J|A|C|S
ARTICLES

Published on Web 12/15/2009

Ensemble Calculations of Unstructured Proteins Constrained by RDC and PRE Data: A Case Study of Urea-Denatured Ubiquitin

Jie-rong Huang and Stephan Grzesiek*

Division of Structural Biology, Biozentrum, University of Basel, Klingelbergstrasse 50, 4056 Basel, Switzerland

Received September 18, 2009; E-mail: stephan.grzesiek@unibas.ch

Abstract: The detailed, quantitative characterization of unfolded proteins is a largely unresolved task due to the enormous experimental and theoretical difficulties in describing the highly dimensional space of their conformational ensembles. Recently, residual dipolar coupling (RDC) and paramagnetic relaxation enhancement (PRE) data have provided large numbers of experimental parameters on unfolded states. To obtain a minimal model of the unfolded state according to such data we have developed new modules for the use of steric alignment RDCs and PREs as constraints in ensemble structure calculations by the program XPLOR-NIH. As an example, ensemble calculations were carried out on urea-denatured ubiquitin using a total of 419 previously obtained RDCs and 253 newly determined PREs from eight cysteine mutants coupled to MTSL. The results show that only a small number of about 10 conformers is necessary to fully reproduce the experimental RDCs, PREs and average radius of gyration. C$^{\alpha}$ contacts determined on a large set (400) of 10-conformer ensembles show significant (10−20%) populations of conformations that are similar to ubiquitin's A-state, i.e. corresponding to an intact native first β-hairpin and α-helix as well as non-native α-helical conformations in the C-terminal half. Thus, methanol/acid (A-state) and urea denaturation lead to similar low energy states of the protein ensemble, presumably due to the weakening of the hydrophobic core. Similar contacts are obtained in calculations using solely RDCs or PREs. The sampling statistics of the C$^{\alpha}$ contacts in the ensembles follow a simple binomial distribution. It follows that the present RDC, PRE, and computational methods allow the statistically significant detection of subconformations in the unfolded ensemble at population levels of a few percent.

Introduction

A detailed, quantitative description of the unfolded state ensemble of proteins is crucial for the understanding of protein folding,[1] protein misfolding and aggregation in amyloidogenic diseases such as Alzheimer's and Parkinson's,[2] and function of intrinsically disordered proteins.[3,4] The astronomical size of the conformational space of an unfolded polypeptide chain makes such a description both experimentally and theoretically very difficult.

NMR is one of the most powerful tools for the experimental characterization of unstructured proteins, since it can give specific information for almost all atoms with minimal interference. Besides chemical shifts, scalar couplings, nuclear Overhauser effects (NOEs) and relaxation rates, recently paramagnetic relaxation enhancements (PREs) and residual dipolar couplings (RDCs) have been added to the arsenal of NMR parameters that describe the unfolded state (for reviews see refs 5, 6). In contrast to other parameters, the size of PRE and RDC effects can be calculated very precisely as ensemble and time averages from the well understood geometry dependence of electron-nucleus or nucleus−nucleus dipolar interactions. PREs can be observed after introduction of a suitable paramagnetic tag and report on long-range contacts from the tag to the protein in a $1/r^6$ dependence. RDCs are induced after making the protein solution anisotropic, in most cases of unfolded proteins by dissolution in mechanically squeezed polyacrylamide gels.[7,8] In this way, very large numbers of RDCs can be observed, which report on the size and direction of internuclear vectors. Unfolded protein ensembles characterized by these methods comprise staphylococcal nuclease,[9,10] acyl-CoA binding protein,[11] apomyoglobin,[12] T4 fibritin foldon,[13] α- and β-synuclein,[14,15] Tau protein,[16] ubiquitin,[17,18] and smaller peptides.[19]

Despite the availability of a large number of such experimental data, theoretical concepts for their analysis are less

[1] Shortle, D. *Faseb J.* **1996**, *10*, 27–34.

[2] Dobson, C. M. *Nature* **2003**, *426*, 884–890.

[3] Dunker, A. K.; Silman, I.; Uversky, V. N.; Sussman, J. L. *Curr. Opin. Struct. Biol.* **2008**, *18*, 756–764.

[4] Wright, P. E.; Dyson, H. J. *Curr. Opin. Struct. Biol.* **2009**, *19*, 31–38.

[5] Mittag, T.; Forman-Kay, J. D. *Curr. Opin. Struct. Biol.* **2007**, *17*, 3–14.

[6] Meier, S.; Blackledge, M.; Grzesiek, S. *J. Chem. Phys.* **2008**, *128*, 052204.

[7] Tycko, R.; Blanco, F. J.; Ishii, Y. *J. Am. Chem. Soc.* **2000**, *122*, 9340–9341.

[8] Sass, H. J.; Musco, G.; Stahl, S. J.; Wingfield, P. T.; Grzesiek, S. *J. Biomol. NMR* **2000**, *18*, 303–309.

[9] Gillespie, J. R.; Shortle, D. *J. Mol. Biol.* **1997**, *268*, 158–169.

[10] Shortle, D.; Ackerman, M. S. *Science* **2001**, *293*, 487–489.

[11] Fieber, W.; Kristjansdottir, S.; Poulsen, F. M. *J. Mol. Biol.* **2004**, *339*, 1191–1199.

[12] Felitsky, D. J.; Lietzow, M. A.; Dyson, H. J.; Wright, P. E. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 6278–6283.

10.1021/ja907974m

developed. Interpretation of PRE data on unfolded states has been pursued by constrained molecular dynamics (MD) simulations or by selection of structures from a pool of pregenerated conformers according to the observed long-range contacts.[9,20−26] For apomyoglobin it was also shown that the observed contacts correlated to the average area buried upon folding.[12] Generation of structural ensembles from RDC data in restrained MD calculations is more difficult, since this requires the computationally costly calculation of the orientation tensor for every single member of the ensemble at every time step. In contrast, comparison of the measured RDC data to predictions from models of the unfolded state is straightforward. Thus, Sosnick, Blackledge, and co-workers could show that structural ensembles created according to amino-acid-specific ϕ/ψ angle propensities in non-α, non-β conformations of PDB structures (PDB coil library) reproduced the trends of RDCs along the polypeptide sequence.[6,27,28] Apparently, this so-called coil model[29,30] is a good, first approximation of the unfolded state ensemble. In turn, deviations from the coil model point to residual order of the unfolded state. Such deviations reveal e.g. highly populated turn conformations in the natively unfolded Tau protein[16] and show that urea binding drives the backbone to more extended conformations for ubiquitin.[17]

Longer-range interactions in unfolded states are outside of the scope of the coil model. Such contacts are clearly detectable not only by NOEs,[31] PREs,[9] but also by RDCs. Thus, long-range RDCs between amide protons gave evidence for a remaining, significant (10−20%) population of a β-hairpin in (8 M) urea-denatured ubiquitin.[18] This subpopulation also contains a number of fully formed H-bonds as evident from scalar couplings between the amide ^{15}N donor and the ^{13}C′ acceptor atoms.[18]

Clearly, it is desirable to use RDCs, PREs, and possibly other observables together as constraints for a single minimal description of the unfolded state ensemble that is compatible with the experimental data. For this reason, we have developed two new modules for steric alignment RDC and PRE ensemble calculations, which are incorporated into the commonly used structure calculation program, XPLOR-NIH.[32,33] The RDC module very efficiently calculates the steric alignment tensor for every member of the ensemble at every time step and derives a potential energy from the difference of the predicted ensemble average relative to the experimental RDCs. The new PRE module is optimized from the existing[34] for use in unstructured protein ensembles. We have applied these algorithms to RDC and PRE data of ubiquitin unfolded under conditions of 8 M urea at pH 2.5. The data comprise 419 short- and long-range RDCs obtained by steric alignment in previous studies.[17,18] In addition, 253 long-range PRE restraints were newly determined from eight ubiquitin cysteine mutants paramagnetically tagged with the nitroxide spin label 1-oxyl-2,2,5,5-tetramethyl-3-pyrroline-3-methyl)-methanethiosulfonate (MTSL). The ensemble structure calculations yield an assessment of the information content of the experimental data and show that averages over only very few conformers suffice for an agreement within experimental errors. An analysis of the C$^{\alpha}$ contacts indicates that the urea-denatured state contain significant amounts (10−20%) of structured subpopulations that resemble ubiquitin's methanol/acid-denatured A-state. According to an evaluation of the sampling statistics these subpopulations are statistically significant. Despite systematic uncertainties such as unknown correlation times for PREs and unknown microscopic details of the alignment interaction for RDCs, separate ensemble calculations using either PREs or RDCs yield very similar results for these more highly populated C$^{\alpha}$ contacts, which are enhanced when both experimental restraints are used together. Thus the results appear rather robust with respect to the type of input data and the underlying uncertainties of interpretation.

Theory and Algorithms

Calculation of RDC Ensemble Averages from Steric Alignment. RDCs of size D observed in solution result from the average over the dipolar interaction between two nuclei i and j:

$$D = -\frac{\gamma_i\gamma_j\hbar\mu_0}{4\pi^2 r^3}\left\langle\frac{(3\cos^2\theta - 1)}{2}\right\rangle = D_{\max}\langle P_2(\cos\theta)\rangle \quad (1)$$

where θ is the instantaneous angle of the internuclear vector with respect to the magnetic field, P_2 is the second-order Legendre polynomial, and the internuclear distance r is assumed as fixed. The average within the angular parentheses corresponds to an ensemble average over the sample and a time average up to the millisecond range corresponding to the total experimental observation time.

[13] Meier, S.; Guthe, S.; Kiefhaber, T.; Grzesiek, S. *J. Mol. Biol.* **2004**, *344*, 1051–1069.

