J|A|C|S ARTICLES

Published on Web 11/12/2009

Quantitative Description of Backbone Conformational Sampling of Unfolded Proteins at Amino Acid Resolution from NMR Residual Dipolar Couplings

Gabrielle Nodet,[†] Loïc Salmon,[†] Valéry Ozenne,[†] Sebastian Meier,[‡] Malene Ringkjøbing Jensen,[†] and Martin Blackledge*[,†]

Protein Dynamics and Flexibility, Institut de Biologie Structurale Jean-Pierre Ebel, CEA, CNRS, UJF UMR 5075, 41 Rue Jules Horowitz, Grenoble 38027, France, and Carlsberg Laboratory, Gamle Carlsberg Vej 10, 2500 Valby, Denmark

Received August 22, 2009; E-mail: martin.blackledge@ibs.fr

Abstract: An atomic resolution characterization of the structural properties of unfolded proteins that explicitly invokes the highly dynamic nature of the unfolded state will be extremely important for the development of a quantitative understanding of the thermodynamic basis of protein folding and stability. Here we develop a novel approach using residual dipolar couplings (RDCs) from unfolded proteins to determine conformational behavior on an amino acid specific basis. Conformational sampling is described in terms of ensembles of structures selected from a large pool of conformers. We test this approach, using extensive simulation, to determine how well the fitting of RDCs to reduced conformational ensembles containing few copies of the molecule can correctly reproduce the backbone conformational behavior of the protein. Having established approaches that allow accurate mapping of backbone dihedral angle conformational space from RDCs, we apply these methods to obtain an amino acid specific description of ubiquitin denatured in 8 M urea at pH 2.5. Cross-validation of data not employed in the fit verifies that an ensemble size of 200 structures is appropriate to characterize the highly fluctuating backbone. This approach allows us to identify local conformational sampling properties of urea-unfolded ubiquitin, which shows that the backbone sampling of certain types of charged or polar amino acids, in particular threonine, glutamic acid, and arginine, is affected more strongly by urea binding than amino acids with hydrophobic side chains. In general, the approach presented here establishes robust procedures for the study of all denatured and intrinsically disordered states.

Introduction

Despite decades of experimental and theoretical advances in the characterization of structure, kinetics, dynamics, and thermodynamics of many thousands of soluble, folded proteins, the mechanism of protein folding, the conformational transition from a flexible unfolded polypeptide chain to a stable folded protein structure, remains largely unexplained.[1] One reason for this is that one side of the protein folding equation is essentially impossible to characterize in atomic detail using classical approaches to structural biology, requiring instead the development of approaches that explicitly invoke the highly dynamic nature of the unfolded state.[2−5] An atomic-resolution characterization of the structural properties of unfolded proteins is therefore an essential prerequisite for a quantitative understanding of the thermodynamic basis of protein folding and stability.

The importance of developing techniques that are capable of describing the conformational sampling of unfolded polypeptide chains in solution has gained further importance with the gradual realization, over the past decade, that a large fraction of eukaryotic genomes codes for proteins that are intrinsically disordered in their native state.[6−9] Of particular relevance is the relationship between intrinsic structural characteristics of the unfolded chain and the mechanisms of protein folding upon binding, underlining the need for a basic understanding of the conformational space that is populated by a protein in the unfolded state.[10,11] The role that intrinsically disordered proteins (IDPs) play in neurodegenerative disease and cancer further emphasizes the importance of understanding conformational transitions from physiological to pathological forms of the same protein.[12]

Nuclear magnetic resonance (NMR) spectroscopy is probably the most powerful biophysical tool for studying IDPs due to

[†] Institut de Biologie Structurale Jean-Pierre Ebel.
[‡] Carlsberg Laboratory.

(1) Dill, K. A.; Shortle, D. *Annu. Rev. Biochem.* **1991**, *60*, 795–825.
(2) Daggett, V.; Fersht, A. R. *Natl. Rev. Mol. Cell Biol.* **2003**, *4*, 497–502.
(3) Vendruscolo, M.; Paci, E.; Karplus, M.; Dobson, C. M. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 14817–14821.
(4) Mittag, T.; Forman-Kay, J. D. *Curr. Opin. Struct. Biol.* **2007**, *17*, 3–14.
(5) Eliezer, D. *Curr. Opin. Struct. Biol.* **2009**, *19*, 23–30.
(6) Uversky, V. N. *Protein Sci.* **2002**, *11*, 739–756.
(7) Dunker, A. K.; Brown, C. J.; Lawson, J. D.; Iakoucheva, L. M.; Obradovic, Z. *Biochemistry* **2002**, *41*, 6573–6582.
(8) Tompa, P. *TIBS.* **2002**, *27*, 527–533.
(9) Fink, A. L. *Curr. Opin. Struct. Biol.* **2005**, *15*, 35–41.
(10) Dyson, H. J.; Wright, P. E. *Curr. Opin. Struct. Biol.* **2002**, *12*, 54–60.
(11) Fuxreiter, M.; Simon, I.; Friedrich, P.; Tompa, P. *J. Mol. Biol.* **2004**, *338*, 1015–1026.
(12) Dobson, C. M. *Trends Biol. Sci.* **1999**, *24*, 329–332.

 10.1021/ja9069024 CCC: $40.75

the remarkable sensitivity of different NMR phenomena to dynamics occurring on time scales varying from picoseconds to hours and the ability to report on both local and long-range structure.[13] In particular, residual dipolar couplings (RDCs), which become measurable when a protein is dissolved in an anisotropic alignment medium or matrix,[14,15] have been shown to be very sensitive reporters of local and long-range structure,[16] even in highly disordered systems.[17] Since the initial demonstration that RDCs can be measured in proteins even under highly denaturing conditions,[18−25] it has been recognized that RDCs provide unique site-specific probes of orientational order in disordered states.[17,26]

A recently developed explicit ensemble description of IDPs, flexible-meccano,[27] constructs multiple copies of the protein in different states, designed to represent all possible conformational states that exchange on time scales relevant to the NMR time scale. Using a statistical coil description that samples amino acid-specific backbone dihedral angle $\{\phi/\psi\}$ propensities, a conformational ensemble is created, and RDCs are calculated for each conformer and then averaged over the ensemble. This approach implicitly assumes that all conformers are in rapid exchange on time scales faster than a millisecond, an assumption based on the presence of a single set of NMR signals detected in ^{1}H and ^{15}N spectra of denatured and intrinsically disordered proteins. The absence of conformational exchange broadening excludes the presence of exchange between significantly populated conformational states occurring on slower time scales. RDCs simulated using these approaches present reasonable agreement with experimental couplings measured in both intrinsically disordered and chemically denatured proteins.[28−32] These studies have been used to provide evidence that site-specific differences in RDCs measured along the primary chain can result from native differences in the rigidity of different amino acid types in an otherwise fully disordered chain,[27] from the presence of transiently populated local secondary structural elements[31] or from the presence of transient interactions between sites distant in the chain.[28]

While $^{15}N-^1H^N$ RDCs are by far the most commonly measured dipolar couplings, for reasons of experimental facility and precision, the advantages of measuring more RDCs from different spin-pairs in the peptide unit were recently demonstrated by Meier et al., who determined up to seven RDCs per amino acid from urea-unfolded ubiquitin at pH 2.5, including $^{15}N-^1H^N$, $^{13}C^\alpha-^1H^\alpha$, and $^{13}C^\alpha-^{13}C'$ RDCs, inter- and intraresidue $^1H^N-^1H^\alpha$ RDCs, and $^1H^N-^1H^N$ RDCs measured using quantitative *J*-type experiments[33] on perdeuterated ubiquitin. In combination, these data indicated that the standard description of the statistical coil behavior was inappropriate for urea unfolded proteins and that a modification of the random coil description was necessary to account simultaneously for all data.[34] On the basis of extensive simulation, the authors proposed that, in the presence of urea, the backbone dihedral angles defining the conformational behavior of the unfolded chain have a significantly higher propensity to sample more extended regions of Ramachandran space ($\psi > 50°$, $\phi < 0°$). This indication is supported by a comparison of extensive experimental small angle scattering (SAS) and pulse field gradient (PFG) dependences measured from urea-denatured proteins, with predicted data from conformational ensembles constructed using statistical coil models sampling increasing levels of this extended region (P. Bernado, personal communication). These independent biophysical techniques concur to substantiate an overall description of conformational bias respected by disordered polypeptide chains in the presence of high concentrations of denaturant.[35−38] RDCs measured between different spins within the peptide unit have also been shown to exhibit complementary dependences on the presence of local structure, an observation that has been shown to be crucial for the quantitative determination of the nature and extent of helical sampling present in molecular recognition elements of intrinsically disordered viral proteins[31] and the disordered N-terminal domain of p53.[39]

These studies have mainly used a rational, hypothesis-based approach, calculating explicit ensembles containing tens of thousands of conformers from different conformational sampling regimes and comparing the ensemble-averaged couplings to experimental data. In this study, we are interested in taking the analysis of RDCs one crucial step further, by investigating the possibility of defining the conformational sampling of the peptide chain directly from the experimental NMR data at amino

