J|A|C|S ARTICLES

Published on Web 03/11/2010

Refinement of Ensembles Describing Unstructured Proteins Using NMR Residual Dipolar Couplings

Santi Esteban-Martín,[†] Robert Bryn Fenwick,[†] and Xavier Salvatella*[,†,‡]

ICREA and Institute for Research in Biomedicine, Barcelona, Baldiri Reixac 10, 08028 Barcelona, Spain

Received August 24, 2009; E-mail: xavier.salvatella@irbbarcelona.org

Abstract: Residual dipolar couplings (RDCs) are unique probes of the structural and dynamical properties of biomolecules on the sub-millisecond time scale that can be used as restraints in ensemble molecular dynamics simulations to study the relationship between macromolecular motion and biological function. To date, however, this powerful strategy is applicable only to molecules that do not undergo shape changes on the time scale sampled by RDCs, thus preventing the study of key biological macromolecules such as multidomain and unstructured proteins. In this work, we circumvent this limitation by using an algorithm that explicitly computes the individual alignment tensors of the different ensemble members from their coordinates at each step in the simulation. As a first application, we determine an ensemble of conformations that accurately describes the structure and dynamics of chemically denatured ubiquitin. In analogy to dynamic refinement of folded, globular proteins, where simulations are initiated from average structures, we use statistical coil models as starting configuration because they represent the best available descriptions of unstructured proteins. We find that refinement with RDCs causes significant structural corrections and yields an ensemble that is in complete agreement with the measured RDCs and presents transient mid-range inter-residue interactions between strands β1 and β2 of the native protein, also observed in other studies based on trans-hydrogen bond $^{3}J_{NC'}$ scalar couplings and paramagnetic relaxation enhancements. Finally, and in spite of the high structural heterogeneity of the refined ensemble, we find that it can be cross-validated against RDCs not used to restrain the simulation. This method increases the range of systems that can be studied using ensemble simulations restrained by RDCs and is likely to yield new insights into how the large-scale motions of macromolecules relate to biological function.

1. Introduction

NMR residual dipolar couplings[1,2] (RDCs), which report on the orientation of bond vectors in molecules that are partially aligned with respect to the laboratory frame, provide unique information on the relative position of distant elements of structure. RDCs have allowed the refinement of the structure of proteins and nucleic acids to unprecedented resolution;[3−5] more recently, their unique sensitivity to motions up to the sub-millisecond time scale has been exploited to describe macromolecular dynamics on a time scale not probed by the analysis of heteronuclear relaxation rates, which reports on motions faster than the rotational correlation time of the molecule,[6,7] and relaxation dispersion methods, which report on relatively slow structural fluctuations that modulate chemical shifts.[8−10]

A variety of computational approaches have been developed to extract the structural and dynamical information contained in RDCs, including the use of motional models that describe the essential structural fluctuations of the protein backbone,[11] restrained ensemble molecular dynamics (MD) simulations that exploit molecular mechanics force fields,[12−14] and methods based on selecting ensembles of conformations from unrestrained MD trajectories.[15,16] The ability of RDCs to report on both the structure and the dynamics of macromolecules at high resolution[17−20] has recently been used to gain new insights into

[†] Institute for Research in Biomedicine, Barcelona.
[‡] ICREA.

(1) Tolman, J. R.; Flanagan, J. M.; Kennedy, M. A.; Prestegard, J. H. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 9279–9283.
(2) Tjandra, N.; Bax, A. *Science* **1997**, *278*, 1111–1114.
(3) Vermeulen, A.; Zhou, H.; Pardi, A. *J. Am. Chem. Soc.* **2000**, *122*, 9638–9647.
(4) Chou, J. J.; Li, S.; Klee, C. B.; Bax, A. *Nat. Struct. Biol.* **2001**, *8*, 990–997.
(5) Mccallum, S.; Pardi, A. *J. Mol. Biol.* **2003**, *326*, 1037–1050.
(6) Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, *104*, 4546–4559.
(7) Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, 4559–4570.
(8) Palmer, A. G.; Kroenke, C. D.; Loria, J. P. *Methods Enzymol.* **2001**, *339*, 204–238.
(9) Mulder, F. A. A.; Mittermaier, A.; Hon, B.; Dahlquist, F. W.; Kay, L. E. *Nat. Struct. Biol.* **2001**, *8*, 932–935.
(10) Akke, M. *Curr. Opin. Struct. Biol.* **2002**, *12*, 642–647.
(11) Bouvignies, G.; Bernadó, P.; Meier, S.; Cho, K.; Grzesiek, S.; Brüschweiler, R.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13885–13890.
(12) Clore, G. M.; Schwieters, C. D. *Biochemistry* **2004**, *43*, 10678–10691.
(13) De Simone, A.; Richter, B.; Salvatella, X.; Vendruscolo, M. *J. Am. Chem. Soc.* **2009**, *131*, 3810–3811.
(14) Lange, O. F.; Lakomek, N. A.; Farès, C.; Schröder, G. F.; Walter, K. F.; Becker, S.; Meiler, J.; Grubmüller, H.; Griesinger, C.; de Groot, B. L. *Science* **2008**, *320*, 1471–1475.
(15) Chen, Y.; Campbell, S.; Dokholyan, N. *Biophys. J.* **2007**, *93*, 2300–2306.
(16) Frank, A. T.; Stelzer, A.; Al-Hashimi, H.; Andricioaei, I. *Nucleic Acids Res.* **2009**, *37*, 3670–3679.
(17) Bax, A.; Grishaev, A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 563–570.

