Computing protein stabilities from their chain lengths

Kingshuk Ghosh^{a,1} and Ken A. Dill^{b,1}

^aDepartment of Physics, University of Denver, Denver, CO 80209; and ^bDepartment of Biophysics, University of California, San Francisco, CA 94143

Contributed by Ken A. Dill, April 10, 2009 (sent for review February 1, 2009)

New amino acid sequences of proteins are being learned at a rapid rate, thanks to modern genomics. The native structures and functions of those proteins can often be inferred using bioinformatics methods. We show here that it is also possible to infer the stabilities and thermal folding properties of proteins, given only simple genomics information: the chain length and the numbers of charged side chains. In particular, our model predicts $\Delta H(T)$, $\Delta S(T)$, ΔC_p , and $\Delta F(T)$ —the folding enthalpy, entropy, heat capacity, and free energy—as functions of temperature T ; the denaturant m values in guanidine and urea; the pH-temperature-salt phase diagrams, and the energy of confinement $F(s)$ of the protein inside a cavity of radius s . All combinations of these phase equilibria can also then be computed from that information. As one illustration, we compute the pH and salt conditions that would denature a protein inside a small confined cavity. Because the model is analytical, it is computationally efficient enough that it could be used to automatically annotate whole proteomes with protein stability information.

macromolecular confinement | protein electrostatics | protein folding | protein stability | protein thermodynamics

We describe a method for predicting a protein's stability from its amino acid sequence. Protein stability is important for various reasons. It underlies structure–function relationships in proteins. It plays a role in properties such as crystallization, aggregation, solubility, fibrilization, protein–protein interactions, and protein adsorption onto surfaces. Protein stability is a key component in formulating therapeutic proteins for biotech and in controlling the biocompatibilities of biomaterials. Moreover, the stabilities of proteins clearly matter to cells, as evidenced by the elaborate mechanisms—such as chaperone proteins, unfolded-protein-response machines, and natural cellular osmolytes—that have evolved to protect cells against the damaging consequences of unfolded proteins.

Thanks to modern genomics, the amino acid sequences of proteins are becoming known at an increasing rate. In many cases, amino acid sequences give sufficient information to infer a protein's native structure, function, and binding partners (1). Less effort has gone into inferring protein folding stabilities and thermal properties. In part, this may be due to a common view that to predict stability, it is necessary to know a protein's native structure, its secondary structures and hydrogen bonding patterns (2), its van der Waals packing (3), and/or its hydrophobic core residues. This view is natural because other properties of proteins, such as their folding kinetics, are known to depend substantially on their native topologies (4).

Interestingly, however, despite considerable study, there is little evidence that stability depends in any systematic way on a protein's native structure. The following evidence indicates that stability just depends primarily on the chain length of a protein and on its numbers of acidic and basic side chains. First, calorimetric measurements show that protein stability does not correlate significantly with the types or amounts of secondary structures in the native structure (5, 6). Second, while it has been long thought that the large positive heat capacity of unfolding came mainly from the hydrophobic amino acids (5–8), calorimetric transfer experiments on the amino acids shows burying polar amino acids in a hydrophobic core also leads to a substantial change in heat capacity (9–11),

indicating less difference between types of amino acids for interpreting thermal properties than originally thought. Third, recent studies of osmolytes show that the burial of backbone hydrogen bonds is an important contributor to protein stability (12), signifying a contribution that scales just with the chain length. And fourth, an extensive review by Robertson and Murphy (13) shows that the predominant dependence of the folding enthalpy, entropy, and heat capacity is simply on the chain length, N , of a protein molecule.

This evidence for such simplicity implies that the universal features of protein stability ought to be capturable within a very simple model. We describe such a model here, most of the basic ideas for which are elementary and date back several years (14–17). What's new here are just the unification into a single quantitative model for all aspects of folding stability and identification of parameters using modern databases of experimental protein stabilities. We introduce the idea of an ideal thermal protein (ITP), a hypothetical protein having the average stability properties of a protein N amino acids long. Parameters for the ITP are taken from the database of Robertson and Murphy on thermal properties of proteins (13). We find that this approach gives a good accounting for the universal properties of protein folding stability, including the linear dependence of the enthalpy, heat capacity, and entropy of folding on chain length, the linear dependence of denaturant m values (slope of energy of folding vs. denaturant concentration) on chain length, the approximate independence of the energy of folding on chain length, the dependence of the energy on the square of the net charge on the side chains, and the inverse square dependence of the energy of confinement on the size of the box in which a protein is confined.

