JACS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Article

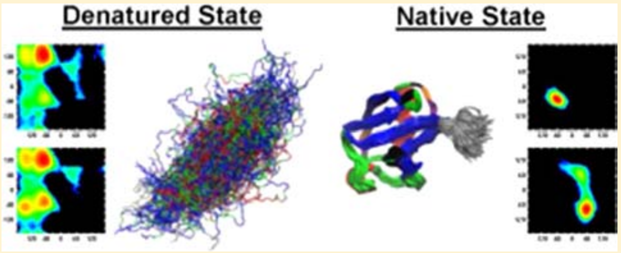
pubs.acs.org/JACS

Context and Force Field Dependence of the Loss of Protein Backbone Entropy upon Folding Using Realistic Denatured and Native State Ensembles

Michael C. Baxa,[†,‡] Esmael J. Haddadian,[†,§] Abhishek K. Jha,[§,∥,#] Karl F. Freed,*[,§,∥,⊥] and Tobin R. Sosnick*[,†,‡,⊥]

[†]Institute for Biophysical Dynamics, [‡]Department of Biochemistry and Molecular Biology, [§]James Franck Institute, [∥]Department of Chemistry, [⊥]Computation Institute, The University of Chicago, 929 East 57th Street, Chicago, Illinois 60637, United States

S Supporting Information



ABSTRACT: The loss of conformational entropy is the largest unfavorable quantity affecting a protein's stability. We calculate the reduction in the number of backbone conformations upon folding using the distribution of backbone dihedral angles (ϕ,ψ) obtained from an experimentally validated denatured state model, along with all-atom simulations for both the denatured and native states. The average loss of entropy per residue is $T\Delta S^{BB}{}_{U-N}$ = 0.7, 0.9, or 1.1 kcal·mol^{-1} at T = 298 K, depending on the force field used, with a 0.6 kcal·mol^{-1} dispersion across the sequence. The average equates to a decrease of a factor of 3–7 in the number of conformations available per residue ($f = \Omega_{Denatured}/\Omega_{Native}$) or to a total of $f_{tot} = 3^n–7^n$ for an n residue protein. Our value is smaller than most previous estimates where f = 7–20, that is, our computed $T\Delta S^{BB}{}_{U-N}$ is smaller by 10–100 kcal mol^{-1} for n = 100. The differences emerge from our use of realistic native and denatured state ensembles as well as from the inclusion of accurate local sequence preferences, neighbor effects, and correlated motions (vibrations), in contrast to some previous studies that invoke gross assumptions about the entropy in either or both states. We find that the loss of entropy primarily depends on the local environment and less on properties of the native state, with the exception of α-helical residues in some force fields.

INTRODUCTION

The reduction in the number of available backbone conformations, $f = \Omega_{Denatured}/\Omega_{Native}$, is directly related to the loss of backbone entropy, $\Delta S^{BB}{}_{U-N} = R \ln f$. As such, an accurate determination of the magnitude of f is essential for a proper and accurate evaluation of ΔG_{U-N}. In principle, the calculation of the backbone entropy and f should be straightforward. The simplest estimates assume that the native state represents a single conformation, while each pair of dihedral ϕ,ψ angles can adopt m rotomeric forms in the denatured state, for a reduction of a total of m^n conformations for an n residue protein.

Although the Ramachandran map contains only 3–5 highly populated regions or basins, it is unclear whether each of these basins can be approximated as a single state. Also, the approximation that the native state corresponds to a single conformation may be inaccurate due to protein dynamics. These issues underscore the broader question of what defines a distinct conformation in either the denatured or native state. Also, little is known about the factor by which the correlated motions of neighboring residues reduce the total number of available conformations.

Many approaches have been employed to calculate the loss of backbone conformational entropy, $\Delta S^{BB}{}_{U-N}$, but none includes all of these aforementioned considerations,[1−4] especially the influence of neighboring residues. Some previous analyses fail to calculate the difference in entropy between the native and unfold states or rely on inaccurate assumptions or gross approximations concerning either of these two states. Not surprisingly, these methods yield values that differ by more than 0.5 kcal·mol^{-1} per residue (at T = 298 K), or 50 kcal·mol^{-1} for a 100 residue protein. Because this uncertainty greatly exceeds a protein's net stability, an accurate determination of $\Delta S^{BB}{}_{U-N}$ is essential to properly quantifying protein thermodynamics and the energetics of water−protein interactions.

We address these issues by calculating the conformational entropy from Ramachandran distributions for realistic ensembles of the folded and denatured state of ubiquitin (Ub) while accounting for correlated motions of adjacent residues.[5,6] We find that the entropy is moderately dependent on force field (FF): $T\Delta S^{BB}{}_{U-N}$ = 0.7 ± 0.3, 0.9 ± 0.3, or 1.1 ± 0.3 kcal·mol^{-1}·residue^{-1} (or f = 3.3, 4.6, or 7.0 lost states per residue), respectively, for the OPLS/AA-L[7,8] and Garcia-Sanbonmatsu modified Amber94 (GS-A94)[9] FF with implicit solvent, and the CHARMM27 FF with explicit solvent.[10−12]