(13) Dyson, H. J.; Wright, P. E. *Chem. Rev.* **2004**, *104*, 3607–3622.
(14) Tjandra, N.; Bax, A. *Science* **1997**, *278*, 1111–1114.
(15) Prestegard, J. H.; al-Hashimi, H. M.; Tolman, J. R. *Q. Rev. Biophys.* **2000**, *33*, 371–424.
(16) Blackledge, M. *Prog. Nucl. Magn. Reson. Spectrosc.* **2005**, *46*, 23–61.
(17) Meier, S.; Blackledge, M.; Grzesiek, S. *J. Chem. Phys.* **2008**, *128*, 052204.
(18) Shortle, D.; Ackerman, M. S. *Science* **2001**, *293*, 487–489.
(19) Alexandrescu, A. T.; Kammerer, R. A. *Protein Sci.* **2003**, *12*, 2132–2140.
(20) Mohana-Borges, R.; Goto, N. K.; Kroon, G. J. A.; Dyson, H. J.; Wright, P. E. *J. Mol. Biol.* **2004**, *340*, 1131–1142.
(21) Fieber, W.; Kristjansdottir, S.; Poulsen, F. M. *J. Mol. Biol.* **2004**, *339*, 1191–1199.
(22) Meier, S.; Güthe, S.; Kiefhaber, T.; Grzesiek, S. *J. Mol. Biol.* **2004**, *344*, 1051–1069.
(23) Ohnishi, S.; Lee, A. L.; Edgell, M. H.; Shortle, D. *Biochemistry* **2004**, *43*, 4064–4070.
(24) Sallum, C. O.; Martel, D. M.; Fournier, R. S.; Matousek, W. M.; Alexandrescu, A. T. *Biochemistry* **2005**, *44*, 6392–6403.
(25) Ding, K.; Louis, J. M.; Gronenborn, A. M. *J. Mol. Biol.* **2004**, *335*, 1299–1307.
(26) Jensen, M. R.; Markwick, P.; Griesinger, C.; Zweckstetter, M.; Meier, S.; Grzesiek, S.; Bernado, P.; Blackledge, M. *Structure* **2009**, *17*, 1169–1185.
(27) Bernado, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R. W. H.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17002–17007.
(28) Bernado, P.; Bertoncini, C.; Griesinger, C.; Zweckstetter, M.; Blackledge, M. *J. Am. Chem. Soc.* **2005**, *127*, 17968–17969.
(29) Mukrasch, M. D.; Markwick, P. R. L.; Biernat, J.; von Bergen, M.; Bernado, P.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 5235–5243.
(30) Dames, S. A.; Aregger, R.; Vajpai, N.; Bernado, P.; Blackledge, M.; Grzesiek, S. *J. Am. Chem. Soc.* **2006**, *128*, 13508–13514.
(31) Jensen, M. R.; Houben, K.; Lescop, E.; Blanchard, L.; Ruigrok, R. W. H.; Blackledge, M. *J. Am. Chem. Soc.* **2008**, *130*, 8055–8061.
(32) Jensen, M. R.; Blackledge, M. *J. Am. Chem. Soc.* **2008**, *130*, 11266–11267.
(33) Meier, S.; Häussinger, D.; Jensen, P.; Rogowski, M.; Grzesiek, S. *J. Am. Chem. Soc.* **2003**, *125*, 44–45.
(34) Meier, S.; Grzesiek, S.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 9799–9807.
(35) Kohn, J. E.; Millett, I. S.; Jacob, J.; Zagrovic, B.; Dillon, T. M.; Cingel, N.; Dothager, R. S.; Seifert, S.; Thiyagarajan, P.; Sosnick, T. R.; Hasan, M. Z.; Pande, V. S.; Ruczinski, I.; Doniach, S.; Plaxco, K. W. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 12491–12496.
(36) Merchant, K. A.; Best, R. B.; Louis, J. M.; Gopich, I. V.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1528–1533.
(37) Möglich, A.; Joder, K.; Kiefhaber, T. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 12394–12399.
(38) Gabel, F.; Jensen, M. R.; Zaccai, G.; Blackledge, M. *J. Am. Chem. Soc.* **2009**, *131*, 8769–8771.
(39) Wells, M.; Tidow, H.; Rutherford, T. J.; Markwick, P.; Jensen, M. R.; Mylonas, E.; Svergun, D. I.; Blackledge, M.; Fersht, A. R. *Proc. Natl. Acad. Sci. (U.S.A.)* **2008**, *105*, 5762–5767.

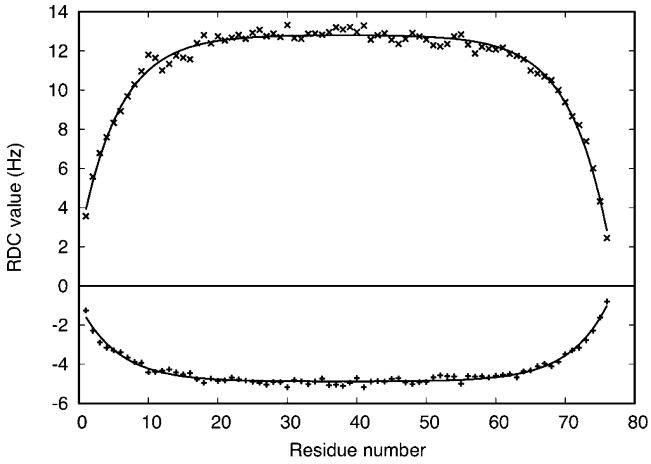

Figure 1. Residual dipolar coupling baselines in unfolded chains. Baseline effects underlying simulated ensemble-averaged RDCs from 100K copies of a polyvaline chain of 76 amino acids in length (crosses) and predicted RDCs following a hyperbolic cosine curve of the form given in eq 1 (line). $^{15}N-^{1}H^{N}$ couplings are shown below zero and $^{13}C^{\alpha}-^{1}H^{\alpha}$ RDCs are shown above zero.

acid-specific or even atomic resolution, as have recently been developed in the Bonvin and Forman-Kay laboratories.[40,41] In order to do this, we develop a novel algorithm to select from a large pool of possible conformers, created using the algorithm flexible-meccano, to best describe the system.

We test this approach, using extensive simulation, to determine how well the fitting of RDCs to reduced conformational ensembles containing few copies of the molecule can correctly reproduce the backbone conformational behavior of the protein. We also use cross-validation of data not employed in the fit to determine the most appropriate ensemble size to characterize the highly fluctuating molecule. Having established approaches that allow accurate mapping of conformational space from RDCs, we apply these methods to the amino acid-specific description of backbone conformational sampling in ubiquitin denatured in 8 M urea at pH 2.5.

Results and Discussion

RDCs from Disordered Proteins Modeled by Multiplication of Local Sampling Profiles and Underlying Baseline. RDCs can be simulated from explicit molecular ensembles of disordered proteins using shape-based considerations of the alignment properties of each copy of the molecule, and the average couplings can be predicted by taking the mean over the entire ensemble.[27,42] Comparison of such predictions with experimental data has revealed the unique sensitivity of RDCs to local and global sampling properties of highly disordered proteins. A key disadvantage of this approach is the number of structures that need to be treated, before the average RDC value converges to a nonfluctuating value. This number can reach many tens of thousands in proteins of 100 amino acids. It has recently been proposed that convergence of RDCs toward experimental data can be achieved with a smaller number of conformers if the protein is divided into short, uncoupled segments (Local Alignment Windows, LAWs) and the RDCs are calculated using the alignment tensor of these segments.[43,44] The ability to describe the conformational properties with ensembles containing fewer structures will of course make any ensemble selection procedure more tractable and is therefore an attractive prospect. In general, however, RDCs are affected both by the local conformational sampling and the chain-like nature of the unfolded protein, which induce an effective baseline reflecting the increasing degrees of freedom available toward the ends of the chain.[45,46] Long-range information is therefore necessarily absent from an approach that only employs LAWs to predict the RDCs. If this approach is employed, the simulated data need to be corrected for the effects of the unfolded chain.

We have simulated ensemble-averaged RDCs for polyvaline chains of differing lengths. The predicted RDCs can be relatively well fitted to a hyperbolic cosine curve of the form (Figure 1)

$$B(i) = 2b\cosh(a(i - d)) - c \qquad (1)$$

where i is the residue number and d is half the number of residues. a, b, and c are optimized for each different coupling type, where $(2b - c)$ is the RDC value at position d. This baseline dependence can be used to correct RDCs calculated using LAWs as described below.

RDCs are simulated for the central residue of LAWs of equal length, sliding the LAW one amino acid at a time along the chain (note that the termini are treated in the same way by adding dummy residues beyond the ends of the chain; see Experimental Section). These RDCs are then averaged over all structures. RDCs simulated for LAWs of m amino acids in length will exhibit a flat baseline, because each calculated RDC is at the center of a fragment of m amino acids and is therefore at the middle of the same local effective baseline. The RDC distribution resulting from the LAWs therefore depends on amino acid type but does not contain the baseline effects. It can be shown (Figure 2) that this amino acid-specific distribution can be multiplied with the baseline predicted in eq 1, to closely reproduce RDCs predicted from the explicit full-length description of the protein, which contains both amino acid-specific effects and the chain nature of the full length protein.