[14] Binolfi, A.; Rasia, R. M.; Bertoncini, C. W.; Ceolin, M.; Zweckstetter, M.; Griesinger, C.; Jovin, T. M.; Fernandez, C. *J. Am. Chem. Soc.* **2006**, *128*, 9893–9901.

[15] Bertoncini, C. W.; Rasia, R. M.; Lamberto, G. R.; Binolfi, A.; Zweckstetter, M.; Griesinger, C.; Fernandez, C. *J. Mol. Biol.* **2007**, *372*, 708–722.

[16] Mukrasch, M. D.; Markwick, P.; Biernat, J.; Bergen, M. V.; Bernado, P.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 5235–5243.

[17] Meier, S.; Grzesiek, S.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 9799–9807.

[18] Meier, S.; Strohmeier, M.; Blackledge, M.; Grzesiek, S. *J. Am. Chem. Soc.* **2007**, *129*, 754–755.

[19] Dames, S. A.; Aregger, R.; Vajpai, N.; Bernado, P.; Blackledge, M.; Grzesiek, S. *J. Am. Chem. Soc.* **2006**, *128*, 13508–13514.

[20] Lindorff-Larsen, K.; Kristjansdottir, S.; Teilum, K.; Fieber, W.; Dobson, C. M.; Poulsen, F. M.; Vendruscolo, M. *J. Am. Chem. Soc.* **2004**, *126*, 3291–3299.

[21] Bertoncini, C. W.; Jung, Y. S.; Fernandez, C. O.; Hoyer, W.; Griesinger, C.; Jovin, T. M.; Zweckstetter, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 1430–1435.

[22] Kristjansdottir, S.; Lindorff-Larsen, K.; Fieber, W.; Dobson, C. M.; Vendruscolo, M.; Poulsen, F. M. *J. Mol. Biol.* **2005**, *347*, 1053–1062.

[23] Dedmon, M. M.; Lindorff-Larsen, K.; Christodoulou, J.; Vendruscolo, M.; Dobson, C. M. *J. Am. Chem. Soc.* **2005**, *127*, 476–477.

[24] Francis, C. J.; Lindorff-Larsen, K.; Best, R. B.; Vendruscolo, M. *Proteins* **2006**, *65*, 145–152.

[25] Marsh, J. A.; Neale, C.; Jack, F. E.; Choy, W. Y.; Lee, A. Y.; Crowhurst, K. A.; Forman-Kay, J. D. *J. Mol. Biol.* **2007**, *367*, 1494–1510.

[26] Song, J.; Guo, L. W.; Muradov, H.; Artemyev, N. O.; Ruoho, A. E.; Markley, J. L. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 1505–1510.

[27] Jha, A. K.; Colubri, A.; Freed, K. F.; Sosnick, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13099–13104.

[28] Bernado, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R. W.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17002–17007.

[29] Serrano, L. *J. Mol. Biol.* **1995**, *254*, 322–333.

[30] Smith, L. J.; Bolin, K. A.; Schwalbe, H.; MacArthur, M. W.; Thornton, J. M.; Dobson, C. M. *J. Mol. Biol.* **1996**, *255*, 494–506.

[31] Neri, D.; Billeter, M.; Wider, G.; Wuthrich, K. *Science* **1992**, *257*, 1559–1563.

[32] Schwieters, C. D.; Kuszewski, J. J.; Tjandra, N.; Clore, G. M. *J. Magn. Reson.* **2003**, *160*, 65–73.

[33] Schwieters, C. D.; Kuszewski, J. J.; Clore, G. M. *Prog. Nucl. Magn. Reson. Spectrosc.* **2006**, *48*, 47–62.

[34] Iwahara, J.; Schwieters, C. D.; Clore, G. M. *J. Am. Chem. Soc.* **2004**, *126*, 5879–5896.

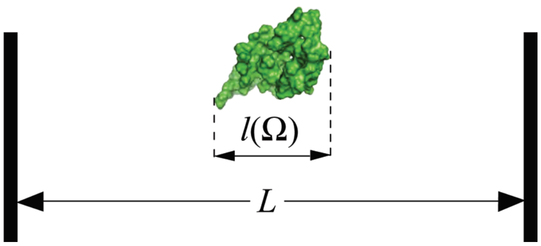

Figure 1. Illustration of excluded volume effects that describe the steric alignment probability at a certain orientation of a molecule. The black bars represent two infinite parallel planes at distance L. The length l is the projected length of the molecule at a certain orientation Ω onto the plane normal and is proportional to the excluded volume.

In the case of folded, rigid proteins, it is usual to express the internuclear vector orientation in local, molecular polar coordinates Θ, Φ and to describe the overall rotation of the molecule by a Wigner rotation matrix with Euler angles $\Omega = (\alpha, \beta, \gamma)$. Thus,

$$\begin{aligned}\langle P_2(\cos\theta)\rangle &= \sqrt{\frac{4\pi}{5}}\sum_{m=-2}^{2}\langle D^2_{m0}(\alpha,\beta,\gamma)\rangle Y_{2m}(\Theta,\Phi)\\ &= \frac{4\pi}{5}\sum_{m=-2}^{2}\langle Y^*_{2m}(\beta,\alpha)\rangle Y_{2m}(\Theta,\Phi)\\ &= \sqrt{\frac{4\pi}{5}}\sum_{m=-2}^{2} S^*_m Y_{2m}(\Theta,\Phi) \qquad (2)\end{aligned}$$

where

$$S_m = \sqrt{\frac{4\pi}{5}}\langle Y_{2m}(\beta,\alpha)\rangle$$

is the orientation tensor of the molecule in irreducible form.[35]

Assuming that the time average equals the ensemble average, S_m is obtained from the probability distribution $P(\Omega)$ to find the molecule at a certain orientation Ω:

$$S_m = \sqrt{\frac{4\pi}{5}}\int Y_{2m}(\beta,\alpha)\cdot P(\Omega)\mathrm{d}\Omega \qquad (3)$$

For steric alignment $P(\Omega)$ can be calculated from excluded volume effects[36,37] (see Figure 1) as

$$P(\Omega) = \frac{L - l(\Omega)}{4\pi L - \int l(\Omega)\mathrm{d}\Omega} \approx \frac{L - l(\Omega)}{4\pi L} \qquad (4)$$

where L is the distance between two infinite parallel planes and l is the maximal length of the molecule in the direction perpendicular to the planes. Thus the parameter L together with the anisotropy of the molecular shape determines the absolute size of the molecular orientation.

In contrast to the folded state, it is obvious for unfolded proteins that the global shape and the corresponding alignment will depend strongly on local conformations, e.g. when a turn residue changes its backbone angles. Thus, it clearly has no sense to assume a common orientation tensor for all the different conformations of an unfolded peptide chain. A viable approximation for an unfolded protein may be to assume that the dipolar coupling results from an average over an ensemble of N molecular conformations, where every conformation k has its individual orientation tensor S_{km}:

$$D = \frac{D_{\max}}{N}\sqrt{\frac{4\pi}{5}}\sum_{k=1}^{N}\sum_{m=-2}^{2} S^*_{km}Y_{2m}(\Theta_k,\Phi_k) \qquad (5)$$

This approach was used successfully for the prediction of RDCs from coil model ensembles of unfolded proteins where a good correlation was found to measured RDCs.[27,28] The approximation may be problematic, if the interconversion between different conformers is faster than the orienting contact event, which is expected to be on the time scale of nanoseconds.[6] In these cases, the alignment and the molecular conformation could average independently. However, the possible complication by such effects does not appear very severe for the present calculations, since structural ensembles generated independently by either RDCs or PREs show very similar results for the most highly populated subconformations (see below).

To constrain model ensembles of unfolded proteins by the measured RDCs against the predictions of eq 5, a target function was implemented into the program XPLOR-NIH:

$$E_{\mathrm{RDC}} = k_{\mathrm{RDC}}\sum_i \frac{(D_i^{\mathrm{obs}} - D_i^{\mathrm{calc}})^2}{\sigma_i^2} \qquad (6)$$

where D^{calc} is the calculated RDC, D^{obs} the experimental value, σ the experimental error, k_{RDC} a force constant, and the summation runs over all measured RDCs. For every evaluation of this target function, D^{calc} is obtained according to eq 5 from a full calculation of the alignment tensor of every member of the ensemble according to eqs 3 and 4. An efficient algorithm for the integration of eq 3 was implemented according to ideas formulated by Scheek and co-workers.[37] Since the exact value of the parameter L is difficult to predict from the experimental conditions, an additional overall scaling factor λ for the calculated couplings was determined from a linear least-squares fit as

$$\lambda = \frac{\sum_i D_i^{\mathrm{obs}}D_i^{\mathrm{calc}}}{\sum_i (D_i^{\mathrm{calc}})^2} \qquad (7)$$

In this XPLOR module, the length of internuclear vectors can be user-defined or calculated from the coordinates. Thus, e.g. for one bond RDCs, the $\mathrm{NH^N}$, $\mathrm{C^\alpha H^\alpha}$, and $\mathrm{C^\alpha C'}$ distances were set to 1.023, 1.10, and 1.52 Å, respectively,[38] whereas HH distances were calculated from the molecular coordinates at each time-step. When the sign of the RDC is not determined, e.g. for D_{HH}, the target function is implemented as

$$E_{\mathrm{RDC}} = k_{\mathrm{RDC}}\sum_i \frac{(|D_i^{\mathrm{obs}}| - |D_i^{\mathrm{calc}}|)^2}{\sigma_i^2} \qquad (8)$$

We name this XPLOR-NIH module sardcPot (for **s**teric **a**lignment **rdc Pot**ential).

Calculation of PRE Ensemble Averages. In order to use PRE information for the calculation of unstructured protein ensembles, we have also developed a new module with ensemble simulation features for XPLOR-NIH. Although an energy potential for PRE (prePot) was described by Iwahara et al.,[34] its original implementation of an ensemble was not compatible

[35] Moltke, S.; Grzesiek, S. *J. Biomol. NMR* **1999**, *15*, 77–82.