Received: July 2, 2012
Published: August 28, 2012

dx.doi.org/10.1021/ja3064028 | J. Am. Chem. Soc. 2012, 134, 15929−15936

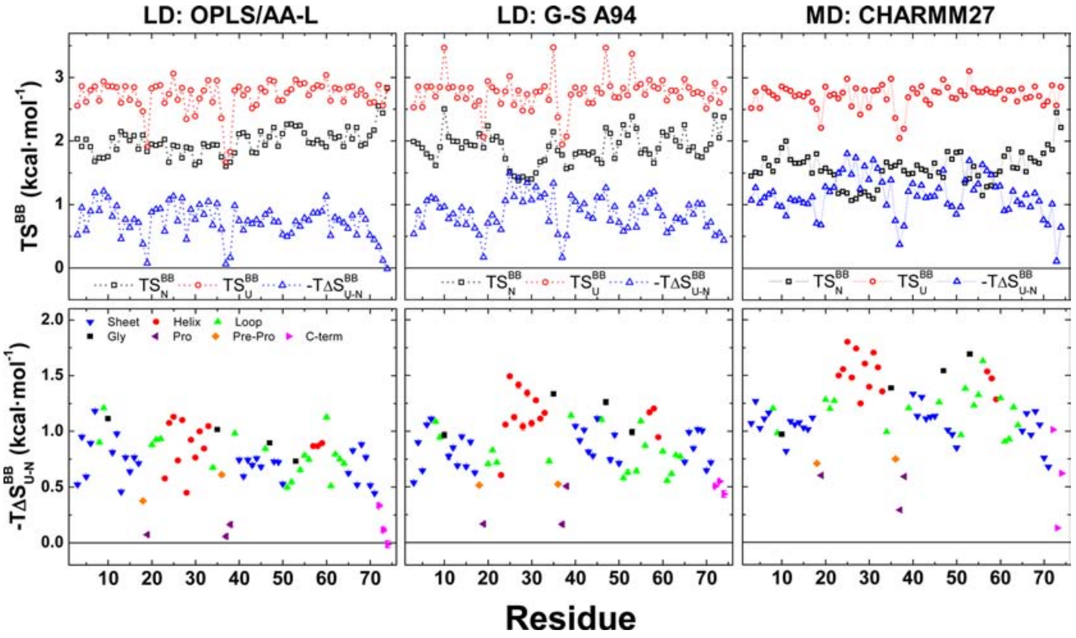


Figure 1. Loss of backbone entropy upon folding for Ubiquitin. (Upper Panels) The backbone entropies corrected for nearest neighbor correlations for the folded and denatured states, along with the differences between the two states, for residues 3−74 calculated using both the OPLS/AA-L (left) and G-S A94 FFs (middle), as well as the CHARMM FF in explicit solvent (right). The entropy calculations for the native and DSE implicitly depend on the pixel resolution used to construct the probability distributions. We eliminate this dependence by computing the entropy for multiple bin widths and fitting the difference in entropy as a function of the ratio of pixel sizes (see Supporting Information Methods, Supporting Information Figure 1). (Lower Panels) The change in backbone entropy during folding is presented with the residues colored according to native secondary structure elements. While the loss of entropy varies across the sequence, no strong dependence on sequence appears, except for the unstructured carboxy-terminal, proline, and preproline residues that incur smaller changes in entropy during folding.

Except for helical residues, the loss of backbone entropy is largely independent of other native state properties, for example, surface burial. Our values are smaller than those calculated in other studies.[3,4,13−20] The influence of neighboring residues indicates that the total chain entropy is not the sum of entropies for individual residues, as usually assumed.

■ RESULTS AND DISCUSSION

The Denatured State Ensemble. The denatured state ensemble (DSE) is generated beginning from dihedral angles obtained from a highly restricted PDB-based coil library. Individual chains created using these angles are then subjected to implicit solvent Langevin Dynamics (LD) or explicit solvent molecular dynamics (MD) simulations. The coil library excludes helices, strands, turns, and any residue adjacent to these three types of hydrogen bonded structures. Our library recapitulates global (radius of gyration, R_g) and local (NMR residual dipolar couplings, RDCs) properties of chemically denatured states.[21] Because the conformational diversity of each residue is affected by the neighboring residues, our entropy calculation for each residue includes the influence of both of the neighboring residues (e.g., Val-Arg-Lys). The finite size of the PDB library restricts the initial DSE to adequately reflecting only the probabilities of occupying each of the major Ramachandran basins (e.g., $P_{\alpha R}$, P_{β}, P_{PPII}, $P_{\alpha L}$, and P_{other}), while the statistics are inadequate for sampling *within* each basin. Hence, the distributions within each basin are determined using LD or MD simulations that constrain each residue to remain within its original basin. Thus, this calculation decomposes the total probability distribution into two components: the interbasin distribution (established by the Ramachandran basin propensities in the coil library) and the distribution for intrabasin motions obtained with all-atom simulations.[22]

To constrain the LD and MD simulations to remain in the original basins, each residue is restricted to a single basin using a harmonic reflecting "wall" at the edge of the basin (Methods). This wall also prevents the denatured chains from collapsing to an unrealistic near-native radius of gyration, as often generated using many FFs.[23−27] This degree of compaction is not observed experimentally for small proteins such as Ub even under native-like conditions, with either small angle scattering[28−31] or fluorescence resonance energy transfer (FRET).[32−34] Both experimental methods indicate the DSE is highly expanded, albeit with relatively minor numerical discrepancies.[31]

The implicit solvent LD simulations are run with two different FFs, the OPLS/AA-L[7,8] and the Garcia and Sanbonmatsu modified version of Amber 94 (G-S A94)[35] FFs. The entropy of this LD-augmented DSE is largely independent of position except for glycine, proline and preproline residues (Figure 1).

Computing the Conformational Entropy. The entropy is calculated from the 2D Ramachandran map for each residue that has been divided into equal sized pixels of area b^2 (Supporting Information Methods). The entropy is calculated according to $S = -R\sum P_i \ln P_i$ where P_i is the probability in the i^{th} pixel and R is Boltzmann's constant. Because neighboring residues have correlated basin probabilities, the influence of neighboring residue is calculated using a 4D Ramachandran space where P_i is the probability for four consecutive angles (ϕ_i, ψ_i, ϕ_{i+1}, ψ_{i+1}) in a voxel of volume b^4. The contribution of the correlation is equally split between the two neighbors, $\Delta S_j = (\Delta S_{j-1,j} + \Delta S_{j,j+1})/2$ (higher order correlations should be

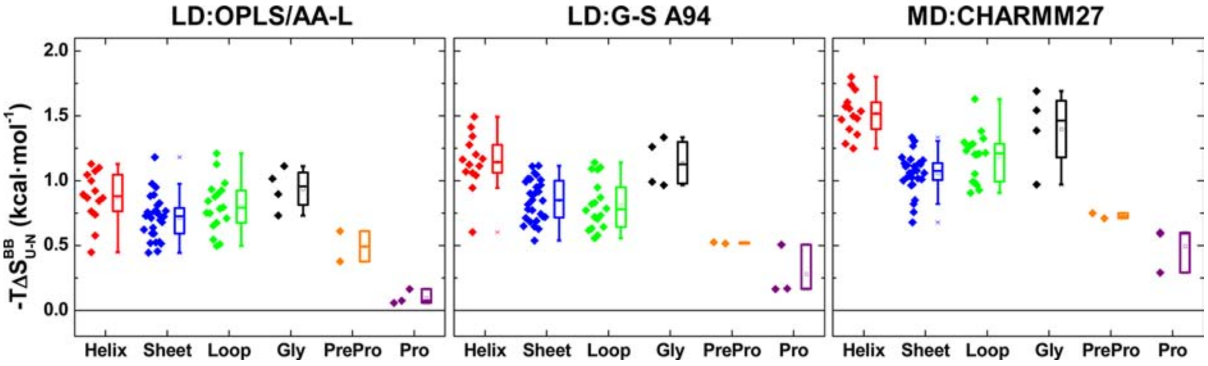


Figure 2. Loss of backbone entropy for secondary structure elements. Calculated changes in backbone entropy are averaged over various secondary structure types. Glycines and helical residues on average yield a slightly larger loss in entropy than coil and sheet residues. Proline residues exhibit little change in entropy between states. Preproline residues likewise have a reduced change in entropy. Individual values are shown for each secondary structure type along with a box-whisker plot covering the interquartile range (IQ = Q2 − Q1) and the upper inner (Q2 + 1.5·IQ) and lower inner (Q1 − 1.5·IQ) fence values, respectively.

relatively insignificant according to our previous peptide simulations[36]). When we partition the Ramachandran space, a choice of b = 10° provides sufficient resolution to converge ΔS^{BB} and adequately distinguish backbone conformations while not being limited by counts (Supporting Information Figure 1). Absolute entropies depend on pixel/voxel size (i.e., "How different do the angles need to be for two conformations to be considered distinct states?"), but entropy differences do not.

