Molecular Origin of Constant *m*-Values, Denatured State Collapse, and Residue-Dependent Transition Midpoints in Globular Proteins[†]

Edward P. O'Brien,[‡,§] Bernard R. Brooks,[§] 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 Laboratory of Computational Biology, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland 20892

Received November 15, 2008; Revised Manuscript Received February 27, 2009

ABSTRACT: Experiments show that for many two-state folders the free energy of the native state, $\Delta G_{ND}([C])$, changes linearly as the denaturant concentration, [C], is varied. The slope $\{m = [d\Delta G_{ND}([C])]/(d[C])\}$, is nearly constant. According to the transfer model, the *m*-value is associated with the difference in the surface area between the native (N) and denatured (D) state, which should be a function of ΔR_g^2, the difference in the square of the radius of gyration between the D and N states. Single-molecule experiments show that the R_g of the structurally heterogeneous denatured state undergoes an equilibrium collapse transition as [C] decreases, which implies *m* also should be [C]-dependent. We resolve the conundrum between constant *m*-values and [C]-dependent changes in R_g using molecular simulations of a coarse-grained representation of protein L, and the molecular transfer model, for which the equilibrium folding can be accurately calculated as a function of denaturant (urea) concentration. In agreement with experiment, we find that over a large range of denaturant concentration (>3 M) the *m*-value is a constant, whereas under strongly renaturing conditions (<3 M), it depends on [C]. The *m*-value is a constant above [C] > 3 M because the [C]-dependent changes in the surface area of the backbone groups, which make the largest contribution to *m*, are relatively narrow in the denatured state. The burial of the backbone and hydrophobic side chains gives rise to substantial surface area changes below [C] < 3 M, leading to collapse in the denatured state of protein L. Dissection of the contribution of various amino acids to the total surface area change with [C] shows that both the sequence context and residual structure are important. There are [C]-dependent variations in the surface area for chemically identical groups such as the backbone or Ala. Consequently, the midpoints of transition of individual residues vary significantly (which we call the Holtzer effect) even though global folding can be described as an all-or-none transition. The collapse is specific in nature, resulting in the formation of compact structures with appreciable populations of nativelike secondary structural elements. The collapse transition is driven by the loss of favorable residue−solvent interactions and a concomitant increase in the strength of intrapeptide interactions with a decreasing [C]. The strength of these interactions is nonuniformly distributed throughout the structure of protein L. Certain secondary structure elements have stronger [C]-dependent interactions than others in the denatured state.

The folding of many small globular proteins is often modeled using the two-state approximation in which a protein is assumed to exist in either the native (N)[1] or denatured (D) state (*1*). The stability of N relative to D, $\Delta G_{ND}(0)$, is typically obtained by measuring $\Delta G_{ND}([C])$ as a function of the denaturant concentration, [C], and extrapolating to [C] = 0 using the linear extrapolation method (*2*). The denaturant-dependent change in native state stability, $\Delta G_{ND}([C])$, for these globular proteins is usually a linear function of [C] (*2*−*9*). Thus, $\Delta G_{ND}([C]) = \Delta G_{ND}(0) + m[C]$, where $m = \partial\Delta G_{ND}([C])/\partial[C]$ is a constant (*5*), which by convention is called the *m*-value. However, deviations from linearity, especially at low [C], have also been found (*10*), indicating that the *m*-value is concentration-dependent. In this paper, we address two inter-related questions: (1) Why are *m*-values constant for some proteins, even though there is a broad distribution of conformations in the denatured state ensemble (DSE)? (2) What is the origin of denatured state collapse, that is, the compaction of the DSE, with a decreasing [C] that is often associated with nonconstant *m*-values (*10*−*12*)?

Potential answers to the first question can be gleaned by considering the empirical transfer model (TM) (*13*−*15*), which has been remarkably successful in accurately predict-

[†] This work was supported in part by a grant from the National Science Foundation (05-14056) and the Air Force Office of Scientific Research (FA9550-07-1-0098) to D.T., by a National Institutes of Health GPP Biophysics Fellowship to E.P.O., and by the Intramural Research Program of the National Heart, Lung and Blood Institute.

* To whom correspondence should be addressed: Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742. Phone: (301) 405-4803. Fax: (301) 314-9404. E-mail: thirum@umd.edu.

[‡] University of Maryland.

[§] National Institutes of Health.

[1] Abbreviations: TM, transfer model; MTM, molecular transfer model; D, denatured state; N, native state; DSE, denatured state ensemble; NSE, native state ensemble; C_α-SCM, C_α side chain model; MREX, multiplexed replica exchange; GdmCl, guanidinium chloride; NMR, nuclear magnetic resonance; FRET, Forster resonance energy transfer; R_{ee}, end-to-end distance.

10.1021/bi8021119 CCC: $40.75
Published on Web 03/11/2009

ing *m*-values for a large number of proteins (*15, 16*). The revival of the TM as a practical tool in analyzing the effect of denaturants (and more generally osmolytes) comes from a series of pioneering studies by Bolen and co-workers (*15*−*17*). Assuming that proteins exist in only two states (*8, 15*), the TM expression for the *m*-value is

$$m = \frac{1}{[C]}\sum_{k=1}^{N_S}\frac{n_k\delta g_k^S([C])}{\alpha_{k,G-k-G}^S}\Delta\alpha_k^S + \frac{1}{[C]}\sum_{k=1}^{N_B}\frac{n_k\delta g_k^B([C])}{\alpha_{k,G-k-G}^B}\Delta\alpha_k^B \quad (1)$$

where the sums are over the side chain (S) and backbone (B) groups of the different amino acid types (Ala, Val, Gly, etc.), n_k is the number of amino acid residues of type k in the protein, and δg_k^S and δg_k^B are the experimentally measured transfer free energies for k (*13, 17, 18*) (Figure 1a). In eq 1, $\Delta\alpha_k^P = \langle\alpha_{k,D}^P\rangle - \langle\alpha_{k,N}^P\rangle$ (P = S or B), where $\langle\alpha_{k,D}^P\rangle$ and $\langle\alpha_{k,N}^P\rangle$ are the average solvent accessible surface areas (*19*) of group k in the D and N states, respectively, and $\alpha_{k,G-k-G}^P$ is the corresponding value in the tripeptide glycine-*k*-glycine. There are two fundamentally questionable assumptions in the TM model: (1) The free energy of transferring a protein from water to aqueous denaturant solution at an arbitrary [C] may be obtained as a sum of transfer energies of individual groups of the protein without regard to the polymeric nature of proteins. (2) The surface area changes $\Delta\alpha_k^P$ are independent of [C], the residual denatured state structure, and the amino acid sequence context in which k is found.

The linear variation of $\Delta G_{ND}([C])$ as [C] changes can be rationalized if (i) $\delta g_k^P([C])$ is directly proportional to [C] and (ii) $\Delta\alpha_k^P$ is [C]-independent. Experiments have shown that $\delta g_k^P([C])$ is a linear function of [C] (*7*) while the near independence of $\Delta\alpha_k^P$ from [C] can be inferred only on the basis of the accuracy of the TM in predicting *m*-values (*15, 16*). In an apparent contradiction to such an inference, small-angle X-ray scattering experiments (*20*−*23*) and single-molecule FRET experiments (*24*−*29*) show that the denatured state properties, such as the radius of gyration R_g and the end-to-end distance (R_{ee}), can change dramatically as a function of [C]. These observations suggest that the total solvent accessible surface area of the protein, $\Delta\alpha_T$ ($=\sum_{k=1}^{N_S}\Delta\alpha_k^S + \sum_{k=1}^{N_B}\Delta\alpha_k^B$), and the various groups must also be a function of [C], since we expect that $\Delta\alpha_T$ must be a monotonically increasing function of ΔR_g^2, which is the difference between the R_g^2 of the D and N states (*26, 30*). For compact objects, $\Delta\alpha_T \propto \Delta R_g^2$, but for fractal structures, the relationship is more complex (*31*). Furthermore, NMR measurements have found that many proteins adopt partially structured or random coil-like conformations at high [C] values (*32*−*35*), which necessarily have large fluctuations in global properties such as $\Delta\alpha_k^P$ and R_g. Thus, the contradiction between the constancy of *m*-values and the sometimes measurable changes in denatured state properties is a puzzle that requires a molecular explanation.

Bolen and collaborators have shown that quantitative estimates of *m* can be made by using measured transfer free energies of model compounds (*15, 16*). More importantly, these studies established the dominant contribution to *m* arises from the backbone (*15, 16*). However, only by characterizing the changes in the distribution of $\Delta\alpha_k^S$ and $\Delta\alpha_k^B$ as a function of [C] can the reasons for the success of the TM in obtaining the global property *m* be fully appreciated. This is one of the goals of the study presented here. In addition, we correlate *m* with denatured state collapse, [C]-dependent changes in residual structure, and the solution forces acting on the denatured state, properties that cannot be analyzed using the TM.

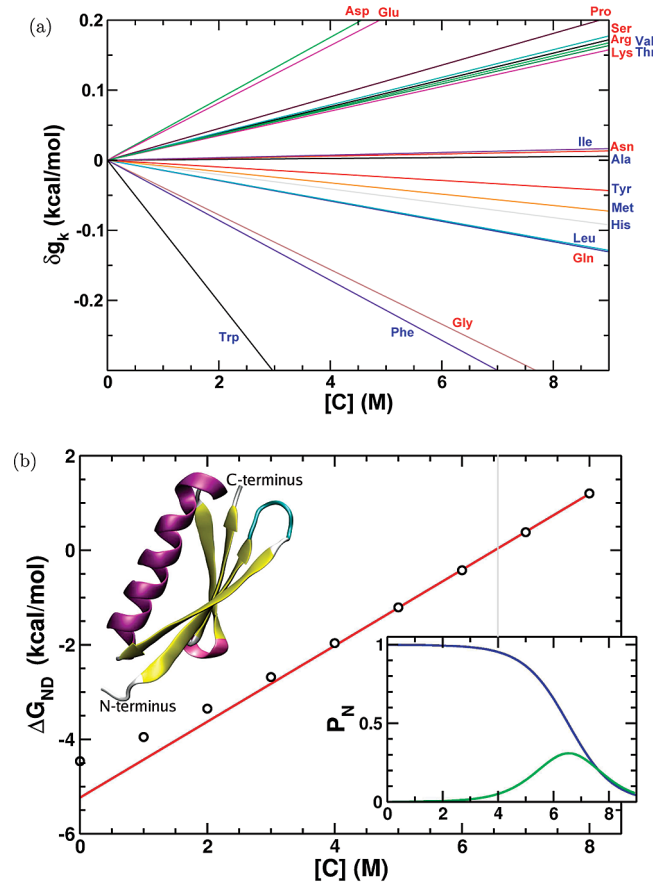

FIGURE 1: (a) Transfer free energy of the backbone (the glycine residue) and side chain groups as a function of urea concentration. The lines are a linear extrapolation of the experimentally measured δg_k upon transfer from 0 to 1 M urea (*15*). The amino acid corresponding to a given line is labeled using a three-letter abbreviation. Blue labels are for hydrophobic side chains, while red labels indicate polar or charged side chains according to the hydrophobicity scale in ref *63*. (b) Native state stability (○) of protein L as a function of urea concentration, [C], at 328 K. $\Delta G_{ND}([C]) = -k_BT\ln\{P_N([C])/[1 - P_N([C])]\}$, where $P_N([C])$ is the probability of being folded as a function of [C]. The midpoint of the transition $C_m = 6.56$ M urea. The red line is a linear fit to the data in the range of 5.1−7.9 M. At [C] < 3 M, there is a departure from linearity (i.e., a [C]-dependent *m*-value). The top left inset is a ribbon diagram of the crystal structure of protein L (*48*). The bottom right inset shows $P_N([C])$ vs [C] at 328 K (blue line). In addition, $|dP_N/d[C]|$, the absolute value of the derivative of P_N vs [C], is shown (green line). The full width at half-maximum of $|dP_N/d[C]|$ (denoted $2\delta C$) is 2.8 M and is defined as the "transition region" given by $C_m \pm \delta C$.