In order to determine the convergent characteristics when RDCs are simulated using LAWs of different lengths, we have compared the average values taken over an increasing number of conformers. Examples are shown in Figure 3a of the same $^{1}D_{NH}$ RDC when the RDC is calculated for the central amino acid of LAWs of different lengths (3, 9, 15, 25, and full length protein of 76 amino acids). Further simulations of $^{1}D_{C\alpha H\alpha}$, $^{1}D_{C\alpha C'}$, $D_{NH H\alpha}$, and D_{NHNH} RDCs show similar convergent characteristics (data not shown). It is clear that for the full-length protein the average is only converged when more than 10 000 structures are taken into account, while for LAWs of 15 amino acids this number falls to a few hundred. Figure 3b shows the strong dependence of the range of sampled RDCs on the length of the LAW. As the LAW gets longer, the individual structures can have larger RDC values, rendering the average less and less stable (vide infra).

(40) 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.
(41) Krzeminski, M.; Fuentes, G.; Boelens, R.; Bonvin, A. M. J. J. *Proteins: Struc. Funct. Bioinform.* **2009**, *74*, 894–905.
(42) Jha, A. K.; Colubri, A.; Freed, K.; Sosnick, T. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13099–13105.
(43) Marsh, J. A.; Baker, J. M. R.; Tollinger, M.; Forman-Kay, J. D. *J. Am. Chem. Soc.* **2008**, *130*, 7804–7805.
(44) Marsh, J. A.; Forman-Kay, J. D. *J. Mol. Biol.* **2009**, *391*, 359–374.
(45) Louhivuori, M.; Pääkkönen, K.; Fredriksson, K.; Permi, P.; Lounila, J.; Annila, A. *J. Am. Chem. Soc.* **2003**, *125*, 15647–15650.
(46) Obolensky, O. I.; Schlepckow, K.; Schwalbe, H.; Solov'yov, A. V. *J. Biomol. NMR* **2007**, *39*, 1–16.

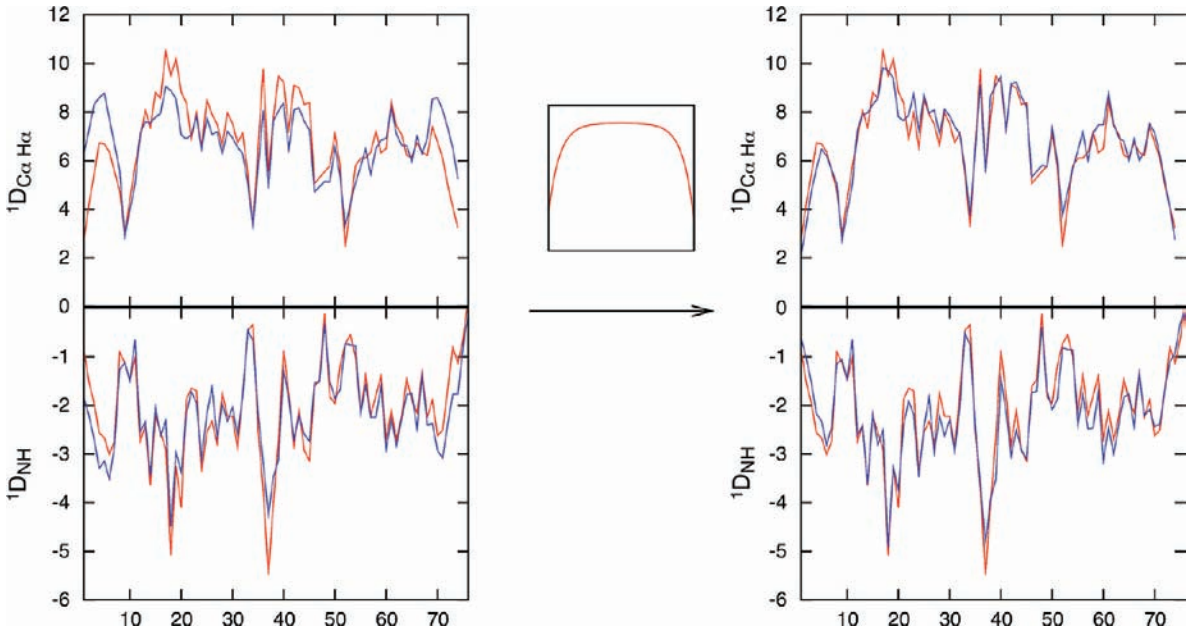

Figure 2. Multiplication of RDCs calculated using LAWs with RDC baselines in unfolded chains. $^{15}N-^1H^N$ and $^{13}C^\alpha-^1H^\alpha$ RDCs calculated from the central amino acid of a 15 amino acid LAW (blue, left) contain no baseline information and therefore diverge from the RDCs calculated from an explicit ensemble using a global alignment tensor (red). When multiplied with the hyperbolic cosine curve (eq 1), RDCs from the LAW (blue, right) more closely resemble the RDCs calculated from the global alignment tensor (red).

Alignment Strand Length Required To Define Accurately Conformational Sampling. In order to further determine the accuracy of describing RDCs using LAWs, we have compared the ability of LAWs of different lengths (after multiplication with the baseline described by eq 1) to reproduce RDCs simulated using a global alignment tensor (Figure 4). Not surprisingly, the shortest LAWs (three amino acids in length) never correctly reproduce average RDCs, due to the effects of neighboring amino acids (beyond nearest neighbors), on the local conformational sampling. The influence of neighboring residues on local conformational sampling is commonly estimated in terms of a so-called "persistence length", beyond which the remainder of the chain can be considered to exert a negligible effect. The persistence length depends on the relative rigidity of the local primary sequence. The relevance of taking full account of the persistence length on the local conformational sampling is further demonstrated by simulations that have been performed using a more rigid statistical coil model for which RDCs simulated using LAWs of nine amino acids fail to reproduce the averaged RDCs calculated using the global alignment tensor (data not shown). These simulations therefore indicate that while convergence characteristics of the predicted RDCs improve with shorter LAWs, the shortest strands can never fully reproduce the correct average, even if a very large number of structures were used in the average. On the basis of these simulations, we consider that a LAW length of 15 amino acids should be an acceptable compromise between efficiency and accuracy for the subsequent analyses.

How Many Structures Are Required for RDCs To Define Accurately Conformational Sampling? The next question concerns the number of structures required to describe correctly the conformational sampling. The averaging of RDCs is particularly demanding in terms of numbers of structures for two main reasons: first, because of the large number of backbone dihedrals whose relevant conformational space must be efficiently sampled before the overall shape and dimensions of the protein, and therefore the associated alignment tensor, average to convergent values. A second consideration is less obvious, but potentially more important: each dipolar coupling calculated from a single conformer of the entire molecule will sample a value within a range that can be orders of magnitude higher than the range spanned by the average values (Figure 3b). This dynamic-range problem can induce significant instability in the fitting procedure when using an ensemble containing too few structural models.

In order to numerically estimate the minimum number of structures that can accurately reproduce the true structural propensities of a conformational equilibrium, we have undertaken the following simulation: Two distinct statistical coil sampling regimes were defined, and entire sets of RDCs were calculated from flexible-meccano using these regimes with the global alignment tensor. The first, regime S, defines the standard statistical coil model employed in flexible-meccano, where amino acid-specific conformational distributions are extracted from populations of coil regions found in the protein structural database. The second sampling regime (E) samples a more extended region of Ramachandran space, populating the region $\{50° < \psi < 180°\}$ with a higher propensity than the S regime (see Experimental Section), while retaining the amino acid specific sampling from the S database. These data sets were then used as targets for the ensemble selection algorithm ASTEROIDS (A Selection Tool for Ensemble Representations Of Intrinsically Disordered States) described in the Experimental Section.

