Biochemical Society TRANSACTIONS

Residual dipolar couplings measured in unfolded proteins are sensitive to amino-acid-specific geometries as well as local conformational sampling

Jie-rong Huang*, Martin Gentner†, Navratna Vajpai†, Stephan Grzesiek† and Martin Blackledge*[1]

*Institut de Biologie Structurale Jean-Pierre Ebel, CEA-CNRS-UJF UMR 5075, 41 Rue Jules Horowitz, Grenoble 38027, France, and †Division of Structural Biology, Biozentrum, University of Basel, Klingelbergstrasse 50, 4056 Basel, Switzerland

Abstract

Many functional proteins do not have well defined folded structures. In recent years, both experimental and computational approaches have been developed to study the conformational behaviour of this type of protein. It has been shown previously that experimental RDCs (residual dipolar couplings) can be used to study the backbone sampling of disordered proteins in some detail. In these studies, the backbone structure was modelled using a common geometry for all amino acids. In the present paper, we demonstrate that experimental RDCs are also sensitive to the specific geometry of each amino acid as defined by energy-minimized internal co-ordinates. We have modified the FM (flexible-Meccano) algorithm that constructs conformational ensembles on the basis of a statistical coil model, to account for these differences. The modified algorithm inherits the advantages of the FM algorithm to efficiently sample the potential energy landscape for coil conformations. The specific geometries incorporated in the new algorithm result in a better reproduction of experimental RDCs and are generally applicable for further studies to characterize the conformational properties of intrinsically disordered proteins. In addition, the internal-co-ordinate-based algorithm is an order of magnitude more efficient, and facilitates side-chain construction, surface osmolyte simulation, spin-label distribution sampling and proline *cis*/*trans* isomer simulation.

Introduction

Approximately 50 % of mammalian proteins are predicted to contain long (more than 30 residues) disordered regions, and approximately 25 % of their proteins are predicted to be fully disordered in the absence of a well-defined three-dimensional structure under physiological conditions [1]. These so-called IDPs (intrinsically disordered proteins) play key roles in a variety of physiological processes, including signalling, cell cycle control, molecular recognition, transcription and replication, and in the development of neurodegenerative diseases, such as Alzheimer's disease and Parkinson's disease [2,3]. Despite the marginal amount of well-defined structure, conformational characterization of these proteins provides insights into the dynamics of their stability, leading to further understanding of disease-related aggregation and fibrillation. Owing to their structural heterogeneity, conventional approaches for structure determination are inappropriate for studying such flexible systems. Novel experimental techniques and computational models therefore become essential for characterizing their rapidly interconverting nature.

Key words: coil library, flexible-Meccano algorithm, intrinsically disordered protein (IDP), molecular dynamics simulation, residual dipolar coupling (RDC), statistical coil model.

Abbreviations used: FM, flexible-Meccano; IC, internal co-ordinate; IDP, intrinsically disordered protein; RDC, residual dipolar coupling.

[1]To whom correspondence should be addressed (email martin.blackledge@ibs.fr).

Although MD (molecular dynamics) simulations can provide an atomic-resolution description of the unfolded protein ensemble [4,5], there are still limitations in the currently available potential energy force fields to correctly describe the conformational sampling [6] and timescale of unfolded proteins in solution [7,8]. Alternatively, we have developed a conformational sampling algorithm termed FM (flexible-Meccano) [9,10] based on the so-called statistical coil model [9,11,12]. FM efficiently samples the backbone dihedral angle energy surface (φ/ψ) derived from highly resolved crystallographic structures excluding secondary elements, and constructs conformers with only sequence information. To verify FM-generated models, NMR spectroscopy provides the most informative experimental parameters, at amino-acid-specific resolution. In addition to using regular parameters, such as chemical shifts, scalar couplings, nuclear Overhauser effects and relaxation rates, to characterize the properties of unfolded proteins [13], RDCs (residual dipolar couplings) have also been demonstrated to be extremely useful to describe the unfolded state ensemble [14–19] owing to their sensitivity to local conformational sampling. The distribution of RDCs can be calculated very precisely as ensemble and time averages from the well-understood geometry-dependence of nucleus–nucleus dipolar interactions [20]. Accordingly, our ensemble model was verified by comparing RDCs calculated from ensemble

structures and experimental measurements [9,18]. In order to improve agreement between the properties of unfolded proteins and our model, increasing the amount of data available from different types of RDCs is essential. Thus eight types of published experimental RDCs for urea-denatured ubiquitin [18,21], as well as five types of newly measured couplings of urea-denatured Protein G, were used to assist in refining our algorithm. In the present paper, we describe a new method, which uses energy-minimized geometry derived from an existing potential energy force field [22], to construct the structural ensemble in a more accurate and efficient way. The predicted RDCs derived from the ensemble generated by the new algorithm improve the agreement with experimental data. This improvement is expected to have direct consequences on the quality of ensembles selected against experimental observables, for example using the ASTEROIDS approach [23–26], from a pool of structures generated using FM.

Protein purification and NMR measurement

The purification and preparation of denatured ubiquitin and Protein G (8 M urea and 10 mM glycine/HCl buffer, pH 2.5, at 25 °C) in isotropic solution or in stretched polyacrylamide gel are described in [18,27]. The RDCs for ubiquitin are taken from our previous publications [18,21]; RDCs for GB1 (the first Ig-binding domain of Protein G) was recorded using the same methods as described in the ubiquitin studies.

Data analysis

Theoretical RDCs were calculated on the assumption of steric exclusion [28,29] using an efficient in-house-written algorithm. The detailed algorithm is described in [30]. Briefly, the maximal extension of a molecule for each direction of a unit sphere is calculated. The probability for finding the molecule in a certain orientation is then derived as the volume that can be occupied by the molecule between two infinitely extended parallel planes relative to the total distance between the planes. The alignment tensor then corresponds to the average over all orientations of second rank spherical harmonics weighted by this probability. The theoretical RDCs are then calculated from the alignment tensor:

$$D_{k,ij}^{\text{calc}} = -\frac{\gamma_i \gamma_j \hbar \mu_0}{4\pi^2}\sqrt{\frac{4\pi}{5}}\sum_{m=-2}^{2} S_{k,m}^{*}\frac{Y_{2m}(\Theta_{k,ij}, \Phi_{k,ij})}{r_{k,ij}^3}$$

where $D_{k,ij}$ represents the RDC between nuclei i and j for ensemble member k with individual alignment tensor $S_{k,m}$ (written in irreducible form [31]), Y_{2m} are spherical harmonics, $r_{k,ij}$, $\Theta_{k,ij}$ and $\Phi_{k,ij}$ are the polar co-ordinates of the internuclear vector, and γ are the nuclear gyromagnetic ratios. The distances used for calculating D_{HN} and $D_{\text{C}\alpha\text{H}\alpha}$ are 1.02 and 1.1 Å (1 Å = 0.1 nm) respectively according to Bax and co-workers [32] otherwise using values calculated from the co-ordinates. RDCs then were averaged over all members of the ensemble to obtain the predicted value. The size of each ensemble throughout the present paper is 50000.