Modeling Protein Stability as a Balance Between Chain Entropy and Amino Acid Burial

We start with the chain entropy. For a chain having N monomers, the number of conformers in the denatured state will be approximately z^N , where z is the number of rotational isomers around a given backbone lattice bond. Hence, for an N -mer, we approximate the chain entropy of the denatured state as

$$S_D = Nk \ln z, \quad [1]$$

where k is Boltzmann constant. For the folded state chain entropy, we take $S_N = 0$. This conformational entropy opposes folding. Next, we assume that the folding of the protein is driven by the desolvation and burial of amino acids, both hydrophobic and polar. On balance, folding must be driven by the monomer–monomer attractions—from hydrophobic interactions, hydrogen bonding, and van der Waals interactions—which are more favorable than the monomer–solvent interactions. Let g represent the average energy of burying one amino acid (backbone plus side chain), that is, of transferring it from the aqueous solvent into the folded hydrophobic core. We assume that each burial of each amino acid

Author contributions: K.G. and K.A.D. designed research; K.G. performed research; K.G. and K.A.D. analyzed data; and K.G. and K.A.D. wrote the paper.

The authors declare no conflict of interest.

¹To whom correspondence may be addressed. E-mail: kghosh@du.edu or dill@maxwell.compbio.ucsf.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0903995106/DCSupplemental.

Fig. 1. The m value depends linearly on chain length for guanidine hydrochloride and urea (Eq. 4).

is independent of the others. For a chain having N amino acids, combining the burial energy and the chain entropy components gives

$$\Delta F_{\text{fold}}(N) = N[g + kT \ln z]. \quad [2]$$

This basic idea, first elucidated by Tanford (14) and Brandts (15), is sufficient for what follows.

Protein Stability Depends Linearly on Denaturant Concentration: The m Value. This model is readily applied to agents that denature or stabilize proteins. Examples of chemical denaturants include guanidine hydrochloride and urea. The opposite effect, of native protein stabilization, results from molecules such as glycerol, sugars, some salts, TMAO (trimethylamine- N -oxide), and sarcosine (12, 18–20). Adding denaturant to water makes the solution more non-polar, weakening the monomer–monomer attractions within the protein chain, shifting the equilibrium toward the unfolded state, ultimately denaturing the protein if the denaturant is sufficiently concentrated. In contrast, stabilizing agents lead to a strengthening of intrachain interactions. Adding guanidine or urea tends to dissociate 2 monomers in water. Consider a denaturing agent having a concentration c in water. At low concentrations, experiments show that denaturants will change the monomer–monomer dissociation energy approximately linearly (21, 22):

$$g(c) = g_0 + m_1 c. \quad [3]$$

In the absence of denaturant ($c = 0$), the energy of dissociation of 2 monomers is g_0 . The reason for the linear dependence, $m_1 c$, is that increasing the amount of denaturant in the solution increases proportionately the amount of denaturant in the first solvation shell around each nonpolar solute molecule, decreasing proportionately the monomer–monomer attraction. Substituting Eq. 3 into Eq. 2 predicts a linear dependence of the folding energy on

denaturant concentration,

$$\Delta F_{\text{fold}}(c) = N[kT \ln z + g_0 + m_1 c]. \quad [4]$$

Eq. 4 predicts that the m value, which is the slope of $\Delta F_{\text{fold}}(c)$ vs. c can be expressed as $m = Nm_1$. This follows from the assumption that m value is a product of the increased surface area of monomers exposed upon denaturation multiplied by the energy per unit surface area. For denaturants, m_1 is positive, while for stabilizers, m_1 is negative. Experiments generally confirm a correlation of m values with estimated changes in surface area of the chain (23), but for some proteins, the denatured state is compact, leading to more complexity (24). Fig. 1 shows a new, but not unexpected result, from the present model: m values depends linearly on chain length N , consistent with experiments (from table 1 of ref. 23 for GdnHCL and urea). Our analysis gives a value of $m_1 = 25$ cal/mol-M for GdnHCL and $m_1 = 13.2$ cal/mol-M for urea.

Protein Stability Is a Curved Function of Temperature. Next, to predict how protein stability depends on temperature, we must take into account how g_0 , the burial energy, depends on temperature. To do this, we consider small-molecule compounds in model transfer experiments. Experimental studies by Makhataadze and Privalov (25) have shown that the folding energy is a curved function of temperature.

Typically, the transfer of an amino acid from water into a medium of amino acids is accompanied by a heat capacity change, Δc_p , that is approximately constant, independent of temperature. Thus, to capture experimental data on model compound transfer data requires three parameters: Δc_p ; T_h , the temperature at which the enthalpy of transfer is zero; and T_s , the temperature at which the entropy of transfer is zero (26, 27). Then the enthalpy of transfer is

$$\Delta h(T) = \int_{T_h}^T \Delta c_p dT = \Delta c_p (T - T_h), \quad [5]$$

and the entropy of transfer is

$$\Delta s(T) = \int_{T_s}^T \frac{\Delta c_p}{T} dT = \Delta c_p \ln(T/T_s). \quad [6]$$