The Change in Conformational Entropy in Folding. The loss of backbone entropy is defined as the difference in entropy between the DSE and the native state ensemble (Figure 1). In general, β sheet residues exhibit smaller entropy loss than the α helical residues. This difference predominantly reflects the reduced entropy of the helical residues in the native state, since the residues in helices sample a much smaller region of the Ramachandran map (Supporting Information Figures 2–4). To test for adequate sampling, the native and DSE are split in half and the entropy is computed for each half separately; the values differ minimally (<~0.1 kcal·mol^{-1}·residue^{-1}).

The conformation and chemical identity of both a residue and its nearest neighbors influence its loss of backbone entropy. Differences between the G-S A94 FF and the OPLS/AA-L FF are evident in the helical regions in the native state. Helical regions exhibit a higher degree of rigidity with the G-S A94 FF, resulting in a slightly larger change in conformational entropy compared to the OPLS/AA-L simulations (1.1 ± 0.2 vs 0.9 ± 0.2 kcal·mol^{-1}, respectively). Also, sheet regions incur a larger loss in entropy with the G-S A94 FF than the OPLS/AA-L FF due to increased rigidity in the native state simulations (0.9 ± 0.2 vs 0.7 ± 0.2 kcal·mol^{-1}). The loss of entropy in the loop regions is comparable for the two FFs (Figure 2, Table 1, Supporting Information Table 1). All standard deviations reported here represent *site-to-site variations* across the Ub sequence and not the statistical error, which generally is smaller (Table 1, Supporting Information Tables 1 and 2).

Glycine residues display different behaviors in the two FFs. Glycines in both the DSE and native state simulations exhibit greater conformational diversity with the G-S A94 FF as is apparent in the entropy profiles in Figure 1, as well as in the probability distributions in Figure 3 and Supporting Information Figures 3 and 4.

However, the resulting difference in backbone entropy is comparable between the two FFs in implicit solvent ($-T\Delta S^{BB}{}_{U-N}$ = 0.9 ± 0.2 and 1.1 ± 0.1 kcal·mol^{-1} for G-S A94 and OPLS/AA-L, respectively). Proline residues yield similar backbone entropies in N and U ($-T\Delta S^{BB}{}_{U-N}$ = 0.1 ± 0.1 and 0.3 ± 0.2 kcal·mol^{-1} for the OPLS/AA-L and G-S A94 FF, respectively). Preproline residues exhibit a lower change in backbone entropy between states (0.5 ± 0.2 kcal·mol^{-1}).

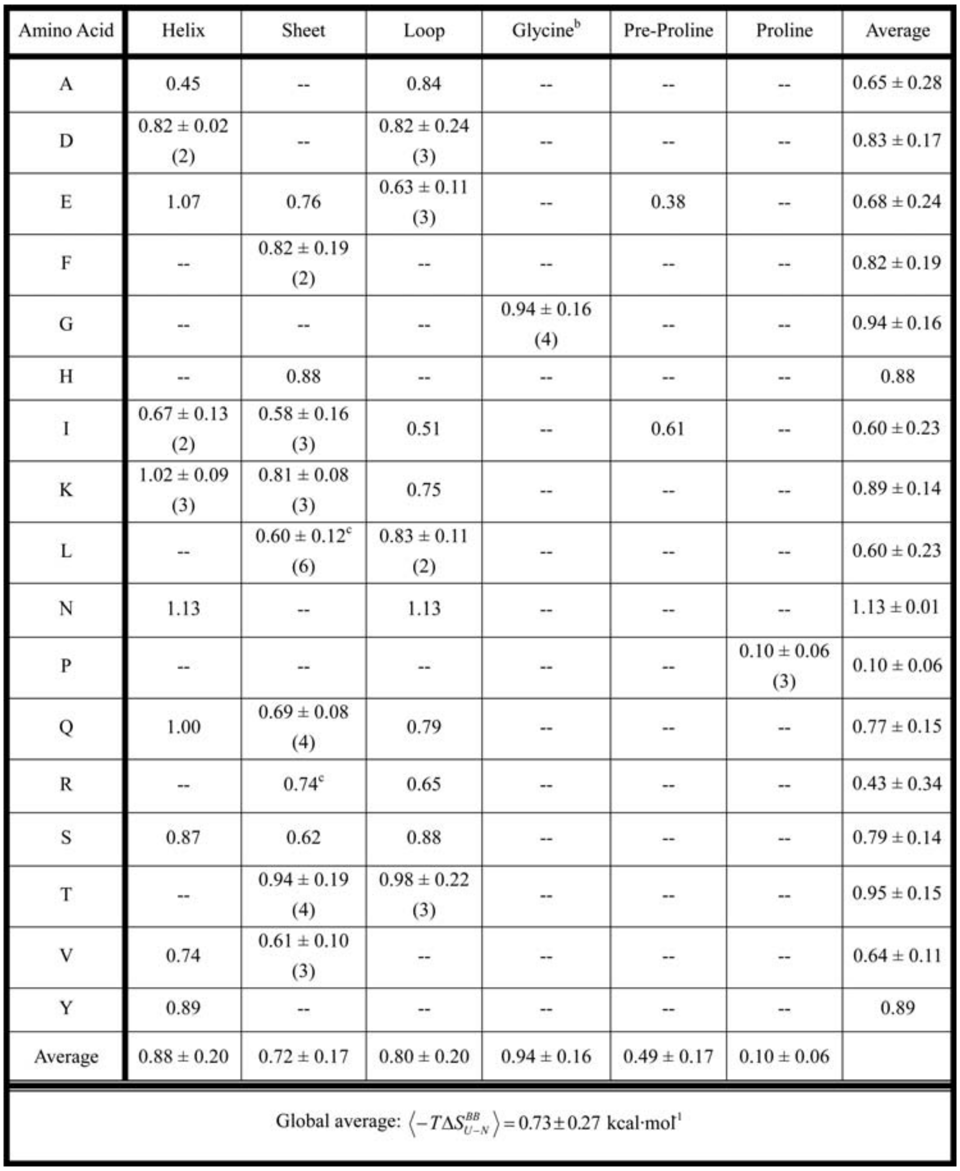
The burial level in the native state is only weakly correlated with the loss in backbone entropy (R ~ −0.2, Supporting Information Figure 5). The fractional change in solvent accessible surface area is uncorrelated to the loss in backbone entropy (Supporting Information Figure 5). Again, the native state properties have little effect on the backbone entropy as compared to the sequence.

