R^*. In terms of the folding landscape, fluctuations in δN_R^* determines the heterogeneity of the transition-state ensemble and most directly determines the transition-state widths. Because generically we expect $\delta N_R^* \approx N^{1/2}$,[37] even a protein with $N = 100$ can have nearly half of the residues as part of the folding nucleus. It is clear that single-molecule experiments that can accurately monitor the folding trajectories will be needed to probe these fine structure details associated with folding nuclei.[38]

The current work also serves as a reminder that surface area changes, with mesoscopic manifestation of surface tension effects, are crucial in determining the nature of the folding much like that in the freezing of clusters of rare gas atoms.[18] It is crucial to consider fluctuations in both the U and F states to accurately estimate surface tension effects (see eq 15). Indeed, the precise exponent that determines N_R^* is due to the dependence of the accessible surface areas in the folded and unfolded states (see eq 7). It appears that accurate measurements of these quantities and their link to the topology (helical, β-sheet, or mixed) of the folded states are needed for reliable estimates of the size of the most probable folding nucleus.

Acknowledgment. We thank Peter Wolynes for several constructive suggestions. Most of the analysis was performed while J. D. Bryngelson was part of the scientific staff at NIH. This work was supported by grants from NSF (CHE 05-14056) and the Air Force Office of Scientific Research (FA9550-07-1-0098).

Appendix

Unfolding Free Energy from Experimental Data. From the thermodynamics relation

$$C_p = T\left(\frac{\partial S}{\partial T}\right) \qquad (A\text{-}1)$$

$$= \left(\frac{\partial H}{\partial T}\right) \qquad (A\text{-}2)$$

the enthalpy and entropy changes upon folding may be written as

$$\Delta H(T) = \Delta H(T_R) + \int_{T_R}^{T} \Delta C_p(T)\mathrm{d}T \qquad (A\text{-}3)$$

$$\Delta S(T) = \Delta S(T_R) + \int_{T_R}^{T} \frac{\Delta C_p(T)}{T}\mathrm{d}T \qquad (A\text{-}4)$$

where T_R represents a reference temperature and all temperatures are in degrees Kelvin. For convenience, we shall take ΔC_p to be independent of temperature, which is a good approximation

for temperatures less than 353.15 K (80 °C). In this case, the above relations become

$$\Delta H(T) = \Delta H(T_R) + (T - T_R)\Delta C_p(T) \quad \text{(A-5)}$$

$$\Delta S(T) = \Delta S(T_R) + \Delta C_p \log\left(\frac{T}{T_R}\right) \quad \text{(A-6)}$$

Moreover, the heat capacity changes can be expressed in terms of solvent-accessible surface areas since ΔC_p can be related to changes in hydration.[39] Data on the change in heat capacity upon unfolding can be fit by the function[26]

$$\Delta C_p = [0.45\Delta A_{np} - 0.26\Delta A_p] \text{ cal mol}^{-1} \quad \text{(A-7)}$$

where ΔA_{np} and ΔA_p refer to the accessible surface area of on-polar and polar surfaces, respectively, and the accessible surface areas are expressed in units of Å^2. Similarly, the change in enthalpy upon unfolding for the reference temperature of T_R = 333.15 K (60 °C) is fit by[26]

$$\Delta H(T_R) = -8.44\Delta A_{np} + 31.4\Delta A_p \quad \text{(A-8)}$$

where T_R = 333.15 K is used because it is a typical median temperature for the thermal denaturation of proteins. Substituting the above expression for the enthalpy (eq A-5) and the expression for the heat capacity (eq A-7) for $T = T_R$ into the expression for the heat capacity for arbitrary temperature yields an empirical expression for the change in enthalpy upon unfolding at arbitrary temperatures

$$\Delta H(T) = [-8.44 + 0.45(T - T_R)]\Delta A_{np} + [31.4 - 0.26(T - T_R)]\Delta A_p \quad \text{(A-9)}$$

in units of cal mol^{-1}.