The ability of the algorithm to reproduce the correct conformational sampling and the correct RDCs for two different LAWs and the global alignment tensor is summarized in Figure 5 as a function of the number of structures constituting the ensemble. Using the target function χ^2_{Ram}, which measures the population of four different regions of Ramachandran space defined in Figure 6, we measure the ability of the protocol to reproduce amino acid-specific conformational sampling throughout the molecule (see Experimental Section). In each of the three considered window lengths, (9, 15, and full length protein), the

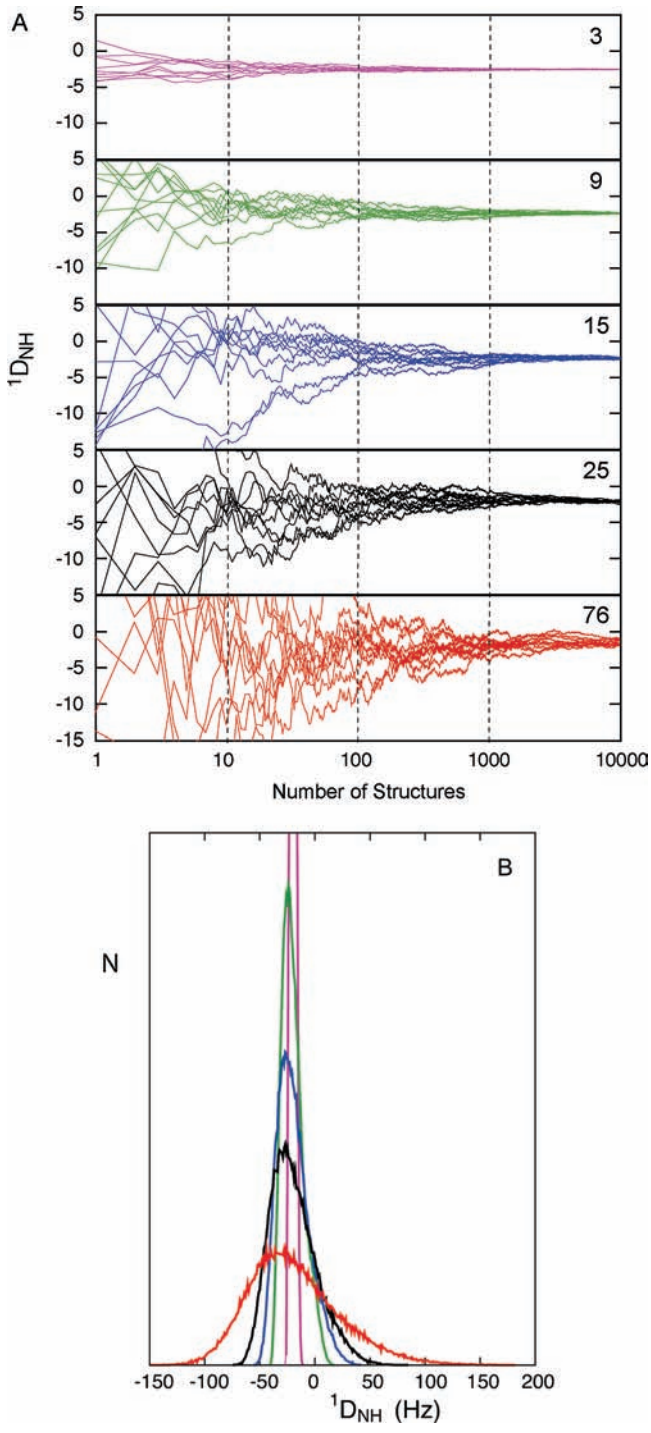

Figure 3. Convergence of $^{15}N-^1H^N$ RDCs calculated using LAWs of different lengths. (a) Comparison of 10 simulations of the central amino acid of an m amino acid LAW. The same $^1D_{NH}$ RDC (amino acid 41 of ubiquitin) is calculated using LAWs of $m = 3$, 9, 15, 25 or from the full length (76 amino acid) protein using a global alignment tensor. The x-axis represents the number of structures used to calculate the average. (b) Range and distribution of RDCs from the simulations shown in part a. Color code is the same in both cases (purple, three amino acid window; green, nine amino acids; blue, 15 amino acids; black, 25 amino acids; red, 76 amino acids).

reproduction of the RDCs improves rapidly with the number of structures included in the ensemble average. Simultaneously, the reproduction of the correct conformational sampling (the sampling used to simulate the RDC data) improves in all cases. These simulations, and those applied to the more extended

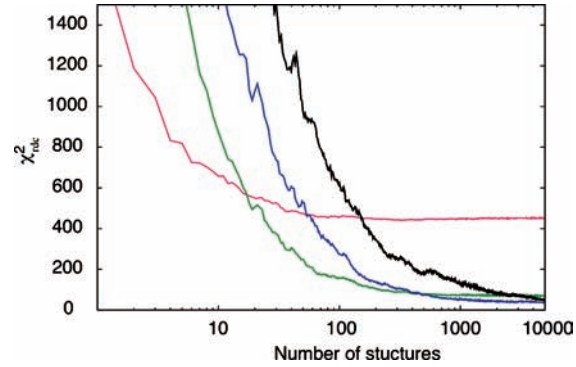

Figure 4. Accuracy of RDCs calculated using LAWs compared to a full length description. Equation 4 was used to directly compare the ability of RDCs calculated using the convolution of baseline and LAWs to reproduce RDCs calculated using an explicit description of the full length protein. The x-axis defines the number of averaged RDCs. χ^2_{RDC} was calculated over the entire protein. Color code: purple, three amino acid window; green, nine amino acids; blue, 15 amino acids; black, 25 amino acids.

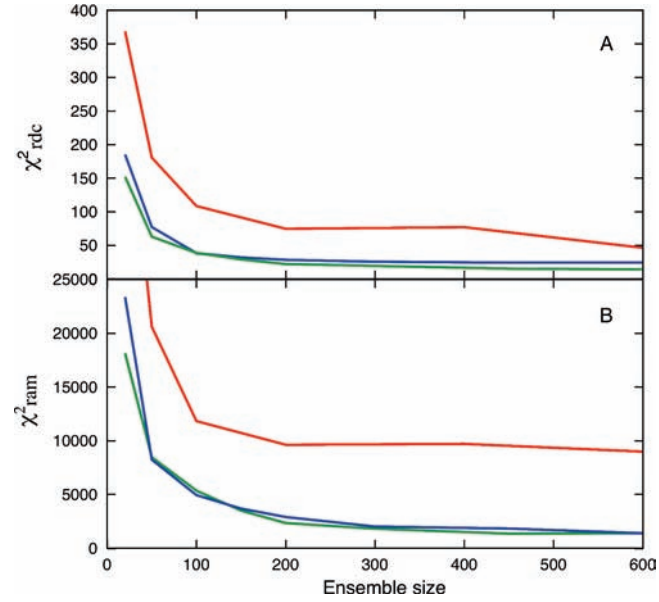

Figure 5. Accuracy of ensembles of structures calculated using LAWs of different lengths. The ability of ASTEROIDS to reproduce the correct conformational sampling and the correct RDCs for LAWs of different lengths is summarized as a function of the number of structures constituting the ensemble. (a) χ^2_{RDC} measures the reproduction of the target RDCs calculated using the full length 50 000-strong explicit description of the global alignment tensor. (b) χ^2_{Ram} measures the ability of the protocol to reproduce conformational sampling throughout the molecule. Color code: green, nine amino acid LAWs; blue, 15 amino acid LAWs; red, 76 amino acids (global alignment tensor). The x-axis defines the number of structures used.

sampling regime (data not shown), indicate that the optimal combination for an accurate description of conformational behavior of the protein backbone requires a window length of at least 15 amino acids and 200 structures.

The site-specific reproduction of the different RDCs comprising the χ^2_{RDC} using an ensemble of 200 and 20 structures is shown in Figure 7, for a LAW of 15 amino acids. Although the fit is significantly poorer in the case of 20 structures, the overall features are actually quite well reproduced, and the quality of the fit would probably be considered acceptable in the presence of commonly encountered levels of experimental noise. The conformational sampling is, however, very poorly reproduced, throughout the protein, when only 20 structures are

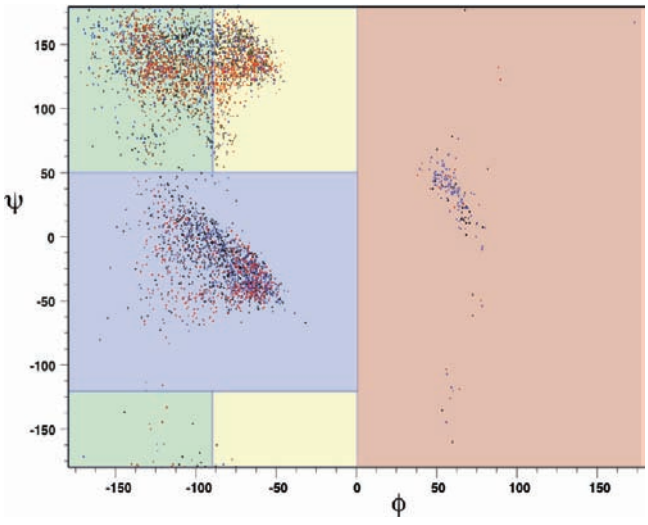

Figure 6. In order to quantify the similarity between conformational sampling between different ensembles, Ramachandran space is divided into four quadrants and defined as follows: α_L, $\{\phi > 0°\}$; α_R, $\{\phi < 0, -120° < \psi < 50°\}$; β_P, $\{-90° < \phi < 0°, \psi > 50°$ or $\psi < -120°\}$; β_S, $\{-180° < \phi < -90°, \psi > 50°$ or $\psi < -120°\}$. The population of these quadrants is indicated as p_{α_L}, p_{α_R}, p_{β_P}, and p_{β_S}. Dots represent standard statistical coil distributions of valine (red), lysine (blue), and leucine (black).

included. This is graphically underlined in Figure 8, where the populations of the four quadrants of conformational space present in the 200- and 20-fold ensembles are compared with those present in the ensemble used to create the simulated data. Discrepancies in the population of the different quadrants of up to 30% compared to the value present in the original ensemble are found throughout the primary sequence for the 20-fold ensemble. These differences do not appear to be correlated to amino acid type. The 200-fold ensembles, on the other hand, closely reproduce the original sampling (figure 8b) for every region of primary sequence. It is therefore evident that, in cases where too few structures are included in the average, achieving acceptable reproduction of experimental data does not guarantee that the resulting ensemble accurately represents the correct conformational distribution.