10.1021/ja906995x

the mechanisms of key biological processes such as the transfer of structural information across macromolecular structures[11,21] and conformational selection in molecular recognition,[14] thus certifying the key role that RDCs can play in understanding how the motions of macromolecules are related to their function.

A number of important applications of RDCs to dynamic refinement have, however, remained relatively unexplored due to challenges in the structural interpretation of this NMR parameter when the motions of the molecule and those of the tensor that describes its alignment with respect to the laboratory are correlated. This can occur when macromolecular dynamics involve important shape changes that modify the degree and main direction of alignment of the molecule and therefore prevent the use of a single, average alignment tensor.[22−24] The effect of such correlations on the analysis of the dynamics of globular, single-domain proteins has been shown to be negligible,[24] but it can be significant—and needs to be addressed—in very flexible systems such as multidomain[25] and unstructured proteins,[26−29] as well as in RNAs,[30] because they can potentially lead to artifacts if the RDCs are analyzed using a single alignment tensor.[24] Given the importance of accurately characterizing the conformational fluctuations of both multidomain and unstructured proteins in structural biology, there is a pressing need to develop methods to extract the structural and dynamical information contained in RDCs that are applicable to such challenging systems.

In this article we present an approach to the structural and dynamical characterization of such macromolecules based on ensemble MD simulations restrained by RDCs measured in steric alignment that, unlike methods of dynamic refinement employed to date,[12−14] does not require that all conformations simulated simultaneously have the same degree and main direction of alignment. Instead, it allows the tensor to vary across the simulated ensemble by explicitly computing it for each ensemble member, from its coordinates, using a fast and independently validated algorithm.[31,32] Since it does not require the use of structural restraints to enforce a single molecular reference frame,[12,13] this approach makes it possible to analyze the dynamics of flexible macromolecules such as RNAs as well as multidomain and unstructured proteins.

To illustrate how the method performs, we present its application to the problem of determining ensembles of conformations that describe the structure and dynamics of unstructured proteins, which has become a key challenge in structural biology due to the increasing awareness that intrinsically disordered proteins (IDPs) can represent a very significant fraction of the human proteome.[33] These proteins, which present no persistent secondary and tertiary structure, often play the role of hubs in protein interaction networks,[34] where in many cases they become structured upon binding their partners[35] in a process where binding and folding are tightly coupled.[36]

Although RDCs contain a large amount of structural information, they are unlikely, in the absence of valid structural models, to be sufficient to uniquely identify ensembles of conformations that accurately reflect the structural heterogeneity of unstructured proteins. In this work, therefore, we approached the determination of such ensembles by refining statistical coil models (SCMs), derived from an analysis of the structural propensities of residues in loops and termini of protein structures deposited in the Protein Databank (PDB), which have been shown to be in good qualitative agreement with large sets of RDCs as well as with small angle X-ray scattering (SAXS) data.[28,29] By using ensemble simulations restrained by RDCs measured in a steric alignment medium, we obtained ensembles in unprecedented agreement with experiment and validated them both by using cross-validation against RDCs not used to restrain the simulation and by comparison with the results of NMR experiments based on different NMR parameters such as trans-hydrogen bond $^3J_{NC'}$ scalar couplings[37] and paramagnetic relaxation enhancements (PREs).[38]