Recognizing that the total folding entropy is the sum of the transfer term plus the chain configurational entropy gives

$$\Delta s(T) = -k \ln z + \Delta c_p \ln(T/T_s). \quad [7]$$

Also, recognizing that the total folding enthalpy is the sum of the term for the transfer into the nonpolar core plus an average enthalpy g_0 for the ordering and packing gives

$$\Delta h(T) = g_0 + \Delta c_p (T - T_h). \quad [8]$$

The total changes in enthalpy $\Delta H(T) = N\Delta h(T)$ and $\Delta S(T) = N\Delta s(T)$ are

$$\Delta H(T) = N[g_0 + \Delta c_p (T - T_h)] \quad [9]$$

and

$$\Delta S(T) = N[-k \ln z + \Delta c_p \ln(T/T_s)]. \quad [10]$$

Our definitions here of T_h and T_s are identical to those of Baldwin and Robertson and Murphy (13, 28); they are the points at which the heat capacity contribution vanishes, so $\Delta H(T_h)$ and $\Delta S(T_s)$ are linear functions of the chain length across different proteins. Now, combining Eq. 9 and Eq. 10 with $\Delta F = \Delta H - T\Delta S$ gives the total folding free energy:

$$\Delta F_{\text{fold}}(T) = N \left[g_0 + kT \ln z + \Delta c_p (T - T_h) - T \Delta c_p \ln \left(\frac{T}{T_s} \right) \right]. \quad [11]$$

Fig. 2. Thermal properties depend linearly on protein chain length. (A) ΔS^* , enthalpy of unfolding at 100 °C. (B) ΔH^* , entropy of unfolding at 112 °C. (C) ΔC_p , temperature independent heat capacity of unfolding. The thermal properties shown were determined for 49 different proteins, based on scanning calorimetry and spectroscopic experiments, and are plotted as a function of chain length, N . It should be noted that these changes are for unfolding, whereas we discuss changes due to folding, which are related through minus signs. The lines are linear regressions. Reprinted with permission from ref. 13. Copyright 1997 American Chemical Society.

Eq. 11 predicts curvature in $\Delta F_{\text{fold}}(T)$ vs. temperature. It predicts 2 denaturation temperatures, i.e., 2 points at which $\Delta F_{\text{fold}} = 0$. One denaturation midpoint occurs at a high temperature: above that point, the conformational entropy dominates the free energy, and the chain unfolds. In addition, proteins can also be denatured by low temperatures. Cold denaturation is a peculiarity of hydrophobic interactions in water: At low temperatures, the protein unfolds because the monomer–monomer attractions become weak. The present model simply illustrates that the entropy, enthalpy, and heat capacity of protein folding should depend linearly on the chain length N , which is consistent with the experiments described throughout the rest of this manuscript.

Thermal Properties Depend Linearly on Chain Length. Eq. 10 indicates that two entropies contribute to protein stability: (i) The chain entropy, $Nk \ln z$, which opposes folding, and (ii) the solvation entropy, a component of the hydrophobic effect, which drives folding at some temperatures. Consider the stability at the temperature $T = T_s$, where model compound solvation entropies are approximately 0 and therefore where the protein's entropy is exclusively due to the chain entropy (29). Small-molecule compound studies give $T_s \approx 112^\circ\text{C}$ (13, 28). Fig. 2 shows the fit with data indicating the folding entropies of many proteins at 112 °C have an average value of $\Delta S/N = -16.8 \text{ J K}^{-1} \text{ mol res}^{-1} = -4.02 \text{ cal K}^{-1} \text{ mol res}^{-1}$ (figure 6b in ref. 13), which converts to $z = 7.54$ in our model.

Eq. 9 also predicts, not surprisingly, that ΔH (total change in enthalpy) for protein folding should increase linearly with protein chain length, N . Fig. 2 shows the comparison with experiments. From these plots, $g_0 = \Delta h(T = 100^\circ\text{C}) = -5.03 \text{ kJ/mol} = -1.20 \text{ kcal/mol}$ (figure 5b in ref. 13). It also establishes that ΔC_p (total change in heat capacity) increases linearly with chain length N , thus $\Delta C_p = N\Delta c_p$ and $\Delta c_p = -62 \text{ J/K mol} = -14.8 \text{ cal/(K-mol)}$ (figure 3 in ref. 13).

The Ideal Thermal Protein. Some proteins fall on the average lines of the enthalpy, entropy, and heat capacity of folding vs. chain length shown in Fig. 2, and others scatter around those lines. We introduce the notion of an ideal thermal protein (ITP) as any protein for which the enthalpy, entropy, and heat capacity falls exactly on those lines. The ITP represents an average protein in its thermal properties for a given chain length N ; it is represented by the parameters $\Delta c_p = -14.8 \text{ cal/(K-mol)}$, $T_h = 100^\circ\text{C}$, $T_s = 112^\circ\text{C}$, $g_0 = -1.20 \text{ kcal/mol}$ and $z = 7.54$. Because the thermal properties depend predominantly on chain length, the ITP simply defines the expectation for thermal properties of a protein of with chain length N . If you knew only the chain length N of a protein, and did not know its value of C_p , T_h , and T_s , then the ITP values would give you the best initial estimate of the expected thermal properties prior to doing the thermal experiment. We use these ITP

values for predictions in some cases, as described in the rest of the article. Of course, if Δc_p , T_h , T_s , g_0 , z have been measured for a particular protein, then substituting those values should give better predictions for a protein's stability than the ITP estimate of it.