Comparison with Explicit Solvent. We regenerate denatured and native state ensembles using explicit solvent simulations using the TIP3P water model and the CHARMM27 FF (Figures 1 and 2). The loss of entropy is systematically higher, but the overall trend is the same; for example, helical residues display the greatest loss of entropy. The native state profile is more sensitive to the choice of FF than the DSE profile. The most pronounced difference is for helical residues, which are conformationally more diverse in the OPLS/AA-L FF than in the CHARMM27 FF. We emphasize that the differences are largely due to the FF and not a consequence of the choice of solvent model. A long time (57 ns; the first 15 ns are excluded) explicit solvent simulation using the OPLS/AA-L FF for the native state is more similar to the implicit solvent simulations of the native state with the same FF (Supporting Information Figure 6). These differences only highlight biases in the various FFs, which has been noted by others,[5,37−40] and the inadequacy of assuming that the native state is a single conformation.[16,18]

Comparison with Other Studies. Many computational and experimental studies have calculated the change in conformational entropy upon folding, and a spectrum of values has been found with varying overlap, as detailed below. Despite any apparent overlap between our calculation and others, we stress that many of the methods are predicated on gross or false assumptions regarding the properties of the two states or the calculation of the entropy.

Although our calculations are very similar in spirit to other Ramachandran-based determinations of the conformational entropy,[3,17−20] our values are smaller by 0.3−1.5 kcal·mol^{-1}. The primary difference in approaches lies in our use of an experimentally validated PDB-based model for the DSE. In

Table 1. Average Loss of Backbone Entropy, $T\Delta S^{BB}$, upon Folding Using the OPLS/AA-L FF[a]



Amino Acid	Helix	Sheet	Loop	Glycine[b]	Pre-Proline	Proline	Average
A	0.45	--	0.84	--	--	--	0.65 ± 0.28
D	0.82 ± 0.02 (2)	--	0.82 ± 0.24 (3)	--	--	--	0.83 ± 0.17
E	1.07	0.76	0.63 ± 0.11 (3)	--	0.38	--	0.68 ± 0.24
F	--	0.82 ± 0.19 (2)	--	--	--	--	0.82 ± 0.19
G	--	--	--	0.94 ± 0.16 (4)	--	--	0.94 ± 0.16
H	--	0.88	--	--	--	--	0.88
I	0.67 ± 0.13 (2)	0.58 ± 0.16 (3)	0.51	--	0.61	--	0.60 ± 0.23
K	1.02 ± 0.09 (3)	0.81 ± 0.08 (3)	0.75	--	--	--	0.89 ± 0.14
L	--	0.60 ± 0.12[c] (6)	0.83 ± 0.11 (2)	--	--	--	0.60 ± 0.23
N	1.13	--	1.13	--	--	--	1.13 ± 0.01
P	--	--	--	--	--	0.10 ± 0.06 (3)	0.10 ± 0.06
Q	1.00	0.69 ± 0.08 (4)	0.79	--	--	--	0.77 ± 0.15
R	--	0.74[c]	0.65	--	--	--	0.43 ± 0.34
S	0.87	0.62	0.88	--	--	--	0.79 ± 0.14
T	--	0.94 ± 0.19 (4)	0.98 ± 0.22 (3)	--	--	--	0.95 ± 0.15
V	0.74	0.61 ± 0.10 (3)	--	--	--	--	0.64 ± 0.11
Y	0.89	--	--	--	--	--	0.89
Average	0.88 ± 0.20	0.72 ± 0.17	0.80 ± 0.20	0.94 ± 0.16	0.49 ± 0.17	0.10 ± 0.06	

Global average: $\left\langle -T\Delta S^{BB}_{U-N} \right\rangle = 0.73 \pm 0.27$ kcal·mol^{1}

[a] Units in kcal·mol^{-1} (T = 298 K). Errors are the standard deviation from averaging over multiple residues. Values in parentheses are the number of instances, if greater than one. [b] Entropy changes are computed for glycines located in loop regions of Ub. [c] These values exclude the largely unstructured C-terminal residues R72, L73, and R74, which have $T\Delta S = 0.33 \pm 0.01$, 0.12 ± 0.02, and -0.01 ± 0.03 kcal·mol^{-1}, respectively.

contrast, the Ramachandran distributions used in prior studies are much broader (e.g., from simplified peptide models), leading to an overestimation of over 0.3−0.5 kcal·mol^{-1}·residue^{-1}. In particular, calculations with distributions determined for dipeptides contain only a single pair (ϕ,ψ) of dihedral angles[3] and intrinsically cannot include the influence of neighboring side chains. The dipeptide model is an inappropriate representation of the DSE as the neighboring residues affect both the basin propensities and the motions of a residue. Fitzkee and Rose estimate that local chain sterics and backbone solvation requirements produce a small, 20% depletion in allowable denatured state conformations per residue ($T\Delta S$ = 0.1 kcal mol^{-1} residue^{-1}).[41]

A second difference arises from our accounting for the contribution for correlated motions. A residue's entropy depends on amino acid type and the chemical identity and conformation of adjacent residues (Figure 4). This dependence yields contributions ranging between −0.4 and 0.5 kcal·mol^{-1}·residue^{-1} in our calculations and accounts for 0.1−0.3 kcal·mol^{-1} per residue in the difference between our calculation and others for the denatured state entropy.

Other calculations of the change in conformational entropy utilize estimations from the covariance matrix for the atomic displacement of the atoms in the proteins under a single quantum harmonic well approximation.[13,14] While probably suitable for the compact helical monomer and trimeric coiled-coil, the use of the covariance matrix is unsuitable for determining accurate conformational entropies for denatured proteins or native proteins with residues that undergo substantial backbone conformational transitions as illustrated by the following simple example. Consider a one-dimensional symmetric double well potential[22,42] with barrier at x = 0 and wells at $x = \pm a$. The covariance matrix has the single element $\langle x^2 \rangle = a^2$, which grossly overestimates the conformational