2. Theory

Ensemble simulations restrained by RDCs are a very powerful approach to the study of macromolecular dynamics on the submillisecond time scale because they provide a physical framework, established by molecular mechanics force fields, for the efficient extraction of the dynamical information contained in this NMR parameter.[12,13] In these methods, several conformations are simulated simultaneously to reflect the structural heterogeneity of the macromolecule, and empirical quadratic potentials E_{RDC} (eq 1) are added to the force field energy to bias the RDCs back-calculated from the simulated ensemble to agree with experiment,

$$E_{RDC} = \alpha_{RDC}(D^{calc} - D^{exp})^2 \quad (1)$$

where α_{RDC} is a force constant that needs to be adjusted for each RDC type, so that the average violation in the RDC—that is, the root-mean-squared deviation (rmsd) in hertz—is of the same order as the error in the measurement of D^{exp}.

Contrary to what is the case for NMR parameters such as scalar couplings[39] and chemical shifts,[40] the value of the RDC not only depends on the geometry of the relevant nuclei but also is a function

(18) Blackledge, M. *Prog. Nucl. Magn. Reson. Spectrosc.* **2005**, *46*, 23–61.
(19) Tolman, J. R.; Ruan, K. *Chem. Rev.* **2006**, *106*, 1720–1736.
(20) Al-Hashimi, H. *Biopolymers* **2007**, *86*, 345–347.
(21) Zhang, Q.; Stelzer, A.; Fisher, C.; Al-Hashimi, H. *Nature* **2007**, *450*, 1263–1267.
(22) Louhivuori, M.; Otten, R.; Lindorff-Larsen, K.; Annila, A. *J. Am. Chem. Soc.* **2006**, *128*, 4371–4376.
(23) Louhivuori, M.; Otten, R.; Salminen, T.; Annila, A. *J. Biomol. NMR* 2007.
(24) Salvatella, X.; Richter, B.; Vendruscolo, M. *J. Biomol. NMR* **2008**, *40*, 71–81.
(25) Goto, N. K.; Skrynnikov, N. R.; Dahlquist, F. W.; Kay, L. E. *J. Mol. Biol.* **2001**, *308*, 745–764.
(26) Louhivuori, M.; Paakkonen, K.; Fredriksson, K.; Permi, P.; Lounila, J.; Annila, A. *J. Am. Chem. Soc.* **2003**, *125*, 15647–15650.
(27) Fredriksson, K.; Louhivuori, M.; Permi, P.; Annila, A. *J. Am. Chem. Soc.* **2004**, *126*, 12646–12650.
(28) Jha, A. K.; Colubri, A.; Freed, K. F.; Sosnick, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13099–13104.
(29) Bernadó, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R. W.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17002–17007.
(30) Zhang, Q.; Sun, X.; Watt, E. D.; Al-Hashimi, H. *Science* **2006**, *311*, 653–656.
(31) Almond, A.; Axelsen, J. B. *J. Am. Chem. Soc.* **2002**, *124*, 9986–9987.
(32) Wu, B.; Petersen, M.; Girard, F.; Tessari, M.; Wijmenga, S. S. *J. Biomol. NMR* **2006**, *35*, 103–115.
(33) Dunker, A. K.; Silman, I.; Uversky, V. N.; Sussman, J. L. *Curr. Opin. Struct. Biol.* **2008**, *18*, 756–764.
(34) Haynes, C.; Oldfield, C. J.; Ji, F.; Klitgord, N.; Cusick, M. E.; Radivojac, P.; Uversky, V. N.; Vidal, M.; Iakoucheva, L. M. *PLoS Comput. Biol.* **2006**, *2*, e100.
(35) Sugase, K.; Dyson, H. J.; Wright, P. E. *Nature* **2007**, *447*, 1021–1025.
(36) Wright, P. E.; Dyson, H. J. *Curr. Opin. Struct. Biol.* **2009**, *19*, 31–38.
(37) Meier, S.; Strohmeier, M.; Blackledge, M.; Grzesiek, S. *J. Am. Chem. Soc.* **2007**, *129*, 754–755.
(38) Huang, J. R.; Grzesiek, S. *J. Am. Chem. Soc.* **2010**, *132*, 694–705.
(39) Karplus, M. *J. Am. Chem. Soc.* **1963**, *85*, 2870–2871.
(40) Neal, S.; Nip, A. M.; Zhang, H.; Wishart, D. S. *J. Biomol. NMR* **2003**, *26*, 215–240.