Application of ASTEROIDS to Experimental RDCs from Urea-Unfolded Ubiquitin. Using the optimal parameters determined on the basis of the simulations described above, we have applied the ASTEROIDS approach to the determination of a representative ensemble to describe the conformational behavior of the protein ubiquitin under denaturing conditions (pH 2.5 and 8 M urea). In the initial analysis, ensembles of 200 structures were selected from a set of 12 000 conformers for which LAWs of 15 amino acids in length were used to calculate the dipolar couplings. The results, shown in Figure 9a, indicate a reasonable reproduction of experimental data but reveal notable systematic effects, in particular that the $D_{NHH\alpha(i\text{-}1)}$, $D_{NHNH(i+1)}$ RDCs are overestimated when the other couplings, effectively the $^1D_{NH}$ and $^1D_{C\alpha H\alpha}$ RDCs agree optimally with simulation. These observations agree qualitatively with identification of differential scaling of $^1H-^1H$ couplings compared to covalently bound spins in the analysis of these RDCs. In order to allow for this possibility in the current analysis, we allowed for two independent scaling factors, K_1 for the $^1D_{NH}$, $^1D_{C\alpha H\alpha}$, and $^1D_{C\alpha C'}$ and K_2 for the $D_{NHH\alpha}$, $D_{NHH\alpha(i\text{-}1)}$, $D_{NHNH(i+1)}$, and $D_{NHNH(i+2)}$. These factors are optimized uniformly for the covalently bound and through-space dipolar interactions, resulting in the data reproduction shown in Figure 9b. The two scaling factors $K_1 = 0.58$ and $K_2 = 0.96$ differ by approximately 0.6, a difference that may result from additional local conformational dynamics that are not taken into account by the statistical coil model and that scale the $D_{NHH\alpha(i\text{-}1)}$, $D_{NHNH(i+1)}$ RDCs differentially to the RDCs between spins whose distances are effectively fixed. This possibility is currently under more detailed investigation.

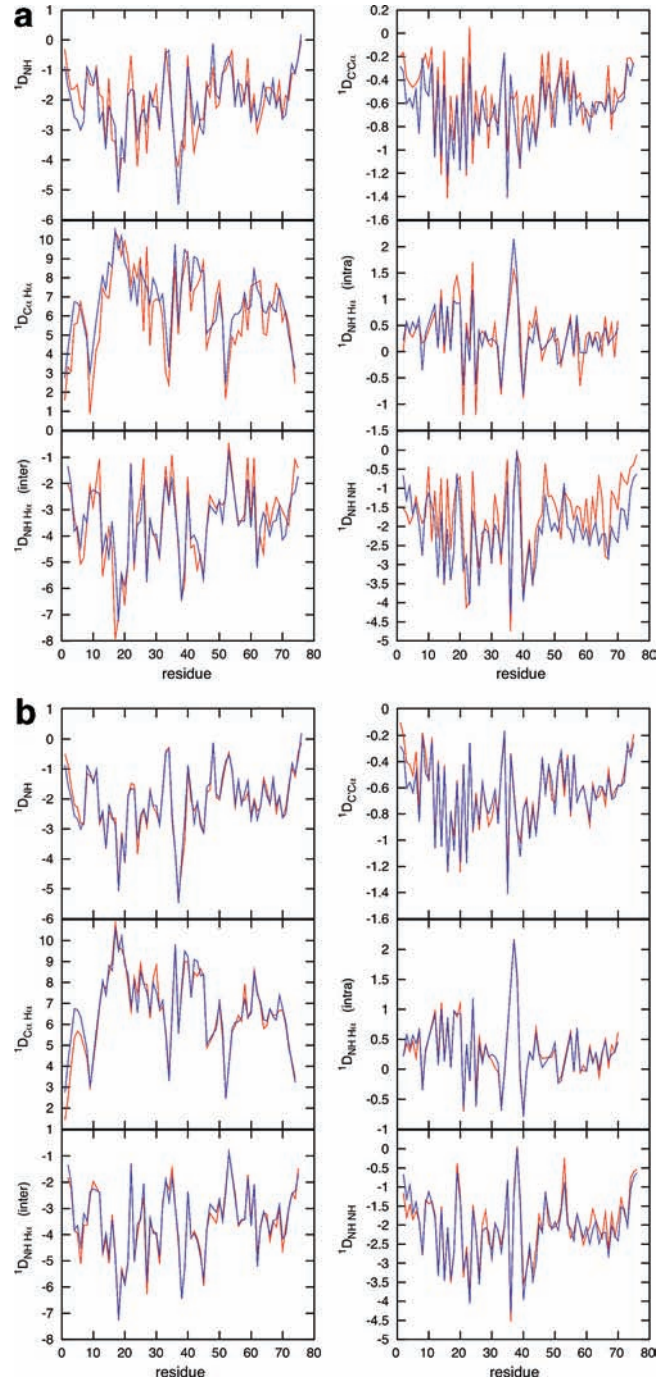

Figure 7. Site-specific reproduction of the RDCs simulated using an explicit ensemble of 50 000 structures. (a) Reproduction of the target data (blue) using an ensemble of 20 structures (red) for a window length of 15 amino acids. (b) Reproduction of the target data (blue) using an ensemble of 200 structures (red) for a window length of 15 amino acids. In both cases, the genetic algorithm ASTEROIDS was used to select the optimal ensemble.

In order to test the validity of the approaches shown here for the analysis of experimental data, we have repeated the ASTEROIDS ensemble selection procedure, taking 10% of the

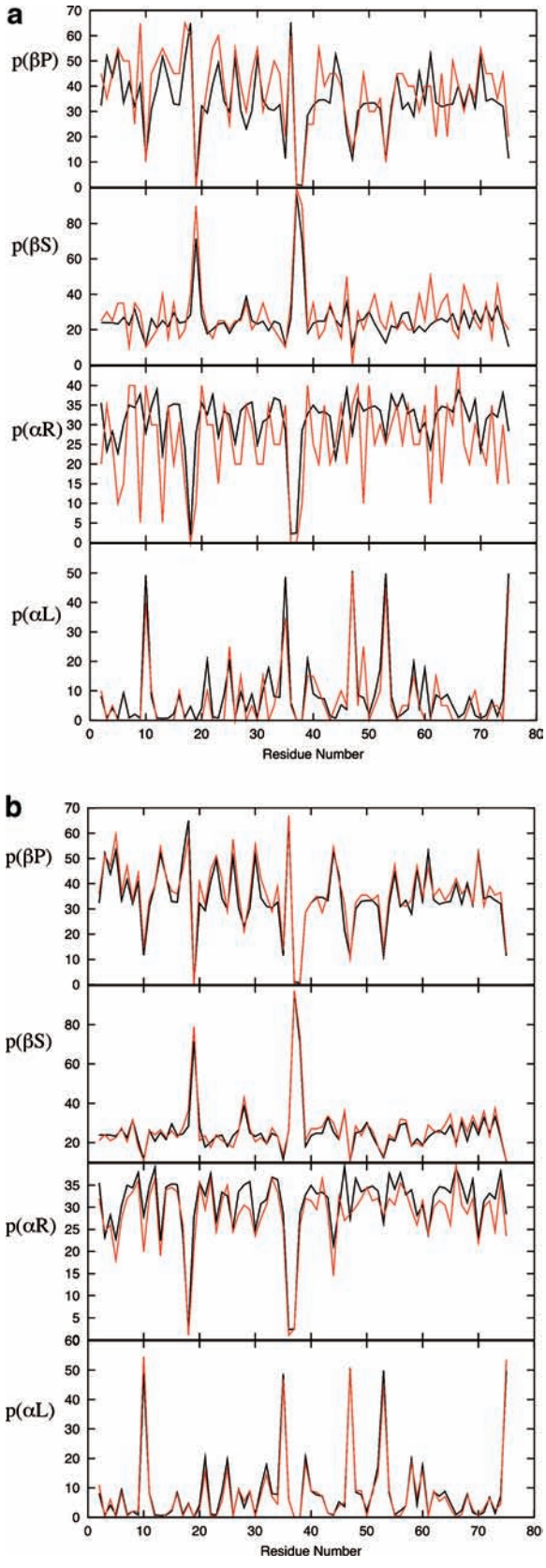

Figure 8. Accuracy of the reproduction of conformational sampling using the ASTEROIDS approach with ensembles of 20 and 200 structures. Populations of the four quadrants of conformational space defined in Figure 6 using the (a) 20-fold and (b) 200-fold ensembles (red) compared with those present in the ensemble used to create the simulated data (black). Discrepancies in the population of the different quadrants of up to 30% compared to the value present in the original ensemble are found for ensembles of size 20.

