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 $\Delta R_g^2$, 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) or denatured (D) state (I). 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-
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

\[
m = \frac{1}{[C]} \sum_{k=1}^{N_k} n_k \delta g_k^S([C]) \Delta \alpha_k^S + \frac{1}{[C]} \sum_{k=1}^{N_k} n_k \delta g_k^B([C]) \Delta \alpha_k^B
\]

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 \(\delta g_k^S\) and \(\delta g_k^B\) are the experimentally measured transfer free energies for \(k\) (13, 17, 18) (Figure 1a). In eq 1, \(\Delta \alpha_k^S = \langle \alpha_k^S \rangle_D - \langle \alpha_k^S \rangle_N\) (P = S or B), where \(\langle \alpha_k^S \rangle_D\) and \(\langle \alpha_k^S \rangle_N\) are the average solvent accessible surface areas (19) of group \(k\) in the D and N states, respectively, and \(\alpha_k^S = k\) 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 \(\Delta \alpha_k^S\) 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) \(\delta g_k^S([C])\) is directly proportional to \([C]\) and (ii) \(\Delta \alpha_k^S\) is \([C]\)-independent. Experiments have shown that \(\delta g_k^S([C])\) is a linear function of \([C]\) (7) while the near independence of \(\Delta \alpha_k^S\) 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, \(\Delta \alpha_S = \sum_{k=1}^{N_S} \Delta \alpha_k^S + \sum_{k=1}^{N_B} \Delta \alpha_k^B\), and the various groups must also be a function of \([C]\), since we expect that \(\Delta \alpha_k\) must be a monotonically increasing function of \(R_{ee}^2\), which is the difference between the \(R_{ee}^2\) of the D and N states (26, 30). For compact objects, \(\Delta \alpha_T \sim \Delta R_{ee}^2\), 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 \(\Delta \alpha_k^F\) and \(R_{ee}\). 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.

The denatured and perhaps even the native state should be described as ensembles of fluctuating conformations and will here be named DSE and NSE (natural 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
regions of the DSE can collapse to varying degrees as [C] correlates with regions of residual structure. Thus, different interactions is nonuniformly distributed in the DSE and area distributions are independent of [C] as we do in eq 1.

Because the conformations and energies are known exactly with small changes in [C], backbone solvent accessible surface area is narrow, (2) Here, we establish that the distribution of the backbone solvent accessible surface area is narrow, with small changes in $\Delta \alpha^B$ 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 secondary structural elements. We decompose the nonbonded parameters that are residue-dependent. We take $\langle \alpha^A_{k,L}(\text{Res}) \rangle$ 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$_\alpha$ 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$_\alpha$ side chain model (C$_\alpha$-SCM) (37) to represent the 64-residue protein L. In the C$_\alpha$-SCM, each residue in the polypeptide chain is represented using two interaction sites, one that is centered on the $\alpha$-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$_\alpha$-SCM is a sum of bond angle ($E_\Lambda$), backbone dihedral ($E_D$), improper dihedral ($E_i$), backbone hydrogen bonding ($E_{HB}$), and nonbonded Lennard-Jones ($E_{LJ}$) terms ($E_P = E_\Lambda + E_D + E_i + E_{HB} + E_{LJ}$). 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$_\alpha$-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 strengths 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 ($\epsilon_{ij}$) in $E_{LJ}$ (36). Thus, the C$_\alpha$-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$_\alpha$-SCM in this study.

Simulation Details. Equilibrium simulations of the folding and unfolding reaction using the C$_\alpha$-SCM are performed using multiplexed replica exchange (MREX) (41, 42) in conjunction with low-friction Langevin dynamics (43) at [C]
$\varrho_{G}$ is the mean position of the 2

\[ \langle A([C], T) \rangle = Z([C], T)^{-1} \sum_{l=1}^{R} \sum_{t=1}^{n_{l}} \alpha_{n} \cdot e^{-\beta \langle E_{l}(t,[C]) + \Delta G_{c}(l,[C]) \rangle} \sum_{m=1}^{R} \sum_{n=1}^{n_{m}} \alpha_{n} \cdot e^{-\beta \langle E_{m}(t,[C]) \rangle} \]

