Statistical coil model of the unfolded state: Resolving the reconciliation problem

Abhishek K. Jha[†‡§], Andrés Colubri[†‡], Karl F. Freed[†§¶], and Tobin R. Sosnick[‡¶‖]

[†]Department of Chemistry, [‡]Institute for Biophysical Dynamics, [§]The James Franck Institute, and [‖]Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, IL 60637

Communicated by R. Stephen Berry, University of Chicago, Chicago, IL, July 27, 2005 (received for review May 13, 2005)

An unfolded state ensemble is generated by using a self-avoiding statistical coil model that is based on backbone conformational frequencies in a coil library, a subset of the Protein Data Bank. The model reproduces two apparently contradicting behaviors observed in the chemically denatured state for a variety of proteins, random coil scaling of the radius of gyration and the presence of significant amounts of local backbone structure (NMR residual dipolar couplings). The most stretched members of our unfolded ensemble dominate the residual dipolar coupling signal, whereas the uniformity of the sign of the couplings follows from the preponderance of polyproline II and β conformers in the coil library. Agreement with the NMR data substantially improves when the backbone conformational preferences include correlations arising from the chemical and conformational identity of neighboring residues. Although the unfolded ensembles match the experimental observables, they do not display evidence of native-like topology. By providing an accurate representation of the unfolded state, our statistical coil model can be used to improve thermodynamic and kinetic modeling of protein folding.

denatured state | protein folding | residual dipolar coupling | nearest neighbor | radius of gyration

Denatured proteins are the initial state for mechanistic and thermodynamic studies of folding. The recent resurgence of interest in the unfolded state is partly motivated by the development of NMR methods that are capable of providing site-resolved structural information (1–7). These measurements indicate that unfolded proteins have far richer structural diversity than earlier believed, possibly even encoding the native topology (1, 8–11).

These works seem at odds with the classic studies by Tanford *et al.* (12, 13), who demonstrated by using hydrodynamic methods that the global dimensions of denatured proteins exhibit the size dependence expected for self-avoiding "random coil" polymers. More recent measurements of the radius of gyration, R_g, using small-angle scattering methods exhibit the same random coil scaling behavior with length $R_g \propto N^{0.585}$ (8, 14). These observations are consistent with denatured proteins being random coils in good solvent conditions (15). This finding leads to the so-called "reconciliation problem" between the random coil scaling behavior and the presence of significant amounts of local structure in unfolded state (14, 16).

However, Rose and Fitzkee (17) demonstrate that even a "deliberately extreme" model of chains composed of native-like segments connected by flexible residues also can reproduce random coil scaling behavior. Hence, the recapitulation of the scaling behavior provides only a weak test for any unfolded state model. Nevertheless, spectroscopic measurements, such as circular dichroism, indicate that most unfolded states, particularly chemical-denatured proteins (8, 13, 18), have little secondary structure. Accordingly, the unrealistic native-like segment model is ruled out. More exacting tests are needed, particularly those that involve site-resolved information.

NMR measurements of $^3J_{HN\alpha}$ coupling constants reflect individual residue's backbone ϕ dihedral angles and also largely support the view that proteins are statistical coils (19–23). However, α and polyproline II (PPII) conformations have similar ϕ values (24), and nearly as good agreement with experiment can be obtained by using conformational preferences based on the entire Protein Data Bank (PDB), which is dominated by α conformers, than by using ones based on unstructured regions in a coil library (25), which is dominated by PPII and β conformations (data not shown). Hence, these measurements likewise do not yield a stringent test for statistical coil behavior.

NMR residual dipolar coupling constants (RDCs) provide a powerful site-resolved tool for probing the structure of denatured proteins (1, 10, 11, 26–30). RDCs probe the orientation of bond vectors (generally backbone amide NH) relative to an alignment tensor fixed in the molecular frame. Although RDCs generally vanish when the molecules freely tumble, the RDCs no longer vanish when proteins are confined in weakly aligning media, such as compressed acrylamide gels (1–7, 31).

Recently, the molecular origins of RDCs in denatured proteins have attracted considerable attention (29, 32). Shortle and coworkers (1, 26–28) argue that denatured proteins retain some native-like topology in the denatured state. However, others propose simpler explanations related to the fundamental nature of random coils in aligned media (33) or to short stretches of extended conformations (29, 34).

Here, we generate a statistical coil model for the unfolded state (19–23). The backbone conformations are determined by their frequencies in regions outside of, and not adjacent to, helices, sheets, and turns in high-resolution crystal structures (25). These frequencies generate a statistical potential that accurately reproduces the known helical, β-sheet, and PPII propensities (25). The statistical potential can be chosen to include correlations between adjacent residues to account for both residue type and conformation, as emphasized in our previous studies (25, 35) and other earlier studies of nearest neighbor (NN) effects (20, 23, 36–40). After assigning backbone dihedral angles according to the statistical potential, unfolded chain conformations are slightly "nudged" to satisfy excluded volume constraints. An ensemble of unfolded chains so generated and individually aligned is shown to predict the experimental RDCs in the chemically denatured state. This agreement requires a preponderance of extended backbone conformers in the library. The inclusion of NN effects enhances the agreement with experimental RDCs and R_g values. No support emerges for the existence of native-like topology in the denatured state. A web server (unfolded.uchicago.edu) is created to generate unfolded ensembles to serve as realistic starting points for folding simulations and as reference states for rigorous thermodynamic calculations.