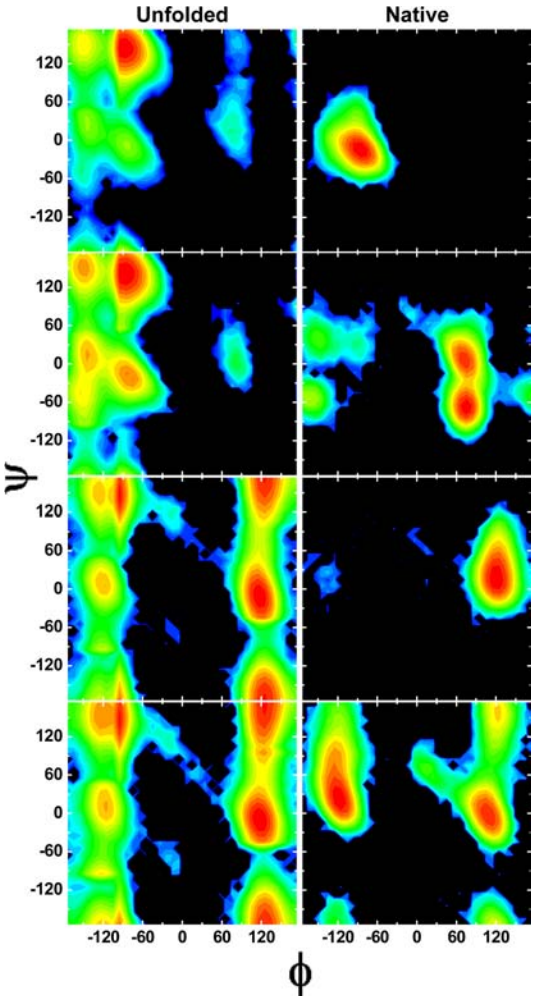


Figure 3. Ramachandran plots of alanine and glycine residues. Free energy landscapes in Ramachandran space are displayed for Ala-28 and Ala-46 (upper panels) and for Gly-35 and Gly-53 (lower panels) in both the denatured and native state ensembles. Data are taken from simulations using the OPLS/AA-L FF. The probability distributions are calculated using a pixel size of 10° × 10° and are converted to free energy distributions using $-RT \ln P$. The color scale ranges from red (ground state) to blue (6 kcal·mol^{-1}). Dihedral angles with free energies larger than 6 kcal·mol^{-1} are represented in black.

flexibility of $<(x + a)^2>$ and $<(x - a)^2>$ in the two separate wells, along with the $k \ln 2$ contribution to the entropy from the partitioning between the two wells. This example illustrates the need for separately treating the distribution between conformational basins and the thermal fluctuations within the individual basins as applied here for the evaluation of the conformational entropy between a disordered denatured state and a native state. Moreover, our treatment considers the specific dependence on amino acid, secondary structure, and neighbor dependence, features which are partly addressed in an average fashion by van Gunsteren et al.[13,14]

Another measure of residue-level changes in entropy has been provided by the Lipari-Szabo S^2 order parameter,[43,44] which probes backbone NH bond vector motion on the pico- to nanosecond time scale. Average changes in backbone entropy inferred using this method range from 0.8 to 1.6 kcal·mol^{-1}·residue^{-1},[4,13] a range overlapping some with our calculations. However, difficulties in calculating entropies from S^2, obtained either from experiment or simulations, arise in part because of the lack of a global reference frame for the denatured state. Furthermore, the NH vector distribution often is assumed to have azimuthal symmetry,[4] but the Ramachandran map lacks this symmetry. Also, individual NH bond vector motions on the nsec time scale probably are poor proxies of the backbone conformational entropy and do not account for correlated motions on any longer length scale. We demonstrate that correlations between neighboring residues are significant, but how these affect the conversion of S^2 values to entropies is unclear. Progress in this area will benefit from our analysis of entropies.

Another method uses data from experiments involving pulling measurements of unfolded polyproteins.[16] The work required to stretch the chain is 1.4 ± 0.1 kcal·mol^{-1}·residue^{-1}, which implicitly includes contributions from correlated motions and neighbor effects as the entire chain is extended. To obtain a value for the loss of conformational entropy upon folding, the backbone entropy of a fully extended chain is assumed to be the same as for the native state.

Our calculation for $T\Delta S$ suggest that the fully extended chain has ~1.4- to 3-fold fewer states, implying that the work required to fully extend a polypeptide exceeds the backbone entropy lost during the folding of the protein. This difference may be explained by the stretched chain only having conformations with both dihedral angles near ±180°, while a native protein may sample a larger region of the Ramachandran map.

Best and Hummer have modified FFs to improve agreement with experimental helix−coil measurements and to calculate the total change in enthalpy and entropy.[45] They also calculate the loss of backbone entropy in a manner similar to a restricted form of our calculation. Their computed entropy loss of 0.4−0.5 kcal·mol^{-1} is lower than ours because their treatment only considers population shifts from within the helical basin to the region specific for authentic helical structure; their calculation focuses on the entropy change upon formation of helical hydrogen bonds when starting from a near-helical geometry, rather than the total loss of entropy upon folding from an initial unfolded state where all basins are well populated.

Applications of landscape theory to simulations of protein folding use a value for the total conformational entropy in the range of $T\Delta S \sim 0.3{-}1$ kcal·mol^{-1}·residue^{-1},[46,47] consistent with our value for the backbone entropy.

Ala → Gly Substitutions. Ala → Gly entropy differences have served as the benchmark for calculations of entropies and helical propensities. Alanines exhibit much higher helix propensity than glycines, $\Delta\Delta G^{\text{helix}}{}_{\text{A}\to\text{G}} = 0.7{-}1$ kcal·mol^{-1}.[3,48] This difference generally has been attributed to the greater conformational entropy in the denatured state of glycine. Our calculation for an A28G substitution in Ub's major α helix is consistent with this view. The difference between the backbone entropy in the denatured state and native state is $\Delta(T\Delta S^{\text{BB}}{}_{\text{U-N}})_{\text{A28G}} = 0.6 \pm 0.1$ and 0.8 ± 0.1 kcal·mol^{-1} in the OPLS/AA-L and G-S A94 FF, respectively, mostly due to changes in the denatured state. In addition, the computed change in backbone entropy using OPLS/AA-L is quite similar to the experimental change in free energy, 0.52 ± 0.04 kcal·mol^{-1}.[49]

Other factors such as solvation or enthalphic effects can contribute to the decrease in helical propensity beyond an increase in the loss of conformational entropy; Jha et al. find that the helical propensities for different amino acids are well explained by the relative probability of being in the helical basin