The denatured and perhaps even the native state should be described as ensembles of fluctuating conformations and will here be named DSE and NSE (native state ensemble), respectively. As a result, it is crucial to characterize the distribution of various molecular properties in these ensembles and how they change with [C] to describe quantitatively the properties of the DSE. Because the D state is an ensemble of conformations with a distribution of accessible surface areas, eq 1 should be considered an approximate expression for the *m*-value. Even if the basic premise of the

TM is valid, we expect that $\Delta\alpha_k^P$ should depend on the conformation of the protein and the denaturant concentration. Consequently, the *m*-value should be written with an explicit concentration dependence as

$$m([C]) = \frac{1}{[C]}\sum_{k=1}^{N_S}\frac{n_k\delta g_k^S([C])}{\alpha_{G-k-G}^S}[\langle\alpha_{k,D}^S([C])\rangle - \langle\alpha_{k,N}^S([C])\rangle] + \frac{1}{[C]}\sum_{k=1}^{N_B}\frac{n_k\delta g_k^B([C])}{\alpha_{G-k-G}^B}[\langle\alpha_{k,D}^B([C])\rangle - \langle\alpha_{k,N}^B([C])\rangle] \quad (2)$$

where $\langle\alpha_{k,j}^P([C])\rangle = \int_0^\infty \alpha_{k,j}^P P(\alpha_{k,j}^P;[C])\, d\alpha_{k,P}$ (j = D or N, and P = S or B). In principle, the denominator in eq 2 should also be [C]-dependent; however, we ignore this for the sake of simplicity. In contrast to eq 1, the conformational fluctuations in the DSE and NSE are taken into account in eq 2 by integrating over the distribution of surface areas [$P(\alpha_{k,j}^P;[C])$]. Moreover, we do not assume that the surface area distributions are independent of [C] as we do in eq 1. Such an assumption can be justified only by evaluating $P(\alpha_{k,j}^P;[C])$ using molecular simulations or experiments.

We use the molecular transfer model (MTM) (*36*) in conjunction with coarse-grained simulations of protein L using the C_α side chain model (C_α-SCM) (see Methods) to test the molecular origin of the constancy of *m*-values. Because the conformations and energies are known exactly in the C_α-SCM simulations, we can determine how an ensemble of denatured conformations, with a distribution of solvent accessible areas in the DSE, gives rise to a constant *m*-value. We show that the *m*-values are nearly constant for two reasons. (1) As previously shown (*15, 16*), the bulk of the contribution to $\Delta G_{ND}([C])$ changes comes from the protein backbone. (2) Here, we establish that the distribution of the backbone solvent accessible surface area is narrow, with small changes in $\Delta\alpha_k^B$ as [C] decreases.

Determination of the molecular origin of denatured state collapse, often associated with a concentration-dependent *m*-value, requires characterization of the DSE of protein L at low [C] (<3 M urea) where the NSE is thermodynamically favored. Under these conditions, we find that the radius of gyration (R_g) of the DSE undergoes significant reduction as [C] decreases. Urea-induced collapse transition of protein L is continuous as a function of [C] and results in nativelike secondary structural elements. We decompose the nonbonded energy into residue−solvent and intrapeptide interactions and show that (1) these two opposing energies govern the behavior of the R_g of the DSE and (2) the strength of these interactions is nonuniformly distributed in the DSE and correlates with regions of residual structure. Thus, different regions of the DSE can collapse to varying degrees as [C] changes.

METHODS

C_α Side Chain Model for Protein L. To ascertain the conditions under which eq 1 is a good approximation to eq 2, we use the coarse-grained C_α side chain model (C_α-SCM) (*37*) to represent the 64-residue protein L. In the C_α-SCM, each residue in the polypeptide chain is represented using two interaction sites, one that is centered on the α-carbon atom and another at the center of mass of the side chain (*37*). The potential energy (E_P) of a given conformation of the C_α-SCM is a sum of bond angle (E_A), backbone dihedral (E_D), improper dihedral (E_I), backbone hydrogen bonding (E_{HB}), and nonbonded Lennard-Jones (E_{LJ}) terms ($E_P = E_A + E_D + E_C + E_{HB} + E_{LJ}$). The functional form of these terms and derivation of the parameters used are explained in the Supporting Information of ref *36*.

Sequence information is included in the C_α-SCM by using nonbonded parameters that are residue-dependent. We take into account the size of a side chain by varying the collision diameter used in the E_{LJ} term. The interaction strength between side chains *i* and *j*, which are in contact in the native structure, depends on the amino acid pair and is modeled by varying the well depth (ε_{ij}) in E_{LJ} (*36*). Thus, the C_α-SCM incorporates both sequence variation and packing effects. Numerous studies have shown that considerable insights into protein folding can be obtained using coarse-grained models (*38*−*40*), thus rationalizing the choice of the C_α-SCM in this study.

Simulation Details. Equilibrium simulations of the folding and unfolding reaction using the C_α-SCM are performed using multiplexed replica exchange (MREX) (*41, 42*) in conjunction with low-friction Langevin dynamics (*43*) at [C]

Table 1: van der Waals Radii of the Side Chain Beads for Various Amino Acids Based in Part on Measured Partial Molar Volumes (*62*)

residue	radius (Å)	residue	radius (Å)
Ala	2.14	Met	2.63
Cys	2.33	Asn	2.33
Asp	2.37	Pro	2.36
Glu	2.52	Gln	2.56
Phe	2.70	Arg	2.79
Gly (backbone)	2.70	Ser	2.20
Hsd[a]	2.63	Thr	2.39
Ile	2.63	Val	2.49
Lys	2.70	Trp	2.88
Leu	2.63	Tyr	2.75

[a] The same value of the radius was used regardless of the protonation state.

Table 2: Solvent Accessibility of the Backbone and Side Chain Groups of Residue *k* in the Tripeptide Gly−*k*−Gly ($\alpha_{Gly-k-Gly}$)

	$\alpha_{Gly-k-Gly}$ (Å^2)	
k	backbone	side chain
Ala	62.5	108.3
Met	50.3	164.7
Arg	46.2	186.0
Gln	52.1	155.4
Asn	55.6	138.7
Gly	85.0	0.0
Tyr	47.3	179.9
Asp	56.7	133.7
Trp	43.8	198.7
Phe	48.3	174.6
Cys	57.7	128.6
Pro	56.9	132.7
Lys	48.3	174.6
Hsd[a]	51.4	159.2
Hse[b]	51.6	159.2
Hsp[c]	51.4	159.2
Ser	60.9	114.9
Thr	56.2	135.7
Val	53.8	147.1
Ile	50.3	164.7
Glu	53.0	150.8
Leu	50.3	164.6

[a] Hsd, neutral histidine, proton on the ND1 atom. [b] Hse, neutral histidine, proton on the NE2 atom. [c] HSP, protonated histidine.

= 0. We used CHARMM to carry out the Langevin dynamics (*44*), while an in-house script handles the replica exchange calculation. In the MREX simulations, multiple independent trajectories are generated at several temperatures. In addition to the conventional replica exchange acceptance/rejection criteria for swapping conformations between different temperatures (*41*), MREX also allows exchange between replicas at the same temperature (*42*). Replicas were run at eight temperatures: 315, 335, 350, 355, 360, 365, 380, and 400 K. At each temperature, four independent trajectories were simultaneously simulated. Every 5000 integration time steps, the system configurations were saved for analysis. Random shuffling occurred between replicas at the same temperature with 50% probability. Exchanges between neighboring temperatures were attempted using the standard replica exchange acceptance criteria (*41*). A Langevin damping coefficient of 1.0 ps^{-1} was used, with a 5 fs integration time step. In all, 90000 exchanges were attempted, of which the first 10000 were discarded to allow for equilibration. All trajectories were simulated in the canonical (NVT) ensemble.

Analysis with the Molecular Transfer Model. We model the denaturation of protein L by urea using the molecular transfer model (*36*). Previous work (*36*) has already shown that the MTM quantitatively reproduces experimentally measured single-molecule FRET efficiencies (*27*−*29*) as a function of [C] [guanidinium chloride (GdmCl)] for protein L and the cold shock protein, thus validating the methodology. The MTM combines simulations at [C] = 0 with the TM (*13, 14*), experimentally measured transfer free energies (*15, 16*), and a reweighting method for predicting protein properties at any urea concentration of interest (*36, 45*−*47*). The MTM equation, which has the form of the weighted histogram analysis method (*46*), is

$$\langle A([\mathrm{C}],T)\rangle = Z([\mathrm{C}],T)^{-1}\sum_{l=1}^{R}\sum_{t=1}^{n_l}\frac{A_{l,t}e^{-\beta[E_P(l,t,[0])+\Delta G_{tr}(l,t,[\mathrm{C}])]}}{\sum_{n=1}^{R} n_n e^{f_n-\beta_n E_P(l,t,[0])}} \quad (3)$$

where $\langle A([\mathrm{C}],T)\rangle$ is the average of a protein property A at urea concentration [C] and temperature T and $Z([\mathrm{C}],T)$ is the partition function. The sums in eq 3 are over the R different replicas from the MREX simulations, which vary in terms of temperature, and n_l protein conformations from the l^{th} replica. The value of A from replica l at time t is $A_{l,t}$, and $E_P(l,t,[0])$ is the potential energy of that conformation at [C] = 0 and $\beta = 1/(k_B T)$, where k_B is Boltzmann's constant. In eq 3, $\Delta G_{tr}(l,t,[\mathrm{C}])$, the reversible work of transferring the l,t protein conformation from 0 to [C] M urea solution, is estimated using a form of the TM and is given by

$$\Delta G_{tr}(l,t,[\mathrm{C}]) = \sum_{k=1}^{N_S}\frac{n_k\delta g_k^S([\mathrm{C}])}{\alpha_{G-k-G}^S}\langle\alpha_k^S(l,t,[\mathrm{C}])\rangle + \sum_{k=1}^{N_B}\frac{n_k\delta g_k^B([\mathrm{C}])}{\alpha_{G-k-G}^B}\langle\alpha_k^B(l,t,[\mathrm{C}])\rangle \quad (4)$$

All terms in eq 4 are the same as in eq 2 except instead of computing a difference in surface areas, only the surface areas from conformation l,t [$\langle\alpha_k^P(l,t,[\mathrm{C}])\rangle$] are included. In the denominator of eq 3, the sum is over the different replicas and n_n, β_n, and f_n are the number of conformations from replica n, $\beta_n = 1/(k_B T_n)$, where T_n is the temperature of the mth replica, and the free energy f_n of replica m is obtained by solving a self-consistent equation, respectively (see ref *45*).