of the degree and direction of alignment of the molecule with respect to the laboratory frame, which depends on its structure and the mechanism by which alignment is induced. This information is contained in the alignment tensor **A**, with elements A_{ij}, which is a traceless and symmetric 3×3 matrix defined by five independent elements[2] that are, in most cases, unknown and need to be determined empirically by identifying the values of A_{ij} that maximize, for proteins of known structure, the agreement with experiment of the RDCs back-calculated using eq 2, expressed as quality factor Q (eq 3), where $i,j = (x,y,z)$, μ_0 is the magnetic susceptibility of vacuum, γ_X is the gyromagnetic ratio of nucleus X, h is Planck's constant, r is the internuclear distance, and ϕ_i is the angle between the internuclear vector and axis i of the molecular reference frame where **A** is defined.

$$D^{\text{calc}} = -\frac{\mu_0 \gamma_X \gamma_Y h}{8\pi^3 r^3} \sum_{ij} A_{ij} \cos\phi_i \cos\phi_j \tag{2}$$

$$Q = \frac{\text{RMS}(D^{\text{calc}} - D^{\text{exp}})}{\text{RMS}(D^{\text{exp}})} \tag{3}$$

For the study of the dynamics of globular proteins using restrained ensemble MD simulations, it is often assumed that the contribution to the ensemble-averaged RDC of all conformations adopted by the protein on the relevant time scale can be described by using a single, effective alignment tensor $\mathbf{A}^{\text{eff}}$ so that $\langle D^{\text{calc}} \rangle$ can be expressed using eq 4,[12,13,24]

$$\langle D^{\text{calc}} \rangle = -\frac{\mu_0 \gamma_X \gamma_Y h}{8\pi^3} \sum_{ij} A_{ij}^{\text{eff}} \left\langle \frac{\cos\phi_i \cos\phi_j}{r^3} \right\rangle \tag{4}$$

where N is the number of conformations simulated simultaneously, k runs through the restrained ensemble, and angular brackets denote ensemble-averaged quantities; since the alignment tensor elements A_{ij}^{eff} are not known, they are optimized simultaneously with the coordinates of the ensemble members.

In ensemble MD simulations, the use of a single effective alignment tensor can be implemented by (i) restraining the tensor elements of the different conformations, $A_{ij,k}$, to be minimally different from the five ensemble-averaged tensor elements $\langle A_{ij} \rangle$, using $5N$ quadratic potentials $E_{A,k}$, as shown in eq 5, where α_A is a force constant determined empirically, and to establish a single molecular frame, (ii) preventing ensemble members from rotating relative to one another.[24,41]

$$E_{A,k} = \alpha_A (A_{ij,k} - \langle A_{ij} \rangle)^2 \tag{5}$$

A number of recent studies on the effect of fast (sub-nanosecond) protein dynamics on simple mechanisms of alignment[22,23] have highlighted that changes in the alignment tensor can take place as the structure of the molecules fluctuates, but, as we recently showed,[24] such changes have no practical consequences for the analysis of the molecular dynamics of macromolecules that do not involve shape changes. The RDC restraint presented in eqs 4 and 5 is therefore appropriate for the determination of ensembles for folded, globular proteins[24] but cannot be used to study very flexible macromolecules that undergo significant shape changes, such as multidomain and unstructured proteins and RNAs. In this more complex case, it is necessary to consider explicitly the alignment tensor of each ensemble member individually, as presented in eq 6.

$$\langle D^{\text{calc}} \rangle = -\frac{\mu_0 \gamma_X \gamma_Y h}{8\pi^3} \sum_{ij} \left\langle \frac{A_{ij} \cos\phi_i \cos\phi_j}{r^3} \right\rangle \tag{6}$$

The availability of methods to rapidly compute the alignment tensor of macromolecules in steric alignment media from knowledge of their structure[31,32,42,43] makes it possible to develop algorithms to analyze the structure and dynamics of flexible systems from steric RDCs by explicitly computing the alignment tensor of all ensemble members at each time step during the simulation. In the implementation that we present, called ERIDU (ensemble refinement of intrinsically disordered and unstructured molecules), the alignment tensor is, similarly to recently reported algorithms,[38,44] computed after each MD step using the procedure introduced by Almond and Axelsen[31] that derives the alignment tensor elements A_{ij} from the gyration tensor elements GT_{ij}, which are computed from the coordinates of the macromolecule by using eq 7,