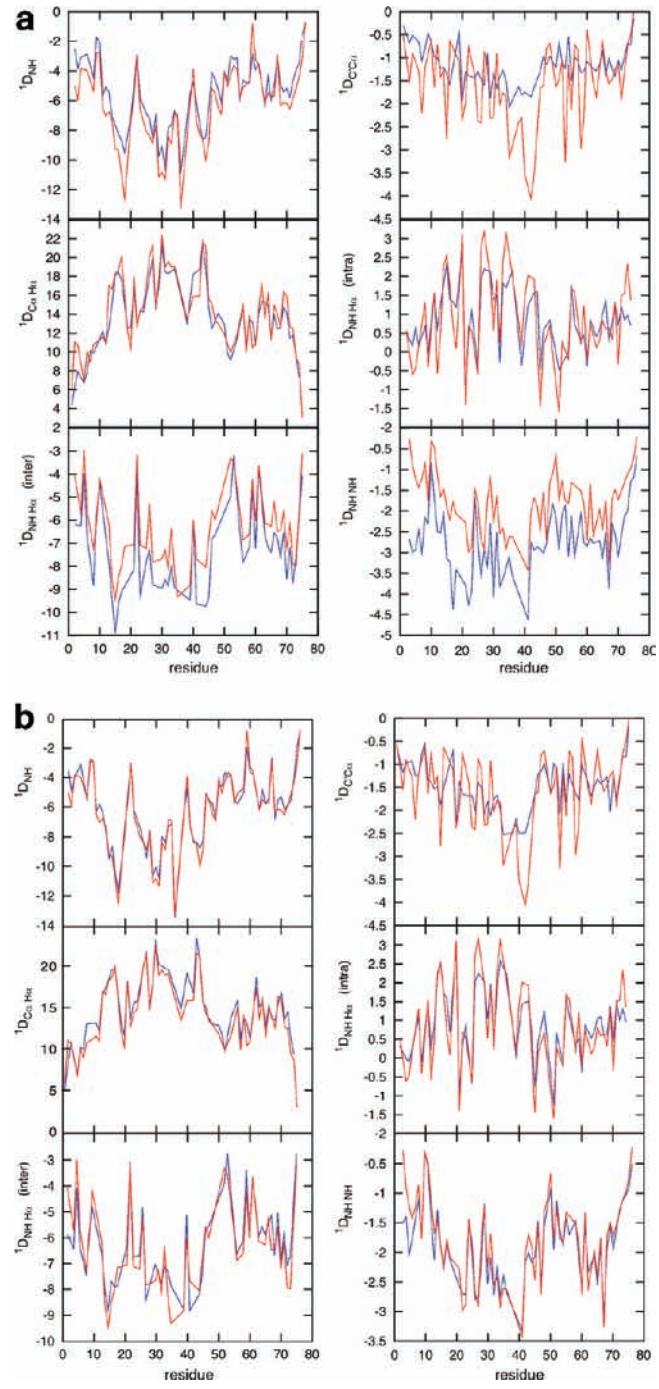

Figure 9. Application of ASTEROIDS to experimental RDCs from urea-unfolded ubiquitin. (a) Reproduction of experimental data (red) using an ensemble of 200 structures (blue). (b) Reproduction of experimental data (red) using an ensemble of 200 structures (blue) with differential scaling of the covalently bound and interproton RDCs.

RDCs out of the analysis and comparing the predicted values using the resulting ensemble with the experimental RDCs. The results are shown in Figure 10, where the back-calculated RDCs are found to be in reasonable agreement with the experimentally determined values. The calculation was repeated 10 times at seven different ensemble sizes. The average cross-validated χ^2 is plotted as a function of ensemble size (Figure 10b). The size of 200-fold ensembles used in the current approach is within the range where the cross validation target function is essentially flat.

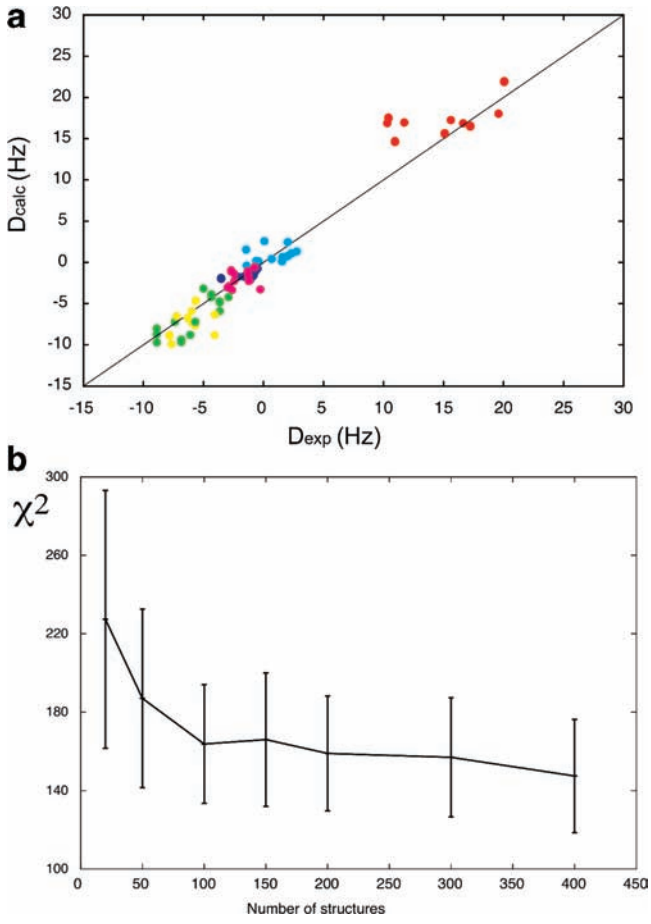

Figure 10. Reproduction of data not used in the fitting procedure. (a) The ASTEROIDS ensemble selection procedure was repeated, taking 10% of the RDCs out of the analysis and comparing the predicted values using the resulting ensemble with the experimental RDCs. Color code: green, $^1D_{NH}$; red, $^1D_{C\alpha H\alpha}$; dark blue, $^1D_{C\alpha C'}$; cyan, $D_{NH H\alpha}$; yellow, $D_{NH H\alpha(i-1)}$; magenta, $D_{NHNH(i+1)}$.(b) Average χ^2 over 10 cross-validation calculations at each of seven different ensemble sizes. The 200-fold ensemble size used in the current approach is within the range where the cross-validation target function is essentially flat.

The precision with which the RDCs can define the conformational behavior of the backbone has been assessed using noise-based Monte Carlo simulations (see Experimental Section) based on estimates of experimental uncertainty. The results are summarized in Figure S1 of the Supporting Information, and show that the average uncertainty in the populations of the different quadrants is approximately ±3%. We have also repeated the entire analysis in the absence of one experimental data set to assess the relative importance of each data set for the conformational description. The results are shown in Figure S2 and summarized in Tables S1 and S2 of the Supporting Information, where the backbone sampling is compared to the populations determined using all data. The root-mean-square deviation of the four populations defined in Figure 6 and the average differences demonstrate that although we find that the most important RDCs are the $D_{NHHai,i+1}$ and D_{NHNH}, the effects are actually not very large when these RDCs are removed (maximum rmsd of 5%, and average difference in populations of 3%). These results suggest that both covalently bound and interproton RDCs are important for an accurate description of conformational sampling but that none of the RDC types are critical for the validity of the description or the conclusions drawn from it.

The amino acid Ramachandran sampling has been used to calculate expected $^3J_{NHH\alpha}$ scalar couplings, reporting on the sampling of the ϕ backbone dihedral angle. These values have been compared to experimentally determined couplings[47] (Figure S3, Supporting Information), in comparison to the reproduction of the data using the standard coil database. The *J*-coupling data reproduction is quite good in both cases, but only slightly better in the case of the selected ensemble ($\chi^2 = 11.5$ compared to 12.6), probably reflecting the fact that the differences in the two descriptions are often found in the distribution of the ψ backbone dihedral angle. However, this analysis does demonstrate that the local analysis of RDCs in terms of Ramachandran distributions does not contradict independent experimental data in a significant way.

Urea Preferentially Affects the Conformational Sampling of Amino Acids with Side Chain Hydrogen-Bonding Moieties. Figure 11 shows the backbone dihedral angle distributions resulting from the analysis of experimental data of urea-unfolded ubiquitin and the normalized difference compared to the distribution of angles derived using an ensemble of structures produced using the standard statistical coil model of the unfolded state. Figure S4 of the Supporting Information shows the amino acid specific populations of all amino acids for the standard statistical coil model. The sampling of the different regions of the Ramachandran space defined in Figure 6 is summarized in Figure 12.

In general, the results indicate that the sampling of backbone dihedral angles in Ramachandran space is more extended, sampling the β_P and β_S regions with higher propensity and the α_R region with lower propensity than the statistical coil database. This result is in agreement with a previous study of the more general characteristics of conformational sampling, using the same experimental data.[34] In this study, a hypothesis-driven approach was used to suggest a general extension of conformational sampling of the peptide chain. With the new techniques developed here, we are able to extract amino acid-specific conformational sampling directly from the RDC data. This approach relies on the supposition that the database from which structures are selected contains enough conformational diversity to allow for a representative description to be constructed from its population. Under these conditions, the method is relatively hypothesis-free in comparison to previous approaches. This reveals that the effects of urea on backbone conformational sampling are far from uniform. The extended nature of the chain is more apparent in localized contiguous segments of primary sequence: the regions 30−36 and 70−73 sample the β_P region more extensively than both the statistical coil and the remainder of the protein, while extended β regions are preferentially sampled in the region 14−18. This latter tendency may be correlated with the previously observed presence of a small (around 20%) residual population of β hairpin in this region of the molecule.[48] Amino acids preceding prolines (18 and 36) are found to better reproduce experimental RDCs with a more uniform sampling of propensities in the β_P and β_S regions, compared to the statistical coil database that preferentially samples the α_R region.