The change in entropy upon unfolding as a function of temperature is most conveniently expressed in terms of the temperatures T^*_{np}, which is the temperature at which the exposure of the nonpolar surface to water causes no change in entropy, and T_p^*, which is the temperature at which the exposure of the polar surface to water causes no change in entropy. Using these temperatures as parameters, and also eq A-6 for the entropy of unfolding and eq A-7 for the heat capacity, yields the empirical expression for the solvation entropy of unfolding as a function of temperature

$$\Delta S_{solv} = 0.45 \log\left(\frac{T}{T^*_{np}}\right)\Delta A_{np} - 0.26 \log\left(\frac{T}{T^*_p}\right)\Delta A_p \quad \text{(A-10)}$$

in units of cal mol^{-1} K^{-1}. The temperatures T^*_{np} and T_p^* have been estimated from experimental data, leading to the approximate values T^*_{np} = 385.15 K and T_p^* = 335.15 K.[26]

In addition to the change in solvation entropy calculated above, the increase in the configuration entropy has been investigated by D'Aquino et al.[27] They divide the change in configurational entropy into three parts. The first part, ΔS_{ub}, is the change in configurational entropy of an amino acid side chain caused by moving a buried amino acid side chain to the exterior of a protein. The second term, ΔS_{uf}, is the change in configurational entropy of an amino acid side chain caused by the backbone unfolding. The third term, ΔS_{bb}, is the change in configurational entropy of a unit of the peptide backbone upon unfolding. D'Aquino et al. estimated values for each of these three terms for the 20 amino acids. We have used these entropy values and the amino acid frequencies reported by McCaldon and Argos[40] to calculate average values for each of the three entropy change terms. We find

$$\overline{\Delta S_{ub}} = 2.52 \text{ cal mol}^{-1} \text{ K}^{-1} \text{ per residue} \quad \text{(A-11)}$$

$$\overline{\Delta S_{uf}} = 0.93 \text{ cal mol}^{-1} \text{ K}^{-1} \text{ per residue} \quad \text{(A-12)}$$

$$\overline{\Delta S_{bb}} = 3.37 \text{ cal mol}^{-1} \text{ K}^{-1} \text{ per residue} \quad \text{(A-13)}$$

and hence yield a total averaged configurational entropy change of

$$\overline{\Delta S_{conf}} = \overline{\Delta S_{ub}} + \overline{\Delta S_{uf}} + \overline{\Delta S_{bb}} \quad \text{(A-14)}$$

$$= 6.82 \text{ cal mol}^{-1} \text{ K}^{-1} \text{ per residue} \quad \text{(A-15)}$$

Combining the equations for the enthalpy (eq A-9), solvation entropy (eqA-10), and configurational entropy (eq A-15) of unfolding gives the empirical expression for the free energy of unfolding as a function of absolute temperature, T, of the number of amino acid residues, N, and of the changes in accessible surface area, ΔA_{np} and ΔA_p

$$\Delta G_U = \Delta H(T) - T(\Delta S_{solv}(T) + \Delta S_{conf}) \quad \text{(A-16)}$$

$$= \left[-8.44 + 0.45(T - T_R) - 0.45T \log\left(\frac{T}{T^*_{np}}\right)\right]\Delta A_{np} + \left[31.4 - 0.26(T - T_R) + 0.26T \log\left(\frac{T}{T^*_p}\right)\right]\Delta A_p - 6.82TN \quad \text{(A-17)}$$

where the accessible surface areas are expressed in units of Å and the free energy is expressed in units of cal mol^{-1}. This free-energy expression applies only to the unfolding of an average fragment of the native protein structure and not to any specific sequence. The free energy of the unfolding free energy of specific sequences would be expected to fluctuate about this average. Equation A-17 is used in the main body of the test.

References and Notes

(1) Jackson, S. E. *Folding Des.* **1998**, *3* (4), R81−R91.
(2) Fersht, A. R.; Daggett, V. *Cell* **2002**, *108* (4), 573–582.
(3) The two-state model merely provides a convenient framework for analyzing data. Because single-domain proteins are relatively small finite-sized systems, fluctuations play a major role in thermodynamics and kinetics. Consequences of the finite size include, but are not limited to, ruggedness of the energy landscape even in the folded state, variations in the collapse (or melting) temperatures of different residues, a broad transition-state ensemble, and diversity of folding pathways especially during the early stages of folding. These effects are not considered here.
(4) Guo, Z. Y.; Thirumalai, D. *Biopolymers* **1995**, *36* (1), 83–102.
(5) Thirumalai, D.; Guo, Z. Y. *Biopolymers* **1995**, *35* (1), 137–140.
(6) Bryngelson, J. D.; Wolynes, P. G. *Biopolymers* **1990**, *30* (1−2), 177–188.
(7) Abkevich, V. I.; Gutin, A. M.; Shakhnovich, E. I. *Biochemistry* **1994**, *33* (33), 10026–10036.
(8) Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92* (24), 10869–10873.
(9) Itzhaki, L. S.; Otzen, D. E.; Fersht, A. R. *J. Mol. Biol.* **1995**, *254* (2), 260–288.