$$\text{GT}_{ij}^2 = \frac{1}{N} \sum_{n=1}^{N} r_i^{(n)} r_j^{(n)} \tag{7}$$

where $i,j = (x,y,z)$, n runs through the N atoms of the molecule, and $r_i^{(n)}$ is the i component of the vector that defines the position of atom n in the molecular frame. As shown by Almond and Axelsen,[31] the alignment tensor **A** and the gyration tensor GT have the same eigenvectors, and the eigenvalues of the former can be derived from those of the latter using the following set of equivalences:

$$A_{xx} \propto 1 - \frac{1}{2}\delta \quad A_{yy} \propto \delta - \frac{1}{2} \quad A_{zz} \propto -\frac{1}{2} - \frac{1}{2}\delta \tag{8}$$

where the value of δ is given by

$$\delta = \frac{\rho_2 - \rho_3}{\rho_1 - \rho_3} \tag{9}$$

where ρ_1, ρ_2, and ρ_3 are in turn derived from the eigenvalues of GT, computed by diagonalization, according to eq 10,

$$\rho_1 = \text{GT}_{xx}^{1/2} \qquad \rho_2 = \text{GT}_{yy}^{1/2} \qquad \rho_3 = GT_{zz}^{1/2} \tag{10}$$

and by considering that the values of A_{ij} thus obtained need to be multiplied by $(\rho_1 - \rho_3)$ to account for the relative degree of alignment of the different ensemble members.

Given that current methods to compute the steric alignment tensor from macromolecular structures cannot determine the absolute degree of alignment, the ensemble-averaged RDCs obtained using eq 6 are globally scaled, prior to the calculation of the penalty E_{RDC}, using eq 1, so as to minimize the rmsd with the experimental RDCs. ERIDU has been implemented in the molecular simulation program CHARMM[45] (version 35b1); the source code and the input files necessary to reproduce this work as well as the resulting ensembles can be obtained from http://lmb.irbbarcelona.org or directly from the authors.

3. Application to the Refinement of Ensembles Describing Unstructured Proteins

To demonstrate the ability of ERIDU to accurately characterize the structure and dynamics of flexible macromolecules, we used this method to determine an ensemble of conformations describing the protein ubiquitin in its chemically denatured state. Similar to the case for the native protein, which has been thoroughly studied by solution NMR[46,47] and is a model system

(41) Clore, G. M.; Schwieters, C. D. *J. Am. Chem. Soc.* **2004**, *126*, 2923–2938.
(42) Zweckstetter, M.; Bax, A. *J. Am. Chem. Soc.* **2000**, *122*, 3791–3792.
(43) Zweckstetter, M.; Hummer, G.; Bax, A. *Biophys. J.* **2004**, *86*, 3444–3460.
(44) Marsh, J. A.; Forman-Kay, J. D. *J. Mol. Biol.* **2009**, *391*, 359–374.
(45) Brooks, B. R.; et al. *J. Comput. Chem.* **2009**, *30*, 1545–1614.
(46) Cornilescu, G.; Marquardt, J. L.; Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 6836–6837.

for the development of approaches to analyze backbone and side-chain dynamics of globular proteins,[14,48−50] chemically denatured ubiquitin has, in recent years, become an extremely useful model system to develop, optimize, and validate experimental and computational approaches to the structural characterization of unstructured proteins.[37,51,52] Interest in such methods has markedly increased due to sequence-based predictions that suggest that a very significant fraction of the eukaryotic proteome is intrinsically disordered,[33] devoid of persistent secondary and tertiary structure and thus not amenable to conventional structural analysis.

NMR spectroscopy is a particularly powerful technique for the study of such challenging systems[53,54] because it can provide residue-specific parameters such as RDCs, scalar couplings,[55] chemical shifts (CSs), and paramagnetic relaxation effects (PREs)[56−58] that can be interpreted structurally and used for the determination of dynamic ensembles. Similar to the case for structured proteins, it is possible to exploit such structural and dynamical information by using different approaches that differ mostly in whether they are hypothesis-driven, i.e. use the experimental data to validate ensembles produced *a priori*[28,29,59] or, on the contrary, use the data to directly bias the conformational search.[57,60]