Our model predicts that ΔF_{fold} should also depend linearly on N but with a slope near 0. Consistent with this prediction, for real proteins, stabilities are not very dependent on N . There is a compensation of enthalpy and entropy. Any small systematic dependence of ΔF_{fold} on N is overwhelmed by a larger noise contribution due to the scatter among proteins.

Confinement Can Increase the Stability of a Protein

Proteins are often in confined environments (i.e., contained within a small restrictive space) or in crowded spaces (i.e., within a dense medium of other molecules or organelles). Crowding and confinement are common in biology: proteins fold inside chaperones and ribosomes, and the interior of a typical cell is nearly solid-like, with organelles and large molecules occupying nearly 25% of the volume. Both crowding and confinement can affect a protein's stability. In the simplest cases, the reason is simple: the native structure is not affected by confinement, but some of the most expanded denatured conformations are no longer sterically viable in a confined space, making the denatured state entropically more unfavorable than when the protein was not confined. Here, generalize the simple model above using the confinement theory of Zhou and Dill, to consider the stabilities of proteins in small constraining volumes. We want to compute $\Delta \Delta F = (F_{N,\text{confine}} - F_{D,\text{confine}}) - (F_{N,\text{bulk}} - F_{D,\text{bulk}})$. $F_N - F_D$ is the energy of folding, native minus denatured, and the subscripts indicate the difference in stability of the protein in a confined environment relative to its stability in the bulk. If you confine a random-flight polymer chain (modeling the denatured state) having end-to-end distance $R_d^2 = Nb^2$ (b is the bond length) into a cube where each linear dimension is $s < R_d$, a good approximation to the steric reduction of the conformational options is given by (27, 30):

$$\frac{F_{D,\text{confine}} - F_{D,\text{bulk}}}{kT} = \left(\frac{\pi^2}{2}\right) \left(\frac{R_d}{s}\right)^2 - 3 \ln \left(\frac{8s}{\pi^2 r}\right) \quad [12]$$

where r is radius of the larger (macroscopic) box that defines the bulk container (which becomes irrelevant when we take differences below). Second, confinement restricts the volume available to the native protein because the protein cannot penetrate any closer to the walls of the cavity than $2R_n$, which is the diameter of the protein in the native state:

$$\frac{F_{N,\text{confine}} - F_{N,\text{bulk}}}{kT} = -3 \ln \left(\frac{s - 2R_n}{r}\right). \quad [13]$$

Combining Eqs. 12 and 13 gives

$$\frac{\Delta\Delta F}{kT} = -\left(\frac{\pi^2}{2}\right)\left(\frac{R_d}{s}\right)^2 - 3\ln\left(1 - \frac{2R_n}{s}\right) + 3\ln\left(\frac{8}{\pi^2}\right). \quad [14]$$

We can compute how this confinement-induced stabilization affects the denaturation temperature of the protein. The energy of folding of a confined protein is $\Delta F_{\text{fold,confine}} = \Delta F_{\text{fold}} + \Delta\Delta F$ where ΔF_{fold} is from Eq. 11 and $\Delta\Delta F$ is from Eq. 14. The melting temperature, T_{mc} of the confined protein is that value of the temperature at which $\Delta F_{\text{fold,confine}} = 0$. Fig. S1 shows the prediction that the protein is stabilized by confinement, because confinement eliminates some of the denatured conformations that are available to the unconfined chain. That is, confinement increases the denaturation temperature. This is indicated by the increased melting temperatures in smaller cages (except for the very smallest ones, which also affect the native state). Fig. S1 shows the behavior of the ITP model for α -lactalbumin [123 amino acids and $R_n = 15.7$ Å and $R_d = 73.5$ Å (32)] to compare with the confinement experiments of Eggers et al. (31). The predicted maximum increase in melting temperature is 40 °C, depending on the confinement box size (not known exactly), close to their observed 25–32 °C increase (31).

We can also put Eq. 14 into more useful terms by expressing R_n and R_d as functions of the chain length N of the protein. We use $R_n = 2.3N^{0.35}$ to capture the native state radius of gyration dependence on chain length (33) and for the value of $R_d^2 = Nb^2$, we use $b = 8$ Å (30). Thus, we obtain

$$\frac{\Delta\Delta F}{kT} = -\left(\frac{\pi^2 N}{2}\right)\left(\frac{b}{s}\right)^2 - 3\ln\left(1 - \frac{4.6N^{0.35}}{s}\right) + 3\ln\left(\frac{8}{\pi^2}\right). \quad [15]$$

Mittal and Best (34) have recently studied confinement in a coarse-grained computer model. We plot our data in the same form as theirs, in terms of the confinement melting temperature normalized by the bulk melting temperature $[(T_m^c - T_m^{\text{bulk}})/T_m^{\text{bulk}}]$ as a function of the normalized cage size in the units of denatured state radius of gyration R_g^u (see Fig. S2). They treated excluded volume but not hydrophobicity. We treat hydrophobicity but not excluded volume. We find a linear regime where our slope falls below 2 due to the hydrophobic interactions. In the limit of $\Delta C_p \rightarrow 0$, our model recovers the usual Gaussian chain confinement scaling, i.e., a slope of 2.