χ^2 analysis is used to indicate the agreement between experimental and theoretical values. It is defined as:

$$\chi^2 = \sum_{i=1}^{N}\left(\frac{D_i^{\text{obs}} - D_i^{\text{calc}}}{\sigma_i}\right)^2$$

where σ_i is the experimental error, and the summation runs over all observed data N.

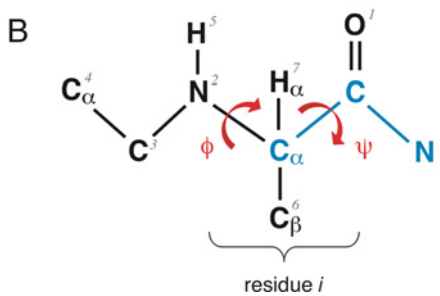
Figure 1 | Example of the IC description

(**A**) Every amino acid type contains seven lines of IC for constructing the backbone. The IC of leucine is shown as an example. The first four columns are names of atoms; the final three columns indicate the dihedral or improper dihedral angle between these four atoms, the angle between the last three atoms and the bond length between the last two atoms respectively. The asterisks on the third atom indicate that the first parameter is an improper dihedral angle. The plus sign specifies the atom for the previous residue, and the minus sign indicates the atom for the next residue. (**B**) Atoms coloured blue represent seed atoms. Atoms in black are built consecutively according to the IC. The numbers next to the atom indicate the order in which these atoms are built.

A

+N	CA	*C	O	180.00	120.56	1.2299
+N	C	CA	N	ψ	106.05	1.4508
C	CA	N	-C	ϕ	124.31	1.3474
CA	N	-C	-CA	180.00	117.93	1.5184
-C	CA	*N	HN	180.00	114.26	0.9979
N	C	*CA	CB	121.52	112.12	1.5543
N	C	*CA	HA	-116.50	107.57	1.0824

B



IC (internal co-ordinate)-based algorithm for constructing unfolded protein ensembles

Geometries, in the form of ICs, were derived from energy-minimized structures in the CHARMM force field topology [22]. Each line in the IC contains the names of four atoms and three parameters (see Figure 1 as example). These three parameters indicate the dihedral angle between these four atoms, the angle between the last three atoms and the bond length between the last two atoms respectively. Therefore, from the co-ordinates of the first three atoms and these three parameters, the co-ordinates of the fourth atom can be derived. Accordingly, the algorithm starts from three seed atoms: C$\alpha(i)$, C(i) and N($i+1$) for residue i, which can be present in a folded domain, for the cases of partially folded

proteins, or a standard three-atom geometry for constructing a fully unfolded polypeptide chain.

Each additional residue (i) is built consecutively according to the order of the topology file: O(i), N(i), C($i-1$), Cα($i-1$), H_N(i), Cβ(i) and Hα(i). While constructing atoms N(i) and C($i-1$), a combination of ψ and φ angles is randomly taken from the coil library database. Once a residue is constructed, an amino-acid-specific sphere [33] is placed at the position of Cβ (or Cα for glycine). If the sphere is overlapped with the other pre-built ones, this residue will be rejected and another φ/ψ angle combination will be selected from the database, until a non-steric clash conformation is found.

Difference between the two algorithms

Instead of using peptide planes derived from highly resolved X-ray structures, as is the case in the previous FM algorithm, this new algorithm applies energy-minimized backbone geometry as building blocks, giving specific conformations for different amino acid types. A detailed comparison between the geometries in terms of IC of these two algorithms is listed in Supplementary Table S1 at http://www.biochemsoctrans.org/bst/040/bst0400989add.htm. The most pronounced differences are the angles between atoms Cα, N and H_N; for some amino acid types, this can differ from the previous FM model by up to 7°. The tetrahedral angles around Cα, instead of using idealized 109°, range from ~105° to ~114° in the energy-minimized geometry. We note that, although the local geometries in terms of angle and bond length are different between these two algorithms, the radius of gyration (R_g) and φ/ψ angle distribution generated from them are very similar (Supplementary Figures S1 and S2 at http://www.biochemsoctrans.org/bst/040/bst0400989add.htm), indicating that the new algorithm is not altering the overall geometry or the local sampling.

In addition to the difference between building blocks, the new algorithm is approximately ten times faster than the original version (50000 structures of a 76-amino-acid protein can be created in 2 min on a single Intel 2.8 GHz CPU) mainly because a Levenberg–Marquardt minimization is no longer used to position the peptide plane [34]. We denote this new algorithm FM2 in the present paper for further comparison.

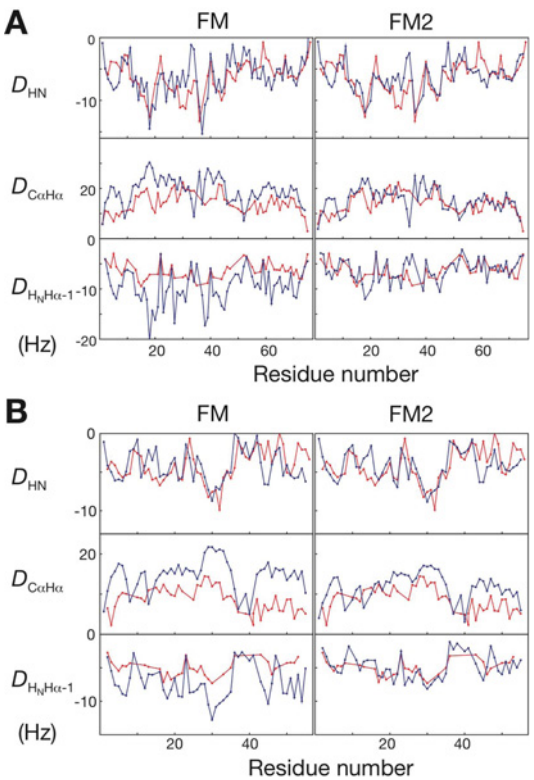
Using extensive sets of RDCs to verify the computational model

RDCs reporting the time and ensemble average of the nuclear dipolar–dipolar interaction provide a quantitative description of the local order in the unfolded state and are therefore probably the most powerful parameters for verifying the conformational sampling of a simulated structural ensemble. It has been demonstrated that RDCs calculated from a statistical coil ensemble can reasonably well reproduce experimental RDCs in unfolded states, as well as identifying transient long-range contact or residual structures in systems that diverge from the unfolded state [9,11,17,19,35]. Owing to the relative difficulty of RDC measurements, most RDCs reported to date for unfolded proteins are limited to a few types, mostly D_{HN} and $D_{C\alpha H\alpha}$. In order to extend our understanding of the unfolded protein conformation, we have collected RDCs for denatured ubiquitin using up to eight different coupling types and five different coupling types for denatured Protein G.