For the determination of an ensemble that accurately describes chemically denatured ubiquitin, here we use a combination of these two approaches by employing RDCs measured in steric alignment to refine a SCM generated *a priori*. SCMs are large ensembles of protein conformations obtained in the absence of experimental data that aim to reproduce the structure and dynamics of proteins devoid of long-range interactions and are instead defined strictly by local conformational preferences. Such preferences can be derived from first principles or, as is most often the case, determined from a statistical analysis of the corresponding 2D (ϕ,φ) histograms of regions of sequence not involved in tertiary contacts, such as loops and termini, in protein structures deposited in the PDB.[28,29,61,62] It has been shown that SCMs built to match such statistics have back-calculated scalar couplings,[61,62] RDCs, and SAXS profiles[28,29] that agree with their experimental counterparts, strongly suggesting that they represent well the main structural features of unstructured proteins. As SCMs built in such a way are the best models of such systems currently available, they are optimal initial configurations for refinement using steric RDCs. Our use of ERIDU for the refinement of SCMs differentiates our approach from that very recently reported by Huang and Grzesiek,[38] in which the authors used as initial configurations for structure determination randomized structures generated by unrestrained MD simulations started from extended structures; in this sense ERIDU is conceptually related to the ENSEMBLE[44] and ASTEROIDS[63] approaches, where a Monte Carlo approach and a genetic algorithm are used respectively to select ensembles that agree with experiment from large databases of configurations generated *a priori*.

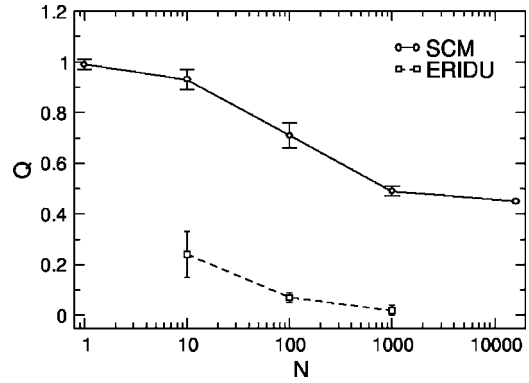

Figure 1. Dependence of Q (eq 3) of the back-calculated RDCs for the SCM (solid line) and the ERIDU ensemble (dashed line) on the number of conformations considered (N). For each case the reported Q is the average of five independent simulations and the error bars represent the standard deviations; lower values of Q can be obtained by pooling the five ensembles.

In order to determine the optimal number of conformations needed to define a good quality starting configuration, we compared the agreement with experiment of ensembles of increasing size ($N = 1$, 10, 10^2, 10^3), taken from the SCM for chemically denatured ubiquitin determined by Jha et al.,[28] with that of the complete ensemble (Figure 1). It is clear that, contrary to the case for ensembles of small size ($N = 10$), ensembles composed of 10^3 and, to a lesser extent, 10^2 conformations give a degree of agreement that approaches that of the complete SCM, indicating that these are valid sizes for the starting configuration.

Ensembles of various sizes were then refined against seven sets of RDCs measured in steric alignment using ERIDU, in implicit solvent[64] and at constant temperature ($T = 298$ K), in a cluster of 2744 PowerPC processors running at 2.2 GHz, where ensemble simulations took ca. 1 h per CPU. The initial values of the force constants used to restrain the different sets of RDCs (eq 1) were chosen empirically so that agreement with experiment was reached at a similar rate for all sets and were multiplied by 1.25 every 10^3 MD steps, which were of 1 fs. The alignment tensors of the N conformers were computed numerically, from their coordinates, at each time step of the simulation by using the algorithm of Almond and Axelsen[31] presented in the Theory section.

The results of refinement presented in Table 1 and Figures 1 and 2 illustrate that ERIDU refinement using steric RDCs yields