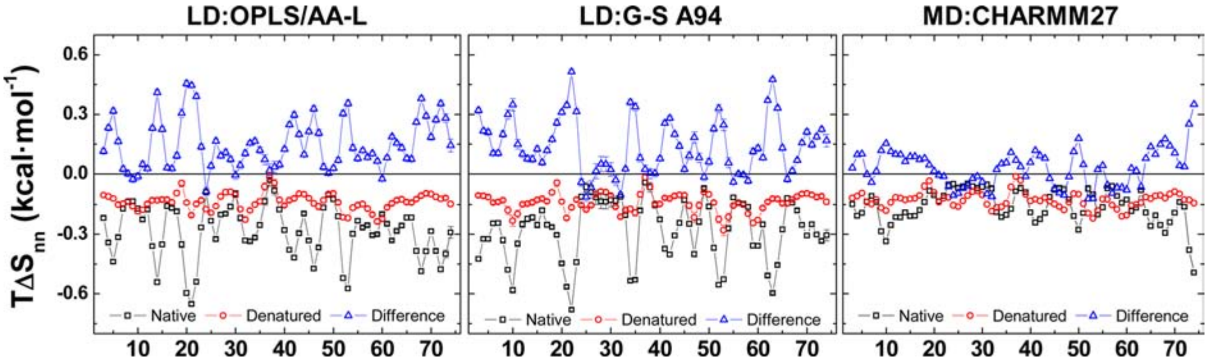


Figure 4. Nearest neighbor contributions to the backbone entropy. The contributions of conformational correlations between nearest neighbors to the backbone entropy are displayed for both the folded and denatured states and for both the OPLS/AA-L and G-S A94 FFs. The contributions are larger in magnitude in the native state ensemble than in the DSE ($T\Delta S^{nn}$ = −0.3 ± 0.1 and −0.2 ± 0.1 kcal·mol^{-1}, respectively). The turn regions between the β1−β2 hairpin and α-helix and the β4−β5 hairpin yield the greatest contributions in the native state, but pronounced contributions occur along other regions of the protein as well. The largest contributions in the denatured state are associated with glycine residues and their nearest neighbors. The OPLS/AA-L FF yields a slightly larger contribution to glycines and preglycine residues ($T\Delta S^{nn}$ = −0.22 ± 0.03 kcal·mol^{-1}), whereas the average for all other residues is $T\Delta S^{nn}$ = −0.17 ± 0.05 kcal·mol^{-1}. However, the contributions in the denatured state are larger for the G-S A94 FF, i.e., the contributions for glycine residues exceed those for pre- and postglycine residues and all other residues ($T\Delta S^{nn}$ = −0.63 ± 0.08, −0.49 ± 0.04, −0.46 ± 0.13, −0.33 ± 0.08 kcal·mol^{-1}, respectively).

in the PDB-based coil library (a similar result holds for β sheet propensities as well).[6] Given our entropy difference of 0.5 kcal·mol^{-1} for Ala → Gly substitutions, the experimental $\Delta\Delta G^{\text{helix}}{}_{A\rightarrow G}$ of 0.7−1.0 kcal·mol^{-1} suggests the presence of a significant enthalphic contribution.

Our $\Delta(T\Delta S^{BB}{}_{U-N})_{A28G}$ exceeds the average value $\Delta(T\Delta S^{BB}{}_{U-N})_{A\rightarrow G}$ ~ 0.1 kcal mol^{-1} calculated by Daggett and co-workers in their Dynameomics project (their change in the denatured state entropy is slightly larger $\Delta(TS^{BB}{}_{U})_{A\rightarrow G}$ ~ 0.4 kcal mol^{-1}).[20] However, their denatured state Ramachandran distribution for Ala, and Gly to a lesser extent, is heavily dominated by helical conformations. In contrast, our distribution is dominated by extended β and polyproline II conformers whose preponderance is necessary to recapitulate experimental RDCs.[21]

Implications. The values presented here for Ub should apply equally to other proteins because native proteins possess similar motions and the DSE is primarily determined by local sequence effects. The total loss of backbone entropy for a given protein can be calculated as the sum of the loss for the individual residues by accounting for the influence of secondary structure content (helical residues lose 0.2−0.5 kcal mol^{-1} more than sheet and coil residues, depending upon FF), and the sequence (e.g., the structured residues in Ub 1−74 include 14 α or 3_{10} helical residues, 3 prolines (two are consecutive), 2 preprolines and 4 glycines, see Table 1, Supporting Information Tables 1 and 2 for numerical values). We believe that delineating according to secondary structure type rather than amino acid type is sufficiently adequate for estimating the total entropy loss because the dispersions tend to be tighter when averaging over secondary structure rather than amino acid type. The data in Table 1 contain the effects of correlated motions. Hence, their sum provides a good estimate of the total entropy loss. For a protein with an unknown structure, the entropy loss can be calculated using the predicted secondary structure content and our values for the average loss for helical, strand and coil residues.

The loss in backbone entropy for helical residues can account for the total free energy penalty for initiating a helix. The formation of four helical residues costs 3.6−6 kcal·mol^{-1} in backbone entropy, depending on the FF. The lower value equates to an equilibrium constant K_{eq} = 0.002, which is similar to reported values for σ, the Zimm−Bragg helix−coil nucleation parameter.[50,51] Therefore, it may be unnecessary to invoke other energetic effects, such as hydrophobic burial, to account for helix initiation.

The energy surface for the early stages of folding is dominated by the entropic penalty associated with forming contacts (loop closure entropy). Our lower value for $T\Delta S^{BB}$ implies that this penalty is reduced. Hence, the energies associated with forming a long-range contact compete against a smaller entropic penalty, and the free energy surface is flatter at the beginning of the folding process.

■ CONCLUSIONS

We have calculated the loss of backbone conformational entropy upon folding using realistic ensembles for the denatured and native states, accounting for amino acid type and secondary structure as well as correlated motions. Because of these correlations and the PDB-based sampling, our denatured state ensemble contains less conformational diversity than most other representations. As a result of this and other factors, our calculated loss of backbone entropy is as much as 2-fold smaller than the commonly reported value for $T\Delta S^{BB}$.

Our entropy loss varies from 0.7 to 1.2 kcal mol^{-1} residue^{-1} and depends primarily on the FF rather than the solvent model. The variance is mostly attributed to differences in native state dynamics. Although this variance appears minor, the cumulative sum for an entire protein is appreciable. This issue greatly affects thermodynamic calculations and, thus, should be considered during FF parametrization.

We find that the decrease in the number of states upon folding, $f = \Omega_U/\Omega_N$ = 3−7, is close to the number of Ramachandran basins (β, α_R, PPII, α_L and ε). This similarity suggests that folding can be grossly approximated as the reduction in the number of basins sampled. This approximation requires that the intrabasin dynamics in both the native and denatured states be similar, with small-scale motions being largely governed by local properties rather than tertiary packing. We find that this assumption is more accurate for residues in β sheet and loops than in helices where the dihedral angles are restricted to a tighter region in the Ramachandran map.

The differences between our and prior studies have other implications, such as the balance of forces in protein folding. The experimentally determined change in total entropy for folding often is near zero, indicating that the loss of conformational entropy is nearly offset by an equal gain in solvent entropy.[52] Therefore, our revised value reduces the estimated gain in solvent entropy by as much as 1 kcal·mol^{-1}·residue^{-1}, because less compensation by solvent entropy is required to account for the loss of backbone entropy. A more complete estimate of the correction requires an analysis of side chain entropy losses, which is in progress.