In computing $\langle\alpha_k^P(l,t,[\mathrm{C}])\rangle$ for use in eq 4, we use the radii listed in Table 1 where the backbone group corresponds to the glycine. These parameters are different from the ones reported in ref *36*. They result in better agreement between predicted m-values using the MTM and predicted m-values from Auton and Bolen's implementation of the TM (*15, 18*). The values for α_{G-k-G}^S, used in eq 4, are reported in Table 2.

We calculate the average of a number of properties of protein L using eq 3. The end-to-end distance (R_{ee}) of a given conformation is the distance between the C_α sites at residues 1 and 64. The radius of gyration, R_g, is computed using $\sqrt{R_g^2} = 1/(2N - N_G)\langle\sum_{i=1}^{2N-N_G}(r_i - r_{CM})^2\rangle$, where N is the number of residues, N_G is the number of glycines in the sequence, r_i is the position of interaction site i, and $r_{CM} = 1/(2N - N_G)\sum_{i=1}^{2N-N_G} r_i$ is the mean position of the $2N - N_G$ interaction sites of the protein. The solvent accessible surface area of a backbone or a side chain (α_k^P) in residue k in a given conformation was computed using the CHARMM program (*44*), which computes the analytic solution for the surface area. A probe radius of 1.4 Å, equivalent to the size of a water molecule, was used.

The extent to which a structural element is formed (denoted f_S) in a conformation of protein L is defined by Q_p, the fraction of native backbone contacts formed by structural element p, where p is β-hairpin S12 or S34 or a β-strand pairing between S1 and S4 (Figure 2b). We define Q_p as

$$Q_p = \sum_{j}^{N-4}\sum_{k=j+4}^{N}\frac{\Theta(R_C - d_{jk})}{C_p} \quad (5)$$

where the sum is over the $N = 64$ C_α sites, R_C (=8 Å) is a cutoff distance, d_{jk} is the distance between interaction sites j and k, and $\Theta(R_C - d_{jk})$ is the Heaviside step function. Strand 1 (S1) corresponds to residues 4−11. S2 is between residues 17 and 24. S3 corresponds to residues 47−52. S4 is between residues 57 and 62 (Figure 2b). In eq 5, C_p is the maximum number of native contacts for structural element p. The extent of helix formation in a conformation r of protein L is computed as the ratio $N_\phi(r)/N_\phi(N)$, where $N_\phi(r)$ is the number of neighboring dihedral pairs, between residues 26 and 44, that have backbone dihedral angles within ±20° of the dihedral's value in the native state, and $N_\phi(N) = 15$.

The nonbonded interaction energy, E_I, in the C_α-SCM is equal to $E_{LJ} + E_{HB}$. We include only the Lennard-Jones (LJ) and hydrogen bond (HB) energies in E_I (*36*). The urea solvation energy, E_S, of a given conformation is set equal to eq 4; E_M is a simple sum of E_I and E_S. The values of E_I and E_S for the various structural elements of protein L were computed by neglecting nonbonded and solvation energies of residues that were not part of the structural element of interest.

The time series of the various properties were inserted into eq 3 to compute their averages as a function of [C]. To compute averages $\langle A_D\rangle$ and $\langle A_N\rangle$ of the DSE and NSE

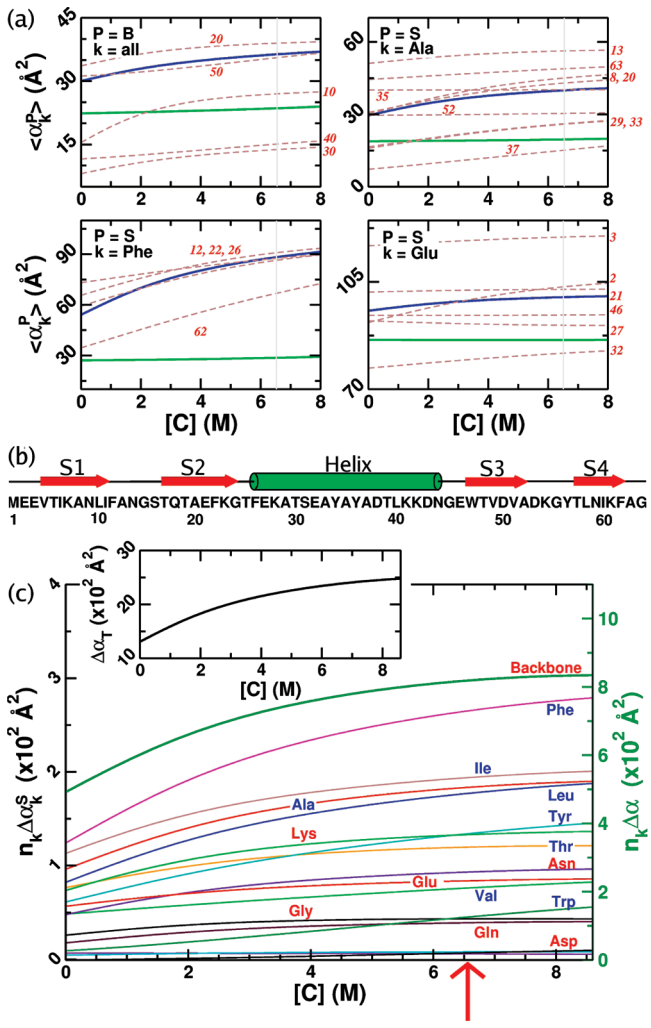

FIGURE 2: (a) $\langle\alpha^P_{k,j}\rangle$ vs urea concentration for the backbone and the side chains alanine, phenylalanine, and glutamate, computed using the relationship $\langle\alpha^P_{k,j}([C])\rangle = \int_0^\infty \alpha^S_{k,j} P(\alpha^P_{k,j};[C])\, d\alpha_{k,S}$ (j = D or N, and P = S or B). For the backbone, $\langle\alpha^B_{k=\text{all},j}([C])\rangle = N^{-1}\sum_{k=1}^{N_B}\langle\alpha^B_{k,j}([C])\rangle$, where $N = 64$, the number of residues in the protein. $\langle\alpha^P_{k,N}\rangle$ and $\langle\alpha^P_{k,D}\rangle$ are displayed as green and blue lines, respectively. Brown dashed lines show $\langle\alpha^P_{k,D}\rangle$ for individual residues of type k; the residue indices are indicated by the numbers in red. For the backbone, only six groups (from residues 1, 10, 20, 30, 40, and 50) out of 64 backbone groups are shown. (b) Linear secondary structure representation of protein L. β-Strands are shown as red arrows; the α-helix is shown as a green cylinder, and unstructured regions are shown as a solid black line. Secondary structure assignments were made using STRIDE (*64*). The residues corresponding to each secondary structure element are listed below the representation. (c) $n_k\Delta\alpha^P_k$ (eq 2) as a function of urea concentration for the backbone (green line, with the corresponding ordinate on the right), and all the other 16 unique amino acid types in protein L (with the corresponding ordinate on the left). For the sake of clarity, labels for Met and Ser residues are not shown. Met and Ser have $n_k\Delta\alpha^P_k$ values close to zero in this graph. $\Delta\alpha^P_k = \langle\alpha^P_{k,D}\rangle - \langle\alpha^P_{k,N}\rangle$ (P = S or B). For the backbone, we plot $\sum_{k=1}^{N_B} n_k\Delta\alpha^B_k$. The inset shows $\Delta\alpha_T$ as a function of urea concentration. The red arrow indicates C_m.

respectively, a modification was made to eq 3. The numerator was multiplied by $\Theta_n(l,t)$, where $\Theta_n(l,t)$ is the Heaviside step function that is equal to $\Theta[5 - \Delta(l,t)]$ when the average of the NSE is computed (i.e., n = NSE) and is equal to $\Theta[5 + \Delta(l,t)]$ when the average of the DSE is computed (i.e., n = DSE). Here, $\Delta(l,t)$ is the root-mean-square deviation between the C_α carbon sites in the C_α-SCM of conformation l,t and the C_α carbon atoms in the crystal structure [Protein Data Bank entry 1HZ6 (*48*)]. When $\Delta(l,t)$ is greater than 5 Å, $\Theta[5 + \Delta(l,t)] = 0$ and $\Theta[5 - \Delta(l,t)] = 1$, and when $\Delta(l,t)$ is less than 5 Å, $\Theta[5 + \Delta(l,t)] = 1$ and $\Theta[5 - \Delta(l,t)] = 0$.

Probability distributions were computed using $P(A\pm\delta_A;[C]) = Z(A\pm\delta_A,[C],T)/Z([C],T)$, where $Z(A\pm\delta_A,[C],T)$ is the restricted partition function as a function of A. Due to the discrete nature of the simulation data, a bin with finite width $\pm\delta_A$, whose value depends on A, is used. $Z(A\pm\delta_A,[C],T) = \sum_{l=1}^{R}\sum_{t=1}^{n_L}\{f_A(l,t)e^{-\beta[E_P(l,t,[0])+\Delta G_{tr}(l,t,[C])]}\}/\{\sum_{n=1}^{R} n_n e^{f_n-\beta_n E_P(l,t,[0])}\}$, where all terms are the same as in eq 3 except for $f_A(l,t)$, which is a function that we define to be equal to 1 when protein conformation l,t has a value of A in the range of $A \pm \delta_A$, and zero otherwise.

RESULTS AND DISCUSSION

$\Delta G_{ND}([C])$ Changes Linearly as the Urea Concentration Increases. We chose the experimentally well characterized B1 IgG binding domain of protein L (*27, 28, 49*) to illustrate the general principles that explain the linear dependence of $\Delta G_{ND}([C])$ on [C] for proteins that fold in an apparent two-state manner. In our earlier study (*36*), we showed that the MTM accurately reproduces several experimental measurements including [C]-dependent Forster resonance energy transfer as a function of guanidinium chloride (GdmCl) concentration. Prompted by the success of the MTM, we now explore urea-induced unfolding of protein L. The MTM predictions for urea effects are expected to be more accurate than for GdmCl, since the experimentally measured $\delta g^P_k([C])$ urea data, used in eq 1, include activity coefficient corrections while the GdmCl data do not (*13, 15*). The calculated $\Delta G_{ND}([C])$ as a function of urea concentration for protein L shows linear dependence above [C] > 4 M (Figure 1b) with an m of 0.80 kcal mol^{-1} M^{-1} and a C_m [obtained using $\Delta G_{ND}(C_m) = 0$)] of ≈6.6 M. The consequences of the deviation from linearity, which is observed for [C] < 3 M, are explored below. It should be stressed that the error in the estimated $\Delta G_{ND}([0])$ is relatively small (~0.8 kcal mol^{-1}) if measurements at [C] > 4 M are extrapolated to [C] = 0 (Figure 1b). Thus, from the perspective of free energy changes, the assumption that $\Delta G_{ND}([C]) = \Delta G_{ND}([0]) + m[C]$, with constant m, is justified for this protein.