Methods

Unfolded Ensembles. The protein sequence-culling server PISCES (41) is used to select x-ray structures of the 2,020 chains of >20

Abbreviations: apoMb, apomyoglobin; AR, aspect ratio; NN, nearest neighbor; PDB, Protein Data Bank; PPII, polyproline II; R_g, radius of gyration; RDC, residual dipolar coupling; Ub, ubiquitin.

¶To whom correspondence may be addressed. E-mail: k-freed@uchicago.edu or trsosnic@uchicago.edu.

residues with a maximum R factor of 0.3 and a resolution of better than 2 Å. A coil library is constructed from these proteins by retaining only internal residues within stretches of four or more residues that lie outside of helices, sheets, and turns (25). End effects are reduced by removing residues adjacent to structured regions. Unfolded conformations are built by initially assigning each residue to one of five Ramachandran basins (α_R, β, PPII, α_L, and γ; see Fig. 6, which is published as supporting information on the PNAS web site) based on their frequencies in the coil library.

A residue's basin frequencies depend on the identity and conformation of the neighboring residues. However, the size of the coil library is insufficient to consider the simultaneous influence of both of the neighbors' sequence and conformation. Hence, we adopt a strategy based on pairs (dimers) of residues. The monomer basin frequencies, $P(a_i, b_i)$, are converted into energy units by

$$U(a_i, b_i) = -RT \ln P(a_i, b_i), \quad [1]$$

where a_i is the identity of the ith amino acid that resides in the b_i Ramachandran basin. Similarly, the joint probability, called the dimer library, of finding two consecutive residues a_i and a_{i+1} in basins b_i and b_{i+1} gives

$$U(a_i, b_i) + U(a_{i+1}, b_{i+1}) + \delta U(a_i, b_i, a_{i+1}, b_{i+1}) = -RT \ln P(a_i, b_i, a_{i+1}, b_{i+1}). \quad [2]$$

Combining Eqs. **1** and **2** enables expressing the NN correlation energy term $\delta U(a_i, b_i, a_{i+1}, b_{i+1})$ in terms of probabilities derived from the coil library

$$\delta U(a_i, b_i, a_{i+1}, b_{i+1}) = -RT \ln \frac{P(a_i, b_i, a_{i+1}, b_{i+1})}{P(a_i, b_i)P(a_{i+1}, b_{i+1})}. \quad [3]$$

The local interactions that dominate the structure of the polypeptide chain can now be modeled by an energy function that includes first neighbor effects. An individual residue contributes $U(a_i, b_i)$, and an additional term $\delta U(a_i, b_i, a_{i+1}, b_{i+1})$ from each of the neighbors combines to give the total statistical potential for a polypeptide with N residues

$$U_{\text{total}} = \sum_i U(a_i, b_i) + \sum_i \delta U(a_i, b_i, a_{i+1}, b_{i+1}). \quad [4]$$

An equilibrium ensemble of peptide chains is generated from Monte Carlo simulations with this energy function. The elementary transitions consist of choosing a Ramachandran basin for randomly determined residues and then accepting or rejecting the transition according to the standard Monte Carlo criteria. Once the basins are assigned, the specific ϕ,ψ backbone angles within the basin are selected from occurrences in the coil library for each residue type, independent of NN effects.

To remove steric overlap, the ϕ,ψ angles are nudged within the basin of each amino acid by minimizing a simple excluded volume energy function for intrabasin relaxation with two terms. The first term is a hard sphere energy contribution for the main-chain heavy atoms (N, C^α, C, and O) and the C^β, side-chain β-carbons. This hard sphere interaction is the same as that used in the program SCWRL3 (42). The second term accounts for steric interactions between main-chain atoms and the remaining side chains and involves a soft-sphere interaction that depends on the side groups through two parameters, the location of the center of mass of the side group with respect to the main-chain atoms and a "virtual" soft-sphere radius that encompasses 90% of the total volume explored by the side-chain atoms when all allowed rotameric states are sampled. We also generate an ensemble by switching off the interaction term associated with the side-chain atoms.

The long-range energy function is minimized by using a simulated annealing algorithm. The elementary transition is a change in ϕ,ψ angles within the already assigned basin. These angles are chosen from the coil library described earlier, taking into account both the chemical and conformational identity of the given residue. Finally, side chains and hydrogen atoms are added to the chain backbone by using SCWRL (42) and REDUCE (43), respectively.

RDC. The RDC between the amide nitrogen and proton is given by

$$D_{\text{NH}} = D_{\max} \langle P_2(\cos\theta) \rangle, \quad [5]$$

where $P_2(x) = (3x^2 - 1)/2$. $D_{\max}$ contains details of the bond lengths and the magnitudes of the magnetic moments. The orientation dependence emerges through θ, the angle between the NH bond vector and the magnetic field. The proportionality factor $D_{\max}$ is constant for the NH bonds considered here, so we focus attention on the average of P_2.

Results and Discussion

We generate ensembles of unfolded structures for which the Ramachandran basin frequencies for each amino acid are assigned according to its identity, followed by the specification of the explicit ψ,ϕ angles obtained from the PDB. The basin assignment either uses or neglects information on the chemical and conformational identity of neighboring residues. A comparison of these ensembles for several protein sequences with experimental data provides insight both into the ability of these ensembles to recapitulate the experimental results as well as the importance of NN effects.