(10) The nucleation−collapse mechanism proposed in ref 4 was independently also suggested by Itzhaki et al.[9] who used the term nucleation−collapse for the same process. To our knowledge, the basic ideas in the two papers[4,9] are the same.
(11) Guo, Z. Y.; Thirumalai, D. *Folding Des.* **1997**, *2* (6), 377–391.
(12) Klimov, D. K.; Thirumalai, D. *J. Mol. Biol.* **1998**, *282* (2), 471–492.
(13) Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94* (12), 6170–6175.
(14) Onuchic, J. N.; Socci, N. D.; LutheySchulten, Z.; Wolynes, P. G. *Folding Des.* **1996**, *1* (6), 441–450.
(15) Brockwell, D. J.; Radford, S. E. *Curr. Opin. Struct. Biol.* **2007**, *17* (1), 30–37.
(16) Matheson, R. R.; Scheraga, H. A. *Macromolecules* **1978**, *11* (4), 819–829.
(17) Rowlinson, J. S.; Widom, B. *Molecular Theory of Capolarity*; Dover Publications: Mineola, NY, 2003.
(18) Reiss, H.; Mirabel, P.; Whetten, R. L. *J. Phys. Chem.* **1988**, *92* (26), 7241–7246.
(19) Qi, X.; Portman, J. J. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (32), 11164–11169.
(20) Klimov, D. K.; Thirumalai, D. *J. Chem. Phys.* **1998**, *109* (10), 4119–4125.
(21) Dima, R. I.; Thirumalai, D. *J. Phys. Chem. B* **2004**, *108* (21), 6564–6570.
(22) Camacho, C. J.; Thirumalai, D. *Phys. Rev. Lett.* **1993**, *71* (15), 2505–2508.
(23) Miller, S.; Janin, J.; Lesk, A. M.; Chothia, C. *J. Mol. Biol.* **1987**, *196* (3), 641–656.
(24) Makhatadze, G. I.; Privalov, P. L. *Adv. Protein Chem.* **1995**, *47*, 307–425.
(25) Lazaridis, T.; Archontis, G.; Karplus, M. *Adv. Protein Chem.* **1995**, *47*, 231–306.
(26) Luque, I.; Mayorga, O. L.; Freire, E. *Biochemistry* **1996**, *35* (42), 13681–13688.
(27) D'Aquino, J. A.; Gomez, J.; Hilser, V. J.; Lee, K. H.; Amzel, L. M.; Freire, E. *Proteins* **1996**, *25* (2), 143–156.
(28) Reuveni, S.; Granek, R.; Klafter, J. *Phys. Rev. Lett.* **2008**, *100* (20), 208101.
(29) Chahine, J.; Nymeyer, H.; Leite, V. B. P.; Socci, N. D.; Onuchic, J. N. *Phys. Rev. Lett.* **2002**, *88* (16), 168101.
(30) Li, M. S.; Klimov, D. K.; Thirumalai, D. *Phys. Rev. Lett.* **2004**, *93* (26), 268107.
(31) Hedberg, L.; Oliveberg, M. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (20), 7606–7611.
(32) Oliveberg, M.; Wolynes, P. G. *Q. Rev. Biol.* **2005**, *38* (3), 245–288.
(33) Li, L.; Mirny, L. A.; Shakhnovich, E. I. *Nat. Struct. Biol.* **2000**, *7* (4), 336–342.
(34) Straub, J. E.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90* (3), 809–813.
(35) Sadqi, M.; Fushman, D.; Munoz, V. *Nature* **2006**, *442* (7100), 317–321.
(36) Liu, F.; Du, D. G.; Fuller, A. A.; Davoren, J. E.; Wipf, P.; Kelly, J. W.; Gruebele, M. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (7), 2369–2374.
(37) Thirumalai, D. *J. Phys. I* **1995**, *5* (11), 1457–1467.
(38) Fernandez, J. M.; Li, H. B. *Science* **2004**, *303* (5664), 1674–1678.
(39) Gomez, J.; Hilser, V. J.; Xie, D.; Freire, E. *Proteins* **1995**, *22* (4), 404–412.
(40) Mccaldon, P.; Argos, P. *Proteins* **1988**, *4* (2), 99–122.

JP806161K