(47) Chang, S. L.; Tjandra, N. *J. Magn. Reson.* **2005**, *174*, 43–53.
(48) Meiler, J.; Prompers, J. J.; Peti, W.; Griesinger, C.; Bruschweiler, R. *J. Am. Chem. Soc.* **2001**, *123*, 6098–6107.
(49) Lindorff-Larsen, K.; Best, R. B.; Depristo, M. A.; Dobson, C. M.; Vendruscolo, M. *Nature* **2005**, *433*, 128–132.
(50) Richter, B.; Gsponer, J.; Várnai, P.; Salvatella, X.; Vendruscolo, M. *J. Biomol. NMR* **2007**, *37*, 117–135.
(51) Wirmer, J.; Peti, W.; Schwalbe, H. *J. Biomol. NMR* 2006.
(52) Meier, S.; Grzesiek, S.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 9799–9807.
(53) Shortle, D. R. *Curr. Opin. Struct. Biol.* **1996**, *6*, 24–30.
(54) Dyson, H. J.; Wright, P. E. *Chem. Rev.* **2004**, *104*, 3607–3622.
(55) Peti, W.; Hennig, M.; Smith, L.; Schwalbe, H. *J. Am. Chem. Soc.* **2000**, *122*, 12017–12018.
(56) Gillespie, J. R.; Shortle, D. *J. Mol. Biol.* **1997**, *268*, 158–169.
(57) Dedmon, M. M.; Lindorff-Larsen, K.; Christodoulou, J.; Vendruscolo, M.; Dobson, C. M. *J. Am. Chem. Soc.* **2005**, *127*, 476–477.
(58) Bertoncini, C. W.; Jung, Y. S.; Fernandez, C. O.; Hoyer, W.; Griesinger, C.; Jovin, T. M.; Zweckstetter, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 1430–1435.
(59) Mukrasch, M. D.; Markwick, P. R.; Biernat, J.; Bergen, M.; Bernadó, P.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 5235–5243.
(60) Lindorff-Larsen, K.; Kristjansdottir, S.; Teilum, K.; Fieber, W.; Dobson, C. M.; Poulsen, F. M.; Vendruscolo, M. *J. Am. Chem. Soc.* **2004**, *126*, 3291–3299.
(61) Serrano, L. *J. Mol. Biol.* **1995**, *254*, 322–333.
(62) Smith, L. J.; Bolin, K. A.; Schwalbe, H.; MacArthur, M. W.; Thornton, J. M.; Dobson, C. M. *J. Mol. Biol.* **1996**, *255*, 494–506.
(63) Nodet, G.; Salmon, L.; Ozenne, V.; Meier, S.; Jensen, M. R.; Blackledge, M. *J. Am. Chem. Soc.* **2009**, *131*, 17908–17918.
(64) Im, W.; Lee, M. S.; Brooks, C. L. *J. Comput. Chem.* **2003**, *24*, 1691–1702.

Table 1. Comparison of the Violations of the Different Ensembles Expressed as Root-Mean-Squared Deviation in Hertz, with Their Experimental Uncertainty[65]

	unrefined		refined		
coupling	SCM[a]	SCM[b]	MD	ERIDU[c]	error[d]
NH	2.8	2.0	3.6	0.14	0.3
CαHα	4.9	2.9	6.6	0.25	0.6
CαC′	1.0	0.9	1.0	0.04	0.1
HαHN	1.1	1.3	1.8	0.15	0.14
$H\alpha^{(i-1)}HN$	3.4	0.6	4.2	0.13	0.14
$HN^{i}HN^{(i+1)}$	0.9	0.16	2.8	0.14	0.14
$HN^{i}HN^{(i+2)}$	0.2	0.9	0.4	0.07	0.14

[a] Calculated for the ensemble reported in Jha et al.[28] [b] Taken from Meier et al.[52] [c] Obtained by using 10^3 conformers. [d] Taken from Meier et al.[65]

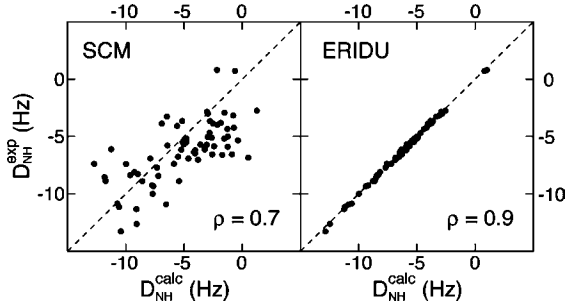

Figure 2. Plot of the correlation between experimental[52] and calculated NH RDCs in the SCM[28] and after refinement with ERIDU using 10^3 conformations; correlation plots of the remaining RDC sets are presented as Supporting Information (Figure S1).

ensembles that are in complete agreement with experiment, as they have RDC violations that are at the level of the experimental error in the measurement of D^{exp}. Very importantly, as shown in Table 1, they also highlight the role played by the RDC restraints, since an unrestrained MD simulation using the SCM as starting configuration worsens agreement with experiment; for example, the violation of the NH RDCs, which is 2.8 Hz for the SCM and is reduced to 0.14 Hz by ERIDU, is increased to 3.6 Hz when the RDCs are not used to restrain the MD simulation.