METHODS

Denatured State Ensemble. An initial ensemble of 13 000 denatured state structures is generated from a coil library of (ϕ,ψ) dihedral angles derived from the PDB for residues in irregular, non-hydrogen bonded conformations.[21] Dihedral angles are selected contingent on both the flanking residues' chemical identity and conformation. To avoid steric overlap, the initially selected angles are "nudged" by minimizing a simple repulsive excluded volume potential. This DSE provides the proper statistics for the distributions of each residue among the five major Ramachandran basins. Using each of two different FFs and saving structures every 1 ps after the first 100 ps, short (300 ps) all-atom intrabasin LD trajectories (described below) are run at 298 K for a randomly chosen subset of 3000 structures to obtain adequate intrabasin sampling for evaluating the backbone conformational entropy. Each of the two ensembles provides 6×10^5 structures.

Native State Ensemble. Ten 28 ns LD trajectories at 298 K are run starting from the energy minimized crystal structure (1UBQ),[53] and structures after the first 10 ns are saved every 1 ps (providing a total of 1.8×10^5 structures). The A28G native state ensemble is calculated from a shorter (10 ns) set of trajectories where structures are retained after the first 1 ns.

Langevin Dynamics Calculations. All-atom dynamic calculations use an enhanced version of the TINKER v3.9 package[54] that has been modified to increase computational efficiency[55] and add various functionality. The simulations utilize an implicit solvent model[56] with a nonlinear distance-dependent electrical permittivity for the calculation of electrostatic interactions.[57] Solute−solvent interactions are described by the Ooi-Scheraga solvent accessible surface area potentials,[58] while the atomic friction coefficients are computed with the Pastor-Karplus scheme.[59]

Initial structures are energy minimized using a limited memory BFGS quasi-Newton nonlinear optimization routine,[60,61] with the dihedral angles restrained using a harmonic potential (k = 1 kcal·mol^{-1}·deg^{-2}). Following energy minimization, the structure is heated from 150 to 298 K by increasing the temperature 10 K every 10 ps with a time step of 1 fs. While raising the temperature, the backbone atomic positions are held fixed with a harmonic potential (k = 10 kcal·mol^{-1}·Å^{-2}) that is successively reduced once the target temperature is reached. The denatured state simulations likewise restrain the dihedral angles during the preparation run (k = 1 kcal·mol^{-1}·deg^{-2}) of total duration 210 ps.

The OPLS/AA-L[7,8] and G-S A94[9] FFs are utilized for calculating atomic interactions within the protein to investigate the robustness of the entropy calculations. The denatured state trajectories are generated using a FF with van der Waals interactions other than those between residues $i,i \pm 1$ replaced by the purely repulsive Weeks-Chandler-Andersen truncation[62] of the Lennard-Jones (LJ) potential, that is,

$$u_0(r) = \begin{cases} u(r) + \varepsilon & r < 2^{1/6}\sigma \\ 0 & r < 2^{1/6}\sigma \end{cases} \qquad (1)$$

where ε and $2^{1/6}\sigma$ are the minimum energy and corresponding critical distance of the LJ potential. Furthermore, electrostatic interactions are ignored other than those between residues $i,i \pm 1$. These energy modifications produce a DSE having the global statistics of chains in good solvents, as deduced from scattering experiments,[63] and thus cannot fold.

Additionally, residues are constrained to remain in their initial Ramachandran basins during the LD simulations to maintain the correct basin statistics inherent in the initial DSE generated from the coil library. This intrabasin restriction is imposed by applying a reflecting harmonic restraining potential (k = 1 kcal·mol^{-1}·deg^{-2}) if the residue's ϕ or ψ angle attempts to cross a basin boundary. The basin definitions are the same as those used in constructing the coil library.[6,21]

Molecular Dynamics Calculations. Molecular dynamics simulations are carried out with the NAMD package[64] for both the native structure and for a representative set of the DSE using the CHARMM27 FF with the TIP3P water model.[10−12]

Plots and data analysis are carried out using Origin (OriginLab, Northampton, MA).

Abbreviations. FF, force field; G-S A94, Garcia and Sanbonmatsu's modified version of Amber 94; R_g, radius of gyration; RDC, residual dipolar couplings; LD, Langevin dynamics; MD, molecular dynamics; ΔS, change in entropy; Ub, ubiquitin; BB, backbone.

ASSOCIATED CONTENT

Supporting Information

Additional figures and methods as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

trsosnic@uchicago.edu; freed@uchicago.edu

Present Address

[#]Agios, 38 Sidney Street, 2nd Floor, Cambridge, MA 02139-4169.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by NIH research grants and NSF Grant CHE-1111918. We thank K. Plaxco, A. Garcia, D. Case, A. Szabo, A. Palmer, R. Bruschweiler, N. Trbovic, G. Makhatadze, J. Wand, R. S. Berry and members of our group for comments and discussions. NAMD was developed by the Theoretical and Computational Biophysics Group in the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana−Champaign. http://www.ks.uiuc.edu/Research/namd/

REFERENCES

(1) Stites, W. E.; Pranata, J. *Proteins* **1995**, *22*, 132−140.

(2) Meirovitch, H. *Curr. Opin. Struct. Biol.* **2007**, *17*, 181−186.

(3) D'Aquino, J. A.; Gomez, J.; Hilser, V. J.; Lee, K. H.; Amzel, L. M.; Freire, E. *Proteins* **1996**, *25*, 143−156.

(4) Yang, D.; Kay, L. E. *J. Mol. Biol.* **1996**, *263*, 369−382.

(5) Zaman, M. H.; Shen, M. Y.; Berry, R. S.; Freed, K. F.; Sosnick, T. R. *J. Mol. Biol.* **2003**, *331*, 693−711.

(6) Jha, A. K.; Colubri, A.; Zaman, M. H.; Koide, S.; Sosnick, T. R.; Freed, K. F. *Biochemistry* **2005**, *44*, 9691−9702.

(7) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *Abstr. Pap.—Am. Chem. Soc.* **2000**, *220*, U279.

(8) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474−6487.

(9) Garcia, A. E.; Sanbonmatsu, K. Y. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 2782−7.

(10) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Comput. Chem.* **1983**, *79*, 926−935.

(11) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.;

Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.

(12) Mackerell, A. D., Jr.; Feig, M.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1400−1415.

(13) Schafer, H.; Daura, X.; Mark, A. E.; van Gunsteren, W. F. *Proteins* **2001**, *43*, 45−56.

(14) Peter, C.; Oostenbrink, C.; van Dorp, A.; van Gunsteren, W. F. *J. Chem. Phys.* **2004**, *120*, 2652−2661.

(15) Alexandrescu, A. T.; Rathgeb-Szabo, K.; Rumpel, K.; Jahnke, W.; Schulthess, T.; Kammerer, R. A. *Protein Sci.* **1998**, *7*, 389−402.