The comparison with the statistical coil model clarifies detail that may be masked by amino acid-specific sampling of backbone dihedral angle and allows the identification of sites

(47) Peti, W.; Henning, M.; Smith, L. J.; Schwalbe, H. *J. Am. Chem. Soc.* **2000**, *122*, 12017–12018.
(48) Meier, S.; Strohmeier, M.; Blackledge, M.; Grzesiek, S. *J. Am. Chem. Soc.* **2007**, *129*, 754–755.

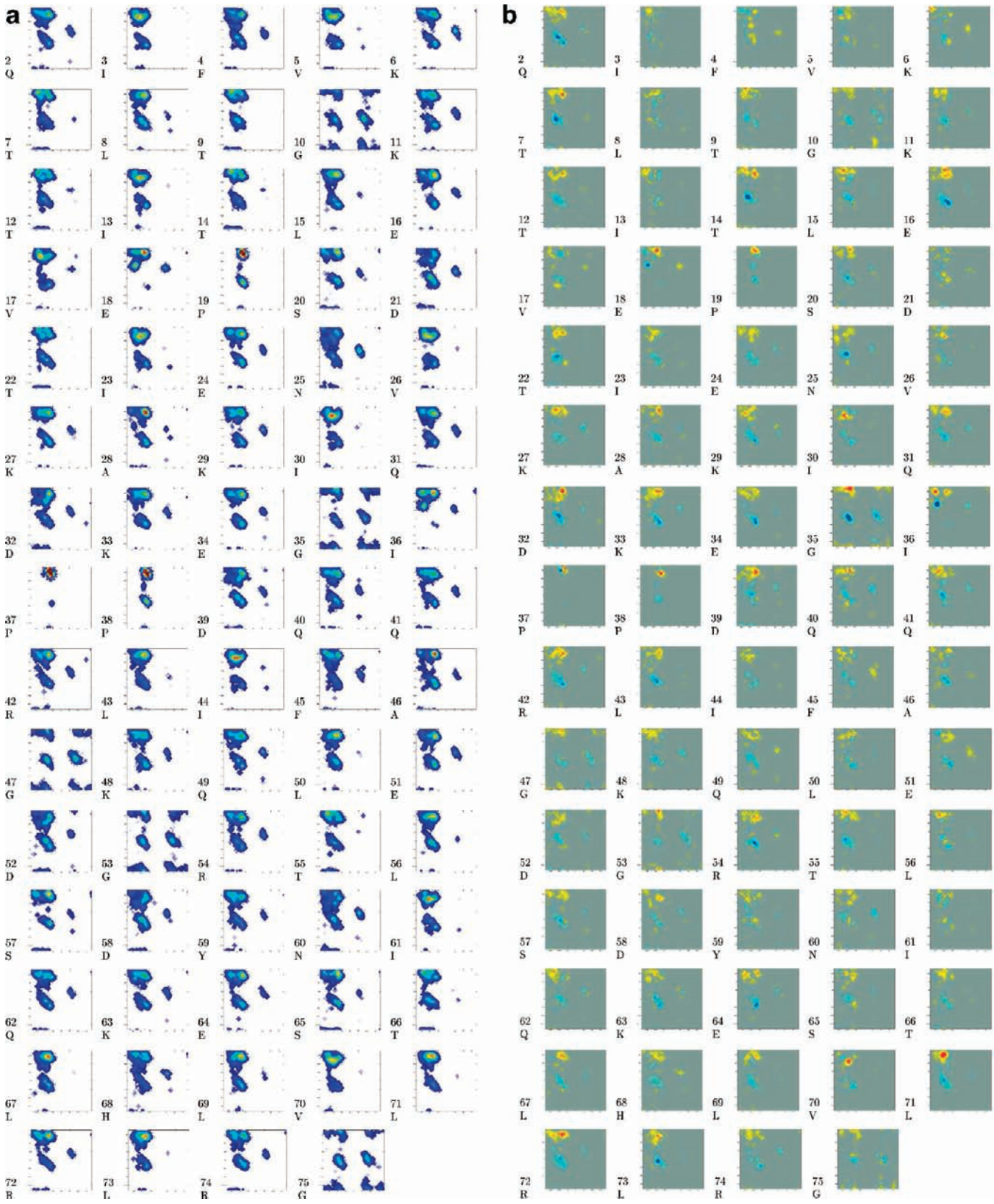

Figure 11. Amino acid-specific Ramachandran distributions for unfolded ubiquitin in 8 M urea at pH 2.5 in comparison with a standard statistical coil distribution. The populations increase from dark blue, via cyan, green, and yellow, to red. (a) Conformational sampling determined from the ASTEROIDS analysis of experimental RDC data (10 calculations were combined to produce 2000 conformers for the sake of figure resolution). (b) Difference between the conformational sampling distributions shown in panel a and the conformational sampling for the flexible-meccano statistical coil distribution. In this case, blue to green corresponds to negative values (population is lower in the urea unfolded sampling than in the statistical coil) and green (via yellow) to red corresponds to positive values (population is higher in the urea-unfolded sampling than in the statistical coil). Gray corresponds to equal populations.

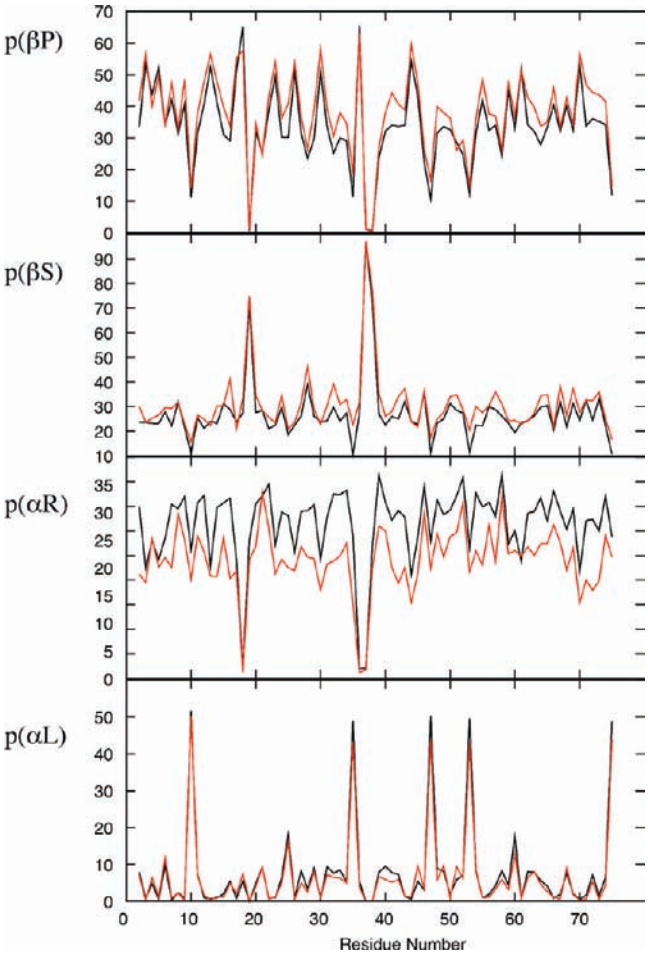

Figure 12. Populations of the four quadrants of conformational space defined in Figure 6 using the amino acid-specific Ramachandran distributions for unfolded ubiquitin in 8 M urea at pH 2.5 shown in Figure 11 (red) in comparison to a standard statistical coil distribution (black).

whose behavior deviates from random coil in the presence of urea. In this context, it is interesting to note that the amino acids whose backbone conformational sampling are most systematically affected by the presence of urea are threonine (four out of seven have a notably more extended backbone sampling than in the statistical coil model), glutamic acid (three out of five are more extended than the statistical coil model), and arginine (three out of four are more extended than the statistical coil model). These amino acids all contain potential hydrogen-bond-donor moieties on their side chains. A recent study using vibrational spectroscopy demonstrated that at low pH urea orients with the carboxyl group pointing toward the protein surface, an observation that supports the suggestion that hydrogen-bond-donor groups may interact preferentially with urea.[49] By contrast, only three of a total of 24 hydrophobic amino acids (valine, leucine, isoleucine, alanine, tyrosine, and phenylalanine) exhibit significantly different conformational sampling between the urea-denatured and the statistical coil states. The specific amino acid composition may therefore be responsible for the apparent localization of differential backbone sampling properties in the different regions of the protein. A recent study used small angle scattering to estimate the number of additional urea molecules that are preferentially recruited during the unfolding transition of ubiquitin from neutral to acidic pH to be approximately 20, a number that correlates qualitatively with the observation here that the backbone behavior of approximately a third of the amino acids are preferentially affected by the presence of urea.[38]

Conclusions

In this study, we have used extensive simulation to optimize an approach that exploits experimental RDCs measured from unfolded proteins to determine conformational sampling on an amino acid-specific basis. Previous applications have used a full-length description of the protein, averaging RDCs over an unrestrained ensemble that is large enough to allow for convergence of the coupling values. Although providing important insight into the behavior of a number of disordered proteins for which conformational information is otherwise difficult to measure, these studies are hypothesis-based, testing different conformational sampling regimes and comparing them to experimental data, an approach that severely limits both the scope and application as well as the potential for discovery. Here we develop a general approach that allows one to select an ensemble directly from the experimental data. Our combination of analytical baseline descriptor and numerical averaging of smaller alignment windows is tested against simulation, and on the basis of these simulations, parameters such as window length and number of structures are calibrated. We find that a combination of LAWs of 15 amino acids in length, with ensemble sizes of 200, accurately describes conformational space, while ensembles of 20 structures reproduce the experimental data but, critically, do not reproduce the correct conformational sampling. Using this approach we can describe conformational sampling at an amino acid resolution.