The data from ubiquitin have been used previously in comparison with the original FM algorithm to characterize the conformational sampling of this unfolded system, and more generally to determine the precision to which highly flexible proteins can be analysed from RDC data [18]. In the present paper, we repeat this analysis using the new conformational sampling algorithm based on amino-acid-specific ICs. The comparisons of D_{HN}, $D_{C\alpha H\alpha}$ and $D_{HNH\alpha\text{-}1}$ from these two algorithms and experimental values for both proteins are shown in Figure 2; the other types of RDCs are shown in Supplementary Figure S3 (at http://www.biochemsoctrans.org/bst/040/bst0400989add.htm). As the overall level of alignment is unknown, the scaling needs to be optimized against the experimental data. In all cases, only one scaling factor is applied, optimized in this case according to D_{HN}.

Figure 2 | Comparison of ensembles generated by FM and FM2 algorithms for denatured ubiquitin (A) and Protein G (B)

Predicted D_{HN}, $D_{C\alpha H\alpha}$ and $D_{HNH\alpha\text{-}1}$ (blue lines) calculated from both algorithms are compared with experimental data (red lines).



The prediction from FM is reproduced as reported previously [18]. In the previous studies using the FM algorithm, agreement between prediction and experiment was improved by using an additional scaling factor for all H–H RDCs, compared with RDCs measured between covalently bound spins. As a result of the comparison, more extended conformational sampling than that present in the statistical coil was evoked, presumably because of the extension of the chain due to the presence of high concentrations of urea [21,36–38]. Using FM2, the agreement between experimental and predicted RDCs is improved significantly for both $D_{C\alpha H\alpha}$ and $D_{HNH\alpha\text{-}1}$ when the single scaling factor is optimized against D_{HN}. This remarkable improvement is due to the difference of angles between backbone geometries, which, although small, nevertheless has a measurable effect on the ability to reproduce the experimental data compared with the common peptide plane geometry that was used for the previous study. A few degrees difference in bond orientation and tetrahedral angle geometry around Cα can significantly change the predicted RDCs, despite the high similarity of overall geometry and local sampling between ensembles generated from these two algorithms. As an example, a 7° difference in Cα-N-H_N can change D_{HN} by approximately 5 % assuming an extended conformation (results not shown).

In the light of these results, we have reassessed our previous conclusions that urea-denatured proteins sample an enhanced population of extended conformation. The predicted values for $D_{C\alpha H\alpha}$ of the FM2 case are still found to be overestimated compared with the experimental data, whereas other RDCs are found to be either underestimated or overestimated (Supplementary Table S3). As in the previous study, different ensembles by means of enhancing the extended region ($50° < \psi < 180°$ and $\varphi < 0°$ in the Ramachandran space) sampling from one to four times more in 0.5 increments. In other words, seven ensembles were generated respectively having 59 % (standard), 68.3 %, 74.2 %, 78.2 %, 81.2 %, 83.4 % and 85.2 % of φ/ψ angles in the extended region. The χ^2 values from the FM2 algorithm for different levels of extension in the case of ubiquitin are shown in Figure 3. This target function converges approximately 80 % of φ/ψ angles distributed in the extended region of Ramachandran space, in excellent agreement with previous results [37]. Crucially, the χ^2 is smaller in all cases for FM2 ensembles (Supplementary Figure S4 at http://www.biochemsoctrans.org/bst/040/bst0400989add.htm). A comparison of D_{HN}, $D_{C\alpha H\alpha}$ and $D_{HNH\alpha\text{-}1}$ predicted from more extended sampling for both proteins (Figure 4) shows that better sampling improves the predicted value in both algorithms for $D_{C\alpha H\alpha}$, but not for $D_{HNH\alpha\text{-}1}$. In fact, $D_{HNH\alpha\text{-}1}$ shows less dependency on the level of sampling (Supplementary Figures S4 and S5 at http://www.biochemsoctrans.org/bst/040/bst0400989add.htm). Therefore the improvement of $D_{C\alpha H\alpha}$ is attributed not only to backbone geometry, but also to more extended sampling, whereas the improvement of $D_{HNH\alpha\text{-}1}$ is mainly contributed by the energy-minimized geometry. Furthermore, as shown previously [18], reproductions of FM2-predicted long-range RDCs, D_{HNHN+1} and D_{HNHN+2}, are significantly improved with more extended sampling (Figure 5) in addition to the fact that the energy-minimized geometry also improves the predicted values (Supplementary Figure S3A).

We have also repeated the analysis using the genetic algorithm ASTEROIDS [39] to select FM2-generated ensembles to describe backbone conformational sampling from RDCs, and predicting side-chain RDCs for unfolded proteins using three-staggered rotamer populations derived from ^3J-couplings [27]. The new algorithm improves results with no contradiction compared with conclusions based on previous analyses (Supplementary

Figure 3 | The χ^2 for all types of RDCs between experimental data and predicted values along with different levels of sampling of the extended region

Levels of sampling were 59 % (standard), 68.3 %, 74.2 %, 78.2 %, 81.2 %, 83.4 % and 85.2 % of φ/ψ angles in the extended region ($\varphi < 0°$ and $50° < \psi < 180°$) of the Ramachandran space.

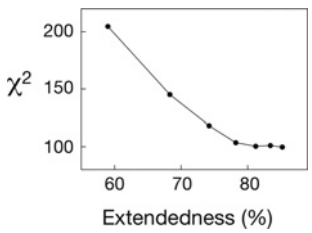


Figure 4 | RDCs predicted from ensembles sample the extended region three times (81.2 %) more than the standard library for ubiquitin (A) and Protein G (B)

Red lines indicate the experimental data and green lines indicate predicted values.

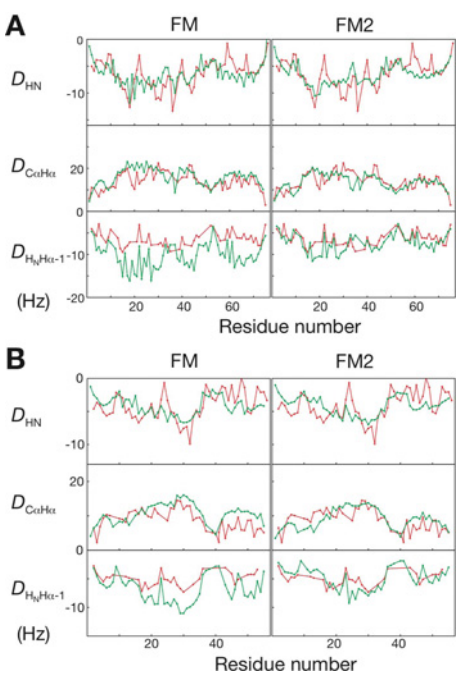
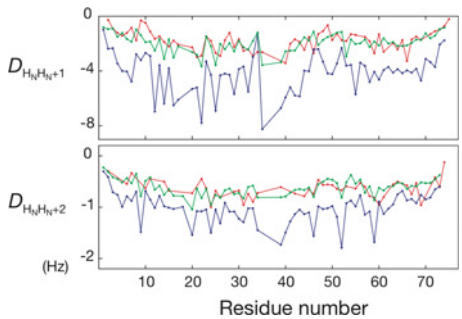


Figure 5 | Comparing long-range RDCs between experimental (red lines) and predicted (blue lines for standard sampling, green lines for extended sampling) values

In this case, only the FM2 algorithm is used for denatured ubiquitin.