where $\langle A([C], T) \rangle$ is the average of a protein property $A$ at urea concentration $[C]$ and temperature $T$ and $Z([C], T)$ 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 $n_{i}$ protein conformations from the $l$th replica. The value of $A$ from replica $l$ at time $t$ is $A_{l,i}$, and $E_{l}(t,[C])$ 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, $\Delta G_{c}(l,[C])$, the reversible work of transferring the $l$th protein conformation from 0 to $[C]$ M urea solution, is estimated using a form of the TM and is given by

\[ \Delta G_{u}(l, t, [C]) = \sum_{k=1}^{N_{u}} \alpha_{G}^{u} \cdot \delta_{G}^{u} \cdot \langle [C] \rangle + \sum_{k=1}^{N_{u}} \alpha_{B}^{u} \cdot \delta_{B}^{u} \cdot \langle [C] \rangle \]

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 $l, t$ ([\{\alpha_{G}^{u}(l, t, [C])\}]) are included.
equal to $\Theta[5 + \Delta(l)R] = 0$ and $\Theta[5 - \Delta(l)R] = 1$, and when $\Delta(l)R$ is less than 5 Å, $\Theta[5 + \Delta(l)R] = 1$ and $\Theta[5 - \Delta(l)R] = 0$.

Probability distributions were computed using $P(\pm \delta_nR[|C|]) = Z(\pm \delta_nR[|C|],T)Z(|C|,T)$, where $Z(A \pm \delta_nR[|C|],T)$ is the restricted partition function as a function of $A$. Due to the discrete nature of the simulation data, a bin with finite width $\pm \delta_nR$, whose value depends on $A$, is used.

$Z(A \pm \delta_nR[|C|],T) = \sum_{N=0}^{\infty} \sum_{P=0}^{N} f_{A}(l,t)e^{-\beta k_{B}T(|C| \pm \delta_nR)}$, 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 $lt$ has a value of $A$ in the range of $A \pm \delta_nR$, 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 $\delta g_{1}(|C|)$ 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 $\approx 6.6$ 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 ($\approx 0.8$ 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 $(\alpha_{0,1}|C|)$ and $(\alpha_{0,0}|C|)$ in eq 2 have the same dependence on $|C|$, making $\Delta \alpha_{1}^{l}$ effectively independent of $|C|$. (2) The distributions $P(\alpha_{0,0,1}|C|)$ in eq 2 are sharply peaked about their mean or most probable values of $\alpha_{0,0,1}|C|$ at all $|C|$ values, thus making $\Delta \alpha_{0}^{l}$ independent of $|C|$. In particular, if the standard deviation in $\alpha_{0,0,1}^{l}$ (denoted $\sigma_{0,0,1}$) is much less than $(\alpha_{0,0,1}|C|)$ for all $|C|$ values, then the $\Delta \alpha_{0}^{l}$ 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 $\Delta \alpha_{0}^{l}$ and $P(\alpha_{0,0,1}|C|)$ matter, thereby making $\Delta \alpha_{0}^{l}$ 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.

Figure 2: (a) $\Delta \alpha_{0,0,1}^{l}$ vs urea concentration for the backbone and the side chains alanine, phenylalanine, and glutamate, computed using the relationship $(\Delta \alpha_{0,0,1}^{l}|C|) = \sum_{P=0}^{N} \Delta \alpha_{0,0,1}^{l}(\alpha_{0,0,1}^{l}|C|)) d\alpha_{0,0,1}^{l}$ (j = D or N, and $P = S$ or B). For the backbone, $(\alpha_{0,0,1}^{l}|C|)) = N^{\sum_{P=0}^{N} \Delta \alpha_{0,0,1}^{l}(\alpha_{0,0,1}^{l}|C|))$, where $N = 64$, the number of residues in the protein. Brown dashed lines show $(\Delta \alpha_{0,0,1}^{l})$ 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) $n_{l} \Delta \alpha_{1}^{l}$ (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 $n_{l} \Delta \alpha_{1}^{l}$ values close to zero in this graph. $\Delta \alpha_{1}^{l} = (\alpha_{0,0,1}^{l}) - (\alpha_{0,0,1}^{l})$ (P = S or B). For the backbone, we plot $\sum_{l} n_{l} \Delta \alpha_{1}^{l}$. The inset shows $\Delta \alpha_{1}^{l}$ 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 $\Theta_{S}(lt)$, where $\Theta_{S}(lt)$ is the Heaviside step function that equals $\Theta[5 + \Delta(l)R]$ when the average of the NSE is computed (i.e., $n = \text{NSE}$) and is equal to $\Theta[5 + \Delta(l)R]$ when the average of the DSE is computed (i.e., $n = \text{DSE}$). Here, $\Delta(l)R$ is the root-mean-square deviation between the $C_{\alpha}$ carbon sites in the $C_{\alpha}$-SCM of conformation $lt$, and the $C_{\alpha}$ carbon atoms in the crystal structure [Protein Data