Molecular Origin of Constant m-Values. Inspection of eq 2 suggests that there are three possibilities that can explain the constancy of m-values, thus making eq 1 a good approximation of eq 2. (1) Both $\langle\alpha^P_{k,D}([C])\rangle$ and $\langle\alpha^P_{k,N}([C])\rangle$ in eq 2 have the same dependence on [C], making $\Delta\alpha^P_k$ effectively independent of [C]. (2) The distributions $P(\alpha^P_{k,D};[C])$ in eq 2 are sharply peaked about their mean or most probable values of $\alpha^P_{k,D}([C])$ at all [C] values, thus making $\Delta\alpha^P_k$ independent of [C]. In particular, if the standard deviation in $\alpha^P_{k,D}$ (denoted σ_{α_k}) is much less than $\langle\alpha^P_{k,D}([C])\rangle$ for all [C] values, then the $\Delta\alpha^P_k$ values would be effectively independent of [C]. (3) One group in the protein, denoted l (backbone in proteins), makes the dominant contribution to the m-value. In this case, only the changes in $\Delta\alpha^P_l$ and $P(\alpha^P_{l,D};[C])$ matter, thereby making $\Delta\alpha^P_l$ insensitive to [C]. The MTM simulations of protein L allow us to test the validity of these plausible explanations for the constancy of m-values, especially when [C] > 3 M (Figure 1b). Only by examining these possibilities, which requires changes in the distribution of various properties as [C] changes, can the observed constancy of m be rationalized.

$\langle\alpha^P_{k,D}([C])\rangle$ and $\langle\alpha^P_{k,N}([C])\rangle$ Do Not Have the Same Dependence on [C]. The changes in $\langle\alpha^P_{k,D}\rangle$ and $\langle\alpha^P_{k,N}\rangle$ as a function of [C] show that as [C] increases, both $\langle\alpha^P_{k,D}\rangle$ and $\langle\alpha^P_{k,N}\rangle$ increase (blue and green lines in Figure 2a). However, $\langle\alpha^P_{k,D}([C])\rangle$ has a stronger dependence on [C] than $\langle\alpha^P_{k,N}([C])\rangle$ for both the backbone and side chains (Figure 2a). Thus, the observed linear dependence of $\Delta G_{ND}([C])$ on [C] cannot be rationalized in terms of similarity in the variation of $\langle\alpha^P_{k,D}([C])\rangle$ and $\langle\alpha^P_{k,N}([C])\rangle$ as [C] changes. The stronger dependence of $\langle\alpha^P_{k,D}([C])\rangle$ on [C] arises from the greater range and magnitude of the solvent accessible surface areas available to the DSE (see below). The greater range allows larger shifts in $\langle\alpha^P_{k,D}([C])\rangle$ than $\langle\alpha^P_{k,N}([C])\rangle$ with [C]. Equally important, the strength of the favorable protein−solvent interactions is positively correlated with the magnitude of the surface area and [C] (see eq 4 and Figure 1a). Thus, the DSE conformations with larger surface areas are stabilized to a greater extent than the NSE conformations with an increasing [C], and subsequently, $\langle\alpha^P_{k,D}([C])\rangle$ shows a stronger dependence on [C].

Surface Area Distributions Are Broad in the DSE. The variation of $\Delta\alpha^S_{k,D}$ and $\Delta\alpha^B_{k,D}$ with [C] suggests that the $P(\alpha^P_{k,D};[C])$ values are not likely to be narrowly peaked and must also depend on [C] (eq 2). As the urea concentration increases, the total backbone surface area distribution in the DSE, $P(\alpha^B_D;[C])$, shifts toward higher values of α^B_D and becomes narrower (Figure 3a). A similar behavior is observed in the distribution of the total surface area (Figure 3b) and for the side chain groups (data not shown). It should be noted that the change in α^B_D with [C] is ~5 times smaller than the corresponding change in α_T (compare panels a and b of Figure 3). Thus, the distribution of surface areas for the various protein components is moderately dependent on [C], and $\Delta\alpha_T$ is more strongly dependent on [C] (Figure 2c, inset). These findings would suggest that *m* should be a function of [C] above 4 M (eq 2), in contradiction to the finding in Figure 1b.

We characterize the width of the denatured state $P(\alpha^P_{k,D})$ distributions by computing the ratio $\rho_k = \sigma_{\alpha_{k,D}}/\langle\alpha^P_{k,D}\rangle$, where $\sigma_{\alpha_{k,D}} = [\langle\alpha^P_{k,D}\rangle^2 - \langle\alpha^P_{k,D}\rangle^2]^{1/2}$. Figure 4a shows ρ_k as a function of [C] for the various protein components (backbone, side chains, and the entire protein). As with the backbone $P(\alpha^B_D)$ distribution (Figure 3a), ρ_k indicates that $P(\alpha^P_{k,D})$ becomes narrower at higher urea concentrations for most values of *k* (Figure 4a). At 8 M urea, the width of $P(\alpha^P_{k,D})$ ranges from 5 to 25% of the average value of $\alpha^P_{k,D}$ for all groups, except when *k* = Trp which has an even larger width. Clearly, ρ_k is large at all [C], which accounts for the dependence of $\Delta\alpha^P_k$ on [C]. The results in Figure 4 show that there are discernible changes in ρ_k which reflects the variations in $P(\alpha^P_{k,D};[C])$ as [C] is changed. Consequently, the constancy of the *m*-value cannot be explained by narrow surface area distributions.

The Weak Dependence of Changes in Accessible Surface Area of the Protein Backbone on [C] Controls the Linear Behavior of $\Delta G_{ND}([C])$. Plots of *m*[C] at several urea concentrations for the entire protein, the backbone groups (second term in eqs 1 and 2), and the hydrophobic side chains Phe, Leu, Ile, and Ala are shown in Figure 4b. The slope of these plots is the *m*-value, which in the transition region (i.e., from 5.1 to 7.9 M urea) is 0.80 kcal mol^{-1} M^{-1} for the entire protein. The contribution from the backbone alone is 0.76 kcal mol^{-1} M^{-1} and from the most prominent hydrophobic side chains (Phe, Leu, Ile, and Ala) is a combined 0.04 kcal mol^{-1} M^{-1}. Thus, the largest contribution to the change in the native state stability, as [C] is varied, comes from the burial or exposure of the protein backbone (95%). The simulations directly support the previous finding that the protein backbone contributes the most to the stability changes with [C] (*16*). Thus, for [C] > 3 M, the magnitude of the *m*-value is largely determined by the backbone groups. However, only by evaluating the [C]-dependent changes in the distribution of surface areas can one assess the extent to which eq 2 be approximated by eq 1.

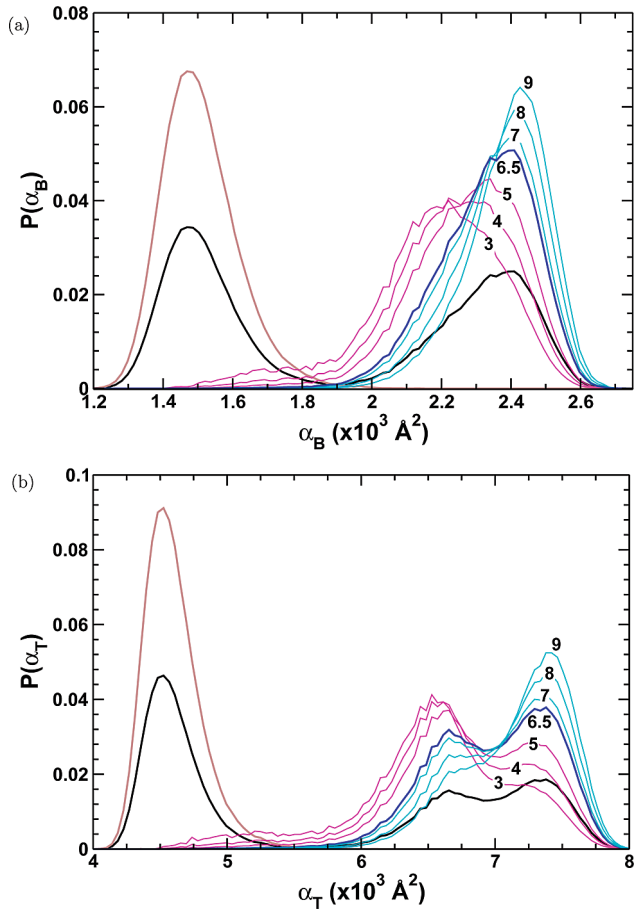

FIGURE 3: (a) Probability distribution of the total backbone surface area in the DSE [$P(\alpha^B_D)$] at various urea concentrations, indicated by the number above each trace. For the sake of comparison, $P(\alpha^B_N)$ for the native state ensemble at 6.5 M urea is shown (solid brown line) as well as the average distribution over both the NSE and DSE at 6.5 M urea (black line). (b) Same as panel a, except the distributions are of the accessible surface area of the entire protein.

The relative change in accessible surface area of the backbone $\Delta\alpha^B_D$ has a relatively weak urea dependence between 4 and 8 M urea, increasing by only 75 Å^2 (Figure 2c). Such a small change in $\Delta\alpha^B_D$ with [C] has a negligible effect on the *m*-value. These results show that *m* is effectively independent of [C] in the transition region because $\Delta\alpha^B_k([C])$ values associated with the backbone groups change by only a small amount as [C] changes, despite the fact that $\Delta\alpha_T$ can change appreciably [$\Delta\alpha_T$(4 M→8 M) ≈ 300 Å^2 (Figure 2c, inset)]. Thus, the third possibility is correct, namely, that the weak dependence of $\Delta\alpha^B_k([C])$ on [C] results in *m* being constant.