Radius of Gyration. Polypeptide dimensions often follow a simple scaling law for a long polymer with N monomers, $R_g = aN^b$, where the prefactor a depends on local structure, and the exponent b depends on solvent quality. The exponent b is $\approx 1/3$ for collapsed polymers (poor solvent conditions), ≈ 0.6 in good solvent conditions, and $\approx 1/2$ in ideal θ solvents. More recent and reliable experimental values are $a = 2.04 \pm 0.17$Å and $b = 0.59 \pm 0.01$, obtained by using small-angle scattering for a set of 28 unfolded proteins (8). Because our model incorporates excluded volume constraints, accurate estimates are expected for the exponent b, but the prefactor a is more sensitive to the details of local structure.

Unfolded ensembles of 10 proteins are generated separately with and without inclusion of NN effects (Fig. 1). Generally, the R_g values from our models are $\approx 10\%$ larger than the experimental values. When NN interactions are neglected, we obtain $a = 2.7 \pm 0.4$ Å and $b = 0.57 \pm 0.03$ (data not shown). The agreement with experiment improves slightly when NN interactions are included ($a = 1.9 \pm 0.3$ Å and $b = 0.64 \pm 0.03$). When side-chain interactions are ignored in the simulated annealing step, more compact conformations are found, with $a = 2.35 \pm 0.3$ Å and $b = 0.58 \pm 0.02$. The exponent b is nearly the same in all of the cases, as anticipated for any self-avoiding random-walk model at large enough length scales.

Comparison with Experimental RDC Patterns. Considered individually, random walks are anisotropic even though the ensemble average is isotropic (44). The principal axes for each unfolded conformation are determined by diagonalizing the moment of inertia tensor. For simplicity, the degree of alignment of all molecules is assumed to be the same, although it may in reality be higher for configurations with higher aspect ratios (ARs). This hypothesis is tested below by calculating RDCs with all of the conformations weighted by varying powers of their ARs, AR^n. The major axis is aligned parallel or perpendicular to the direction of stretched or compressed gel channels, respectively. The orientation of the NH bond vector is evaluated relative to the magnetic field.

In practice, evaluating the RDCs for any residue involves two averages. The first average is over all orientations of the mo-

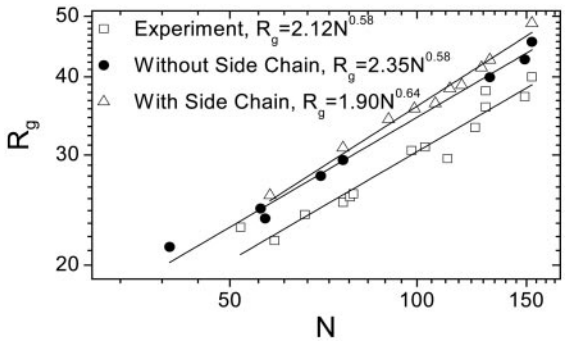

Fig. 1. Global conformation. The radius of gyration (R_g) vs. the number of residues (N) for 10 proteins is computed for an unfolded ensemble generated with and without NN effects. The data in both cases are fit to $R_g = aN^b$. Values are shown for the following: Ub (76 aa, PDB ID code 1UBQ, 16,000 structures), apoMb (153 aa, PDB ID code 1BVC, 5,000 structures), eglin C (70aa, PDB ID code 1CSE, 5,000 structures), and the staphylococcal nuclease (Snase, PDB ID code 1SNQ, 5,000 structures) fragment Δ131Δ, composed of residues 10–140, Calcyclin (Rabbit, Ca^{2+}, PDB ID code 1a03, chain A, 90 aa, 1,000 structures), Ig light and heavy chains (PDB ID code 43c9, chains C and B, 113 and 118 aa, 1,000 structures of each), ileal lipid-binding protein (PDB ID code 1eio, chain A, 127, 1,000 structures), bovine pancreatic trypsin inhibitor (PDB ID code 1BPI, 58 aa, 1,000 structures), and B*5301(PDB ID code 1a1o, chain B, 99 aa, 1,000 structures). Dividing the ensembles in half yields essentially identical results.

lecular axis with respect to the magnetic field for each conformation, whereas the second average is over the ensemble of different chain conformations. The first average yields the value of P_2 for a perfectly aligned chain, multiplied by a factor Δ that describes the average degree of departure from perfect alignment of the principal axis with respect to the magnetic field for each individual chain conformation. Because Δ depends on the nature of the aligning medium, which is poorly understood, we assume Δ to be a constant overall scale factor that is chosen empirically in comparisons between calculated and experimental RDCs. For the proteins investigated, we obtain $100 \leq D_{\max}\Delta \leq 800$, and Δ increases with protein length, suggesting a common alignment mechanism.

We provide a detailed analysis for apomyoglobin (apoMb), ubiquitin (Ub), the Snase fragment Δ131Δ, and eglin C, whose RDCs have been measured in weakly aligning media under chemically denaturing conditions (1, 10, 29). Although computations have been performed for magnetic fields parallel and perpendicular to the principal molecular axis, the results for Ub and eglin C are only presented for perpendicular alignment, whereas apoMb and Δ131Δ are described by using parallel alignment, corresponding to the experimental conditions of stretched or compressed gels, respectively. Geometrically, upon averaging over all possible molecular rotations about the major axis, the RDCs for a parallel molecular alignment are $-1/2$ times those for a perpendicular alignment, as observed in apoMb (29).