[36] Zweckstetter, M.; Bax, A. *J. Am. Chem. Soc.* **2000**, *122*, 3791–3792.

[37] van Lune, F.; Manning, L.; Dijkstra, K.; Berendsen, H. J.; Scheek, R. M. *J. Biomol. NMR* **2002**, *23*, 169–179.

[38] Yao, L.; Vögeli, B.; Ying, J.; Bax, A. *J. Am. Chem. Soc.* **2008**, *130*, 16518–16520.

with the XPLOR-NIH ensemble simulation feature and other potential terms, such as sardcPot, which use this facility. In our implementation of the PRE potential, a single conformer represents a complete biomolecule including its spin label(s). The ensemble of several of such conformers is taken as a model of the unfolded protein ensemble. The module uses the XPLOR-NIH ensemble calculation mechanism, is fully compatible with other ensemble energy functions, and does not require special input coordinate files. The module should be especially useful and easy to use for cases of unstructured proteins, and we denote it as prePotD (for **D**enatured) to distinguish it from the original one (prePot). A very recent, updated implementation of the original prePot module (XPLOR-NIH version 2.22) also contains ensemble simulation features.

In prePotD, calculated PREs are obtained as ensemble averages across the multiple conformers and compared to measured PREs using the target function[34]

$$E_{\mathrm{PRE}} = k_{\mathrm{PRE}} \sum_i w_i [\Delta R_2^{\mathrm{obs}}(i) - \langle \Delta R_2^{\mathrm{calc}}(i) \rangle]^2 \qquad (9)$$

where k_{PRE} is a force constant, $\Delta R_2{}^{\mathrm{obs}}(i)$ and $\Delta R_2{}^{\mathrm{calc}}(i)$ are the observed and calculated relaxation rate differences respectively, the bracket denotes ensemble averaging, and w_i is a weighting factor which is defined for each restraint as[34]

$$w_i = \frac{1}{\sigma_i^2} \frac{\Delta R_2^{\mathrm{obs}}(i)}{\Delta R_2^{\mathrm{obs,max}}} \qquad (10)$$

where σ_i is the experimental error, $\Delta R_2{}^{\mathrm{obs,max}}$ is the maximum observed value of ΔR_2 in the same restraint set. ΔR_2 is calculated according to the Solomon−Bloembergen equation[34,39]

$$\Delta R_2 = \left(\frac{\mu_0}{4\pi}\right)^2 \frac{\gamma_I^2 g^2 \mu_B^2 S(S+1)}{15 r^6} \left(4\tau_c + \frac{3\tau_c}{1 + (\omega\tau_c)^2}\right) \qquad (11)$$

where μ_0 is the permeability of vacuum, γ_I the gyromagnetic ratio of the nucleus, g the electron g-factor, μ_B the electron Bohr magneton, S the electron spin quantum number, r the distance between the electron and nucleus, ω the nuclear Larmor frequency, and τ_c the PRE correlation time given as $1/\tau_c = 1/\tau_r + 1/\tau_e$, with τ_r being the rotational correlation time of the electron−nucleus vector and τ_e the electron spin relaxation time. In the case of nitroxide spin labels, τ_e ($>10^{-7}$ s) is much longer than τ_r, and therefore τ_c equals approximately τ_r.[40] For a number of unfolded proteins τ_c has been assumed to be about 4 ns, since this yielded reasonable radii of gyration in ensemble structure calculations.[12,21,41,42]

Results

Locally Compact Patterns Are Observed for Unfolded Ubiquitin from PREs of Eight MTSL-Labeled Single Cysteine Mutants. The dipolar interaction with a paramagnetic label causes an increase in transverse nuclear relaxation rates, and thus line-broadening of the NMR signal. This effect can be used to probe transient contacts over distances as large as 20 Å[9,20,41] in the case of the nitroxide spin label MTSL. To couple this spin label to ubiquitin, eight highly solvent-exposed residues (K6, S20, K33, G35, K48, S57, K63, R74) were mutated to cysteine (Figure 2). All of these mutants expressed well and showed ^{1}H−^{15}N heteronuclear single-quantum coherence (HSQC) spectra for untagged protein under reducing conditions that were nearly identical to those of the wild-type both in the native and in the unfolded state (data not shown). This indicates that the conformations of these mutants in both states are very close to those of wild-type ubiquitin. Similarly, spectra of quenched MTSL-labeled and untagged urea-denatured ubiquitin mutants were nearly identical, showing that also the presence of the label had no major effect on the unfolded state. Assignments in the folded and unfolded states were then achieved easily by comparison to wild-type protein. Larger chemical shift changes due to cysteine mutations or MTSL-tagging were only found in the immediate vicinity of the spin-label sites (±2 residues). Possible changes in conformation for these few residues have no effect on our analysis, since these residues were all completely bleached out from the large PRE of the spin label, and hence, no PRE rates could be determined.

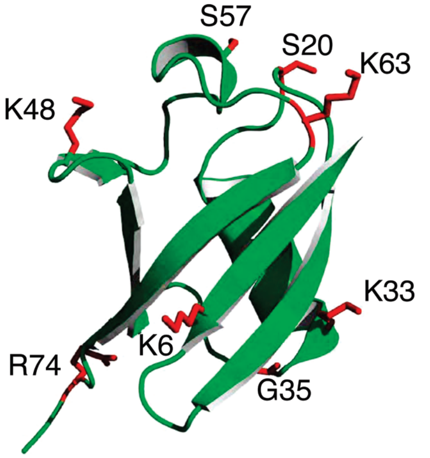

Figure 2. Schematic representation for the structure of ubiquitin. Residues that have been mutated to cysteine for MTSL attachment are shown in red stick mode.

The PRE effect in the eight unfolded ubiquitin mutants was quantified from the increase in ^{15}N R_2 relaxation rates detected by conventional ^{15}N relaxation experiments. These measured PREs are not affected by differential ^{1}H T_1 effects caused by the spin label or by additional attenuation from scalar couplings. Figure 3 shows the experimental relaxation rates for the paramagnetically labeled ubiquitin mutants and their diamagnetic reference. The obtained ^{15}N ΔR_2 rates are in the range of up to 4 Hz. The sensitivity of the experiments was very good, such that statistical errors (obtained from a repetition of the experiments) are mostly smaller than about 0.5 Hz even for residues with weak intensities in the vicinity of the spin label.

The measured ΔR_2 values can be compared to predictions from a Gaussian random chain model with a stiffness of 3.8 Å[42] (red curves Figure 3). Various nonlocal interactions of the unfolded ubiquitin are detected on the basis of the deviations from random chain behavior. Thus, for the K6C mutant, the PREs are stronger than expected for almost all residues (1−19) in the native first β-hairpin, indicating nonrandom conformations in this region. Consistent with this observation, also the spin label of the S20C mutant induces stronger than expected PREs for residues 2−15 toward the N-terminus. Interestingly, such stronger PREs are also detected from S20C in the C-terminal direction (residues 28−36), which comprises the native α-helix. Indeed, a transiently formed α-helix has been reported in this

[39] Solomon, I.; Bloembergen, N. *J. Chem. Phys.* **1956**, *25*, 261–266.
[40] Iwahara, J.; Tang, C.; Clore, G. M. *J. Magn. Reson.* **2007**, *184*, 185–195.
[41] Gillespie, J. R.; Shortle, D. *J. Mol. Biol.* **1997**, *268*, 170–184.
[42] Sung, Y. H.; Eliezer, D. *J. Mol. Biol.* **2007**, *372*, 689–707.

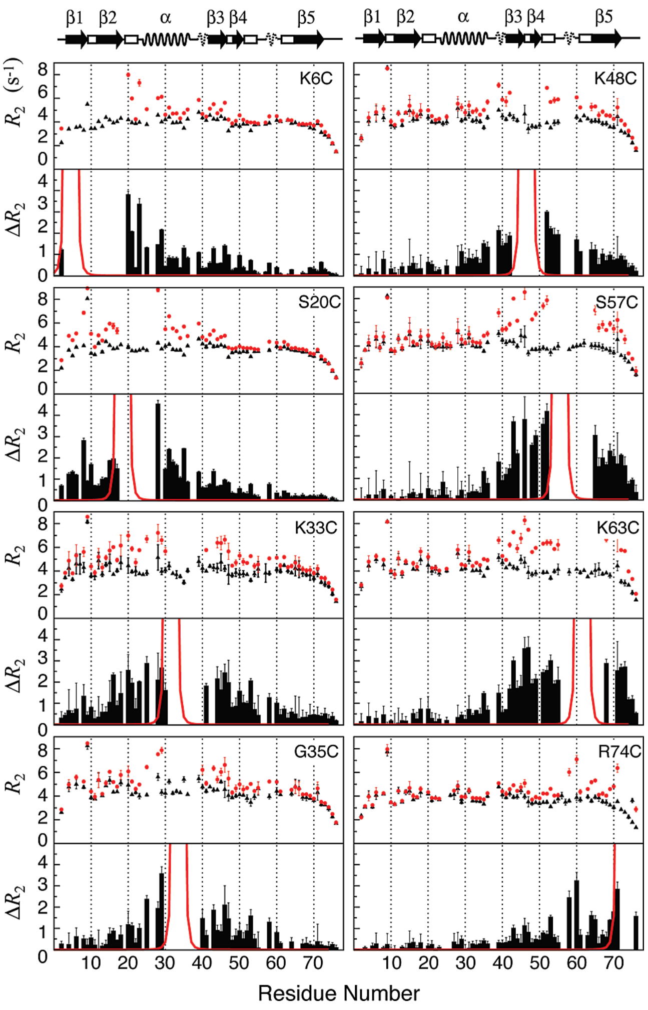

Figure 3. Experimental ^{15}N R_2 and ΔR_2 data. Upper-panel of each subfigure: ^{15}N-transverse relaxation rates R_2 of MTSL-labeled ubiquitin mutants (red) and their diamagnetic reference (black). Lower-panel: Differences of relaxation rate ΔR_2 between spin-labeled and reference samples. The errors in R_2 were obtained from two repeated measurements, and the errors of ΔR_2 are propagated from the R_2 measurements. The red lines represent the theoretical PRE curves calculated from an ideal random coil. The secondary structure of native ubquitin is shown at the top.