Figures S6 and S7 at http://www.biochemsoctrans.org/bst/040/bst0400989add.htm.

In order to determine the general improvement of the amino-acid-specific geometry, we have applied the same approach to GB1. Extensive RDCs were again measured under conditions of urea denaturation (see above). Figure 2(B) shows the comparison of the ability of the two algorithms to reproduce the experimental data, again indicating a better reproduction using the FM2 approach, whereas Figure 4(B) shows the same level of reproduction of the data when the same level of extended conformational sampling is used for GB1 as for ubiquitin. These results indicate that the results are transferrable between the two systems and further underline the remarkable sensitivity of RDCs to the details of local amino acid geometry, as well as to the conformational sampling regime.

Conclusion

It is now generally accepted that many functional proteins do not have well-defined folded structures. In recent years, both experimental and computational approaches were developed to study this type of protein [40–42]. On the computational side, several methods based on sampling-then-selecting were applied on different biologically important systems to have structural insight into IDPs, e.g. protein phosphatase 1 regulators [43], α-synuclein [24] and Sic1 protein [44]. To construct a geometrically correct ensemble of structures for further analysis is critical. In the present paper, we have described a new algorithm to construct such ensembles based on a statistical coil model. This algorithm inherits the advantage of the FM method that sufficiently samples the energy landscape for coil conformation and combines this with more accurate amino-acid-specific geometries from energy-minimized calculations. This new algorithm results in a better reproduction of experimental RDCs and is generally applicable for further studies to characterize the conformational properties of IDPs. In addition, the IC-based algorithm also facilitates side-chain construction [21], surface osmolyte simulation [21], spin-label distribution sampling, and proline *cis–trans* isomer simulation.

Funding

We acknowledge financial support from FINOVI, a MALZ TAU-STRUCT grant from the Agence Nationale de Recherche (France) and the National Science Council of Taiwan (to J.-r.H.)

References

1 Dunker, A.K., Silman, I., Uversky, V.N. and Sussman, J.L. (2008) Function and structure of inherently disordered proteins. Curr. Opin. Struct. Biol. **18**, 756–764
2 Dobson, C.M. (2003) Protein folding and misfolding. Nature **426**, 884–890
3 Wright, P.E. and Dyson, H.J. (2009) Linking folding and binding. Curr. Opin. Struct. Biol. **19**, 31–38
4 Xue, Y. and Skrynnikov, N.R. (2011) Motion of a disordered polypeptide chain as studied by paramagnetic relaxation enhancements, ^{15}N relaxation, and molecular dynamics simulations: how fast is segmental diffusion in denatured ubiquitin? J. Am. Chem. Soc. **133**, 14614–14628
5 Lindorff-Larsen, K., Trbovic, N., Maragakis, P., Piana, S. and Shaw, D.E. (2012) Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. J. Am. Chem. Soc. **134**, 3787–3791
6 Best, R.B., Buchete, N.V. and Hummer, G. (2008) Are current molecular dynamics force fields too helical? Biophys. J. **95**, L07–L09
7 Shaw, D.E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R.O., Eastwood, M.P., Bank, J.A., Jumper, J.M., Salmon, J.K., Shan, Y. and Wriggers, W. (2010) Atomic-level characterization of the structural dynamics of proteins. Science **330**, 341–346
8 Lindorff-Larsen, K., Piana, S., Dror, R.O. and Shaw, D.E. (2011) How fast-folding proteins fold. Science **334**, 517–520
9 Bernadó, P., Blanchard, L., Timmins, P., Marion, D., Ruigrok, R.W. and Blackledge, M. (2005) A structural model for unfolded proteins from residual dipolar couplings and small-angle X-ray scattering. Proc. Natl. Acad. Sci. U.S.A. **102**, 17002–17007
10 Ozenne, V., Bauer, F., Salmon, L., Huang, J.R., Jensen, M.R., Segard, S., Bernadó, P., Charavay, C. and Blackledge, M. (2012) Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. Bioinformatics **28**, 1463–1470
11 Jha, A.K., Colubri, A., Freed, K.F. and Sosnick, T.R. (2005) Statistical coil model of the unfolded state: resolving the reconciliation problem. Proc. Natl. Acad. Sci. U.S.A. **102**, 13099–13104
12 Smith, L.J., Bolin, K.A., Schwalbe, H., MacArthur, M.W., Thornton, J.M. and Dobson, C.M. (1996) Analysis of main chain torsion angles in proteins: prediction of NMR coupling constants for native and random coil conformations. J. Mol. Biol. **255**, 494–506
13 Dyson, H.J. and Wright, P.E. (2004) Unfolded proteins and protein folding studied by NMR. Chem. Rev. **104**, 3607–3622
14 Shortle, D. and Ackerman, M.S. (2001) Persistence of native-like topology in a denatured protein in 8 M urea. Science **293**, 487–489
15 Fieber, W., Kristjansdottir, S. and Poulsen, F.M. (2004) Short-range, long-range and transition state interactions in the denatured state of ACBP from residual dipolar couplings. J. Mol. Biol. **339**, 1191–1199
16 Mohana-Borges, R., Goto, N.K., Kroon, G.J., Dyson, H.J. and Wright, P.E. (2004) Structural characterization of unfolded states of apomyoglobin using residual dipolar couplings. J. Mol. Biol. **340**, 1131–1142
17 Bernadó, P., Bertoncini, C.W., Griesinger, C., Zweckstetter, M. and Blackledge, M. (2005) Defining long-range order and local disorder in native α-synuclein using residual dipolar couplings. J. Am. Chem. Soc. **127**, 17968–17969
18 Meier, S., Grzesiek, S. and Blackledge, M. (2007) Mapping the conformational landscape of urea-denatured ubiquitin using residual dipolar couplings. J. Am. Chem. Soc. **129**, 9799–9807
19 Jensen, M.R., Houben, K., Lescop, E., Blanchard, L., Ruigrok, R.W. and Blackledge, M. (2008) Quantitative conformational analysis of partially folded proteins from residual dipolar couplings: application to the molecular recognition element of Sendai virus nucleoprotein. J. Am. Chem. Soc. **130**, 8055–8061