Bank entry 1HZ6 (48)]. When $\Delta(l)R$ is greater than 5 Å, $\Theta[5 + \Delta(l)R] = 0$ and $\Theta[5 - \Delta(l)R] = 1$, and when $\Delta(l)R$ is less than 5 Å, $\Theta[5 + \Delta(l)R] = 1$ and $\Theta[5 - \Delta(l)R] = 0$. $\Delta(l)R$ regions are shown as a solid black line. Secondary structure representation. (c) Corresponding to each secondary structure element are listed below the arrows; the $R$ values would be effectively independent of $|C|$. (2) The distributions $P(\alpha_{0,0,1}|C|)$ in eq 2 are sharply peaked about their mean or most probable values of $\alpha_{0,0,1}|C|$ at all $|C|$ values, thus making $\Delta \alpha_{0}^{l}$ independent of $|C|$. In particular, if the standard deviation in $\alpha_{0,0,1}^{l}$ (denoted $\sigma_{0,0,1}$) is much less than $(\alpha_{0,0,1}|C|)$ for all $|C|$ values, then the $\Delta \alpha_{0}^{l}$ 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 $\Delta \alpha_{0}^{l}$ and $P(\alpha_{0,0,1}|C|)$ matter, thereby making $\Delta \alpha_{0}^{l}$ 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.
\( \langle \alpha_{LD}(C) \rangle \) and \( \langle \alpha_{SM}(C) \rangle \) Do Not Have the Same Dependence on \([C]\). The changes in \( \langle \alpha_{LD} \rangle \) and \( \langle \alpha_{SM} \rangle \) as a function of \([C]\) show that as \([C]\) increases, both \( \langle \alpha_{LD} \rangle \) and \( \langle \alpha_{SM} \rangle \) increase (blue and green lines in Figure 2a). However, \( \langle \alpha_{LD}(C) \rangle \) has a stronger dependence on \([C]\) than \( \langle \alpha_{SM}(C) \rangle \) for both the backbone and side chains (Figure 2a). Thus, the observed linear dependence of \( \Delta G_{N D}(C) \) on \([C]\) cannot be rationalized in terms of similarity in the variation of \( \langle \alpha_{P}(C) \rangle \) and \( \langle \alpha_{S}(C) \rangle \) as \([C]\) changes. The stronger dependence of \( \langle \alpha_{LD}(C) \rangle \) 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 \( \langle \alpha_{LD}(C) \rangle \) than \( \langle \alpha_{SM}(C) \rangle \) 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, \( \langle \alpha_{SM}(C) \rangle \) shows a stronger dependence on \([C]\).

Surface Area Distributions Are Broad in the DSE. The variation of \( \Delta \alpha_{LD} \) and \( \Delta \alpha_{P} \), with \([C]\) suggests that the probability \( P(\alpha_{LD}^0|C) \) for the backbone surface area distribution in the DSE, \( P(\alpha_{LD}^0|C) \), shifts toward higher values of \( \alpha_{LD}^0 \) 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 \( \alpha_{LD}^0 \) with \([C]\) is \( \sim 5 \) times smaller than the corresponding change in \( \alpha_{P} \) (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 \( \Delta \alpha_{P} \) 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 probability \( P(\alpha_{LD}^0) \) distributions by computing the ratio \( \rho = \sigma_{\alpha_{LD}}/\langle \alpha_{LD}^0 \rangle \), where \( \sigma_{\alpha_{LD}} = (\langle \alpha_{LD}^0 \rangle - \langle \alpha_{LD}^0 \rangle)^{1/2} \). Figure 4a shows \( \rho \) as a function of \([C]\) for the various protein components (backbone, side chains, and the entire protein). As with the backbone probability \( P(\alpha_{LD}^0) \) distribution (Figure 3a), \( \rho \) indicates that \( P(\alpha_{LD}^0) \) becomes narrower at higher urea concentrations for most values of \( \rho \) (Figure 4a). At 8 M urea, the width of \( P(\alpha_{LD}^0) \) ranges from 5 to 25% of the average value of \( \alpha_{LD}^0 \) for all groups, except when \( k = \text{Trp} \) which has an even larger width. Clearly, \( \rho \) is large at all \([C]\), which accounts for the dependence of \( \Delta \alpha_{LD} \) on \([C]\). The results in Figure 4 show that there are discernible changes in \( \rho \) which reflects the variations in \( P(\alpha_{LD}^0|C) \) as \([C]\) changes. 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 \( \Delta G_{N D}(C) \). 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 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 most to the stability changes with \([C]\) (16). Thus, for \([C] \geq 3\) 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.