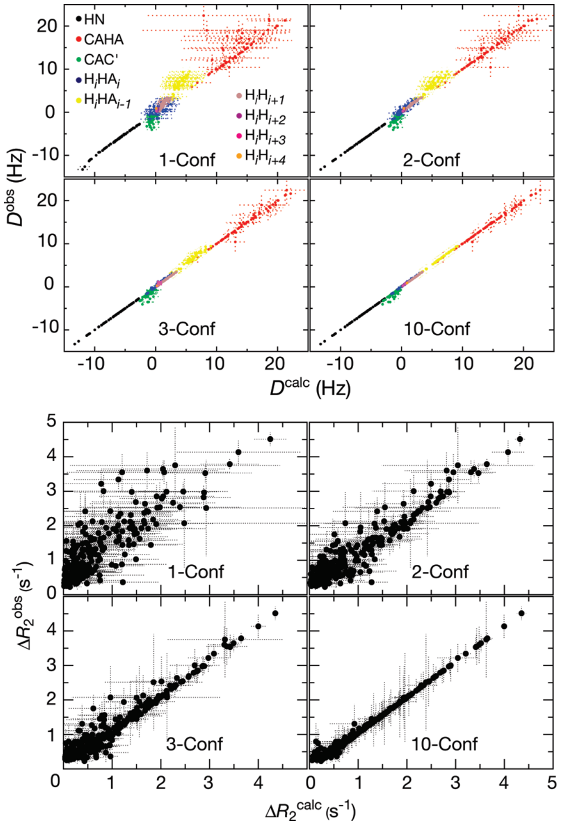

Figure 4. Agreement between calculated and observed RDC (top) and PRE (bottom) data for different sizes of the ensemble. The experimental data and calculated values for 1-, 2-, 3-, and 10-conformer ensembles are compared. The color code for the different types of RDCs is indicated in the upper left panel. The errors of the calculated values are obtained as standard deviations over 100 calculated ensembles according to Table 1.

region in peptide fragment studies[43,44] (see below). The K33C and G35C mutants, located at the end of the native α-helix, also yield slightly stronger than expected PREs in both N- and C-terminal directions. For the spin-labeled K48C, S57C, and K63C mutants stronger PREs effects are observed throughout the entire region between residues ~40−70. Therefore, this region, composed of the native β-strands 3−5, is significantly more compact than expected for a random coil. Finally the PREs for mutant R74C mostly match the random coil curve, consistent with the expected high flexibility at the C-terminal end. Nevertheless, also some deviations are observed at residues 58 and 60.

Ensemble Averages over Few Conformers Suffice to Match the Experimental RDC and PRE Data for Urea-Denatured Ubiquitin. Based on purely local structural accessibility, e.g. of torsion angles,[45] a polypeptide chain has access to an astronomically large number of states. This is clearly not realistic due to the fast folding times of proteins,[45] and thus a bias toward native state must exist even in the unfolded ensemble.[46] With the recent increase of NMR high-resolution RDC and PRE data on the unfolded state, the question arises as to what extent these experimental observables restrict the total number of accessible conformations.

To analyze this information content for the urea-denatured state of ubiquitin, 419 RDCs (nine different types) determined previously[47,48] and the 253 PREs shown in Figure 3 were incorporated as constraints in XPLOR-NIH calculations representing ensembles of varying size and using the newly developed sardcPot and prePotD modules. Each calculation was carried out at least ~100 times for ensemble sizes between 1 to 12 with randomly assigned initial velocities. The resulting 50 lowest energy structural ensembles were selected for further analysis. The top and bottom panels of Figure 4 show the agreement between measured and predicted RDC and PRE values, respectively. It is obvious that for ensemble sizes of 10, perfect agreement is obtained within experimental error. Details of the calculations are listed in Table 1 together with the normalized χ^2 values of the deviations between measured and calculated RDC and PRE data according to

$$\chi^2_{\mathrm{RDC/PRE}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i^{\mathrm{obs}} - y_i^{\mathrm{calc}}}{\sigma_i}\right)^2 \qquad (12)$$

[43] Zerella, R.; Evans, P. A.; Ionides, J. M.; Packman, L. C.; Trotter, B. W.; Mackay, J. P.; Williams, D. H. *Protein Sci.* **1999**, *8*, 1320–1331.

[44] Jourdan, M.; Searle, M. S. *Biochemistry* **2000**, *39*, 12355–12364.

[45] Levinthal, C. *Mossbauer Spectroscopy in Biological Systems: Proceedings of the University of Illinois Bulletin* **1969**, *67*, 22–24.

Table 1. Statistics of Ensemble Calculations with RDC and PRE Restraints Using Different Ensemble Sizes[a]

size[b]	structures[c]	χ^2_{RDC}[d]	χ^2_{PRE}[d]	R_g (Å)[e]
1	50	10.83 ± 1.19	25.90 ± 3.90	15.92
2	100	3.75 ± 0.68	8.15 ± 1.22	15.86
3	150	1.50 ± 0.47	3.77 ± 0.51	16.47
4	200	0.94 ± 0.40	2.34 ± 0.37	16.59
5	250	0.77 ± 0.29	1.67 ± 0.49	18.36
6	300	0.66 ± 0.23	1.27 ± 0.73	19.52
7	350	0.54 ± 0.22	1.00 ± 0.54	20.77
8	400	0.53 ± 0.16	1.07 ± 0.98	22.57
9	450	0.39 ± 0.21	1.18 ± 0.92	25.15
10	500	0.40 ± 0.23	1.07 ± 0.90	26.32
11	550	0.41 ± 0.23	0.85 ± 0.47	26.03
12	600	0.43 ± 0.18	0.80 ± 0.82	26.55
RANDOM[f]	500			37.14

[a] With the exception of the RANDOM ensemble, for every ensemble size a total of 100 ensembles was calculated. Analysis was then performed on the 50 lowest energy ensembles. [b] Number of conformers in the ensemble. [c] Total number of individual conformer structures used. [d] Entries correspond to averages and standard deviations of the ensemble values. [e] R_g is determined from the root-mean-square average over all structures. [f] The RANDOM ensemble was derived from 100 ensembles of 10 conformers calculated without RDC and PRE restraints.

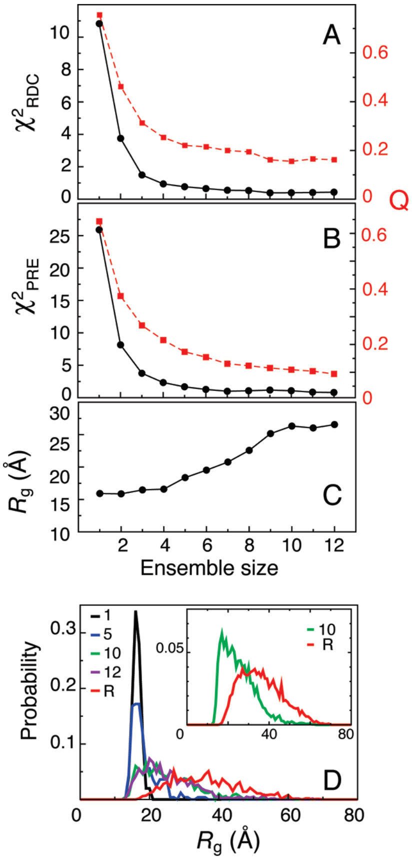

Figure 5. Convergence of ensemble calculations as a function of ensemble size. Average χ^2 (black solid lines) and Q-factors (red dashed lines) of RDCs (A), PREs (B), and radius of gyration R_g (C) vs the number of conformers in the ensemble. Data correspond to entries in Table 1 obtained for 100 calculated ensembles using RDC+PRE restraints. Average Q-factors for all RDCs are calculated as the rms from Q-factors for individual RDC types (Supporting Information, Figure S1). The normalized distributions of R_g are shown in panel D. The numbers indicate the size of ensemble; 'R' represents the restraint-free (RANDOM) ensemble. The inset shows the R_g distribution of the 10-conformer ensemble obtained with RDC+PRE (green) or without any (red, RANDOM) experimental restraints for sets of a total of 4000 calculated structures.

where y_i is the experimental observable, σ_i the experimental error, and the summation runs over all observed data N. Panels A and B of Figure 5 show these normalized χ^2 values for the RDC and PRE data as a function of ensemble size. Initially the χ^2 values decrease very rapidly both for RDCs and PREs when the ensemble size is increased from 1 to 4. Beyond this size, a slower continuous decrease is observed. For the RDCs, the normalized χ^2 value equals about 1 for an ensemble size of 4, whereas for PREs such low values are reached at slightly larger ensemble sizes of about 6−7. Thus, for such an ensemble size the average deviation equals the experimental error, and larger ensemble sizes with lower χ^2 values would overfit the data.