20 Blackledge, M. (2005) Recent progress in the study of biomolecular structure and dynamics in solution from residual dipolar couplings. Prog. Nucl. Magn. Reson. Spectrosc. **46**, 23–61
21 Huang, J.R., Gabel, F., Jensen, M.R., Grzesiek, S. and Blackledge, M. (2012) Sequence-specific mapping of the interaction between urea and unfolded ubiquitin from ensemble analysis of NMR and small angle scattering data. J. Am. Chem. Soc. **134**, 4429–4436
22 Brooks, B.R., Brooks, 3rd, C.L., Mackerell, Jr, A.D., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S. et al. (2009) CHARMM: the biomolecular simulation program. J. Comput. Chem. **30**, 1545–1614
23 Mukrasch, M.D., Markwick, P., Biernat, J., Bergen, M., Bernadó, P., Griesinger, C., Mandelkow, E., Zweckstetter, M. and Blackledge, M. (2007) Highly populated turn conformations in natively unfolded tau protein identified from residual dipolar couplings and molecular simulation. J. Am. Chem. Soc. **129**, 5235–5243
24 Salmon, L., Nodet, G., Ozenne, V., Yin, G., Jensen, M.R., Zweckstetter, M. and Blackledge, M. (2010) NMR characterization of long-range order in intrinsically disordered proteins. J. Am. Chem. Soc. **132**, 8407–8418
25 Jensen, M.R., Salmon, L., Nodet, G. and Blackledge, M. (2010) Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts. J. Am. Chem. Soc. **132**, 1270–1272
26 Bernadó, P., Mylonas, E., Petoukhov, M.V., Blackledge, M. and Svergun, D.I. (2007) Structural characterization of flexible proteins using small-angle X-ray scattering. J. Am. Chem. Soc. **129**, 5656–5664
27 Vajpai, N., Gentner, M., Huang, J.R., Blackledge, M. and Grzesiek, S. (2010) Side-chain χ_1 conformations in urea-denatured ubiquitin and protein G from ^3J coupling constants and residual dipolar couplings. J. Am. Chem. Soc. **132**, 3196–3203
28 Zweckstetter, M. and Bax, A. (2000) Prediction of sterically induced alignment in a dilute liquid crystalline phase: aid to protein structure determination by NMR. J. Am. Chem. Soc. **122**, 3791–3792
29 van Lune, F., Manning, L., Dijkstra, K., Berendsen, H.J. and Scheek, R.M. (2002) Order-parameter tensor description of HPr in a medium of oriented bicelles. J. Biomol. NMR **23**, 169–179
30 Huang, J.R. and Grzesiek, S. (2010) Ensemble calculations of unstructured proteins constrained by RDC and PRE data: a case study of urea-denatured ubiquitin. J. Am. Chem. Soc. **132**, 694–705
31 Moltke, S. and Grzesiek, S. (1999) Structural constraints from residual tensorial couplings in high resolution NMR without an explicit term for the alignment tensor. J. Biomol. NMR **15**, 77–82
32 Yao, L., Vögeli, B., Ying, J. and Bax, A. (2008) NMR determination of amide N-H equilibrium bond length from concerted dipolar coupling measurements. J. Am. Chem. Soc. **130**, 16518–16520
33 Levitt, M. (1976) A simplified representation of protein conformations for rapid simulation of protein folding. J. Mol. Biol. **104**, 59–107
34 Hus, J.C., Marion, D. and Blackledge, M. (2001) Determination of protein backbone structure using only residual dipolar couplings. J. Am. Chem. Soc. **123**, 1541–1542
35 Jensen, M.R., Communie, G., Ribeiro, Jr, E.A., Martinez, N., Desfosses, A., Salmon, L., Mollica, L., Gabel, F., Jamin, M., Longhi, S. et al. (2011) Intrinsic disorder in measles virus nucleocapsids. Proc. Natl. Acad. Sci. U.S.A. **108**, 9839–9844
36 Gabel, F., Jensen, M.R., Zaccaï, G. and Blackledge, M. (2009) Quantitative modelfree analysis of urea binding to unfolded ubiquitin using a combination of small angle X-ray and neutron scattering. J. Am. Chem. Soc. **131**, 8769–8771
37 Bernado, P. and Blackledge, M. (2009) A self-consistent description of the conformational behavior of chemically denatured proteins from NMR and small angle scattering. Biophys. J. **97**, 2839–2845
38 Kokubo, H., Hu, C.Y. and Pettitt, B.M. (2011) Peptide conformational preferences in osmolyte solutions: transfer free energies of decaalanine. J. Am. Chem. Soc. **133**, 1849–1858
39 Nodet, G., Salmon, L., Ozenne, V., Meier, S., Jensen, M.R. and Blackledge, M. (2009) Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings. J. Am. Chem. Soc. **131**, 17908–17918
40 Schneider, R., Huang, J.R., Yao, M., Communie, G., Ozenne, V., Mollica, L., Salmon, L., Jensen, M.R. and Blackledge, M. (2012) Towards a robust description of intrinsic protein disorder using nuclear magnetic resonance spectroscopy. Mol. Biosyst. **8**, 58–68
41 Tompa, P. (2011) Unstructural biology coming of age. Curr. Opin. Struct. Biol. **21**, 419–425
42 Uversky, V.N. and Dunker, A.K. (2010) Understanding protein non-folding. Biochim. Biophys. Acta **1804**, 1231–1264
43 Marsh, J.A., Dancheck, B., Ragusa, M.J., Allaire, M., Forman-Kay, J.D. and Peti, W. (2010) Structural diversity in free and bound states of intrinsically disordered protein phosphatase 1 regulators. Structure **18**, 1094–1103
44 Mittag, T., Marsh, J., Grishaev, A., Orlicky, S., Lin, H., Sicheri, F., Tyers, M. and Forman-Kay, J.D. (2010) Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase. Structure **18**, 494–506

Received 19 July 2012
doi:10.1042/BST20120187

SUPPLEMENTARY ONLINE DATA

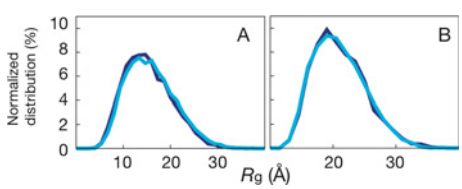
Residual dipolar couplings measured in unfolded proteins are sensitive to amino-acid-specific geometries as well as local conformational sampling

Jie-rong Huang*, Martin Gentner†, Navratna Vajpai†, Stephan Grzesiek† and Martin Blackledge*[1]

*Institut de Biologie Structurale Jean-Pierre Ebel, CEA-CNRS-UJF UMR 5075, 41 Rue Jules Horowitz, Grenoble 38027, France, and †Division of Structural Biology, Biozentrum, University of Basel, Klingelbergstrasse 50, 4056 Basel, Switzerland

Figure S1 | **The R_g distributions generated by FM (light blue) and FM2 (dark blue) for 8 M urea denatured ubiquitin (A) and Protein G (B) at pH 2.5**

Averaged R_g for ubiquitin are 25.64 Å (FM) compared with 25.13 Å (FM2) and for Protein G are 20.65 Å (FM) compared with 20.72 Å (FM2).



[1]To whom correspondence should be addressed (email martin.blackledge@ibs.fr).

Figure S2 | The φ/ψ angle distribution (from highly populated: red to null population: blue) calculated from FM- or FM2-generated structures for ubiquitin (exemplified from residues 2–28) shows no significant difference of the sampling

Single-letter codes are used for amino acids.