Electrostatics: Predicting How the Stability of a Protein Depends on pH and Salt

In this section, we consider how net electrostatic charges on a protein affect its stability, as functions of pH and salt (35, 36). We define the total folding energy between the folded and the unfolded state as $\Delta F_{\text{tot}} = F_N - F_D$ where F_N is the energy of the native state and F_D is that of the denatured state. We parse the folding energy into two components as

$$\Delta F_{\text{tot}} = \Delta F_{\text{neutral}} + \Delta F_{\text{elec}}(T, pH, c_s), \quad [16]$$

where $\Delta F_{\text{neutral}}$ is the energy in the absence of any charge effect and $\Delta F_{\text{elec}}(T, pH, c_s)$ is the electrostatic contribution to the energy as a function of pH and salt concentration (c_s). We model the neutral part of the energy as before by Eq. 11. We treat electrostatics using a simple Born and Debye-Huckel treatment. Within this model, the electrostatic energy is expressed as

$$\frac{\Delta F_{\text{elec}}(T, pH, c_s)}{kT} = \frac{Q_n^2 l_b}{2R_n(1 + \kappa R_n)} - \frac{Q_d^2 l_b}{2R_d(1 + \kappa R_d)}, \quad [17]$$

where Q_n is the total charge on the native protein and Q_d is the total charge on the denatured protein. l_b is the Bjerrum length defined as

$$l_b = \frac{Ce^2}{\epsilon kT} = \frac{1.39 \times 10^{-4} \text{ J m mol}^{-1}}{\epsilon RT}, \quad [18]$$

e is the charge on an electron, R is the gas constant, k is Boltzmann's constant, T is the absolute temperature, and ϵ is the dielectric constant of water. At room temperature, $l_b = 7.13$ Å (27). The justification for using the dielectric constant of water for the native protein is because charged side chains are mainly on the protein's surface and are largely solvated in both native and denatured states. We also define κ as

$$\kappa^2 = 2c_s l_b, \quad [19]$$

where c_s is the salt concentration. The value of R_n (radius of gyration in the native state) is determined from the native structure PDB files except for myoglobin. For myoglobin, we use the value of $R_n = 18.3$ Å previously reported (36, 37). The radii of gyration for the denatured states were obtained from the experimental measurements (table 2 in ref. 32). Finally the net charge, Q_n , is computed as

$$Q_n = Q_b - Q_a, \quad [20]$$

where the net charge on the basic (b) groups, Q_b , is expressed as

$$Q_b = \sum_{i=1}^b \frac{10^{pK_i - pH}}{1 + 10^{pK_i - pH}}, \quad [21]$$

where i is the index that counts the basic groups on side chains and pK_i is the proton dissociation constant for those side chains measured under native conditions. Similarly, for the acid groups (a), the net charge is Q_a

$$Q_a = \sum_{i=1}^a \frac{10^{pH - pK_i}}{1 + 10^{pH - pK_i}}. \quad [22]$$

We obtain the values of the pK_i values of the native states from table 1 of Hellinga et al. (38) for lysozyme and RNase A. For myoglobin we use table II from Breslow et al. (39). The pK_i values for the denatured states are obtained from standard pK values of the side chains for the given protein sequence. The total energy of the protein as a function of temperature and pH is now expressed as (using Eqs. 11, 16 and 17)

$$\frac{\Delta F_{\text{tot}}(T, pH, c_s)}{kT_0} = N \left[\frac{g_0}{kT_0} + \frac{l_b}{2N} \left(\frac{Q_n^2}{R_n(1 + \kappa R_n)} - \frac{Q_d^2}{R_d(1 + \kappa R_d)} \right) + \frac{\Delta c_p}{kT_0} (T - T_h) + \frac{T}{T_0} \ln z - \frac{T}{kT_0} \Delta c_p \ln \frac{T}{T_s} \right]. \quad [23]$$

We assume T_0 to be the room temperature and express energy in the units of kT_0 . Our aim now is to express this free energy in a form that is practical for analyzing experiments. Experiments are often done at values of temperature, pH, and salt that are experimentally convenient, rather than at some standardized values at which proteins could be optimally compared. If experiments are not performed at the isoelectric pH, it is not simple to obtain g_0 and z needed for our model. So, here we give a practical approach for obtaining the key quantities needed. Recognizing that the temperature dependence of the electrostatics term, $F_{\text{elec}}(T, pH, c_s)$, is

Table 1. Values of extracted parameters from calorimetric measurements and other parameters used

Protein name	g_0 , k cal/mol	z	ΔC_p , cal/K mol	R_n , Å	R_d , Å
Myoglobin	-1.19	6.55 (ref. 40)	-13.75 (ref. 40)	18.3 (ref. 37)	34.2 (ref. 32)
RNAse A	-1.27	7.2 (ref. 41)	-9.3 (ref. 41)	14.5	27.3 (ref. 32)
Lysozyme hen egg	-1.45	9.20 (ref. 42)	-12.39 (ref. 42)	15.1	28.7 (ref. 32)

generally weak (only entering through κ), we collect terms into an effective interaction quantity,

$$\frac{g_{\text{eff}}(pH, c_s)}{kT_0} = \frac{g_0}{kT_0} + \frac{l_b}{2N} \left(\frac{Q_n^2}{R_n(1 + \kappa R_n)} - \frac{Q_d^2}{R_d(1 + \kappa R_d)} \right). \quad [24]$$