Residual Denatured State Structure Leads to the Inequivalence of Amino Acids. In applying eq 1 to predict *m*-values, it is generally assumed that all residues of type *k*, regardless

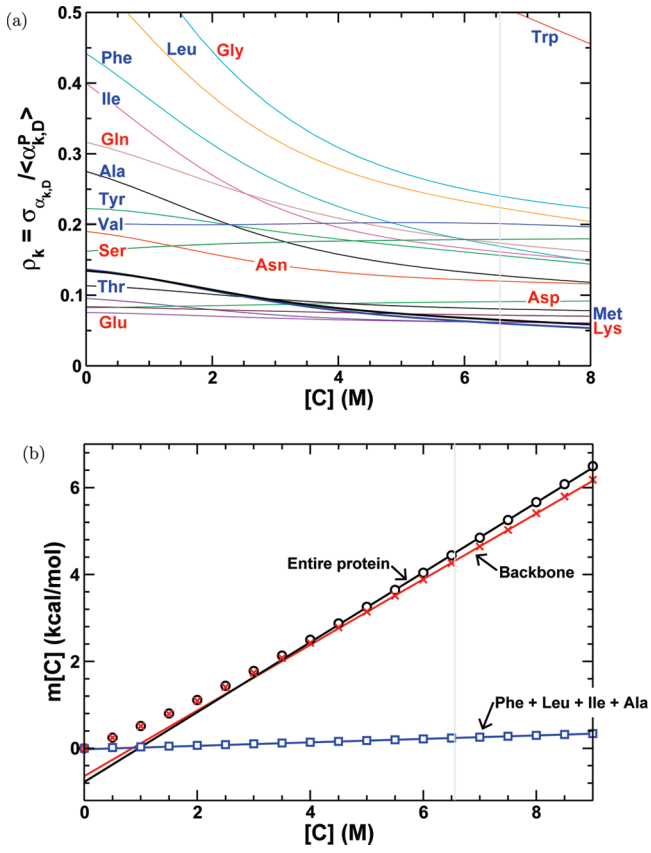

FIGURE 4: (a) Ratio $\rho_k = \sigma_{\alpha_{k,D}}/\langle\alpha^P_{k,D}\rangle$ (see the text for an explanation) as a function of urea concentration for the entire protein (black line), the backbone (blue line), and all other amino acid types found in protein L. (b) Quantity *m*[C] vs urea concentration for the full protein (black circles), the backbone groups (red squares), and the Phe, Leu, Ile, and Ala side chains. Solid lines correspond to linear fits to the data in the range of 5.1−7.9 M urea.

of their sequence context, have the same solvent accessible surface area in the DSE (*15, 16*). Our simulations show that this assumption is incorrect. Comparison of $\alpha^P_{k,D}$ for individual residues of type *k* and the average $\langle\alpha^P_{k,D}\rangle$ as a function of urea concentration (Figure 2a) show that both sequence context and the distribution of conformations in the DSE determine the behavior of a specific residue. Large differences between $\alpha^P_{k,D}$ values are observed between residues of the same type, including alanine, phenylalanine, and glutamate groups, even at high urea concentrations (Figure 2a). The inequivalence of a specific residue in the DSE is similar to NMR chemical shifts that are determined by the local environment. As a result of variations in the local environment, not all alanines in a protein are equivalent. Thus, ignoring the unique surface area behavior of individual residues in the DSE could lead to errors in the predicted *m*-value. Because the backbone dominates the transfer free energy of the protein (Figure 4b), errors arising from this assumption may be small. However, the dispersions in the backbone $\alpha^B_{k,D}$ suggest that different regions of the protein may collapse in the DSE at different urea concentrations, driven by differences in $\Delta\alpha^B_{k,D}$ from residue to residue (see below).

The simulations can be used to calculate [C]-dependent changes in surface areas of the individual backbone groups as well as side chains. Interestingly, even for the chemically homogeneous backbone group, significant dispersion about $\langle\alpha^B_{k,D}\rangle$ is observed when individual residues are considered (Figure 2a). For example, $\alpha^B_{k,D}$ for residue 10 changes more drastically as [C] decreases than it does for residue 20 or 50. Thus, the connectivity of the backbone group can alter not only the conformations as [C] is varied but also the contribution to the free energy.

Even more surprisingly, the changes in $\alpha^S_{k=Ala,D}$ depend on the sequence location of a given alanine residue and the associated secondary structure adopted in the native conformation. The changes in $\alpha^S_{k=Ala,D}$ for residues 8 and 20, both of which adopt a β-strand conformation in the native structure (Figure 2b), are similar with a decrease in [C] (Figure 2a). By comparison, surface area changes in alanine residues 29 and 33, which are helical in the native state (Figure 2b), are similar as [C] varies, while the changes in $\alpha^S_{k=Ala,D}$ for alanines that are in the loops (residues 13 and 63) are relatively small. Examining the probability distribution of surface areas for the individual alanines [$P(\alpha^S_{Ala,D})$ in Figure 5], which is related to the average surface area and higher-order moments, we observe a wide variability between different residues. Similar conclusions can be drawn by analyzing the results for the larger hydrophobic residue Phe and the charged Glu (Figure 2a). Thus, for a given amino acid type, both sequence context and the heterogeneous nature of structures in the DSE lead to a dispersion about the average $\langle\alpha^S_{k,D}\rangle$ and higher-order moments of $P(\alpha^S_{k,D})$ as the urea concentration changes. Much like the chemical shifts in NMR, the distribution functions of chemically identical individual residues bear signatures of their environment and the local structures they adopt as [C] is varied!

The total surface area difference between N and D ($\Delta\alpha_T$) changes by ~1200 Å^2 as [C] decreases from 8 to 0 M (see the inset of Figure 2c). Decomposition of $\Delta\alpha_T$ into contributions from backbone and side chains (eqs 1 and 2) shows that the burial of the backbone groups contributes the most (up to 38%) to $\Delta\alpha_T$ (Figure 2c). Not unexpectedly, hydrophobic residues (Phe, Ile, Ala, and Leu), which are buried in the native structure, also contribute significantly to $\Delta\alpha_T$, which agrees with the recent all-atom molecular dynamics simulations (*50*). Among them, Phe, a bulky hydrophobic residue, makes the largest side chain contribution to $\Delta\alpha_T$ (Figure 2c). For example, as the urea concentration increases from 4 to 8 M, the total backbone $\Delta\alpha^B_D$ increases by 75 Å^2, and $n_k\Delta\alpha^S_{k,D}$ for k = Phe, Leu, Ala, or Ile increases by 21−42 Å^2.

The dispersion in $\alpha^P_{k,D}$ could be caused by residual structure in the DSE (*51, 52*). We test this proposal quantitatively by plotting $\alpha^S_{k,D}/\alpha^{SM}_{k,D}$ for each residue, where $\alpha^{SM}_{k,D}$ is the maximum $\alpha^S_{k,D}$ value for residue type *k* in 8 M urea. If residual structure causes the dispersion in $\alpha^P_{k,D}$ then we expect that $\alpha^S_{k,D}/\alpha^{SM}_{k,D}$ should depend on the secondary structure element that residue *k* adopts in the native state. We find that there is a correlation between $\alpha^S_{k,D}$ and the helical secondary structure element [residues 26−44 (Figure 6)]. The helical region tends to have smaller $\alpha^S_{k,D}/\alpha^{SM}_{k,D}$ values compared to other regions of the protein. Of the nine alanines in protein L, four are found in the helical region of the protein. These four residues have some of the smallest $\alpha^S_{k,D}/\alpha^{SM}_{k,D}$ values of the nine alanines. The [C]-dependent fraction of residual secondary structure in the DSE shows that at 8 M urea the helical content is 32% of its value in the native state (Figure 7a). Taken together, these data show that $\alpha^P_{k,D}$ depends not only on the residue type but also on the residual structure present

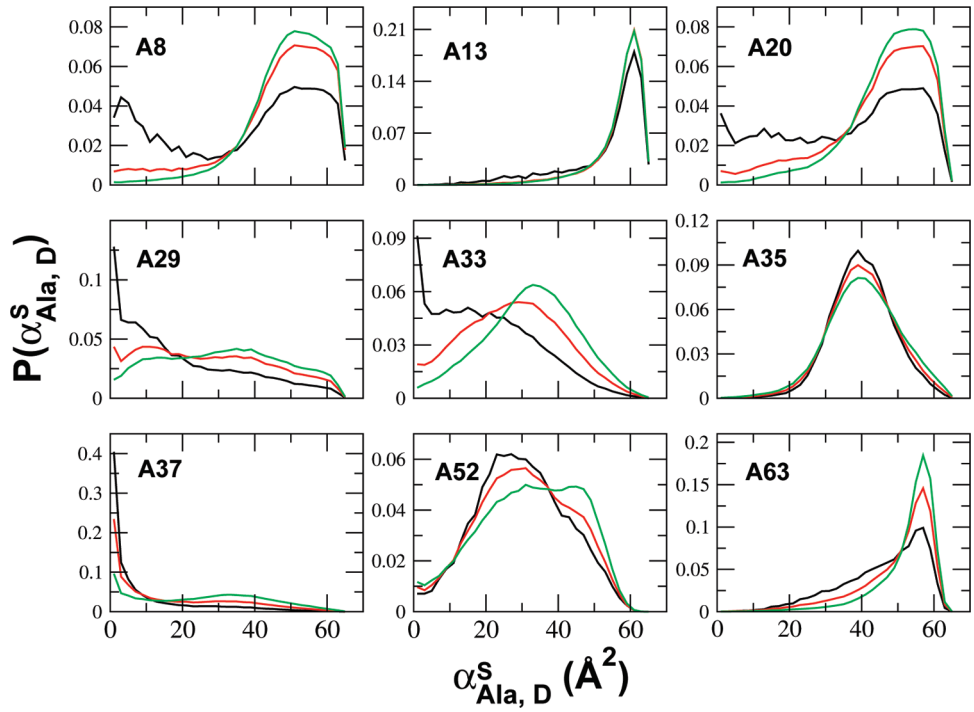

FIGURE 5: Distribution $[P(\alpha^S_{Ala,D})]$ of the solvent accessible surface area of side chains from the nine individual alanine residues in the denatured state ensemble of protein L at various urea concentrations. Black, red, and green lines correspond to 1, 4, and 8 M urea, respectively. The corresponding alanine for each graph is given by its residue number. The large changes in $P(\alpha^S_{Ala,D})$ for the chemically identical residue show that the environment and local structures affect the structures and energetics of the side chains.

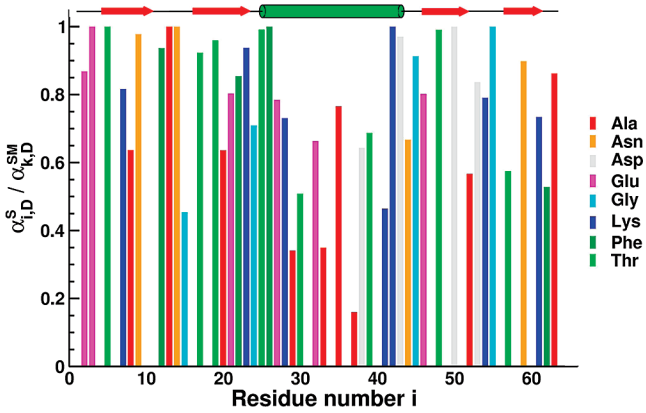

FIGURE 6: $\alpha^S_{i,D}/\alpha^{SM}_{k,D}$ ratio (see the text for an explanation) as a function of residue number *i* at 8 M urea. The legend indicates the amino acid type for each residue. Only amino acid types that occur at least four times in protein L and have at least two of those residues separated by more than 25 residues along sequence space are plotted. For reference, the linear secondary structure representation of protein L is shown above the graph.

in the DSE, which at all values of [C], is determined by the polymeric nature of proteins.