(16) Thompson, J. B.; Hansma, H. G.; Hansma, P. K.; Plaxco, K. W. *J. Mol. Biol.* **2002**, *322*, 645−652.

(17) Yang, A. S.; Honig, B. *J. Mol. Biol.* **1995**, *252*, 351−365.

(18) Nemethy, G.; Scheraga, H. *Biopolymers* **1965**, *3*, 155.

(19) Wang, J.; Szewczuk, Z.; Yue, S. Y.; Tsuda, Y.; Konishi, Y.; Purisima, E. O. *J. Mol. Biol.* **1995**, *253*, 473−492.

(20) Scott, K. A.; Alonso, D. O.; Sato, S.; Fersht, A. R.; Daggett, V. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 2661−2666.

(21) Jha, A. K.; Colubri, A.; Freed, K. F.; Sosnick, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13099−13104.

(22) Perico, A.; Pratolongo, R.; Freed, K. F.; Pastor, R. W.; Szabo, A. *J. Chem. Phys.* **1993**, *98*, 564−573.

(23) Lindorff-Larsen, K.; Trbovic, N.; Maragakis, P.; Piana, S.; Shaw, D. E. *J. Am. Chem. Soc.* **2012**, *134*, 3787−3791.

(24) Guo, Z.; Brooks, C. L.,, 3rd; Boczko, E. M. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 10161−10166.

(25) Kim, S. Y.; Lee, J.; Lee, J. *Biophys. Chem.* **2005**, *115*, 195−200.

(26) Kim, S. Y.; Lee, J.; Lee, J. *J. Chem. Phys.* **2004**, *120*, 8271−8276.

(27) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, *132*, 1526−1528.

(28) Jacob, J.; Krantz, B.; Dothager, R. S.; Thiyagarajan, P.; Sosnick, T. R. *J. Mol. Biol.* **2004**, *338*, 369−382.

(29) Plaxco, K. W.; Millett, I. S.; Segel, D. J.; Doniach, S.; Baker, D. *Nat. Struct. Biol.* **1999**, *6*, 554−556.

(30) Jacob, J.; Dothager, R. S.; Thiyagarajan, P.; Sosnick, T. R. *J. Mol. Biol.* **2007**, *367*, 609−615.

(31) Yoo, T. Y.; Meisburger, S. P.; Hinshaw, J.; Pollack, L.; Haran, G.; Sosnick, T. R.; Plaxco, K. *J. Mol. Biol.* **2012**, *418*, 226−236.

(32) Muller-Spath, S.; Soranno, A.; Hirschfeld, V.; Hofmann, H.; Ruegger, S.; Reymond, L.; Nettels, D.; Schuler, B. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 14609−14614.

(33) Schuler, B.; Lipman, E. A.; Eaton, W. A. *Nature* **2002**, *419*, 743−747.

(34) Sherman, E.; Haran, G. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 11539−11543.

(35) Cheung, M. S.; Garcia, A. E.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 685−690.

(36) Fitzgerald, J. E.; Jha, A. K.; Sosnick, T. R.; Freed, K. F. *Biochemistry* **2007**, *46*, 669−682.

(37) Zaman, M. H.; Shen, M. Y.; Berry, R. S.; Freed, K. F. *J. Phys. Chem. B* **2003**, *107*, 1685−1691.

(38) Freddolino, P. L.; Park, S.; Roux, B.; Schulten, K. *Biophys. J.* **2009**, *96*, 3772−3780.

(39) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. *Biophys. J.* **2011**, *100*, L47−149.

(40) Beauchamp, K. A.; Lin, Y. S.; Das, R.; Pande, V. S. *J. Chem. Theory Comput.* **2012**, *8*, 1409−1414.

(41) Fitzkee, N. C.; Rose, G. D. *J. Mol. Biol.* **2005**, *353*, 873−887.

(42) Perico, A.; Pratolongo, R.; Freed, K. F.; Szabo, A. *J. Chem. Phys.* **1994**, *101*, 2554−2561.

(43) Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, *104*, 4546−4559.

(44) Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, *104*, 4559−4570.

(45) Best, R. B.; Hummer, G. *J. Phys. Chem. B* **2009**, *113*, 9004−9015.

(46) Garcia, A. E.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13898−13903.

(47) Itoh, K.; Sasai, M. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 7298−7303.

(48) Creamer, T. P.; Rose, G. D. *Proteins* **1994**, *19*, 85−97.

(49) Went, H. M.; Jackson, S. E. *Protein Eng., Des. Sel.* **2005**, *18*, 229−237.

(50) Mayne, L.; Englander, S.; Qiu, R.; Yang, J.; Gong, Y.; Spek, E.; Kallenbach, N. *J. Am. Chem. Soc.* **1998**, *120*, 10643−10645.

(51) Yang, J.; Zhao, K.; Gong, Y.; Vologodskii, A.; Kallenbach, N. *J. Am. Chem. Soc.* **1998**, *120*, 10646−10652.

(52) Makhatadze, G. I.; Privalov, P. L. *Protein Sci.* **1996**, *5*, 507−510.

(53) Vijay-Kumar, S.; Bugg, C. E.; Wilkinson, K. D.; Vierstra, R. D.; Hatfield, P. M.; Cook, W. J. *J. Biol. Chem.* **1987**, *262*, 6396−6399.

(54) Ponder, J. W. R., S.; Kundrot, C.; Huston, S.; Dudek, M.; Kong, Y.;Hart, R.; Hodson, M.; Pappu, R.; Mooiji, W.; Loeffler, G.; 3.7 ed.; Washington University: St. Louis, MO, 1999.

(55) Shen, M. Y.; Freed, K. F. *J. Comput. Chem.* **2005**, *26*, 691−698.

(56) Shen, M. Y.; Freed, K. F. *Biophys. J.* **2002**, *82*, 1791−1808.

(57) Jha, A. K.; Freed, K. F. *J. Chem. Phys.* **2008**, *128*, 034501.

(58) Ooi, T.; Oobatake, M.; Nemethy, G.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 3086−3090.

(59) Pastor, R. W.; Karplus, M. *J. Phys. Chem.* **1988**, *92*, 2636−2641.

(60) Liu, D.; Nocedal, J. *Math. Prog.* **1989**, *45*, 503−528.

(61) Nocedal, J. *Math. Comp.* **1980**, *35*, 773−782.

(62) Weeks, J.; Chandler, D.; Andersen, H. *J. Chem. Phys.* **1971**, *54*, 5237.

(63) Kohn, J. E.; Millett, I. S.; Jacob, J.; Zagrovic, B.; Dillon, T. M.; Cingel, N.; Dothager, R. S.; Seifert, S.; Thiyagarajan, P.; Sosnick, T. R.; Hasan, M. Z.; Pande, V. S.; Ruczinski, I.; Doniach, S.; Plaxco, K. W. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 12491−12496.

(64) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781−1802.