The computed RDCs for the amide NH vectors of the four proteins are almost always of a single sign (Figs. 2 and 3). Particularly for apoMb and Ub, the computed variation agrees with the irregular pattern observed experimentally. In contrast, the D_{NH} values for an idealized, uniform random-flight polypeptide are similar for all residues, except for a decreased correlation with the molecular axis for residues closer to the ends (33), which produces a V-shaped curve (Fig. 2*C*). The calculated and measured RDC patterns of real proteins, however, deviate considerably from this smooth, idealized curve, exhibiting significant residue-to-residue variation. This variation indicates the existence of local structural correlations in unfolded proteins that an idealized random-flight polypeptide fails to capture.

The ensemble generated by using our coil library captures the local interactions remarkably well as it successfully mimics the relative residue-to-residue variation for apoMb and Ub and, to a lesser extent, elgin C and Δ131Δ (Figs. 2 and 3; the scatter plot of calculated and observed RDC values is provided in Fig. 7*A*, which is published as supporting information on the PNAS web site). In addition to providing the Pearson coefficient for the scatter plots, we provide a measure that reflects correlations in the alignment of neighboring residues with respect to the global alignment. This measure is the dot product of the calculated and observed RDC patterns

$$S = \frac{\sum_i \mathrm{RDC}_{\mathrm{Calc},i} RDC_{\mathrm{Exp},i}}{\left(\sum_i \mathrm{RDC}_{\mathrm{Calc},i}{}^2 \sum_i \mathrm{RDC}_{\mathrm{Exp},i}{}^2\right)^{1/2}}, \qquad [6]$$

where i is the residue index. A quantitative estimate of the correlations between the calculated and experimental data are obtained by comparing S values with the mean and standard deviation (SD) for a randomized distribution obtained by independently shuffling the order for both the observed and calculated RDCs. The observed S values for apoMb, Ub, elgin, and Δ131Δ are 8, 5, 2, and 1.5 SDs above the mean of their respective random distributions. It is unclear why apoMb exhibits the best agreement, although differences in the acrylamide percentage are known to alter RDC

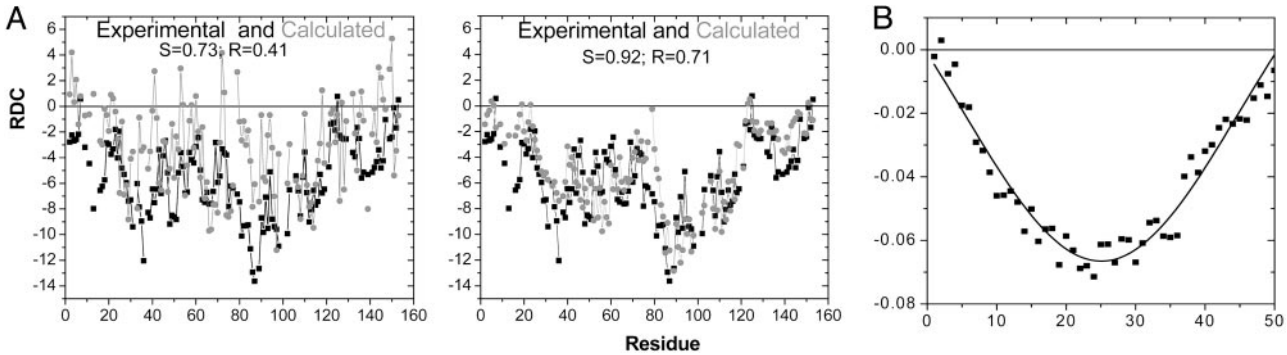

Fig. 2. Local conformation. (*A*) Calculated (gray) and experimental (black) RDCs for apoMb in 10% acrylamide (29). The calculated RDCs are obtained by averaging 5,000 structures generated by using a PDB-based statistical potential derived from the coil library and excluded volume constraints. RDCs are calculated without (*Left*) and with (*Right*) NN effects on backbone conformations. RDCs are presented for all of the residues except prolines and those for which there are no NH resonances. Scale factor is defined as $D_{\max}\Delta$. S is defined as dot product of the two RDC patterns (Eq. **6**). (*B*) RDCs for an idealized random ensemble generated without NN effects for ALA_{50} polypeptide [with equal probabilities (1/3) for the three major basins].

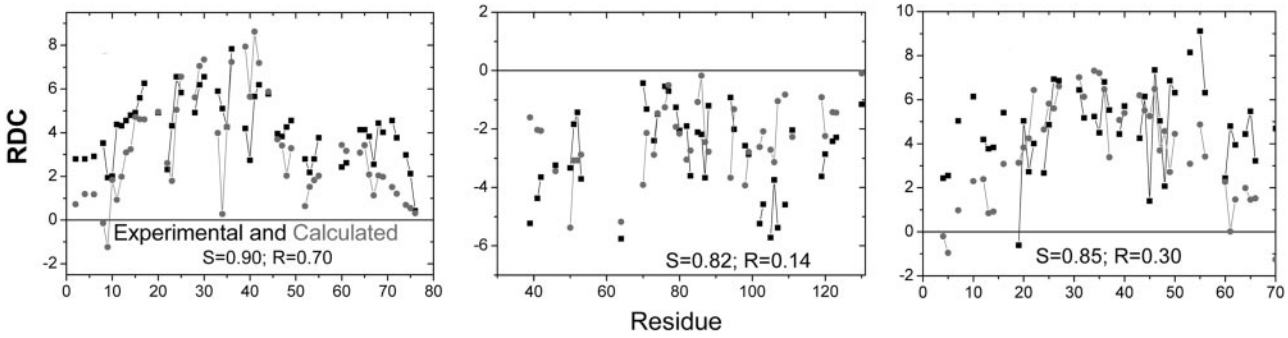

Fig. 3. Comparison of calculated (gray) and experimental (black) RDCs for ensembles with NN correlations for chemically denatured Ub (*Left*), eglin C (12% acrylamide; *Right*) (10), and Δ131Δ (12% acrylamide; *Center*) (1).

patterns (28), suggesting that the experimental conditions of apoMb best match our model.