For comparison, we have also analyzed the deviations between measured and predicted RDCs and PREs by means of the Q-factor [= rms($y^{obs} - y^{calc}$)/rms(y^{obs})],[49] which is often used for folded NMR structures (Figures 5A and B). Both Q-factors have very similar behavior, i.e. dropping rapidly to a plateau of about 0.2, when the ensemble size is increased from 1 to about 5, and slowly decreasing further. The behavior of the individual Q-factors for RDCs (Supporting Information, Figure S1) varies to some extent: whereas $^1D_{HN}$, $^1D_{C\alpha H\alpha}$, $D_{HNiH\alpha i-1}$, $D_{HNiHNi+1}$ quickly converge to values below 0.2 for ensemble sizes of 1−3 conformers, $^1D_{C\alpha C'}$, $D_{HNiH\alpha i}$, $D_{HNiHNi+2}$, $D_{HNiHNi+3}$, $D_{HNiHNi+4}$ reach such low values only for larger ensemble sizes. Empirical Q-values derived from SVD-calculated orientation tensors for typical structures range between 0.2 and 0.3.[49] Since the experimental errors of RDC and PRE detection are very small, and their dependence on geometry is straightforward, such SVD-derived Q-values are dominated by errors in the description of a folded structure by a single static structural model. This may be due to intrinsic dynamics or to lesser extent due to inaccuracies of the assumed idealized geometries. In contrast, a Q-value lower than 0.1 has been observed for protein G using an orientation tensor predicted from steric alignment in a bicelle medium.[36] This indicates that the steric alignment model can be very accurate in favorable cases of very rigid proteins and an ideal, nonspecific interaction with the alignment medium. In our model description of an unfolded protein by multiple conformers, the dynamical averaging should be modeled by distinct instances of the structure, whereas errors in covalent geometry should be small, since our RDCs comprise a large number of long-range and heavy atom RDCs (e.g., $D_{HNH\alpha}$, D_{HNiHNj}, $^1D_{C\alpha C'}$). Thus, possible inaccuracies of our model may be due rather to unknown details of the alignment mechanism for RDCs and the incomplete knowledge of dynamics for PREs.

Figure 5C also shows the average radii of gyration (R_g) for the different ensemble sizes. In contrast to the PRE and RDC χ^2 values, R_g converges more slowly from initial values of about 16 Å for ensemble sizes of 1−3 to values of 26−27 Å for sizes ≥10. Experimental R_g values for urea-denatured ubiquitin have been determined from SAXS (28.0 ± 3.5 Å;[50] 25.2 ± 0.2[51]) and SANS (D_2O, 32.5 ± 2.0 Å[50]). These values contain

(46) Zwanzig, R.; Szabo, A.; Bagchi, B. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 20–22.
(47) Meier, S.; Grzesiek, S.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 9799–9807.
(48) Meier, S.; Strohmeier, M.; Blackledge, M.; Grzesiek, S. *J. Am. Chem. Soc.* **2007**, *129*, 754–755.
(49) Cornilescu, G.; Marquardt, J.; Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 6836–6837.
(50) Gabel, F.; Jensen, M. R.; Zaccaï, G.; Blackledge, M. *J. Am. Chem. Soc.* **2009**, *131*, 8769–8771.

Table 2. Population of C^α-C^α Contacts from Constrained Ensemble Calculations[a]

data set	all contacts[b]	all/1D3Z contacts (%)[c]	native contacts[d]	native/1D3Z contacts (%)[e]	native/all contacts (%)[f]	R_g (Å)
RANDOM	5.80	3.9	1.22	0.8	21.0	36.95
RDC	14.75	10.0	3.59	2.4	24.3	30.82
PRE	25.99	17.6	3.35	2.3	12.9	33.56
RDC+PRE	29.73	20.1	4.49	3.0	15.1	25.99
RDC+PRE (16 conf.)[g]	26.87	18.2	4.49	3.0	16.7	30.82
NATIVE (1D3Z)[h]	148.00	100.0	148.00	100.0	100.0	11.73

[a] Unless indicated, data are obtained from the 400 lowest energy 10-conformer ensembles of 800 ensembles calculated. They thus comprise 4000 individual structures. Contacts are defined for $C^\alpha-C^\alpha$ distances smaller than 8 Å. [b] Average total number of contacts observed, defined as the number of all contacts in all individual structures divided by the number of structures. [c] Average total number of contacts relative to the total number of contacts in the native state structure 1D3Z. [d] Average number of native contacts observed, defined as the number of all native contacts in all individual structures divided by the number of structures. [e] Average number of native contacts relative to the total number of contacts in the native state structure 1D3Z. [f] Number of native contacts relative to the total number of contacts in an individual data set. [g] Data are derived from the 250 lowest energy 16-conformer ensembles of 500 ensembles calculated. They thus comprise 4000 individual structures. [h] Data are derived from the first entry of the folded native NMR structure 1D3Z.

contributions of 1−2 Å from the hydration shell, which are positive for SAXS, but negative for SANS in D_2O.[52,53] This indicates that our R_g values calculated for the smaller ensemble sizes are clearly too compact, whereas R_g values for the larger ensemble sizes are in reasonable agreement with the scattering data. We attribute this finding to the attractive forces exerted by the PREs. The experimental PREs are $1/r^6$ averages over many different conformations of the unfolded protein, and conformations with short distances will dominate the observed PRE effect, even when their population is low.[54] However, in the unfolded ensemble these short distances may not belong to the same molecule. When this situation is simulated by a model ensemble with too few conformers, overtightening occurs, because noncompatible PREs are fitted in the same molecule.

Distributions of the radii of gyration for the different ensemble sizes are shown in Figure 5D. Although the radii of gyration of structures from the 10-conformer calculations using PRE and RDC constraints are broadly distributed, they are still considerably more compact than the distribution for a structural ensemble calculated without these restraints (Figure 5D, Table 1, average R_g 37.1 Å). To obtain a better statistical sampling for the 10-conformer ensemble, its R_g distributions with and without experimental restraints were also determined from a much larger set of 4000 calculated ensemble structures (Figure 5D, inset). It is gratifying to see that the distribution fulfilling the experimental constraints has as a form that is very similar to a distribution obtained from a combined analysis of SAXS and NMR diffusion data on the unfolded drkN SH3 domain.[53]

$C^\alpha-C^\alpha$ Contact Maps Reveal Statistically Significant Conformational Propensities for the Native State. To obtain insights into the structural distributions represented by the various ensembles, the results were analyzed in terms of $C^\alpha-C^\alpha$ contacts. Using the 400 lowest energy 10-conformer ensembles from a total of 800 calculated, the contact probability p_{ij} between residues i and j was determined as the total number of contacts observed with $C^\alpha-C^\alpha$ distances smaller than 8 Å divided by the total number of structures. Figure 6 shows these contact maps for ensembles calculated by using only RDC, only PRE or the combined RDC+PRE restraints. For comparison also contact maps are shown for an ensemble calculated without restraints (RANDOM) as well as from the 1D3Z NMR structure (NATIVE). The upper left parts of the contact maps depict all contact probabilities, whereas in the lower right part only the contacts are shown that belong to the native state. Statistics of the contacts are given in Table 2.

Excluding terminal effects, significant $C^\alpha-C^\alpha$ contacts ($p_{ij} >$ 0.5%) are observed for the RANDOM ensemble only for residue separations $\Delta_{ij} = |i - j|$ smaller than 6. About 21% of these contacts correspond to 0.8% of the 148 native-state $C^\alpha-C^\alpha$ contacts (Table 2). Using only RDCs as constraints, the total number of $C^\alpha-C^\alpha$ contacts increases considerably by about 2.5 times, and contacts become significantly populated to residue separations of up to $\Delta_{ij} = 9$. In particular, also the population of native-state contacts increases to 2.4%. It is striking that the native contacts become rather strongly populated in the region of the first β-turn (up to 21%, residues 4−12) and also in the α-helix (up to 14%, residues 21−38). Using only PREs as constraints, the total number of $C^\alpha-C^\alpha$ contacts is about 1.8 times larger than in the RDC ensemble. This is not surprising due to restriction in distances afforded by the PRE potential. The contacts extend over a much wider range of residue distances, and the population of native-state contacts is similar to the RDC ensemble (2.3%). The native contacts are again strongly populated in the first β-hairpin (up to 7%), in the α-helix (up to 13%), but also at turns at residues 41, 49, 54, 58, and 65 (up to 14%). Many of the longer-range contacts (Δ_{ij} > 10) are non-native. However, native long-range contacts are also populated to about 1%, e.g. between the irregular region around residue 53 and the start of the α-helix around residue 23. Finally, in the combined RDC+PRE ensemble, the number of $C^\alpha-C^\alpha$ contacts is still slightly increased by about 14% relative to the PRE ensemble. The total population of native contacts now amounts to 3.0%, and many of the previously observed native contacts are increased and reach up to 20% in the first β-turn and up to 17% in the middle of the α-helix (red contacts).

In summary, it is clearly evident that with an increasing number of experimental restraints, the number of native contacts increases strongly, i.e. from 0.8% for the RANDOM ensemble to 3.0% for the RDC+PRE ensemble. In this respect it is interesting to observe that there is considerable overlap between the native contacts induced solely either by PRE or RDC constraints. Therefore, both observables give independent

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

(52) Svergun, D. I.; Richard, S.; Koch, M. H.; Sayers, Z.; Kuprin, S.; Zaccai, G. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 2267–2272.

(53) Choy, W. Y.; Mulder, F. A.; Crowhurst, K. A.; Muhandiram, D. R.; Millett, I. S.; Doniach, S.; Forman-Kay, J. D.; Kay, L. E. *J. Mol. Biol.* **2002**, *316*, 101–112.

(54) Tang, C.; Iwahara, J.; Clore, G. M. *Nature* **2006**, *444*, 383–386.