The relative change in accessible surface area of the backbone \( \Delta \alpha_{LD} \) 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 \( \Delta \alpha_{LD} \) 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 \( \Delta \alpha_{LD} \) 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 \( \Delta \alpha_{T}(4 \text{ M} \rightarrow 8 \text{ M}) \approx 300 \text{ Å}^2 \) (Figure 2c, inset). Thus, the third possibility is correct, namely, that the weak dependence of \( \Delta \alpha_{LD} \) 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
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 $\alpha_{k}^{D}$ for individual residues of type $k$ and the average $\langle \alpha_{k}^{D} \rangle$ 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 $\alpha_{k}^{D}$ 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 $\alpha_{k}^{B}$ suggest that different regions of the protein may collapse in the DSE at different urea concentrations, driven by differences in $\Delta \alpha_{k}^{D}$ 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 $\langle \alpha_{k}^{B} \rangle$ is observed when individual residues are considered (Figure 2a). For example, $\alpha_{k}^{B}$ 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 $\alpha_{k}^{D}$ for residues 8 and 20, both of which adopt a $\beta$-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 $\alpha_{k}^{D}$ 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 $P(\alpha_{k}^{D})$ 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 $\langle \alpha_{k}^{D} \rangle$ and higher-order moments of $P(\alpha_{k}^{D})$ 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 $\sim 1200 \text{ Å}^2$ 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 $\Delta \alpha_{B}$ increases by 75 Å$^2$, and $n_{i} \Delta \alpha_{k}^{D}$ for $k =$ Phe, Leu, Ala, or Ile increases by 21–42 Å$^2$.