The agreement between the calculated and observed RDC pattern gets poorer when NN effects are neglected for apoMb ($R = 0.71 \rightarrow 0.41, S = 0.92 \rightarrow 0.73$; Figs. 2 and 7*B*) and the other proteins (data not shown). This result stresses the importance of correlations in the unfolded state associated with the chemical and conformational identities of neighboring residues.

The preference for extended conformations and the global alignment along the long axis of the chain results in nearly all of the RDCs having the same sign (Figs. 2 and 3). The sign depends only on whether the gel is compressed or stretched. Backbone NH vectors of residues in the β and PPII conformations are oriented nearly perpendicular to the long axis, whereas those in the α conformation are oriented nearly parallel. The magnitude of the RDC is proportional to $P_2(\cos\theta)$, which reflects the orientation of the bond vector relative to the applied magnetic field, assumed for now to be parallel to the long molecular axis. In such geometries, the extended and α conformations have $P_2(\cos\theta) \sim -1/2$ and 1, respectively. Accordingly, the average value of the angular term can be roughly approximated by using their relative probabilities, $P_2(\cos\theta) \sim -1/2(\langle P_\beta + P_{\mathrm{PPII}}\rangle) + (\langle P_\alpha\rangle)$. Because the frequencies of the two extended conformations generally exceed those for helical conformers by >2:1 for nearly all residues in our coil library ($\langle P_\beta\rangle$:$\langle P_{\mathrm{PPII}}\rangle$:$\langle P_\alpha\rangle$ = 33:36:27), RDCs are mostly negative in this geometry. Conversely, in compressed gels where the alignment is perpendicular to the principal axis, nearly all RDCs are positive, as observed and computed.

Different Libraries. To test the robustness of our model, we compute the RDCs for apoMb by using four additional libraries with very different conformational frequencies. Agreement with experiment worsens when the ensemble is based on entire PDB library, which is dominated by α conformers (Fig. 4*A*; $\langle P_\beta\rangle$:$\langle P_{\mathrm{PPII}}\rangle$:$\langle P_\alpha\rangle$ = 27:18:53). Even the signs of the computed RDCs no longer agree because of the dominance of α conformers. We also computed the RDCs from a library of all residues outside helices (Fig. 4*B*; $\langle P_\beta\rangle$:$\langle P_{\mathrm{PPII}}\rangle$:$\langle P_\alpha\rangle$ = 43:27:26). Although the relative frequencies of β and PPII conformers have nearly inverted, their sum is similar to that in our coil library. As a result, the correlation between computed and measured RDCs is similar to the original correlation. The correlation degrades slightly if turns are also included (Fig. 4*C*; $\langle P_\beta\rangle$:$\langle P_{\mathrm{PPII}}\rangle$:$\langle P_\alpha\rangle$ = 23:33:37) because of higher α frequencies. The agreement improves for the fourth library that comprises residues outside helices, sheets, and turns but includes the biased terminal residues of the remaining fragments (Fig. 4*D*; $\langle P_\beta\rangle$:$\langle P_{\mathrm{PPII}}\rangle$:$\langle P_\alpha\rangle$ = 32:42:22). Overall, the best correlation is obtained with our original coil library, but reasonable good agreement also emerges for the other libraries in which the fraction of extended conformers is similar.

The NMR experiments are conducted with proteins residing in a concentrated denaturant solution, which stabilizes nonnative conformations. Our coil library is culled from high-resolution crystal structures and refers to irregular conformations in a "protein-like environment." In the coil library, extended PPII and β backbone conformations dominate over the bent α conformers, in accord with experimental results for peptides where substantial amounts of PPII are observed (24, 45–47). The preference for extended over α conformations might even be enhanced under denaturing conditions where the protein is in good solvent conditions because of the increased backbone exposure of β and PPII conformers (18).