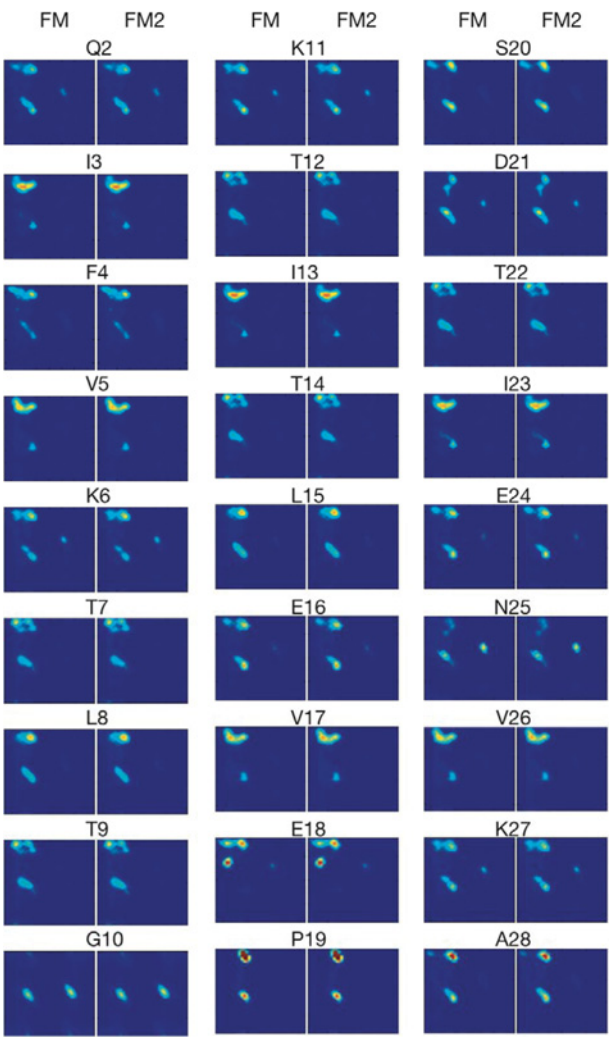


Figure S3 | Comparison of all types of measured RDCs (red) with predicted values (FM, light blue; FM2, dark blue) for denatured ubiquitin (**A**) and Protein G (**B**)

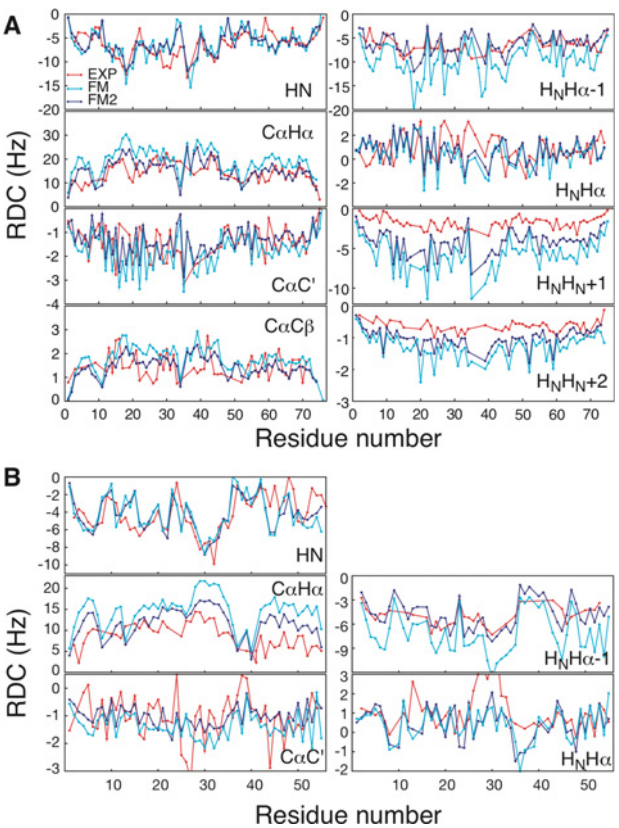


Figure S4 | Comparison of χ^2 between experimental data and predicted values from FM or FM2 for different type of RDCs in the case of denatured ubiquitin along with the level of extension

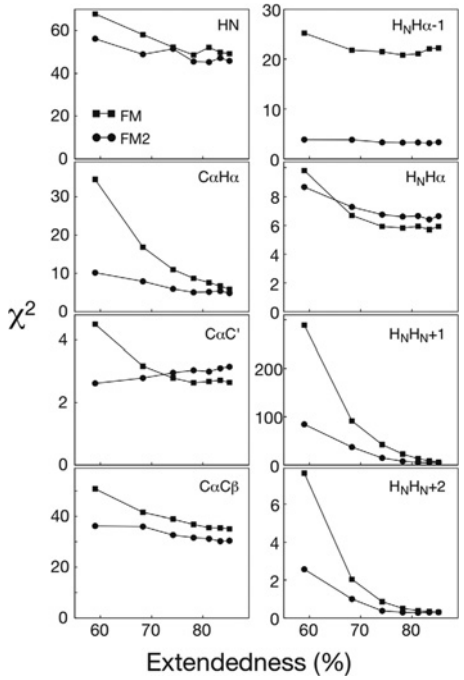


Figure S5 | RDCs for denatured ubiquitin predict from different levels of extension (dark blue, 59.0 %; light blue, 74.2 %; light green, 81.2 %; dark green, 85.2 %) by FM2 algorithm (all scaled to experimental D_{HN}) in comparison with experimental data (red)

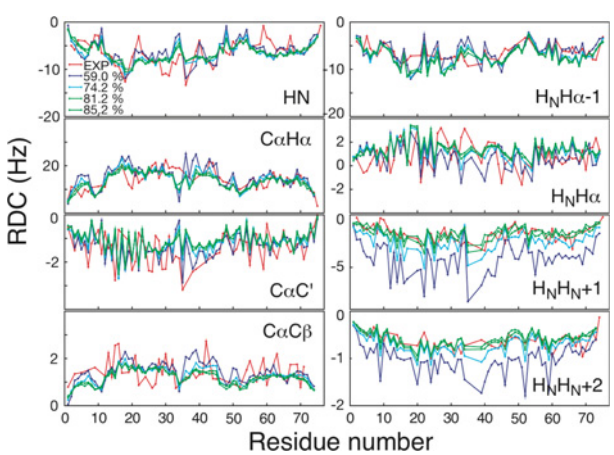


Figure S6 | Analysis of FM2 structures with ASTEROIDS selection [1] RDCs calculated from 200 selected structures (blue lines) are compared with experimental data (red lines). Two cases were performed: (**A**) selection with one scaling factor, and (**B**) selection with a second scaling factor for $D_{HNH\alpha-1}$, D_{HNHN+1} and D_{HNHN+2}. Although the values back-calculated from the selected structures with one scaling factor is improved compared with the original research, the results from two scaling factors are still better to reproduce experimental data (the χ^2 is reduced from 1527.076 with a scaling factor of 0.2 to 1166.484 with two scaling factors of 0.223 and 0.172). This implies that additional local conformational dynamics may be still overlooked using the current FM method as with the previous hypothesis. The χ^2 differs from that in Figure S3 because different weighting factors were applied in ASTEROIDS selection instead of simply from experimental errors.