We now express the free energy of folding as

$$\frac{F_{\text{tot}}(T, pH, c_s)}{kT_0} = N \left[\frac{g_{\text{eff}}(pH, c_s)}{kT_0} + \frac{\Delta c_p}{kT_0} (T - T_h) + \frac{T}{T_0} \ln z - \frac{T}{kT_0} \Delta c_p \ln \frac{T}{T_s} \right]. \quad [25]$$

At the melting temperature (T_m), we have $F_{\text{tot}}(T_m, pH, c_s) = 0$. The enthalpy and entropy components are now

$$\Delta H(T_m, pH, c_s) = N[g_{\text{eff}}(pH, c_s) + \Delta c_p(T_m - T_h)] \quad [26]$$

and

$$\Delta S(T_m, pH, c_s) = -N \left[k \ln z - \Delta c_p \ln \frac{T_m}{T_s} \right]. \quad [27]$$

The virtue of the present formulation is that the thermal measurement of free energy of folding at any convenient experimental pH (call it pH_e) and salt, not necessarily at the isoelectric pH (where the net charge would be zero), can be used to predict free energy of folding at any other pH and salt. This can be achieved by following the standard experimental procedure where

one measures free energy of folding as a function of temperature $\Delta F(T, pH_e, c_s)$ written alternatively as

$$\Delta F(T, pH_e, c_s) = \Delta H(T_m, pH_e, c_s) + \Delta C_p(T - T_m) - T \Delta S(T_m, pH_e, c_s) - T \Delta C_p \ln \frac{T}{T_m}. \quad [28]$$

First, we fit the experimental data with Eq. 28 and extract $\Delta H(T_m, pH_e, c_s)$, $\Delta S(T_m, pH_e, c_s)$ and ΔC_p . We assume ΔC_p (13, 40–42) is approximately independent of temperature or pH and salt. We also keep the salt concentration fixed. Then, we find $g_{\text{eff}}(pH_e, c_s)$ and z from Eqs. 26 and 27. Finally, we use Eq. 24 to compute the two charge states of the protein, Q_n and Q_d at pH_e and we extract g_0 . Having obtained g_0 and z , we can now use Eqs. 24 and 25 to compute the folding free energy $F_{\text{tot}}(T, pH, c_s)$ for any arbitrary conditions. In particular, this allows us to predict melting temperature of the protein at any pH other than pH_e . This model approach provides a way to standardize the thermal properties of protein folding. This is how we use thermodynamic data for myoglobin (40), lysozyme (41) and RNAseA (42) to predict melting temperature vs pH phase diagrams and compare with experiments (5, 7). Table 1 reports the values of g_0 , z , ΔC_p , R_n and R_d that we used for predicting myoglobin, lysozyme and RNAseA phase diagrams (Fig. 3). We also tested the Tanford expression for electrostatics (43), but our approach better treats charge–charge interactions and ionic strength and also gives better fits to the data. We believe the remaining discrepancies between experimental and theoretical values are due to subtle effects: (i) of side chain pK values, (ii) that denatured-state conformations depend on pK 's (44), and (iii) the denatured state radius depends on pH and buffer, not treated here. Nevertheless, given the absence of fitting parameters and the simplicity of the model, we consider the agreement quite good.

We also compute the phase diagram for the denaturation as functions of pH and ionic strength (c_s) by calculating zeros of $F_{\text{tot}}(T, pH, c_s)$ for a given temperature T . We compute this for myoglobin [sperm whale (45)], using the same values of g_0 , z , ΔC_p , R_n and R_d reported in Table 1. At low pH , the agreement

Fig. 3. T_m vs. pH for 3 different proteins with no fitting parameter. The parameter values used have been reported in Table 1. For myoglobin we used pK values from ref. 39 and thermodynamic values from ref. 40. For Lysozyme and RNAse A we used pK values from Hellinga et al. (38) and thermodynamic values reported in refs. 41 and 42, respectively. Solid lines show theoretical predictions, and filled circles represent experimental data (5, 7). Red dotted line shows our theoretical prediction of the melting temperature as a function of pH for myoglobin in the presence of confinement ($s = 50$ Å). The parameters used for the confinement effect are the same as the unconfined case.

Fig. 4. pH vs. Ionic strength phase diagram for Sperm whale myoglobin at 25 °C. Filled circles are data from Friend et al. (36, 45), and solid lines are our theoretical prediction without any fit parameter. Red solid lines denote our prediction for the same in the presence of a confinement with $s = 50$ Å.

is generally good (see Fig. 4), except that the slope is predicted slightly less accurately than in a more sophisticated treatment (36), where the small dependence on salt of the radius of the denatured state is also treated. We show here also the predictions for stabilities at high pH, although we know of no data for this protein under those conditions.