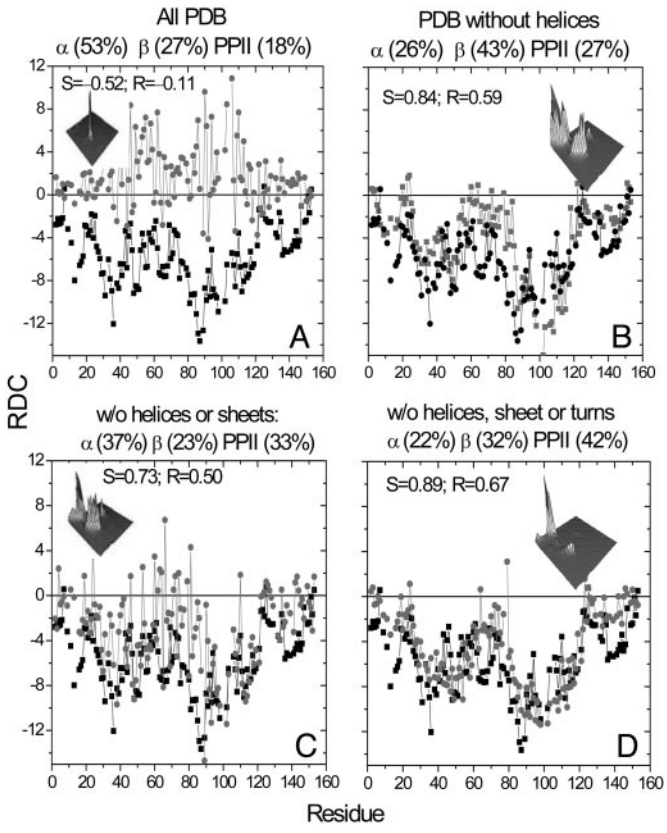

Fig. 4. Alternative libraries for generating unfolded ensembles. Calculated (gray) and experimental (black) RDCs for apoMb. The calculated RDCs are obtained by averaging 5,000 structures generated by using four different PDB-based statistical potentials that incorporate NN effects and excluded volume constraints. The 3D plot in each image is the probability distribution in the Ramachandran plane for each of the libraries. The axes pointing to left and right are ψ and ϕ angles, respectively.

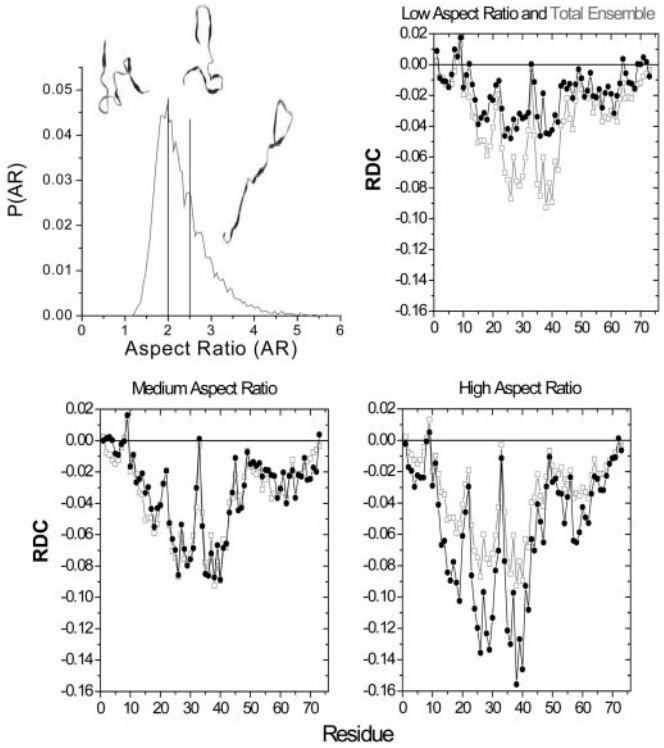

Fig. 5. AR and RDCs. Shown are AR distribution for the 16,000 chains comprising the unfolding ensemble of Ub. The ensemble is partitioned into three nearly equal fractions, AR ≤ 2, $2 < AR < 2.5$ and AR ≥ 2.5. Representative conformations are shown for each partition. Overlay of the computed RDCs for each of the three portions of the ensemble (black) and the average for the total ensemble (gray) are shown.

The unfolded ensemble produced with the coil library lacks hydrogen-bonded structures. There are very few stretches of four or more α conformers, which could constitute authentic helical structures. This situation is well-matched to experiment; little residual structure likely exists in the chemically denatured states (18) of the four proteins. Under acidic denaturing conditions, however, some helical structure is present in apoMb (29) and acylcoenzyme A binding protein (ACBP) (11). The change from extended to helical geometry equates to an $\approx$90° rotation in the average N–H direction relative to the long axis, as a result of which the observed RDCs change, sometimes even switching sign. Therefore, our coil library best represents the denatured state of a protein in high denaturant concentrations where the protein is devoid of significant residual hydrogen-bonded structure.

Stretched Conformations Dominate the RDC Signal. We investigate whether a correlation exists between the magnitude of the RDCs and the AR of the individual members in the ensemble of unfolded state proteins. High ARs are expected for individual chains in any random-walk model even though the ensemble average is spherical (33, 44). The AR is defined as $(I_z)/\sqrt{I_x I_y}$, where I_x, I_y, and I_z are the moments of inertia for the given conformation in ascending order.

For Ub's unfolded ensemble, we partitioned the AR distribution into three nearly equal fractions, AR ≤ 2, $2 < AR < 2.5$, and AR $\geq$ 2.5 (Fig. 5). As the AR increases, the RDCs increase in magnitude, although the residue-to-residue pattern remains largely unchanged. The most extended third contributes $\approx$60% of the ensemble's average signal. These most stretched members of the ensemble have the largest β and PPII percentages.

We calculate the RDCs assuming a common degree of alignment for all molecules. However, molecules with higher ARs should align more strongly. Accordingly, we apply an empirical weighting scheme, alignment $\propto AR^n$, but find a negligible improvement between calculated and experimental RDCs (e.g., apoMB, $R = 0.71 \rightarrow 0.73$ for $n = 0 \rightarrow 5$).

The computed RDCs are of a single sign because of the prevalence of extended conformations in the coil library and the global alignment of the entire chain. Wright, Dyson, and coworkers (29) proposed a similar view for apoMb in that the RDCs originate from transient alignment of short segments composed of extended conformations, although no detailed calculations are presented. To test their assertion, we calculate RDCs by using the local alignment for each of the four major segments of unfolded apoMb identified in their study (results not shown) or alternatively by averaging over every possible seven residue window (the average persistence length of a polypeptide). The agreement with experiment is worse than when calculations assume alignment of the entire chain (see Fig. 8, which is published as supporting information on the PNAS web site). Therefore, the global chain alignment used our study is more probable.