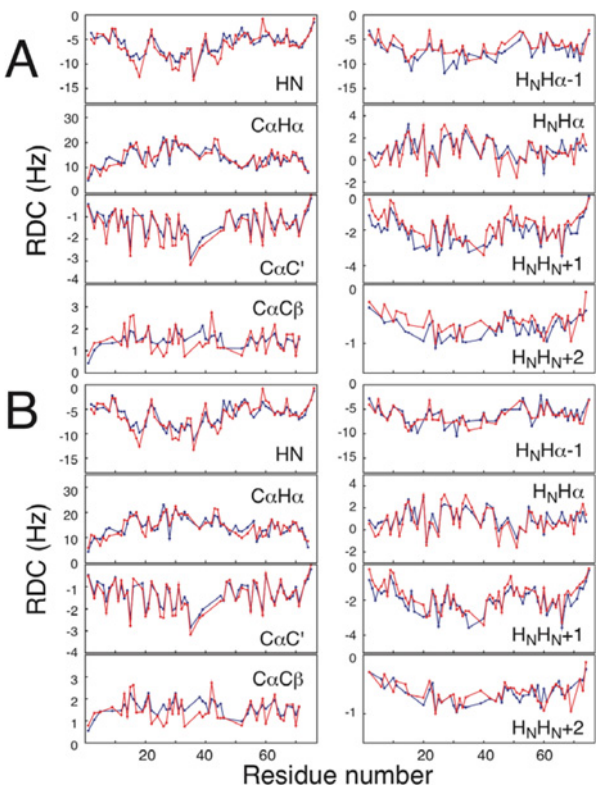


Figure S7 | **Since the side-chain Cβ geometry is modified in FM2, we reanalysed the side-chain $D_{C\beta H\beta}$ of ubiquitin (A) and Protein G (B) [2]**

The predicted $D_{C\beta H\beta}$ is based on a pseudo-side-chain three-stagger rotamer model with best fit to the *J*-coupling data. The same analysis was performed, but instead of using the Pearson correlation coefficient, which provides the correlation but is less sensitive to the scaling factor problem, the χ^2 was compared. The χ^2 of side-chain RDCs is improved in the ensembles generated by the FM2 algorithm (grey bars, FM; black bars, FM2). The predicted side-chain RDCs are also calculated for ensembles with different levels of extendedness, which was not performed in the previous study. It shows that with more extended conformational sampling, side-chain RDCs also fit better with experimental data consistent with backbone RDCs.

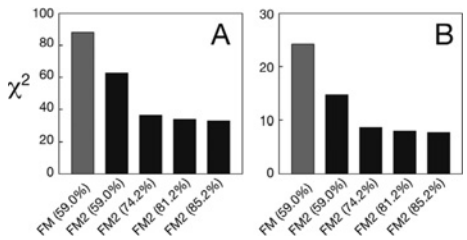


Table S1 | **Internal co-ordinates for each amino acid**

ICs used in the FM2 algorithm are adapted from CHARMM force field topology file. IC is sorted according to the angles between the last three atoms (the second parameter column) to ease comparison. The geometry for the FM model is converted into the same format for comparison.

(a)

FM	**+N**	**CA**	***C**	**O**	**180.00**	**121.50**	**1.2397**
S	+N	CA	*C	O	180.00	120.25	1.2290
T	+N	CA	*C	O	180.00	120.30	1.2294
N	+N	CA	*C	O	180.00	120.32	1.2282
P	+N	CA	*C	O	177.15	120.46	1.2316
F	+N	CA	*C	O	180.00	120.49	1.2287
L	+N	CA	*C	O	180.00	120.56	1.2299
Q	+N	CA	*C	O	180.00	120.59	1.2291
I	+N	CA	*C	O	180.00	120.59	1.2300
C	+N	CA	*C	O	180.00	120.59	1.2306
M	+N	CA	*C	O	180.00	120.64	1.2288
Y	+N	CA	*C	O	180.00	120.67	1.2287
V	+N	CA	*C	O	180.00	120.70	1.2297
D	+N	CA	*C	O	180.00	120.71	1.2330
K	+N	CA	*C	O	180.00	120.79	1.2277
G	+N	CA	*C	O	180.00	120.85	1.2289
E	+N	CA	*C	O	180.00	121.07	1.2306
W	+N	CA	*C	O	180.00	121.08	1.2304
H	+N	CA	*C	O	180.00	121.20	1.2284
R	+N	CA	*C	O	180.00	121.40	1.2271
A	+N	CA	*C	O	180.00	122.52	1.2297

(b)

FM	**+N**	**C**	**CA**	**N**	**180.00**	**109.00**	**1.4543**
N	+N	C	CA	N	180.00	105.23	1.4510
V	+N	C	CA	N	180.00	105.54	1.4570
D	+N	C	CA	N	180.00	105.63	1.4490
S	+N	C	CA	N	180.00	105.81	1.4579
C	+N	C	CA	N	180.00	105.89	1.4533
L	+N	C	CA	N	180.00	106.05	1.4508
T	+N	C	CA	N	180.00	106.09	1.4607
M	+N	C	CA	N	180.00	106.31	1.4510
I	+N	C	CA	N	180.00	106.35	1.4542
F	+N	C	CA	N	180.00	106.38	1.4504
Y	+N	C	CA	N	180.00	106.52	1.4501
Q	+N	C	CA	N	180.00	106.57	1.4506
E	+N	C	CA	N	180.00	107.27	1.4512
K	+N	C	CA	N	180.00	107.29	1.4504
W	+N	C	CA	N	180.00	107.69	1.4507
G	+N	C	CA	N	180.00	108.94	1.4553
R	+N	C	CA	N	180.00	109.86	1.4544
P	+N	C	CA	N	180.00	110.86	1.4585
H	+N	C	CA	N	180.00	112.03	1.4548
A	+N	C	CA	N	180.00	114.44	1.4592

Table S1 | **Continued**

(c)

FM	C	CA	N	– C	180.00	121.40	1.3338
R	C	CA	N	– C	180.00	122.45	1.3496
G	C	CA	N	– C	180.00	122.82	1.3475
P	C	CA	N	– C	– 76.12	122.94	1.3366
W	C	CA	N	– C	180.00	123.51	1.3482
K	C	CA	N	– C	180.00	123.57	1.3482
Y	C	CA	N	– C	180.00	123.81	1.3476
F	C	CA	N	– C	180.00	123.89	1.3476
Q	C	CA	N	– C	180.00	123.93	1.3477
C	C	CA	N	– C	180.00	123.93	1.3479
H	C	CA	N	– C	180.00	123.93	1.3489
N	C	CA	N	– C	180.00	124.05	1.3480
T	C	CA	N	– C	180.00	124.12	1.3471
I	C	CA	N	– C	180.00	124.16	1.3470
M	C	CA	N	– C	180.00	124.21	1.3478
L	C	CA	N	– C	180.00	124.31	1.3474
S	C	CA	N	– C	180.00	124.37	1.3474
E	C	CA	N	– C	180.00	124.45	1.3471
V	C	CA	N	– C	180.00	124.57	1.3482
D	C	CA	N	– C	180.00	125.31	1.3465
A	C	CA	N	– C	180.00	126.49	1.3551