The dispersion in $\alpha_{k}^{D}$ could be caused by residual structure in the DSE (51, 52). We test this proposal quantitatively by plotting $\alpha_{k}^{D} / \alpha_{k}^{N}$ for each residue, where $\alpha_{k}^{N}$ is the maximum $\alpha_{k}^{D}$ value for residue type $k$ in 8 M urea. If residual structure causes the dispersion in $\alpha_{k}^{D}$, then we expect that $\alpha_{k}^{D} / \alpha_{k}^{N}$ should depend on the secondary structure element that residue $k$ adopts in the native state. We find that there is a correlation between $\alpha_{k}^{D}$ and the helical secondary structure element [residues 26–44 (Figure 6)]. The helical region tends to have smaller $\alpha_{k}^{D}$/$\alpha_{k}^{N}$ 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 $\alpha_{k}^{D}$/$\alpha_{k}^{N}$ 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 $\alpha_{k}^{D}$ depends not only on the residue type but also on the residual structure present.
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 \( m,^i \) can be measured. For strict two-state behavior, \( C_{m,i} = C_m \) for all \( i \), where \( C_m \) is the urea concentration below which the \( i \)th residue adopts its native conformation. The ineqivalence 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-IzK 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-IzK, 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 \( Q(C_{m,i}) = 0.5 \), where \( Q \) 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 \( Q(T_{m,i}) = 0.5 \), 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-
erative than denaturant-induced unfolding (58). It should be emphasized that the Holtzer effect is fairly general, and only as $N$ increases will $\Delta C$ and $\Delta T$ 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 $\Delta G_{ND}([C])$ (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, $R_g$, and $\Delta R_T$ 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 $\alpha_{LD}$ values change only modestly (Figure 2a). However, below 4 M, much larger changes in $\alpha_{LD}$ 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 $\approx$300 Å$^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 $n_k \Delta \alpha_k (=n_k(\alpha_{LD}([C])) - \langle \alpha_N([C]) \rangle)$, 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 \( R_g^2 \) 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 \( R_g^2 \) shows that a collapse transition occurs in the DSE. We find no evidence (e.g., a sigmoidal transition in \( R_g^2 \) vs [C]) that the DSE at 0 M \((R_g^2 = 15.5 \, \text{Å})\) and the DSE at 8 M urea \((R_g^2 = 21.5 \, \text{Å})\) 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 \( (\psi) \) in the DSE versus [C] (Figure 7a). Above 4 M urea, only \( \beta \)-hairpin 3–4 and the helix are formed to any appreciable extent. However, below 4 M urea, \( \beta \)-hairpin 1–2 and \( \beta \)-sheet interactions between strands 1 and 4 can be found in the DSE. For example, at 1 M urea, \( \beta \)-hairpin 1–2 and strands 1 and 4 are formed 21 and 16% of the time while there is 56% helical and 74% \( \beta \)-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 Intrapeptide Interactions. If we neglect changes in protein conformational entropy, we find two opposing energies control the [C]-dependent behavior of \( R_g^2 \); the interaction of the peptide residues with solvent (the solvation energy, denoted \( E_S \)) and the intrapeptide nonbonded interactions between the residues (denoted \( E_P \)). For denaturants, such as urea, \( E_S \) favors an increase in \( R_g^2 \) and a concomitant increase in solvent accessible surface area, while \( E_P \) typically is attractive and hence favors a decrease in \( R_g^2 \). Because \( E_S \) in the TM model is proportional to a surface area term and \( E_P \) is likely to be approximately proportional to the number of residues in contact (which increases as the residue density increases upon collapse), we expect \( E_S([C]) \propto -[C](R_g^2([C]))^2 \) and \( E_P([C]) \propto -1/(R_g^2([C]))^3 \). The behavior of these two functions [increasing \( R_u^2([C]) \)] leads to a more favorable \( E_S([C]) \) and unfavorable \( E_P([C]) \) suggests that there should always be some contraction (expansion) of the DSE with decreasing (increasing) [C]. The molecular details in the \( C_{eq} \)-SCM allow us to exactly determine \( E_S([C]) \) and \( E_P([C]) \) 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 \( E_S([C]) \), \( E_P([C]) \), and \( E_{M}([C]) \) \( =E_S([C]) + E_P([C]) \) in the DSE. As indicated by the Flory-like argument given above, \( E_S([C]) \) becomes more favorable with increasing [C], and \( E_P([C]) \) 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 \( R_u^2([C]) \). 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 \( R_u^2 \), \( R_{ee} \), and \( \Delta \alpha_x \) 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 \( R_u^2 \).

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 kcal/mol, except for that of \( \beta \)-hairpin 3–4 which changes by as much as \( \sim 0.9 \) kcal/mol. Below 4 M, the monomer interaction energies change by as much as 1.5 kcal/mol, 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 \( \Delta G_{NC}([C]) \) or \( \Delta G_{NC}([C]) \)] (the greatest to the least) associated with \( \beta \)-hairpin 3–4 > \( \beta \)-strands 1–4 > \( \beta \)-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 \( \Delta G_{NC}([C]) \) 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 \( \Delta G_{NC}([C]) \) or \( \Delta G_{NC}([C]) \) significantly. Our work shows that as the urea concentration (or more generally any denaturant) changes there are substantial changes in \( P(\alpha) \) (Figure 3b), \( R_g \), and \( R_{ee} \) (Figure 9). However, because backbone groups, whose \( \alpha_{DP} \) 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 \( \Delta G_{NC}([C]) \) on [C] has been explained using an approximate treatment of the chain entropy using \( R_u^2 \) 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 \( \Delta \alpha_x \) 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 native conditions are different from those at [C] \( \gg C_m \). The values of \( \Delta G^U_{2D} \) are relatively flat when [C] > \( C_m \) (Figure 2b) but decrease below \( C_m \) because of protein collapse. Because \( \delta G^m([C]) \) dominates even below \( C_m \) (Figure 4b), it follows that departure from linearity in \( \Delta G_{N0}([C]) \) 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 \( \Delta G^U_{2D} \) 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 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.

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 \( R_g \) above \( C_m \) observed in these simulations is within the approximately \( \pm 1.8 \) Å error bars of the experimentally measured \( R_g \) above \( C_m \) (60). Our findings also suggest that the largest change in \( R_g \) occurs well below \( C_m \) (\( \leq 3 \) 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**