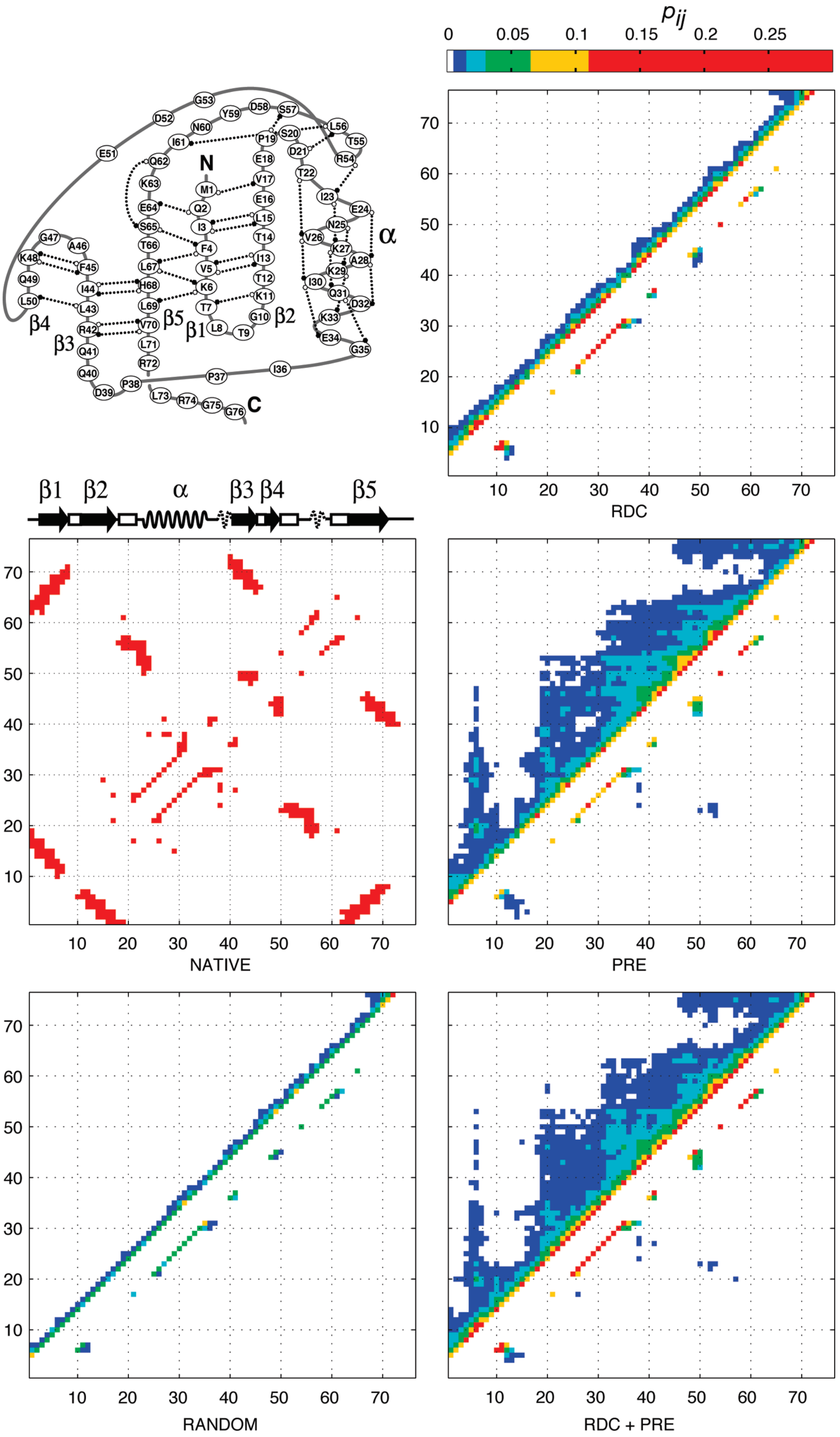

Figure 6. $C^{\alpha}-C^{\alpha}$ contact probability maps of urea-denatured ubiquitin derived from the ensemble calculations. The contact probability p_{ij} between residue i and j was determined as the total number of contacts observed with $C^{\alpha}-C^{\alpha}$ distances smaller than 8 Å divided by the total number of structures. The size of p_{ij} is color-coded as: white 0−0.5%, dark blue 0.5−1.5%, light blue 1.5−3%, green 3−6.5%, yellow 6.5−11%, and red ≥11%. Contact maps are shown for ensemble calculations (10 conformers, 400 ensembles) with no experimental restraints (RANDOM), RDC only, PRE only, and RDC+PRE. For comparison, the contact map of the native structure (NATIVE) is also indicated. The upper-left parts of the contact maps represent all observed contacts; the lower-right parts indicate only contacts also observed in the native state. Ubiquitin's secondary structure and topology are shown schematically in the upper left panel.

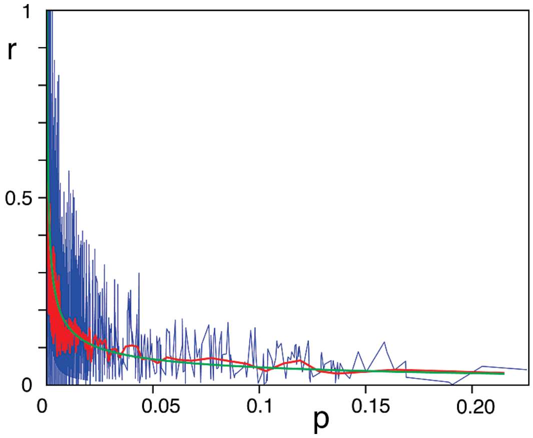

Figure 7. Statistics of $C^{\alpha}-C^{\alpha}$ contacts observed in the 400 lowest energy ensemble calculations of 10-conformer with PRE and RDC restraints. To obtain the relative spread *r* of the observed contact probability *p* was calculated from the standard deviation of the contact probability of the first and the second subsets of 2000 structures. Blue: relative spread *r* of all observed contacts, red: average over 20 observed contacts, green: theoretical curve according to eq 13.

evidence of the formation of certain partial native structures. However, when both constraints are combined, they even act synergistically to drive the ensemble to an even higher number of native contacts.

Since these results are obtained on a total of 400 10-conformer ensembles, the question arises as to whether the limitation to 10-conformers does not artificially restrict the search for more diverse states by the mutual dependence of the substructures. We have tried to assess this question by the comparison of the 10-conformer results to the results from a further set of 250 16-conformer ensembles selected for lowest energy from a total of 500 calculated under RDC+PRE restraints. The $C^{\alpha}-C^{\alpha}$ contact map is almost unchanged relative to the 10-conformer results (Supporting Information, Figure S2). However, the total number of contacts per structure is slightly decreased from 29.73 to 26.87 (Table 2). This is caused by an increase in the radius of gyration from 25.99 to 30.82 Å. Apparently, the larger number of conformers "dilutes" the constricting effect of the PRE restraints, such that the structural ensemble becomes wider. Nevertheless, the number of native contacts of 4.49 per structure remains unchanged as compared to the 10-conformer result of 4.49. We thus conclude that our results for the 10-conformer ensembles are relatively robust with respect to moderate variations of the conformer numbers. Much larger numbers of conformers may drive the ensembles toward too large radii of gyration, since this parameter is not controlled in the calculations.

Statistics of Contacts. It is of interest to consider the statistical significance of the contacts obtained from the ensemble calculations. Assuming that a certain contact is formed randomly within any of the *N* calculated structures with a probability *p*, then the relative spread *r* of the observed contacts, defined as the root-mean-square deviation of the observed contacts *n* divided by the average number of observed contacts *n*, is given according to the binomial distribution (see, e.g., Reif[55]):

$$r = \frac{\langle(\Delta n)^2\rangle^{1/2}}{\langle n\rangle} = \sqrt{\frac{1-p}{pN}} \qquad (13)$$

Figure 7 shows the observed relative spread *r* as a function of the probability *p* for all contacts of the PRE+RDC ensemble (400 lowest energy 10-conformer ensembles). To obtain *r*, the total of 4000 structures was divided into two subsets of 2000 structures each, and the standard deviation of *n* was calculated from the two subsets. The theoretical spread of contacts *r* (green) according to eq 13 agrees well with the observation (blue, red). Thus, e.g. the relative error for a contact population of 10% is about 5%, but increases to about 15% for a contact population of 1%. From eq 13, the necessary number of structures for observing a contact with a certain confidence can be estimated: for example a 30% relative error for a 1% contact population would require the calculation of 1100 structures.

Discussion

New RDC and PRE Modules for Ensemble Structure Calculation. RDCs so far have not been used as direct constraints in structure calculations of unfolded proteins due to the difficulty of obtaining accurate orientation tensors at every step of the calculation. An attempt to interpret measured RDCs for unstructured proteins was proposed by Sosnick, Blackledge, and their co-workers.[27,28] In this approach, a large ensemble of structures (tens of thousands) of the protein of interest is created according to the ϕ/ψ propensities in the coil part of the PDB, while steric clashes are excluded. Theoretical RDCs are then calculated as an average over all structures by using steric alignment. These average values agree in their trends with experimental $^1D_{HN}$ RDCs, and local deviations have been taken as evidence for residual structure.[16] Using a larger set of RDCs comprising also $^1D_{C\alpha H\alpha}$, $^1D_{C\alpha C'}$, $D_{HNiH\alpha i}$, $D_{HNiH\alpha_{i-1}}$, and D_{HNiHNj} data for urea-denatured ubiquitin,[17] it was found that different scaling factors were necessary for the different types of RDCs. This was explained by more extended conformations of the protein backbone than predicted by the coil model.

Here, we have overcome the computational difficulty in structure calculations of unfolded proteins using RDCs by an efficient algorithm for determining the steric alignment tensor,[37] such that restrained MD trajectories even for ensembles of the size of about 10 can be calculated on a single CPU in a few hours. This makes it possible to assess in an objective way the information content of the measured RDCs, i.e. their ability to restrain an unfolded protein model ensemble. In addition, the RDCs can now easily be combined in the simulations with all other experimental data that give information on the unfolded state comprising PREs, scalar couplings, NOEs, diffusion coefficients, radii of gyration, chemical shifts, and others.