(d)

FM	CA	N	– C	– CA	180.00	116.60	1.5241
P	CA	N	– C	– CA	180.00	116.12	1.5399
H	CA	N	– C	– CA	180.00	116.49	1.5225
A	CA	N	– C	– CA	180.00	116.84	1.5390
D	CA	N	– C	– CA	180.00	117.06	1.5315
R	CA	N	– C	– CA	180.00	117.12	1.5227
E	CA	N	– C	– CA	180.00	117.25	1.5216
K	CA	N	– C	– CA	180.00	117.27	1.5187
Y	CA	N	– C	– CA	180.00	117.33	1.5232
N	CA	N	– C	– CA	180.00	117.38	1.5245
W	CA	N	– C	– CA	180.00	117.57	1.5202
G	CA	N	– C	– CA	180.00	117.60	1.4971
F	CA	N	– C	– CA	180.00	117.65	1.5229
T	CA	N	– C	– CA	180.00	117.69	1.5162
S	CA	N	– C	– CA	180.00	117.72	1.5166
Q	CA	N	– C	– CA	180.00	117.72	1.5180
M	CA	N	– C	– CA	180.00	117.74	1.5195
V	CA	N	– C	– CA	180.00	117.83	1.5180
L	CA	N	– C	– CA	180.00	117.93	1.5184
I	CA	N	– C	– CA	180.00	117.97	1.5190
C	CA	N	– C	– CA	180.00	118.30	1.5202

(e)

FM	– C	CA	*N	HN	180.00	119.30	1.0199
P	– C	CA	*N	CD	178.51	112.75	1.4624
D	– C	CA	*N	HN	180.00	112.94	0.9966
E	– C	CA	*N	HN	180.00	113.99	0.9961
S	– C	CA	*N	HN	180.00	114.18	0.9999
I	– C	CA	*N	HN	180.00	114.19	0.9978
L	– C	CA	*N	HN	180.00	114.26	0.9979
T	– C	CA	*N	HN	180.00	114.26	0.9995
M	– C	CA	*N	HN	180.00	114.39	0.9978
V	– C	CA	*N	HN	180.00	114.41	0.9966
Q	– C	CA	*N	HN	180.00	114.45	0.9984
F	– C	CA	*N	HN	180.00	114.47	0.9987
N	– C	CA	*N	HN	180.00	114.49	0.9992
Y	– C	CA	*N	HN	180.00	114.54	0.9986
C	– C	CA	*N	HN	180.00	114.77	0.9982
W	– C	CA	*N	HN	180.00	115.02	0.9972
K	– C	CA	*N	HN	180.00	115.11	0.9988
A	– C	CA	*N	HN	180.00	115.42	0.9996
G	– C	CA	*N	HN	180.00	115.62	0.9992
R	– C	CA	*N	HN	180.00	116.67	0.9973
H	– C	CA	*N	HN	180.00	118.80	1.0041

(f)

FM	N	C	*CA	CB	120.00	109.00	1.5341
G	N	C	*CA	HA1	117.86	108.03	1.0814
H	N	C	*CA	CB	125.13	109.38	1.5533
A	N	C	*CA	CB	123.23	111.09	1.5461
W	N	C	*CA	CB	122.68	111.23	1.5560
V	N	C	*CA	CB	122.95	111.23	1.5660
K	N	C	*CA	CB	122.23	111.36	1.5568
S	N	C	*CA	CB	124.75	111.40	1.5585
Q	N	C	*CA	CB	121.91	111.68	1.5538
E	N	C	*CA	CB	121.90	111.71	1.5516
P	N	C	*CA	CB	113.74	111.74	1.5399
M	N	C	*CA	CB	121.62	111.88	1.5546
C	N	C	*CA	CB	121.79	111.98	1.5584
L	N	C	*CA	CB	121.52	112.12	1.5543
R	N	C	*CA	CB	123.64	112.26	1.5552
Y	N	C	*CA	CB	122.27	112.34	1.5606
F	N	C	*CA	CB	122.49	112.45	1.5594
T	N	C	*CA	CB	126.46	112.74	1.5693
I	N	C	*CA	CB	124.22	112.93	1.5681
N	N	C	*CA	CB	121.18	113.04	1.5627
D	N	C	*CA	CB	122.33	114.10	1.5619

Table S1 | **Continued**

(g)

FM	N	C	*CA	HA	— 120.00	109.00	1.1098
A	N	C	*CA	HA	— 120.45	106.39	1.0840
T	N	C	*CA	HA	— 114.92	106.53	1.0817
R	N	C	*CA	HA	— 117.93	106.61	1.0836
H	N	C	*CA	HA	— 119.20	106.72	1.0832
D	N	C	*CA	HA	— 116.40	106.77	1.0841
I	N	C	*CA	HA	— 115.63	106.81	1.0826
W	N	C	*CA	HA	— 117.02	106.92	1.0835
F	N	C	*CA	HA	— 115.63	107.05	1.0832
Y	N	C	*CA	HA	— 116.04	107.15	1.0833
E	N	C	*CA	HA	— 118.06	107.26	1.0828
S	N	C	*CA	HA	— 115.56	107.30	1.0821
K	N	C	*CA	HA	— 116.88	107.36	1.0833
V	N	C	*CA	HA	— 117.24	107.46	1.0828
Q	N	C	*CA	HA	— 116.82	107.53	1.0832
L	N	C	*CA	HA	— 116.50	107.57	1.0824
M	N	C	*CA	HA	— 116.98	107.57	1.0832
N	N	C	*CA	HA	— 115.52	107.63	1.0848
C	N	C	*CA	HA	— 116.34	107.71	1.0837
G	N	C	*CA	HA2	— 118.12	107.95	1.0817
P	N	C	*CA	HA	— 122.40	109.09	1.0837

References

1 Nodet, G., Salmon, L., Ozenne, V., Meier, S., Jensen, M.R. and Blackledge, M. (2009) Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings. J. Am. Chem. Soc. **131**, 17908–17918

2 Vajpai, N., Gentner, M., Huang, J.R., Blackledge, M. and Grzesiek, S. (2010) Side-chain χ_1 conformations in urea-denatured ubiquitin and protein G from ^3J coupling constants and residual dipolar couplings. J. Am. Chem. Soc. **132**, 3196–3203

Received 19 July 2012
doi:10.1042/BST20120187