Our model above allows for the calculation of a large number of different phase diagrams for proteins, with various combinations of variables, including temperature, denaturants, pH, salts, net charge, or confinement cavity size, for example. For illustration, we compute how the pH-vs.-ionic strength phase diagram for myoglobin would change upon transferring myoglobin from the bulk to confinement within a cavity of radius 50 Å.

Confinement of a Protein Affects Its Acid-Base Stability. We combine Eqs. 15 and 23 and we solve for T_m . The results are shown in Fig. 3. Because confinement stabilizes proteins, the model predicts that the pH that would be required inside a cavity to denature the protein would be lower (or the salt concentration must be lower) than would be required for the same protein in the bulk.

A General Expression for Stability. We now combine the various factors above that contribute to protein stability in the bulk into a single general expression for the stability of a protein as functions of temperature, denaturant, pH, and salt. Adding the denaturant term to Eq. 23 gives

$$\frac{\Delta F_{\text{tot}}(T, pH, c_s, c)}{kT_0} = N \left[\frac{g_0 + m_1 c}{kT_0} + \frac{l_b}{2N} \left(\frac{Q_n^2}{R_n(1 + \kappa R_n)} - \frac{Q_d^2}{R_d(1 + \kappa R_d)} \right) \right] + \frac{\Delta c_p}{kT_0} (T - T_h) + \frac{T}{T_0} \ln z - \frac{T}{kT_0} \Delta c_p \ln \frac{T}{T_s} \quad [29]$$

where c is the concentration of denaturant. Confinement effects can be added by combining Eq. 15 with Eq. 29.

Summary

We describe here a method for computing the stabilities of globular proteins. If you know only the chain length (the number of amino acids N) and the numbers of acidic and basic groups in a protein, then the present model can be used to compute the enthalpy, entropy, heat capacity and free energy of folding vs. temperature, denaturants, pH, salts, and as a function of confinement (if the protein is contained within a small constraining volume). The model predicts properties for the corresponding ITP, i.e., the average protein of this chain length. If, instead, you have direct measurements of the thermal properties, such as the free energy of folding vs. temperature, then the model shows how to compute an approximation that should be better than the ITP. One caveat is that the folding enthalpy and entropy are large quantities having small variances, while their sum, the free energy, is a small difference having a large variance. So the latter property may be less accurately predicted by the ITP approximation. A key finding here is that a very simple model for the electrostatics predicts well the effects of pH and salts without adjustable parameters. We believe this theoretical model may be a useful practical tool for rapidly and efficiently estimating the energetic properties of unknown proteins, such as are becoming widely available from genomics efforts.

ACKNOWLEDGMENTS. We thank Prof. Nick Pace and Gerald R. Grimsley for providing various pK values and Prof. Asoke Chandra Ghose for helpful discussion. This work was supported by Amgen and National Institutes of Health Grant GM 34993.