Residue-Dependent Variations in the Transition Midpoint: The Holtzer Effect. Globally, the denaturant-induced unfolding of protein L may be described using the two-state model (Figure 1b). However, deviations from an all-or-none transition can be discerned if the residue-dependent transitions $C_{m,i}$ can be measured. For strict two-state behavior, $C_{m,i} = C_m$ for all *i*, where $C_{m,i}$ is the urea concentration below which the *i*th residue adopts its native conformation. The inequivalence of the amino acids, described above (Figure 2a), should lead to a dispersion in $C_{m,i}$. The values of $C_{m,i}$ are determined by specific interactions, while the dispersion in $C_{m,i}$ is a finite-size effect (*53, 54*). In other words, because the number of amino acids (*N*) in a protein is finite, all thermodynamic transitions are rounded instead of being infinitely sharp. Finite-size effects on phase transitions have been systematically studied in spin systems (*55*) but have received much less attention in biopolymer folding (*54*). Klimov and Thirumalai (*53*) showed that the dispersion in the residue-dependent melting temperatures $T_{m,i}$, denoted $\Delta T(\Delta C)$, for temperature (denaturant)-induced unfolding scales as $\Delta T/T_m \sim 1/N$ ($\Delta C/C_m \sim 1/N$). The expected dispersion in $C_{m,i}$ or $T_{m,i}$ is the Holtzer effect.

In the context of proteins, Holtzer and co-workers (*56*) were the first to observe that although globally thermal folding of the 33-residue GCN4-lzK peptides can be described using the two-state model, there is dispersion in the melting temperature throughout the protein's structure. In accord with expectations based on the finite size of GCN4-lzK, it was found, using one-dimensional NMR experiments, that $T_{m,i}$ depends on the sequence position. The deviation of $T_{m,i}$ from the global melting temperature is as large as 20% (*56*). More recently, large deviations in $T_{m,i}$ from T_m have been observed for other proteins (*57*).

We have determined, for protein L, the values of $C_{m,i}$ using $Q_i(C_{m,i}) = 0.5$, where Q_i is the fraction of native contacts for the *i*th residue. The distribution of $C_{m,i}$ shows the expected dispersion (Figure 8a), which implies different residues can order at different values of [C]. The precise $C_{m,i}$ values are dependent on the extent of residual structure adopted by the *i*th residue, which will clearly depend on the protein. Similarly, the distribution of the melting temperature of individual residues $T_{m,i}$, calculated using $Q_i(T_{m,i}) = 0.5$, also shows variations from T_m. However, the width of the thermal dispersion is narrower than that obtained from denaturant-induced unfolding (Figure 8b). This result is in accord with the general observation that thermal melting is more coop-

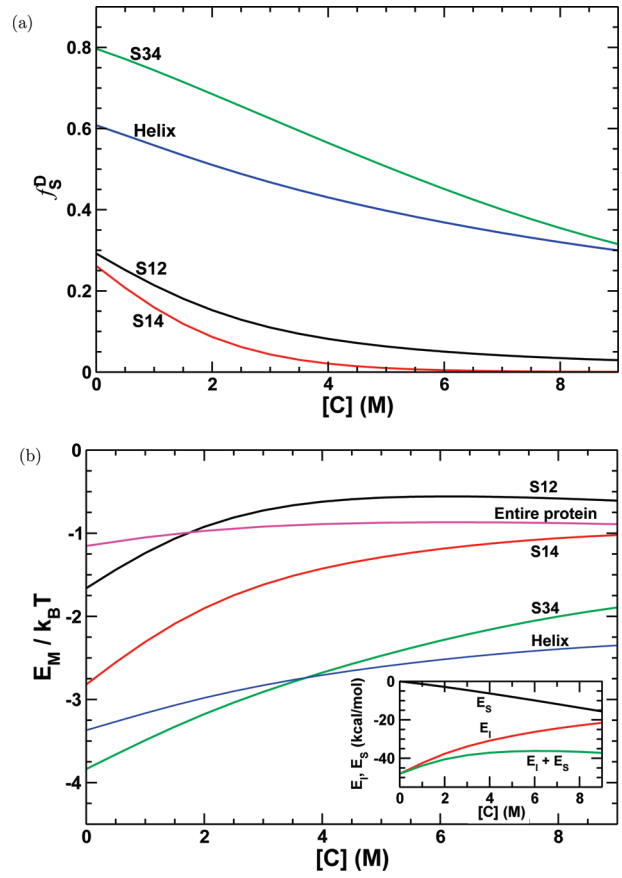

FIGURE 7: (a) Residual secondary structure content in the DSE vs urea concentration. (b) Interaction energy (E_M) in the DSE divided by the number of residues in the secondary structural element, in units of k_BT, vs urea concentration for the entire protein and various secondary structural elements. The inset shows E_I, E_S, and E_M for the entire protein vs urea concentration in units of kilocalories per mole.

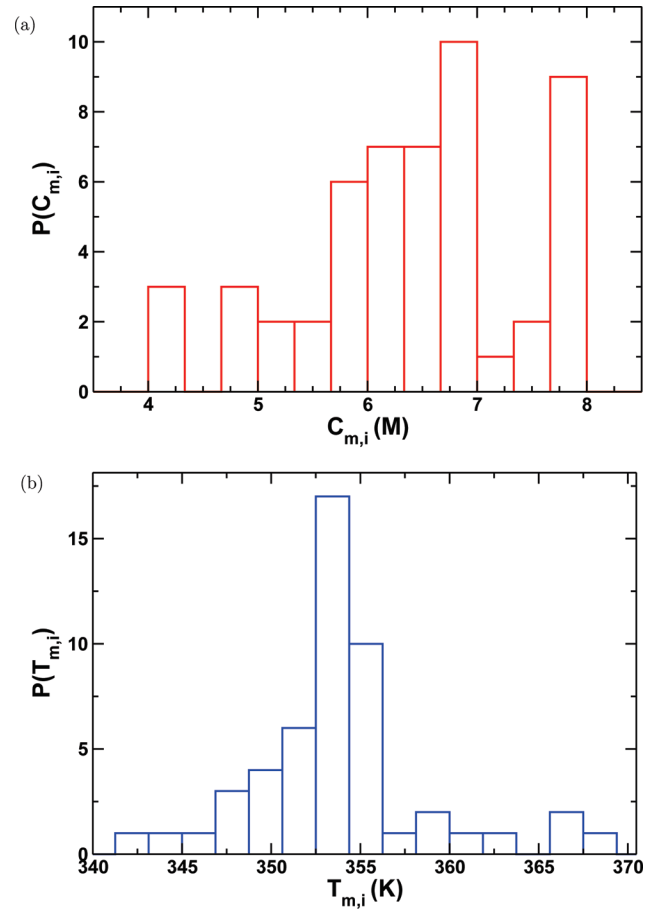

FIGURE 8: Histogram of residue-dependent midpoints of unfolding as a function of (a) urea concentration at 328 K and (b) temperature at 0 M urea. The C_m for the entire protein is ~6.6 M, while the melting temperature is 356 K at 0 M urea.

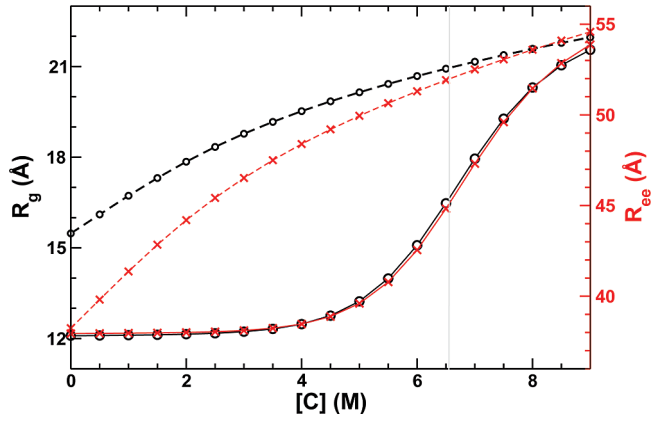

FIGURE 9: Average R_g (○) and R_{ee} (×) as a function of [C] for protein L at 328 K. The values of R_g^{DSE} (○, dashed line, left axis) and R_{ee}^{DSE} (×, dashed line, right axis) as a function of urea concentration are also shown. Lines are a guide to the eye. The gray vertical line at 6.56 M urea denotes the C_m.

erative than denaturant-induced unfolding (*58*). It should be emphasized that the Holtzer effect is fairly general, and only as N increases will ΔC and ΔT decrease.

Specific Protein Collapse at Low [C] and the Balance between Solvation and Intraprotein Interaction Energies. As [C] is decreased below 3 M, there is a deviation in linearity of ΔG_{ND}([C]) (Figure 1b) and the *m*-value depends on [C]. At low [C] values, the characteristics of the denatured state change significantly relative to those of the denatured state at 8 M. The radius of gyration, R_g^D, and $\Delta\alpha_T$ change by up to 6 Å (Figure 9) and 1150 Å^2 (Figure 2c) respectively, indicating that the denatured state undergoes a collapse transition. We detail the consequences of the [C]-dependent changes and examine the nature and origin of the collapse transition.

(i) Surface Area Changes. Above 4 M urea, the $\alpha_{k,D}$ values change only modestly (Figure 2a). However, below 4 M, much larger changes in $\alpha_{k,D}$ occur (Figure 2a). In particular, $\Delta\alpha_T$ decreases by 850 Å^2 with a decrease in [C] from 4 to 0 M urea, compared to ≈300 Å^2 upon [C] decreasing from 8 to 4 M urea (Figure 2c inset). The backbone is the single greatest contributor to $\Delta\alpha_T$, accounting for 24−38% of $\Delta\alpha_T$ at various [C] values. Thus, a significant amount of backbone surface area in the DSE is buried from solvent as [C] is decreased, and the protein becomes compact (Figure 2c). The next largest contribution to $\Delta\alpha_T$, as measured by $n_k\Delta\alpha_k$ $\{=n_k[\langle\alpha_{k,D}([C])\rangle - \langle\alpha_{k,N}([C])\rangle]\}$, arises from the hydrophobic residues Phe, Ile, and Ala (Figure 2c). These residues also exhibit relatively large changes in the DSE surface area as [C] is decreased. The large change in surface area of Phe as [C] decreases shows that dispersion interactions also contribute to the energetics of folding (*50*). On the other hand, for side chains that are solvent-exposed in the native state, such as the charged residue Asp, $n_k\Delta\alpha_k$ is small and does not change significantly with [C] (Figure 2c). The results in Figure 2, and the surface area dependence of the TM, suggest

that the changes in surface area at low [C] values are related to changes in the solvation energy of the backbone (see below).

(ii) R_g and R_{ee} Changes. Decreasing [C] below 4 M leads to a R_g^D change of up to 4 Å and an end-to-end distance (R_{ee}) change of up to 10 Å (Figure 9). Such a large change in R_g^D shows that a collapse transition occurs in the DSE. We find no evidence (e.g., a sigmoidal transition in R_g^D vs [C]) that the DSE at 0 M ($\langle R_g^D \rangle = 15.5$ Å) and the DSE at 8 M urea ($\langle R_g^D \rangle = 21.5$ Å) are distinct thermodynamic states. This suggests that the urea-induced DSE undergoes a continuous second-order collapse transition as the urea concentration decreases.

(iii) Residual Structure Changes. To gain insight into secondary structure changes that occur during the collapse transition, we plot the residual secondary structure (f_S^D) in the DSE versus [C] (Figure 7a). Above 4 M urea, only β-hairpin 3−4 and the helix are formed to any appreciable extent. However, below 4 M urea, β-hairpin 1−2 and β-sheet interactions between strands 1 and 4 can be found in the DSE. For example, at 1 M urea, β-hairpin 1−2 and strands 1 and 4 are formed 21 and 16% of the time while there is 56% helical and 74% β-hairpin 3−4 content in the DSE (Figure 7a). Thus, as [C] is decreased, the level of residual structure in the DSE increases, contributing to changes in R_g, R_{ee}, and the surface areas. This finding suggests that the collapse transition is specific in nature, leading to compact structures with nativelike secondary structure elements.