Lack of Native-Topology in Unfolded State. For the 5,000 members of Ub's unfolded ensemble, the average contact matrix lacks the native topology (see Fig. 9, which is published as supporting information on the PNAS web site). Visual inspection of the Ub conformations highlights their lack of native topology (Fig. 5). Therefore, it is unlikely that native topology persists in the unfolded state for proteins whose experimental RDC pattern our model can recapitulate.

Furthermore, chain configurations with the highest ARs contribute most of the computed RDC signal. These extended configurations have more than twice the number of β plus PPII vs. α conformers. This argument probably applies to the experiments as well. In the presence of denaturant and confinement to the narrow channels of the aligning medium, extended conformations may be further favored. Hence, RDC values arise from conformations that are less likely to reflect the topology of a (compact) native protein.

The previously observed correlation between experimental RDCs in the native state and in the unfolded state is weak for elgin (10). More importantly, the RDCs in these two states represent contributions from completely different sets of backbone angles. The native protein yields both positive and negative RDCs because the helical and sheet elements do not necessarily align with the molecular axis. However, RDCs in the unfolded state are almost always of a single sign, originating from the preponderance of extended conformations that align along the molecular axis. Thus, it is difficult to conceive that the native and unfolded states could have similar topologies.

Shortle *et al.* (1, 10, 28) have analyzed RDCs to argue for the presence of a native-like organization of chain segments in unfolded proteins. Their original study (1) presents a correlation between RDCs of Δ131Δ in 8 M urea and in water. Although the correlation does not extend to the native state, Δ131Δ was deduced to have native topology based on independent distance constraints between numerous extrinsic spin labels (48, 49). Of the four proteins investigated, Δ131Δ exhibits the poorest agreement with our model. The deviation is greater for Δ131Δ in aqueous solutions (data not shown), suggesting that Δ131Δ deviates from statistical coil behavior. Hence, this molecule may have native topology. However, its RDCs are mostly negative, indicating the backbone predominantly is in either a β or PPII conformation, whereas the native structure contains both helices and sheets across this portion of the Snase sequence.

Implications for Computer Simulations. The two common initial conditions for folding simulations are a completely extended or thermally denatured chain. Because both initial choices are unphysical, significant amounts of computer time must be spent to thermalize the system. This issue is most relevant to distributed computing methods, where numerous trajectories must be individ-

ually equilibrated, taking up a large fraction of each trajectory. Furthermore, the initial configuration could even drive the protein down an otherwise unpopulated folding pathway. Thus, a physically realistic configuration for the unfolded state, such as our model, should save computational resources while increasing accuracy.

Conclusions

Various models agree with experimentally determined dimensions of the unfolded state. However, this criterion is insufficient to conclude that the unfolded state is a statistical coil. RDCs provide a local, and more sensitive, measure of structure and, therefore, a much more stringent test. Our statistical coil model using the backbone frequencies observed in a restricted coil library, along with excluded volume repulsions, reproduces both the known coil scaling behavior for the R_g and, more importantly, compares remarkably well with the experimental RDCs for chemically denatured proteins. The RDC values are dominated by high AR chain configurations, which are predominantly composed of PPII and β conformers. No evidence for native-like topology exists in this configuration or any subset of the ensemble. Hence, the model resolves the so-called "reconciliation problem," because the model accurately accounts for the presence of local structure in unfolded states while retaining random coil scaling (14).

The statistical coil model implies that a Levinthal-like number of denatured conformations exists, although correlations between neighboring residues reduce this number to some extent (35, 36). Local sterics generally are too weak to generate persistent structure or native-like topology or to reduce significantly the conformational entropy in the unfolded state. Nevertheless, the locally encoded backbone can bias the conformational search toward native-like elements and thereby reduce the Levinthal paradox. In combination with tertiary context, local sterics determine the structure and stability of the native state.

We thank Drs. S. Koide, D. Shortle, G. Rose, S. Ohnishi, S. W. Englander, N. Kallenbach, R. S. Berry, G. Rose, and R. Pappu and members of our group for comments and discussions; and Drs. S. Ohnishi and D. Shortle for providing the unpublished Ub data. This work was supported by grants from the National Institutes of Health and the National Science Foundation. A.K.J. was supported by Burroughs Wellcome Fund Interfaces 1001774. A.C. is supported by a Burroughs Wellcome Fund Interfaces Postdoctoral Fellowship in Mathematics and Molecular Biology.