For the combination with RDC data, we have also implemented a new module for PRE constraints in ensemble calculations that is optimized for unfolded proteins. Several computational methods have been developed in the past for the interpretation of PREs. Shortle and co-workers[9] have proposed a simulated annealing protocol on single conformers, which uses distances from PRE intensity ratios with loose upper and lower bounds to prevent over-restraint. Structural propensities are then obtained by clustering the lowest energy structures with similar local conformations. This method has been applied to denatured staphylococcal nuclease Δ131Δ,[9] α-synuclein,[21] and the γ-subunit of cGMP phosphodiesterase.[26] Alternatively, Vendruscolo and co-workers have introduced MD ensemble averaging with PRE distance restraints from peak intensity ratios to characterize the denatured state of the bovine acyl-coenzyme A binding protein,[20,22] α-synuclein,,[23] and denatured Δ131Δ.[24] A different approach has been pursued by Forman-Kay and co-workers who selected an ensemble of conformations for unfolded drkN SH3

[55] Reif, F. *Fundamentals of Statistical and Thermal Physics*; McGraw-Hill: New York, 1965.

according to PRE data from a pregenerated pool of structures.[25] Our use of PRE constraints in MD calculations of unfolded protein ensembles is very similar to the Vendruscolo procedure with the difference that structures are constrained against measured ΔR_2 values and not against derived distances. Similar to the RDC module, this PRE ensemble energy function achieves a physically reasonable approximation of the unfolded protein ensemble, easily integrates with other constraints and allows for a simple evaluation of the information content of the experimental data.

Convergence Properties and Statistics of the Simulated Ensembles. Despite the high number of 419 experimental RDCs, ensemble averages over about four conformers can reproduce the data within experimental error. Similarly, averages over about seven conformers are completely sufficient to reproduce the 253 PREs. This behavior is not surprising since the degrees of freedom of a single conformer of ubiquitin (76 aa) counting only ϕ and ψ angles would be 150. Thus, the degrees of freedom for three conformers already exceed the number of RDC restraints. The slightly larger ensemble sizes required for PREs may be caused not only by the longer-range nature of these constraints and their increased restriction of the topology, but also by the uncertainties of the method from the unknown rotational correlation time of individual PRE contacts. In this context, it is important to observe that the convergence of the radius of gyration, requiring about 10 conformers for a stable value, is slower than the convergence toward the RDC or PRE constraints (Figure 5). A smaller number of conformers apparently leads to overtightening caused by conflicting PREs. Thus, the radius of gyration is clearly a very important indicator of the structural quality of the ensemble, and the good agreement of this parameter for the 10-conformer ensemble and the SAXS value gives additional credibility to the ensemble calculations.

The fact that all experimental constraints can be satisfied by one single ensemble containing such a low number of conformers does not mean that such an ensemble is a unique and complete description of the unfolded state; it only describes the information content of the experimental data. In fact, many different ensembles are possible that satisfy the constraints equally well. Apparently the conformational space, which is accessible to these small, constrained ensembles, is not overly restricted by an artificial, mutual dependence of substructures, since the contact maps of the 10- and 16-conformer ensembles are very similar. However, larger conformer numbers lead to more extended ensembles, since the constricting force of the PRE is diluted and no other attractive forces were used in the calculations, which would confine the radius of gyration.

The nature of the sampling problem of the entire unfolded state ensemble is revealed by the statistics of C^{α} contacts from a large number of ensembles. The contact statistics agree with simple random sampling according to eq 13 at least up to populations of 20%. This makes it possible to predict the ensemble size needed to sample a certain contact population. Thus, about 1000 calculated structures are necessary to observe a 1% population with confidence. Conversely, in our pool of 4000 calculated structures of urea-denatured ubiquitin, the large number of native C^{α} contacts with populations of 1% or higher is clearly highly significant.

It is interesting to compare these sampling properties with the coil model approach to interpret RDCs implemented e.g. in the Flexible-Meccano program.[28] In the latter case, typically[17] averages over tens of thousands of structures are needed to obtain sufficient convergence. This is not in contradiction to the results presented here. In the coil model, no constraints are used to drive the ensemble toward the experimental data, and thus a much wider portion of conformational space is being sampled. Recent data (M. Blackledge, personal communication) indicate that smaller subsets of the coil ensembles can be selected that agree with the experimental data to an extent similar to that of the full-size ensembles. Our results here show that the minimal size of such subsets should be on the order of 10 for the present RDC, PRE, and R_g information on urea-denatured ubiquitin.

Comparison of RDC and PRE Restraints. Both RDC and PRE restraints are affected by inherent uncertainties. RDCs of the unfolded ensemble could be biased by the interactions with the acrylamide gel alignment medium,[6] and the model of a fixed steric alignment tensor may be inaccurate for conformers that interchange very rapidly during the contact event with the medium. The present approach will try to approximate such experimental data by a distinct set of conformers, similar to a series expansion of a function. For PREs, the labeling by the paramagnetic tag can induce artificial intraprotein interactions, and the dependence of PREs on the rotational correlation time of the electron−nucleus vector (eq 11) introduces an additional uncertainty. In principle, the correlation time could be different for every PRE and also for every member of the ensemble. No method is available to determine this parameter reliably for an unfolded protein ensemble. As in other studies[12,21,41,42] we have used the compromise of a uniform value of 4 ns, which yields radii of gyration that are close to the SAXS value.

Despite these uncertainties, the more highly populated contacts in the ensembles derived from either RDC or PRE constraints largely coincide, and these contacts are even increased for the RDC+PRE ensemble (Figure 6). Thus, the present experimental and computational methods appear to converge toward a consistent picture of the unfolded ensemble at least for subpopulations that exceed about 5% of the total ensemble.

The Urea-Denatured State of Ubiquitin and the Folding to the Native Structure. A 10−25% native-state conformation of the first β-hairpin (residues 1−20) has been detected previously in urea-denatured ubiquitin from long-range RDC contacts, chemical shifts, and hydrogen bond scalar couplings.[18] These findings are corroborated by the long-range PRE contacts across the first β-hairpin in the K6C mutant to residues in the second β-strand, and vice versa in the S20C mutant to residues at the N-terminus (Figure 3). The ensemble calculations fully reproduce this population estimate of the native conformation in this region. Using only RDCs, only PREs, or both together, the population of native β-hairpin C^{α} contacts amounts to about 10−20% (Figure 6).

For the following native-state α-helix (residues 22−35), no particularly strong structural propensities had been noticed in the previous RDC study.[18] The new PRE results give an indication of a locally more compact structure in this region. Thus, PREs of the spin-labeled S20C mutant to residues in this region are stronger than expected for a random coil, and also the mutants K35C and G35C show stronger contacts toward the N-terminal region of the α-helix (Figure 6). The ensemble calculations with RDC, PRE, or RDC+PRE restraints now clearly reveal native-state α-helical *i,i*-4-contacts on the order of 10−15% in this region.

The PRE results in the C-terminal half of urea-denatured ubiquitin, i.e., mutants K48C, S57C, and K63C, indicate a somewhat more compact conformation than expected for a random coil (Figure 3). With the exception of a helical turn around residue 58, calculated C^{α} contacts in the C-terminal half of ubiquitin are much less native-like than in the N-terminal half (Figure 6). In particular, a large number of significant (p_{ij} > 11%) non-native, helical *i,i*-4-contacts are observed between residues 40 to 60.

Throughout the backbone, the most strongly populated (p_{ij} > 11%) C^{α} contacts correlate closely with structural propensities of isolated peptide fragments and of ubiquitin's A-state. Thus, a peptide comprising the first 17 residues of ubiquitin was found to fold in water at low temperatures to a β-hairpin structure with a population in the range of tens of percent.[43] Similarly, a fragment of residues 21−35 (α-helix) had an α-helical population of about 3% in water.[44] The addition of methanol at low pH increased significantly this content of native β-hairpin and α-helix for fragments of residues 1−21 and 1−35.[56] In contrast, an isolated C-terminal peptide (residues 36−76) showed no particular secondary structure propensity in water, while the addition of methanol induced non-native α-helical structures.[44] In many aspects, the behavior of these fragments is similar to the A-state of full-length ubiquitin observed at low pH and high methanol content (~60%).[57] NMR data for the A-state[58,59] show that the first β-hairpin and the α-helix are preserved and highly populated, whereas the C-terminal part adopts an all α-helical structure, and flexible sequence elements connect these three partial structures. Therefore, in essence, the most strongly populated conformations of the calculated urea-denatured ensemble, which indicate the β-hairpin and α-helix at native positions in the N-terminal and many α-helical turns in the C-terminal half, are identical to those of the A-state.

Besides these highly populated conformations, also many other weaker contacts are apparent that do not belong to either the native or the A-state (Figure 6). The simultaneous presence of native and non-native contacts has also been observed in previous PRE studies on unfolded proteins.[20,60] This behavior is a simple consequence of the limited volume available: since R_g is smaller than that for a completely extended protein, many non-native contacts must be present as long as the protein is not in its native folded state. For urea-denatured ubiquitin, the NMR data also provide information on the dynamics of all these conformations: since NMR spectra only show the presence of one single spectral species, the native, A-state and non-native conformations are in fast exchange with each other on the time scale of the chemical shift (~milliseconds). Thus, the search for the lowest energy conformer can be achieved very efficiently.