These approaches have been applied to the amino acid-specific description of backbone conformational sampling in ubiquitin denatured in 8 M urea at pH 2.5. Having established the precision that the approach is expected to offer, we are able to analyze in fine detail the local conformational differences between the standard statistical coil description and the sampling defined by the experimental data measured in the presence of urea, and we interpret this in the context of urea binding or interacting with specific types of amino acids in the peptide chain.

Experimental Section

Experimental methods for measuring the RDCs included in the analysis have been presented elsewhere. All data were taken from the earlier study by Meier et al.[34]

Flexible-Meccano Calculations. Simulated RDCs were calculated using the program flexible-meccano interfaced to the program PALES[50] as described. The program was run in two modes: For calculations using a global alignment tensor for the entire molecule, the standard procedure was used. For calculations using the local alignment windows (LAWs) the RDC for the central amino acid of the local *m* amino acid segment (3, 9, 15, or 25) was calculated for each individual structure. For the terminal amino acids, alanine amino acids were added to the N or C terminus during the building of the protein, such that the *m* amino acid segment was always present. The resulting RDC profile along the primary sequence is calculated by averaging each value and multiplying with the effective baseline given in eq 1. If RDCs were calculated using the full length protein, they were averaged over all conformers as previously described.

(49) Chen, X.; Sagle, L. B.; Cremer, P. S. *J. Am. Chem. Soc.* **2007**, *129*, 15104–15105.

(50) Zweckstetter, M.; Bax, A. *J. Am. Chem. Soc.* **2000**, *122*, 3791–3792.

A pool of 12 000 structures is generated with flexible-meccano. Half of the structures were calculated using the standard statistical coil model S, and the other half using a more extended regime E. The sampling regime (E) samples a more extended region of Ramachandran space, populating the region $\{50° < \psi < 180°\}$ with a higher propensity than the S regime (78% compared to 59%).

ASTEROIDS Ensemble Selection. ASTEROIDS uses a genetic algorithm[51−53] to build a representative ensemble of structures of fixed size N from a large database. The algorithm selects an ensemble of N structures using the following fitness function compared to the experimental data.

$$\chi^2_{\text{asteroids}} = \sum_i w_i{}^2 (D_{i,\text{calc}} - D_{i,\text{exp}})^2 \quad (2)$$

where w_i is the weight of coupling D_i. The weights were set according to coupling type and determined by the range of each type of coupling in hertz. Values of w were set to 1.0 for $^1D_{\text{NH}}$ and $D_{\text{NHH}\alpha(i-1)}$, 0.5 for $^1D_{\text{C}\alpha\text{H}\alpha}$, 2.0 for $^1D_{\text{C}\alpha\text{C}'}$, $D_{\text{NHH}\alpha}$, and $D_{\text{NHNH}(i+1)}$, and 3.0 for $D_{\text{NHNH}(i+2)}$. The final ensemble is obtained from generations of ensembles that undergo evolution and selection using this fitness function. Each generation comprises 100 different ensembles of size N.

Evolution can proceed in three different ways: random, mutation, and crossing. At each evolution step, the protocol ensures that a structure does not appear more than once in a given ensemble and that a given ensemble is not repeated in a generation. Random evolution proceeds by randomly selecting structures in the complete database. Mutation occurs by taking an ensemble and replacing 1% of the structures (or at least one structure) by structures randomly selected from the complete database (external mutation) or from a new database containing all the structures selected at least once in the previous generation (internal mutation). Crossing is achieved by randomly pairing ensembles from the previous generation. New ensembles are generated by selecting N structures in a pool made of the structures present in the previously defined pairs.

The first generation is always obtained using random evolution. Evolution of this generation is achieved by the following procedure. New ensembles are generated (100 by random evolution, 100 by external mutation, 100 by internal mutation and 100 by crossing). Among these new ensembles and the previous generation, 100 different ensembles representing minima with respect to the fitness function are selected using tournaments to provide the next generation. Ensembles are randomly split into groups and then ordered using the fitness function to determine the winners of the tournament. The best ensembles of each tournament are retained to form the next generation. The number of tournaments and the number of winners of each tournament are adjusted such that 100 ensembles are selected. Selection pressure increases as the number of tournaments decreases. To avoid premature convergence in local minima, the selection pressure is gradually increased during evolution. The number of tournaments therefore successively goes from 100 to 50, 25, 20, 10, 2, and to 1. To ensure robustness of the fitting procedure, the evolution and selection processes are repeated over 2000 successive generations.

Ramachandran Segment Division. In order to describe the sampling of conformational space in the different ensembles and their agreement with known distributions, Ramachandran space is divided into four quadrants indicated in Figure 6 and defined as follows: α_L, $\{\phi > 0°\}$; α_R, $\{\phi < 0, -120° < \psi < 50°\}$; β_P, $\{-90° < \phi < 0°, \psi > 50°$ or $\psi < -120°\}$; β_S, $\{-180° < \phi < -90°, \psi > 50°$ or $\psi < -120°\}$.

The population of these quadrants is indicated as p_{α_L}, p_{α_R}, p_{β_P}, and p_{β_S}. The Ramachandran similarity factor χ^2_{Ram} of the entire molecule is measured by the following function:

$$\chi^2_{\text{Ram}} = \sum_i \sum_q (p_{i,q,\text{ref}} - p_{i,q,\text{fit}})^2 \quad (3)$$

where p_q are the four different populations of the quadrants q, i are the different amino acids, and $_{\text{ref}}$ and $_{\text{fit}}$ signify the target and fitted Ramachandran distributions.

Comparison of RDCs. In order to compare RDCs calculated using different window lengths with those calculated using 50 000 conformers from the full length description of the protein, the following function χ^2_{RDC} is used:

$$\chi^2_{\text{RDC}} = \sum_i (D_{i,\text{LAW}} - D_{i,\text{fl}})^2 \quad (4)$$

where $D_{i,\text{LAW}}$ represents the RDC calculated using LAWs, after multiplication with the baseline function given in eq 1, and $D_{i,\text{fl}}$ is the RDC calculated using the full length description.

Monte Carlo Simulations and Error Analysis. In order to estimate the precision with which the conformational sampling can be defined on the basis of experimental RDCs, we have run noise-based Monte Carlo simulations, using random sampling of Gaussian distributions whose width is based on experimentally estimated uncertainties for each RDC. Fifty Monte Carlo simulations were run, and the effective uncertainty of the Ramachandran quadrant population was calculated on the basis of this.

In order to estimate the importance of the different RDC types, we have repeated the analysis of experimental data with one entire data set removed from the ASTEROIDS approach.

***J*-Coupling Analysis.** $^3J_{\text{NHH}\alpha}$ scalar couplings were calculated by averaging over the amino acid-specific ϕ backbone dihedral angle distributions and compared to experimentally measured values, using recently derived Karplus relationships.[54]

Acknowledgment. L.S. received a grant from the French Ministry of Education. This work was supported by the French Research Ministry through ANR-PCV07_194985. M.R.J. benefited from an EMBO fellowship and Lundbeckfonden support.

Supporting Information Available: A figure showing the standard statistical coil distribution on a residue-specific basis. Residue-specific populations of Ramachandran space resulting from Monte Carlo simulations. A figure and tables showing conformational sampling of the different quadrants of Ramachandran space when specific RDC types are removed. A figure showing calculated and experimental ^3J scalar couplings. This material is available free of charge via the Internet at http://pubs.acs.org.

JA9069024

(51) Fraser, A. S. *Austr. J. Biol. Sci.* **1957**, *10*, 484–491.
(52) Holland, J. H. *Adaptation in Natural and Artificial Systems*; University of Michigan Press: Ann Arbor, 1975.
(53) Jones, G. *Genetic and Evolutionary Algorithms. Encyclopedia of Computational Chemistry*; Wiley: Chichester, U.K., 1998.
(54) Markwick, P. R. L.; Showalter, S. A.; Bouvignies, G.; Brüschweiler, R.; Blackledge, M. *J Biomol NMR* **2009**, *45*, 17–21.