(iv) Solvation vs Intraprotein Interactions. If we neglect changes in protein conformational entropy, we find two opposing energies control the [C]-dependent behavior of R_g^D: the interaction of the peptide residues with solvent (the solvation energy, denoted E_S) and the intraprotein nonbonded interactions between the residues (denoted E_I). For denaturants, such as urea, E_S favors an increase in R_g^D and a concomitant increase in solvent accessible surface area, while E_I typically is attractive and hence favors a decrease in R_g^D. Because E_S in the TM model is proportional to a surface area term and E_I is likely to be approximately proportional to the number of residues in contact (which increases as the residue density increases upon collapse), we expect $E_S([C]) \propto -[C]\langle R_g^D([C])\rangle^2$ and $E_I([C]) \propto -1/\langle R_g^D([C])\rangle^3$. The behavior of these two functions [increasing $\langle R_g^D([C])\rangle$ leads to a more favorable $E_S([C])$ and unfavorable $E_I([C])$] suggests that there should always be some contraction (expansion) of the DSE with decreasing (increasing) [C]. The molecular details in the C_α-SCM allow us to exactly determine $E_S([C])$ and $E_I([C])$ as a function of [C] and thereby gain an understanding of the energy scales involved in the specific collapse of the DSE.

In the inset of Figure 7b, we plot $E_S([C])$, $E_I([C])$, and $E_M([C])$ [$\equiv E_S([C]) + E_I([C])$] in the DSE. As indicated by the Flory-like argument given above, $E_S([C])$ becomes more favorable with increasing [C], and $E_I([C])$ becomes more unfavorable with increasing [C] (Figure 7b, inset). It is important to examine the behavior of $E_M([C])$, as this quantity governs the behavior of $R_g^D([C])$. Above 4 M, $E_M([C])$ is relatively constant, varying by no more than 1 kcal/mol. This finding is consistent with the small changes in R_g^D, R_{ee}, and $\Delta\alpha_k^P$ above 4 M urea (Figures 2c and 9). Below 4 M, the $E_M([C])$ strength increases and is dominated by the attractive intrapeptide interactions [$E_I([C])$] at low [C] (Figure 7b, inset), driving the collapse of the protein as measured by R_g^D.

We dissect the monomer interaction energies further by computing the average monomer interaction energy per secondary structural element (Figure 7b). Above 4 M urea, the monomer interaction energies change by less than 0.4 k_BT, except for that of β-hairpin 3−4 which changes by as much as ~0.9 k_BT. Below 4 M, the monomer interaction energies change by as much as 1.5 k_BT, with the helix exhibiting the smallest change with [C]. These findings, which are in accord with changes in residual secondary structure (Figure 7b), indicate that the magnitudes of the driving forces for specific collapse [defined as $dE_M([C])/d[C]$] are (the greatest to the least) associated with β-hairpin 3−4 > β-strands 1−4 > β-hairpin 1−2 > helix. Thus, the forces driving collapse are nonuniformly distributed throughout the native state topology.

Concluding Remarks. The major findings in this paper reconcile the two-state interpretation of denaturant *m*-values with the broad ensemble of conformations in the unfolded state, and this resolves an apparent conundrum between protein collapse and the linear variation of $\Delta G_{ND}([C])$ with [C]. The success of the TM model in estimating *m*-values (*15, 16*) suggests that the free energy of the protein can be decomposed into a sum of independent transfer energies of backbone and side chain groups (eq 1). However, to connect the measured *m*-values to the heterogeneity in the molecular conformations, it is necessary to examine how the distribution of the DSE changes as [C] changes. This requires an examination of the validity of the second, more tenuous assumption in the TM, according to which the denatured ensemble surface area exposures of the backbone and side chains do not change as [C] changes. This assumption, whose validity has not been examined until now, implies that neither the polymeric nature of proteins, the presence of residual structure in the DSE, nor the extent of protein collapse alters $\langle \alpha_{k,D}^P([C])\rangle$ or $\langle \alpha_{k,N}^P([C])\rangle$ significantly. Our work shows that as the urea concentration (or more generally any denaturant) changes there are substantial changes in $P(\alpha_T)$ (Figure 3b), R_g, and R_{ee} (Figure 9). However, because backbone groups, whose $\alpha_{k,D}^B$ values are more narrowly distributed than those of almost all other groups (see Figure 4a), make the dominant contribution to the *m*-value (see Figure 4b), the *m*-value is constant in the transition region. Therefore, approximating eq 2 using eq 1 causes only small errors in the range of 3−8 M urea for protein L. In a recent study (*65*), the linear dependence of $\Delta G_{ND}[C]$ on [C] has been explained using an approximate treatment of the chain entropy using R_g^2 as an order parameter and a single energy scale to characterize the interaction between the residues and the aqueous denaturant solution.

The utility of the TM in yielding accurate values of *m* using measured transfer free energies of isolated groups, without taking the polymer nature of proteins into account, has been established in a series of papers (*15, 16*). The success of the empirical TM (eq 1), with its obvious limitations, has been rationalized (*15, 16*) by noting that the backbone makes the dominant contribution to *m*. This work expands further on this perspective by explicitly showing that the total backbone surface area ($\Delta\alpha_B$) changes weakly with [C] (for [C] > 3 M for protein L). This finding, to our

knowledge, has not been demonstrated previously. We conclude that eq 1, with the assumption that changes in surface areas are approximately [C]-independent, is reasonable. We ought to emphasize that *m*, a single parameter, is only a global descriptor of the properties of a protein at [C] $\neq 0$. Full characterization of the DSE requires calculation of changes in the distribution functions of a number of quantities (see Figure 3a,b) as a function of [C]. This can only be accomplished using MTM-like simulations and/or NMR experiments, which are by no means routine. The paucity of NMR studies that have characterized [C]-dependent changes in the DSE, at the residue level, shows the difficulty in performing such experiments.

The MTM simulations show discernible deviations from linear behavior at [C] < 3 M (Figure 1b), which can be traced to changes in the backbone surface area in the DSE. The structural characteristics of the unfolded state under such native conditions are different from those at [C] $\gg C_m$. The values of $\Delta\alpha^P_{k,D}$ are relatively flat when [C] > C_m (Figure 2b) but decrease below C_m because of protein collapse. Because δg^B_k([C]) dominates even below C_m (Figure 4b), it follows that departure from linearity in ΔG_{ND}([C]) is largely due to burial of the protein backbone. The often-observed drift in baselines of spectroscopic probes of protein folding may well be indicative of the changes in $\Delta\alpha^B_{k,D}$ and reflect the changing distribution of unfolded states (*5, 59*). Single-molecule experiments (*24*−*27, 29*), which directly probe changes in the DSE even below C_m, exhibit large shifts in the distribution of FRET efficiencies with [C]. Our simulations are consistent with these observations. The logical interpretation is that the DSE and, in particular, the distribution of α_T, α_B, and the radius of gyration, R_g, must be [C]-dependent. These simulations suggest that only by carefully probing these distributions can the replacement of eq 2 by eq 1 be quantitatively justified. In particular, large changes in the DSE occur under native conditions. Therefore, it is important to characterize the DSE under native conditions to monitor the collapse of proteins.

Equilibrium SAXS experiments on protein L at various guanidinium chloride concentrations showed that R_g does not change significantly above C_m (*60*). The ~2 Å change in R^D_g above C_m observed in these simulations is within the approximately ±1.8 Å error bars of the experimentally measured R_g above C_m (*60*). Our findings also suggest that the largest change in R^D_g occurs well below C_m (≤3 M urea). Under these conditions, the fraction of unfolded molecules is less than 1% (Figure 1b, inset), which implies it is difficult to accurately measure the R_g of the DSE using current SAXS experiments and explains why the equilibrium collapse transitions are not readily observed in scattering experiments. This work and a growing body of evidence from single-molecule FRET experiments show that the denatured state can undergo a continuous collapse transition that is modulated by changing solution conditions. This finding underscores the importance of quantitatively characterizing the DSE to describe the folding reaction. Establishing whether the collapse transition is second-order, which is most likely the case, will require tests similar to that proposed by Pappu and co-workers (*61*).

ACKNOWLEDGMENT

We thank Govardhan Reddy for a critical reading of this manuscript. We thank Prof. Buzz Baldwin for his interest and comments and for a tutorial on the historical aspects of the transfer model.

REFERENCES

1. Jackson, S. E. (1998) How do small single-domain proteins fold? *Folding Des. 3*, 81–91.
2. Santoro, M. M., and Bolen, D. W. (1992) A test of the linear extrapolation of unfolding free-energy changes over an extended denaturant concentration range. *Biochemistry 31*, 4901–4907.
3. Greene, R. F., and Pace, C. N. (1974) Urea and guanidine-hydrochloride denaturation of ribonuclease, lysozyme, α-chymotrypsin, and β-lactoglobulin. *J. Biol. Chem. 249*, 5388–5393.
4. Pace, C. N. (1986) Determination and Analysis of Urea and Gunidine Hydrochloride Denaturation Curves. *Methods Enzymol. 131*, 266–280.
5. Santoro, M. M., and Bolen, D. W. (1988) Unfolding free-energy changes determined by the linear extrapolation method. 1. Unfolding of phenylmethanesulfonyl α-chymotrypsin using different denaturants. *Biochemistry 27*, 8063–8068.
6. Jackson, S. E., and Fersht, A. R. (1991) Folding of chymotrypsin inhibitor. 2.1. Evidence for a 2-state transition. *Biochemistry 30*, 10428–10435.
7. Makhatadze, G. I. (1999) Thermodynamics of Protein Interactions with Urea and Guanidinium Hydrochloride. *J. Phys. Chem. B 103*, 4781–4785.
8. Fersht, A. R. (1999) *Structure and Mechanism in Protein Science: A guide to enzyme catalysis and protein folding*, 2nd ed., W. H. Freeman and Co., New York.
9. Street, T. O., Bolen, D. W., and Rose, G. D. (2006) A molecular mechanism for osmolyte-induced protein stability. *Proc. Natl. Acad. Sci. U.S.A. 103*, 13997–14002.
10. Yi, Q., Scalley, M. L., Simons, K. T., Gladwin, S. T., and Baker, D. (1997) Characterization of the free energy spectrum of peptostreptococcal protein L. *Folding Des. 2*, 271–280.
11. Khorasanizadeh, S., Peters, I. D., Butt, T. R., and Roder, H. (1993) Folding and stability of a tryptophan-containing mutant of ubiquitin. *Biochemistry 32*, 7054–7063.
12. Scalley, M. L., Yi, Q., Gu, H. D., McCormack, A., Yates, J. R., and Baker, D. (1997) Kinetics of folding of the IgG binding domain of peptostreptoccocal protein L. *Biochemistry 36*, 3373–3382.
13. Nozaki, Y., and Tanford, C. (1963) Solubility of amino acids and related compounds in aqueous urea solutions. *J. Biol. Chem. 238*, 4074–4081.
14. Tanford, C. (1964) Isothermal unfolding of globular proteins in aqueous urea solutions. *J. Am. Chem. Soc. 86*, 2050–2059.
15. Auton, M., Holthauzen, L. M. F., and Bolen, D. W. (2007) Anatomy of energetic changes accompanying urea-induced protein denaturation. *Proc. Natl. Acad. Sci. U.S.A. 104*, 15317–15322.
16. Auton, M., and Bolen, D. W. (2005) Predicting the energetics of osmolyte-induced protein folding/unfolding. *Proc. Natl. Acad. Sci. U.S.A. 102*, 15065–15068.
17. Auton, M., and Bolen, D. W. (2004) Additive transfer free energies of the peptide backbone unit that are independent of the model compound and the choice of concentration scale. *Biochemistry 43*, 1329–1342.
18. Auton, M., and Bolen, D. W. (2007) Application of the transfer model to understand how naturally occuring osmolytes affect protein stability. *Methods Enzymol. 428*, 397–418.
19. Lee, B., and Richards, F. M. (1971) Interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol. 55*, 379.
20. Smith, C. K., Bu, Z. M., Anderson, K. S., Sturtevant, J. M., Engelman, D. M., and Regan, L. (1996) Surface point mutations that significantly alter the structure and stability of a protein's denatured state. *Protein Sci. 5*, 2009–2019.
21. Doniach, S., Bascle, J., Garel, T., and Orland, H. (1995) Partially folded states of proteins: Characterization by X-ray scattering. *J. Mol. Biol. 254*, 960–967.
22. Bilsel, O., and Matthews, C. R. (2006) Molecular dimensions and their distributions in early folding intermediates. *Curr. Opin. Struct. Biol. 16*, 86–93.
23. Arai, M., Kondrashkina, E., Kayatekin, C., Matthews, C. R., Iwakura, M., and Bilsel, O. (2007) Microsecond hydrophobic