1. Shell MS, et al. (2009) Blind test of physics-based prediction of protein structures. *Biophys J* 96:917–924.
2. Myers JK, et al. (1996) Hydrogen bonding stabilizes globular proteins. *Biophys J* 71:2033–2039.
3. Chen JM, et al. (2001) Packing is a key selection factor in the evolution of protein hydrophobic cores. *Biochemistry* 40:14004–14011.
4. Plaxco KW, et al. (2000) Topology, stability, sequence, and length: Defining the determinants of two-state protein folding kinetics. *Biochemistry* 39:11177–11183.
5. Privalov PL, et al. (1974) Thermodynamic approach to problem of stabilization of globular protein structure—calorimetric study. *J Mol Biol* 86:665–684.
6. Privalov PL, et al. (1988) Stability of protein-structure and hydrophobic interaction. *Adv Prot Chem* 39:191–234.
7. Privalov PL, et al. (1971) Thermodynamic analysis of thermal transitions in globular proteins. 1. Calorimetric study of ribotrypsinogen, ribonuclease and myoglobin. *Biopolymers* 10:1865–1890.
8. Makhatazde GI, et al. (1990) Heat-capacity of proteins. 1. Partial molar heat-capacity of individual amino-acid-residues in aqueous-solution—hydration effect. *J Mol Biol* 213:375–384.
9. Makhatazde GI, et al. (1993) Contribution of hydration to protein-folding thermodynamics. 1. The enthalpy of hydration. *J Mol Biol* 232:639–659.
10. Baldwin RL (2007) Energetics of protein folding. *J Mol Biol* 371:283–301.
11. Zhou HX (2002) Toward the physical basis of thermophilic proteins: Linking of enriched polar interactions and reduced heat capacity of unfolding. *Biophys J* 83:3126–3133.
12. Bolen DW, et al. (2008) Structure and energetics of the hydrogen-bonded backbone in protein folding. *Annu Rev Biochem* 77:339–362.
13. Robertson AD, Murphy KP (1997) Protein structure and the energetics of protein stability. *Chem Rev* 97:1251–1267.
14. Tanford C (1968) Protein denaturation. *Adv Prot Chem* 23:121–282.
15. Brandts JF (1964) Thermodynamics of protein denaturation. 2. Model of reversible denaturation + interpretations regarding stability of chymotrypsinogen. *J Am Chem Soc* 86:4302–4314.
16. Dill KA (1985) Theory for the folding and stability of globular-proteins. *Biochemistry* 24:1501–1509.
17. Dill KA, et al. (1995) Modeling protein stability as heteropolymer collapse. *Adv Prot Chem* 46:59–104.
18. Timasheff SN (1993) The control of protein stability and association by weak-interactions with water—How do solvents affect these processes. *Ann Rev Biophys Biomol Struct* 22:67–97.
19. Davis-Searles PR, et al. (2001) Interpreting the effects of small uncharged solutes on protein-folding equilibria. *Ann Rev Biophys Biomol Struct* 30:271–306.
20. Auton M, et al. (2005) Predicting the energetics of osmolyte-induced protein folding/unfolding. *Proc Natl Acad Sci USA* 102:15065–15068.
21. Alonso DOV, et al. (1991) Solvent denaturation and stabilization of globular-proteins. *Biochemistry* 30:5974–5985.
22. Greene RF, et al. (1974) Urea and guanidine-hydrochloride denaturation of ribonuclease, lysozyme, alpha-chymotrypsin, and beta-lactoglobulin. *J Biol Chem* 249:5388–5393.
23. Myers JK, et al. (1995) Denaturant m-values and heat-capacity changes—Relation to changes in accessible surface-areas of protein unfolding. *Prot Sci* 4: 2138–2148.
24. Baskakov IV, et al. (1998) Monitoring the sizes of denatured ensembles of staphylococcal nuclease proteins: Implications regarding m values, intermediates, and thermodynamics. *Biochem* 37:18010–18017.
25. Makhatazde GI, et al. (1995) Energetics of protein structure. *Adv Prot Chem* 47:307–425.
26. Dill KA (1990) Dominant forces in protein folding. *Biochemistry* 29:7133–7155.
27. Dill KA, Bromberg S (2002) *Molecular Driving Forces* (Garland Science, London).
28. Baldwin RL (1986) Temperature-dependence of the hydrophobic interaction in protein folding. *Proc Natl Acad Sci USA* 83:8069–8072.
29. Makhatazde GI, et al. (1996) On the entropy of protein folding. *Prot Sci* 5:507–510.
30. Zhou HX, et al. (2001) Stabilization of proteins in confined spaces. *Biochemistry* 40:11289–11293.
31. Eggers DK, et al. (2001) Molecular confinement influences protein structure and enhances thermal protein stability. *Prot Sci* 10:250–261.
32. Kamatari YO, et al. (1999) The compact and expanded denatured conformations of apomyoglobin in the methanol-water solvent. *Prot Sci* 8:873–882.
33. Dewey TG, et al. (1993) Protein-structure and polymer collapse. *J Chem Phys* 98:2250–2257.
34. Mittal J, et al. (2008) Thermodynamics and kinetics of protein folding under confinement. *Proc Natl Acad Sci USA* 105:20233–20238.
35. Linderstrom-Lang KU (1924) On the ionisation of proteins. *Compt Rend Trav Lab Carlsberg* 15:1–29.
36. Stigter D, et al. (1991) Protein stability—Electrostatics and compact denatured states. *Proc Natl Acad Sci USA* 88:4176–4180.
37. Breslow E, et al. (1962) Reactivity of sperm whale metmyoglobin towards hydrogen ions and P-nitrophenyl acetate. *J Biol Chem* 237:371–381.
38. Wisz MS, et al. (2003) An empirical model for electrostatic interactions in proteins incorporating multiple geometry-dependent dielectric constant. *Protein Struct Funct Genet* 51:360–377.
39. Breslow E (1964) Changes in side chain reactivity accompanying binding of heme to sperm whale apomyoglobin. *J Biol Chem* 239:486–496.
40. Pivalov PL, et al. (1986) Cold denaturation of myoglobin. *J Mol Biol* 190:487–498.
41. Straume M, et al. (1992) 2-dimensional differential scanning calorimetry—Simultaneous resolution of intrinsic protein structural energetics and ligand-binding interactions by global linkage analysis. *Annal Biochem* 203:259–268.
42. Pfeil W, et al. (1976) Thermodynamic investigations of proteins. 1. Standard functions for proteins with lysozyme as an example. *Biophys Chem* 4:23–32.
43. Tanford C (1970) Protein denaturation. C. Theoretical models for the mechanism of denaturation. *Adv Prot Chem* 24:1–95.
44. Zhou HX (2002) A Gaussian-chain model for treating residual charge-charge interactions in the unfolded state of proteins. *Proc Natl Acad Sci USA* 99: 3569–3574.
45. Friend SH, et al. (1979) Electrostatic stabilization in myoglobin—Ph-dependence of summed electrostatic contributions. *Biochemistry* 18:4612–4619.