It is important to observe that the highest contact populations in the calculated ensemble are A-state-like and already encompass larger secondary structure elements. This is consistent with a hierarchical principle in which 'folding begins with structures that are local in sequence and marginal in stability; these local structures interact to produce intermediates of ever-increasing complexity and grow, ultimately, into the native conformation'.[61] The present ensemble calculations make it possible to observe details of this hierarchical folding principle. The secondary structures of isolated peptides, of the A-state, and of the highest contact populations in the urea-denatured form are native-like at the N-terminus, but non-native at the C-terminus. Apparently, for the complete formation of the native structure, the contacts in the hydrophobic core are missing. This is the case for the peptide fragments and for the A-state, where the hydrophobic contacts are weakened by the addition of methanol. For mechanisms of denaturation by urea, both disturbance of electrostatic/hydrogen bond interactions and weakening of hydrophobic interactions are being discussed. Recent experimental[62] and MD dynamics[63,64] data give some evidence that urea indeed weakens the hydrophobic interactions by disturbing the hydration shell of the protein. The preferences for A-state-like conformations in the calculated urea-denatured ensembles of ubqitin are in agreement with such a weakening of hydrophobic interactions. Conversely, the removal of urea and the concomitant increase of the hydrophobic contacts will then ultimately favor the formation of a β-sheet over the A-state α-helix in ubiquitin's C-terminal half and close the entire structure into the very compact native form. This sequence of folding events is also suggested by a ϕ value analysis, which shows that the first β-hairpin and the α-helix are highly structured in the transition state ensemble.[65]

Conclusion

We have developed two new modules that are able to incorporate steric alignment RDC and PRE data into ensemble structure calculations of unfolded proteins. This enables us to assess the information content of experimental RDC and PRE data in the sense of constraining an unfolded protein ensemble. Despite a high number of experimental data for urea-denatured ubiquitin, only very small ensembles are needed to achieve good agreement. As exemplified by the C^{α} contacts, the statistics of a large set of calculated ensembles follow simple random sampling. The analysis of these ensembles shows that significant subpopulations on the order of few percent exist, which have the characteristics of ubiquitin's A-state. Thus, the urea-denatured state is clearly not a simple random coil. This is consistent with a hierarchical folding model,[61] where larger, more stable substructures emerge from favorable local interactions. The present RDC, PRE, and computational methods allow the reliable detection of such subconformations at population levels of a few percent.

Material and Methods

Sample Preparation. Eight cysteine mutants (K6C, S20C, K33C, G35C, K48C, S57C, K63C, and R74C) of ubiquitin were constructed by using the QuikChange site-directed mutagenesis kit (Stratagene, La Jolla, CA). All mutants were verified by DNA sequencing. ^{15}N-labeled ubiquitin mutants were purified as described before[66] and kept as frozen stock solutions of 300 μM protein, 10 mM phosphate, 5 mM TCEP. The nitroxide spin label MTSL (1-oxy-2,2,5,5-tetramethyl-D-pyrroline-3-methyl)methanethiosulfonate (Toronto Research Chemicals, Toronto) was attached to the

[56] Cox, J. P.; Evans, P. A.; Packman, L. C.; Williams, D. H.; Woolfson, D. N. *J. Mol. Biol.* **1993**, *234*, 483–492.

[57] Wilkinson, K. D.; Mayer, A. N. *Arch. Biochem. Biophys.* **1986**, *250*, 390–399.

[58] Brutscher, B.; Brüschweiler, R.; Ernst, R. R. *Biochemistry* **1997**, *36*, 13043–13053.

[59] Cordier, F.; Grzesiek, S. *Biochemistry* **2004**, *43*, 11295–11301.

[60] Lietzow, M. A.; Jamin, M.; Jane Dyson, H. J.; Wright, P. E. *J. Mol. Biol.* **2002**, *322*, 655–662.

[61] Fitzkee, N. C.; Fleming, P. J.; Gong, H.; Panasik, N.; Street, T. O.; Rose, G. D. *Trends Biochem. Sci.* **2005**, *30*, 73–80.

[62] Chen, X.; Sagle, L. B.; Cremer, P. S. *J. Am. Chem. Soc.* **2007**, *129*, 15104–15105.

[63] Hua, L.; Zhou, R.; Thirumalai, D.; Berne, B. J. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 16928–16933.

[64] Zangi, R.; Zhou, R.; Berne, B. J. *J. Am. Chem. Soc.* **2009**, *131*, 1535–1541.

[65] Went, H. M.; Jackson, S. E. *Protein Eng., Des. Sel.* **2005**, *18*, 229–237.

[66] Sass, J.; Cordier, F.; Hoffmann, A.; Cousin, A.; Omichinski, J. G.; Lowen, H.; Grzesiek, S. *J. Am. Chem. Soc.* **1999**, *121*, 2047–2055.

thiol group of the cysteine by the following procedure. After buffer exchange by ultrafiltration (Vivaspin, Satorius Stedim Biotech) and size exclusion chromatography (PD-10 Column, Amersham Pharmacia Bioscience) to 10 mM phosphate, pH 5.0, typically 3.5 mL of 40 μM protein was reacted with 100 μL of 15 mM MTSL (10-fold molar excess) dissolved in acetonitrile for 30 min. All MTSL-labeled samples were verified by mass spectrometry (MICROTOF system, Bruker), and no mass of dimeric or nonlabeled protein was detected. The buffer of the samples was then exchanged by ultrafiltration to 8 M urea, 10 mM Glycine-HCl, pH 2.5, 10% D_2O/90% H_2O, yielding a final ubiquitin concentration of 300 μM for the NMR samples (450 μL). To obtain diamagnetic references, MTSL was quenched by the addition of 5 μL of a solution of 150 mM ascorbic acid, 10 mM glycine-HCl, 8 M urea, pH 2.5 to the MTSL-labeled samples. Samples were then stored overnight at room temperature to ensure complete reduction of the spin label. Alternatively, ubiquitin mutant samples were prepared in the absence of MTSL under identical buffer conditions, except that 1 mM TCEP was added to prevent oxidation. ^{15}N R_2 rates determined for quenched MTSL-labeled ubiquitin mutants and untagged mutants were found to be identical within less than 0.1 Hz for all residues for which PRE could be observed, indicating that the MTSL-label had no significant influence on the structural dynamics of the unfolded state.

NMR Experiments. All NMR data were acquired at 25 °C on a Bruker DRX 800 MHz instrument equipped with a TCI cryogenic probe. ^{15}N-transverse relaxation rates (R_2) were measured by ^{1}H/^{15}N-edited standard experiments[67] using a total experimental time of 12 h for a series of relaxation time delays of 4, 12, 24, 36, 48, and 60 ms. Data were processed by the NMRPIPE suite of programs,[68] and peak intensities derived by the NMRPIPE-NLINLS program were fitted to exponential decays using in-house written MATLAB (MathWorks, Inc.) Monte Carlo procedures. Statistical errors in relaxation rates were obtained from repeated measurements ($N = 2$).

Structure Calculation and Analysis. The sardcPot and prepotD modules for ensemble structure calculation were coded in C++ and incorporated into XPLOR-NIH version 2.23 with a python interface.[33] These modules are available upon request and will be included in future XPLOR-NIH distributions (http://nmr.cit.nih.gov/xplor-nih). An extended starting structure for MTSL-labeled ubiquitin was created and parametrized by using a standard script contained in the XPLOR-NIH distribution. Using the ensemble calculation feature, this starting structure was copied N times according to the ensemble size and then individually randomized by torsion angle dynamics at 3000 K for 50 ps with bond, angle, and improper energy potentials. The following simulated annealing protocol using the PRE and RDC target functions was modified from the one used by Iwahara et al.[34] starting with a 10 ps dynamics run at 3000 K, followed by cooling from 3000 to 300 K in 12.5 K decrements. At each decrement a 1.5 ps dynamics run was inserted for equilibration. In addition to the RDC and PRE energies, standard potentials were used for describing the covalent structure (bonds, angles, improper torsions, and steric repulsion from van der Waals) and the Ramachandran energy surface. Final Powell minimizations were performed first in torsion-angle space and then in Cartesian space. All torsion-angle dynamics was performed using the internal variable module (IVM)[69] of XPLOR-NIH. A typical complete simulated annealing run took 15−30 min per single conformer on an Opteron 2.6 GHz CPU. The calculation time scaled approximately in a linear way with the ensemble size. Thus, the calculation of a total of 800 ensemble structures with 10 conformers took approximately 2 days using ∼80 CPUs of a Linux Beowulf-Cluster. Obtained ensembles were analyzed by MATLAB (MathWorks, Inc.) scripts for calculation of the radius of gyration R_g and C$^\alpha$ contact maps. R_g values are calculated as the root-mean-square average over all N structures of the ensemble[53]

$$R_g = \sqrt{\frac{1}{N}\sum_i R_{g,i}^2}$$

This takes into account that the Guinier analysis of SAXS data corresponds to an average over R_g^2 of the individual molecules in the experimental ensemble.

Acknowledgment. We thank K. Rathgeb-Szabo, R. Bobby, and M. Rogowski for protein preparation, Dr. D. Häussinger for help with MTSL-labeling, Dr. C.D. Schwieters for extensive help with XPLOR-NIH programming, Drs. H.-J. Sass and M. Blackledge for many helpful discussions, and the Basel Computational Biology Center as well as Prof. S. Bernèche for providing access to their computer cluster. This work was supported by SNF Grant 31-109712.

Supporting Information Available: Figures showing RDC and PRE Q-values as a function of ensemble size and the C$^\alpha$−C$^\alpha$ contact probability map derived from a set of 250 calculated ensemble structures containing 16 conformers. This material is available free of charge via the Internet at http://pubs.acs.org.

JA907974M

[67] Kay, L. E.; Torchia, D. A.; Bax, A. *Biochemistry* **1989**, *28*, 8972–8979.

[68] Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR* **1995**, *6*, 277–293.

[69] Schwieters, C. D.; Clore, G. M. *J. Magn. Reson.* **2001**, *152*, 288–302.