collapse in the folding of *Escherichia coli* dihydrofolate reductase an α/β-type protein. *J. Mol. Biol. 368*, 219–229.
24. Navon, A., Ittah, V., Landsman, P., Scheraga, H. A., and Haas, E. (2001) Distributions of intramolecular distances in the reduced and denatured states of bovine pancreatic ribonuclease A. Folding initiation structures in the C-terminal portions of the reduced protein. *Biochemistry 40*, 105–118.
25. Sinha, K. K., and Udgaonkar, J. B. (2005) Dependence of the size of the initially collapsed form during the refolding of barstar on denaturant concentration: Evidence for a continuous transition. *J. Mol. Biol. 353*, 704–718.
26. Kuzmenkina, E. V., Heyes, C. D., and Nienhaus, G. U. (2005) Single-molecule Forster resonance energy transfer study of protein dynamics under denaturing conditions. *Proc. Natl. Acad. Sci. U.S.A. 102*, 15471–15476.
27. Sherman, E., and Haran, G. (2006) Coil-globule transition in the denatured state of a small protein. *Proc. Natl. Acad. Sci. U.S.A. 103*, 11539–11543.
28. Merchant, K. A., Best, R. B., Louis, J. M., Gopich, I. V., and Eaton, W. (2007) Characterizing the unfolded states of proteins using single-molecule FRET spectroscopy and molecular simulations. *Proc. Natl. Acad. Sci. U.S.A. 104*, 1528–1533.
29. Hoffman, A., Kane, A., Nettels, D., Hertzog, D. E., Baumgartel, P., Lengefeld, J., Reichardt, G., Horsley, D., Seckler, R., Bakajin, O., and Schuler, B. (2007) Mapping protein collapse with single-molecule fluorescence and kinetic synchrotron radiation circular dichroism spectroscopy. *Proc. Natl. Acad. Sci. U.S.A. 104*, 105–110.
30. Geierhaas, C. D., Nickson, A. A., Lindorff-Larsen, K., Clarke, J., and Vendruscolo, M. (2007) BPPred: A Web-based computational tool for predicting biophysical parameters of proteins. *Protein Sci. 16*, 125–134.
31. Tran, H. T., and Pappu, R. V. (2006) Toward an accurate theoretical framework for describing ensembles for proteins under strongly denaturing conditions. *Biophys. J. 91*, 1868–1886.
32. Logan, T. M., Theriault, Y., and Fesik, S. W. (1994) Structural characterization of the FK506 binding protein unfolded in urea and guanidine hydrochloride. *J. Mol. Biol. 236*, 637.
33. Klein-Seetharaman, J., Oikawa, M., Grimshaw, S. B., Wirmer, J., Duchardt, E., Ueda, T., Imoto, T., Smith, L. J., Dobson, C. M., and Schwalbe, H. (2002) Long-range interactions within a non-native protein. *Science 295*, 1719–1722.
34. Muthukrishnan, K., and Nall, B. T. (1991) Effective concentrations of amino-acid side-chains in an unfolded protein. *Biochemistry 30*, 4706–4710.
35. Kuznetsov, S. V., Hilario, J., Keiderling, T. A., and Ansari, A. (2003) Spectroscopic studies of structural changes in two β-sheet-forming peptides show an ensemble of structures that unfold noncooperatively. *Biochemistry 42*, 4321–4332.
36. O'Brien, E. P., Ziv, G., Haran, G., Brooks, B. R., and Thirumalai, D. (2008) Denaturant and osmolyte effects on proteins are accurately predicted using the molecular transfer model. *Proc. Natl. Acad. Sci. U.S.A. 105*, 13403–13408.
37. Klimov, D. K., and Thirumalai, D. (2000) Mechanisms and kinetics of β-hairpin formation. *Proc. Natl. Acad. Sci. U.S.A. 97*, 2544–2549.
38. Klimov, D., and Thirumalai, D. (1999) Deciphering the timescales and mechanisms of protein folding using minimal off-lattice models. *Curr. Opin. Struct. Biol. 9*, 197–207.
39. Shakhnovich, E. (2006) Protein Folding Thermodynamics and Dynamics: Where Physics, Chemistry, and Biology Meet. *Chem. Rev. 106*, 1559–1588.
40. Cheung, M. S., Finke, J. M., Callahan, B., and Onuchic, J. N. (2003) Exploring the interplay between topology and secondary structural formation in the protein folding problem. *J. Phys. Chem. B 107*, 11193–11200.
41. Sugita, Y., and Okamoto, Y. (1999) Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett. 314*, 141–151.
42. Rhee, Y. M., and Pande, V. S. (2003) Multiplexed-replica exchange molecular dynamics method for protein folding simulation. *Biophys. J. 84*, 775–786.
43. Veitshans, T., Klimov, D., and Thirumalai, D. (1997) Protein folding kinetics: Timescales, pathways and energy landscapes in terms of sequence-dependent properties. *Folding Des. 2*, 1–22.
44. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem. 4*, 187–217.
45. Ferrenberg, A. M., and Swendsen, R. H. (1989) Optimized Monte Carlo data analysis. *Phys. Rev. Lett. 63*, 1195–1198.
46. Kumar, S., Bouzida, D., Swendsen, R. H., Kollman, P. A., and Rosenberg, J. M. (1992) The Weighted Histogram Analysis Method for free-energy calculations on biomolecules. 1. The method. *J. Comput. Chem. 13*, 1011–1021.
47. Shea, J., Nochomovitz, Y. D., Guo, Z., and Brooks, C. L. (1998) Exploring the space of protein folding Hamiltonians: The balance of forces in a minimalist β-barrel model. *J. Chem. Phys. 109*, 2895–2903.
48. O'Neill, J. W., Kim, D. E., Baker, D., and Zhang, K. Y. (2001) Structures of the B1 domain of protein L from *Peptostreptococcus magnus* with a tyrosine to tryptophan substitution. *Acta Crystallogr. D57*, 480–487.
49. Kim, D. E., Fisher, C., and Baker, D. (2000) A breakdown of symmetry in the folding transition state of protein L. *J. Mol. Biol. 298*, 971–984.
50. Hua, L., Zhou, R. H., Thirumalai, D., and Berne, B. J. (2008) Urea denaturation by stronger dispersion interactions with proteins than water implies a 2-stage unfolding. *Proc. Natl. Acad. Sci. U.S.A. 105*, 16928–16933.
51. Baskakov, I. V., and Bolen, D. W. (1998) Monitoring the sizes of denatured ensembles of staphylococcal nuclease proteins: Implications regarding m values, intermediates, and thermodynamics. *Biochemistry 37*, 18010–18017.
52. Yang, M., Ferreon, A. C. M., and Bolen, D. W. (2000) Structural thermodynamics of a random coil protein in guanidine hydrochloride. *Proteins: Struct., Funct., Genet. 4*, 44–49.
53. Klimov, D. K., and Thirumalai, D. (2002) Is there a unique melting temperature for two-state proteins? *J. Comput. Chem. 23*, 161–165.
54. Li, M. S., Klimov, D. K., and Thirumalai, D. (2004) Finite size effects on thermal denaturation of globular proteins. *Phys. Rev. Lett. 93*, 268107.
55. Fisher, M. E., and Barber, M. N. (1972) Scaling theory for finite-size effects in critical region. *Phys. Rev. Lett. 28*, 1516–1519.
56. Holtzer, M. E., Lovett, E. G., d'Avignon, D. A., and Holtzer, A. (1997) Thermal unfolding in a GCN4-like leucine zipper: C-13 α NMR chemical shifts and local unfolding. *Biophys. J. 73*, 1031.
57. Sadqi, M., Fushman, D., and Munoz, V. (2006) Atom-by-atom analysis of global downhill protein folding. *Nature 442*, 317–321.
58. Klimov, D. K., and Thirumalai, D. (1998) Cooperativity in protein folding: From lattice models with sidechains to real proteins. *Folding Des. 3*, 127–139.
59. Mello, C. C., and Barrick, D. (2003) Measuring the stability of partly folded proteins using TMAO. *Protein Sci. 12*, 1522–1529.
60. Plaxco, K. W., Millett, I. S., Segel, D. J., Doniach, S., and Baker, D. (1999) Chain collapse can occur concomitantly with the rate-limiting step in protein folding. *Nat. Struct. Biol. 6*, 554–556.
61. Vitalis, A., Wang, X. L., and Pappu, R. V. (2008) Atomistic Simulations of the Effects of Polyglutamine Chain Length and Solvent Quality on Conformational Equilibria and Spontaneous Homodimerization. *J. Mol. Biol. 384*, 279–297.
62. Zamyatnin, A. A. (1984) Amino-Acid, peptide, and protein volume in solution. *Annu. Rev. Biophys. Bioeng. 13*, 145–165.
63. Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H., and Zehfus, M. H. (1985) Hydrophobicity of amino-acid residues in globular proteins. *Science 229*, 834–838.
64. Frishman, D., and Argos, P. (1995) Knowledge-based protein secondary structure assignment. *Proteins 23*, 566–579.
65. Ziv, G., and Haran, G. (2009) Protein folding, protein collapse, and Tanford's transfer model: Lessons from single-molecule FRET. *J. Am. Chem. Soc. 131*, 2942–2947.

BI8021119