1. Shortle, D. & Ackerman, M. S. (2001) *Science* **293,** 487–489.
2. Prestegard, J. H., Al-Hashimi, H. M. & Tolman, J. R. (2000) *Q. Rev. Biophys.* **33,** 371–424.
3. Bax, A. (2003) *Protein Sci.* **12,** 1–16.
4. Tjandra, N. (1999) *Struct. Folding Des.* **7,** R205–R211.
5. Meiler, J., Blomberg, N., Nilges, M. & Griesinger, C. (2000) *J. Biomol. NMR* **16,** 245–252.
6. Skrynnikov, N. R. & Kay, L. E. (2000) *J. Biomol. NMR* **18,** 239–252.
7. Delaglio, F., Kontaxis, G. & Bax, A. (2000) *J. Am. Chem. Soc.* **122,** 2142–2143.
8. 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., *et al.* (2004) *Proc. Natl. Acad. Sci. USA* **101,** 12491–12496.
9. Ohnishi, S. & Shortle, D. (2003) *Protein Sci.* **12,** 1530–1537.
10. Ohnishi, S., Lee, A. L., Edgell, M. H. & Shortle, D. (2004) *Biochemistry* **43,** 4064–4070.
11. Fieber, W., Kristjansdottir, S. & Poulsen, F. M. (2004) *J. Mol. Biol.* **339,** 1191–1199.
12. Tanford, C., Kawahara, K. & Lapanje, S. (1966) *J. Biol. Chem.* **241,** 1921–1923.
13. Tanford, C. (1968) *Adv. Protein Chem.* **23,** 121–282.
14. Millet, I. S., Doniach, S. & Plaxco, K. W. (2002) *Adv. Protein Chem.* **62,** 241–262.
15. Freed, K. F. (1987) *Renormalization Group Theory of Macromolecules* (Wiley Interscience, New York).
16. Tran, H. T., Wang, X. & Pappu, R. V. (2005) *Biochemistry*, 10.1021/bi0501961.
17. Fitzkee, N. C. & Rose, G. D. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 12497–12502.
18. Jacob, J., Krantz, B., Dothager, R. S., Thiyagarajan, P. & Sosnick, T. R. (2004) *J. Mol. Biol.* **338,** 369–382.
19. Swindells, M. B., MacArthur, M. W. & Thornton, J. M. (1995) *Nat. Struct. Biol.* **2,** 596–603.
20. Smith, L. J., Bolin, K. A., Schwalbe, H., MacArthur, M. W., Thornton, J. M. & Dobson, C. M. (1996) *J. Mol. Biol.* **255,** 494–506.
21. Munoz, V. & Serrano, L. (1994) *Proteins* **20,** 301–311.
22. Smith, L. J., Fiebig, K. M., Schwalbe, H. & Dobson, C. M. (1996) *Fold Des.* **1,** R95–R106.
23. Penkett, C. J., Redfield, C., Dodd, I., Hubbard, J., McBay, D. L., Mossakowska, D. E., Smith, R. A., Dobson, C. M. & Smith, L. J. (1997) *J. Mol. Biol.* **274,** 152–159.
24. Shi, Z., Olson, C. A., Rose, G. D., Baldwin, R. L. & Kallenbach, N. R. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 9190–9195.
25. Jha, A. K., Colubri, A., Zaman, M. H., Koide, S., Sosnick, T. R. & Freed, K. F. (2005) *Biochemistry* **44,** 9691–9702.
26. Ohnishi, S. & Shortle, D. (2003) *Proteins* **50,** 546–551.
27. Ackerman, M. S. & Shortle, D. (2002) *Biochemistry* **41,** 3089–3095.
28. Ackerman, M. S. & Shortle, D. (2002) *Biochemistry* **41,** 13791–13797.
29. Mohana-Borges, R., Goto, N. K., Kroon, G. J., Dyson, H. J. & Wright, P. E. (2004) *J. Mol. Biol.* **340,** 1131–1142.
30. Ding, K., Louis, J. M. & Gronenborn, A. M. (2004) *J. Mol. Biol.* **335,** 1299–1307.
31. Tycko, R., Blanco, F. J. & Ishii, Y. (2000) *J. Am. Chem. Soc.* **122,** 9340–9341.
32. Plaxco, K. W. & Gross, M. (2001) *Nat. Struct. Biol.* **8,** 659–660.
33. Louhivuori, M., Paakkonen, K., Fredriksson, K., Permi, P., Lounila, J. & Annila, A. (2003) *J. Am. Chem. Soc.* **125,** 15647–15650.
34. Sallum, C. O., Martel, D. M., Fournier, R. S., Matousek, W. M. & Alexandrescu, A. T. (2005) *Biochemistry* **44,** 6392–6403.
35. Zaman, M. H., Shen, M. Y., Berry, R. S., Freed, K. F. & Sosnick, T. R. (2003) *J. Mol. Biol.* **331,** 693–711.
36. Pappu, R. V., Srinivasan, R. & Rose, G. D. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 12565–12570.
37. Keskin, O., Yuret, D., Gursoy, A., Turkay, M. & Erman, B. (2004) *Proteins* **55,** 992–998.
38. Gibrat, J. F., Garnier, J. & Robson, B. (1987) *J. Mol. Biol.* **198,** 425–443.
39. Kang, H. S., Kurochkina, N. A. & Lee, B. (1993) *J. Mol. Biol.* **229,** 448–460.
40. Avbelj, F. & Baldwin, R. L. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 10967–10972.
41. Wang, G. & Dunbrack, R. L., Jr. (2003) *Bioinformatics* **19,** 1589–1591.
42. Canutescu, A. A., Shelenkov, A. A. & Dunbrack, R. L., Jr. (2003) *Protein Sci.* **12,** 2001–2014.
43. Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. (1999) *J. Mol. Biol.* **285,** 1735–1747.
44. Haber, C., Ruiz, S. A. & Wirtz, D. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 10792–10795.
45. Eker, F., Griebenow, K. & Schweitzer-Stenner, R. (2003) *J. Am. Chem. Soc.* **125,** 8178–8185.
46. Chellgren, B. W. & Creamer, T. P. (2004) *Biochemistry* **43,** 5864–5869.
47. Dukor, R. K. & Keiderling, T. A. (1991) *Biopolymers* **31,** 1747–1761.
48. Gillespie, J. R. & Shortle, D. (1997) *J. Mol. Biol.* **268,** 170–184.
49. Gillespie, J. R. & Shortle, D. (1997) *J. Mol. Biol.* **268,** 158–169.