As previously mentioned, SCMs are also good predictors of backbone scalar couplings; since this NMR parameter was not used to determine the ERIDU ensemble, we analyzed its evolution during refinement to ascertain that the improvement in agreement with the steric RDCs was not taking place at the expense of agreement with such couplings. The results, presented in Figure 3, indicate that ERIDU does not affect the agreement between the experimental[55] and the back-calculated values ($\rho \approx 0.7$) when ensembles of size 10^2 or 10^3 are used but that it can significantly worsen when samples of size 10 are taken from the SCM; this finding highlights that it is possible to generate ensembles of small size that show agreement with experimental RDCs ($Q \approx 0.2$, as shown in Figure 1,) but, as very recently discussed by Nodet et al.,[63] this does not guarantee that the resulting ensemble is accurate, as show here by validation using scalar couplings.

Given the marked improvement in agreement with experiment produced by refinement with ERIDU, it is important to determine the type of changes produced on the starting configuration. To analyze, in particular, to what extent the

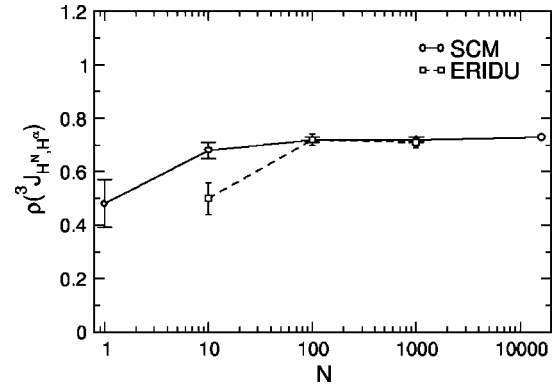

Figure 3. Dependence of the Pearson correlation coefficient between the experimental $^3J_{HNHa}$ backbone scalar couplings and those back-calculated using the Karplus equation with the coefficients suggested by Pardi et al.[66] for the SCM (solid line) and the ERIDU ensemble (dashed line). The error bars represent the standard deviation of five independent runs.

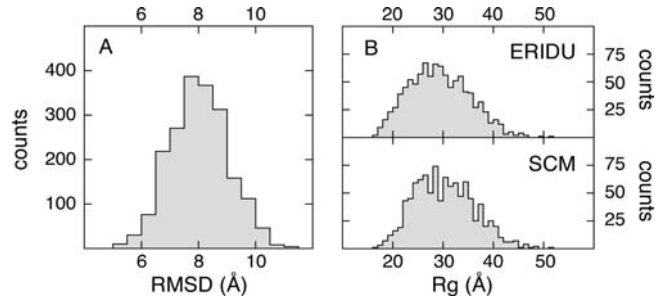

Figure 4. (A) Histogram of the structural corrections (backbone rmsd) introduced by ERIDU in the refinement of the SCM ($N = 10^3$) of chemically denatured ubiquitin reported by Jha et al.[28] (B) Histogram of the radius of gyration (R_g) of the SCM[28] before and after refinement, consistent with recently reported results.[67]

ensemble members modified their structure, we computed the histogram of the backbone rmsd between the starting conformation, belonging to the SCM, and the final conformation, belonging to the refined ensemble. The results, presented in Figure 4A, show that essentially all conformations underwent substantial structural changes, of on average 8 Å. To exclude the possibility that refinement had forced all trajectories to converge to one conformation with low RDC violations or collapse to a structurally heterogeneous compact state, we also analyzed the histogram of the radius of gyration (R_g) for the unrefined and refined ensembles (Figure 4B). We observed that no significant changes to the histogram occurred and that the final ensemble had an R_g distribution in good agreement with the average R_g of 28.0 ± 3.5 Å recently determined by Gabel et al. from SAXS and SANS data.[67] To show the range of conformational changes caused by ERIDU, we present, in Figure 5, three ensemble members that illustrate how refinement did not cause important topological changes but, on the contrary, operated by inducing local changes of structure.

Having established that refinement modified the structure of the vast majority of conformations while at the same time preserving the overall features and structural heterogeneity of the SCM, we analyzed the specific changes with the help of contact maps that reveal the fraction of ensemble members

(65) Meier, S.; Häussinger, D.; Jensen, P.; Rogowski, M.; Grzesiek, S. *J. Am. Chem. Soc.* **2003**, *125*, 44–45.

(66) Pardi, A.; Billeter, M.; Wüthrich, K. *J. Mol. Biol.* **1984**, *180*, 741–751.

(67) Gabel, F.; Jensen, M. R.; Zaccaï, G.; Blackledge, M. *J. Am. Chem. Soc.* **2009**, *131*, 8769–8771.

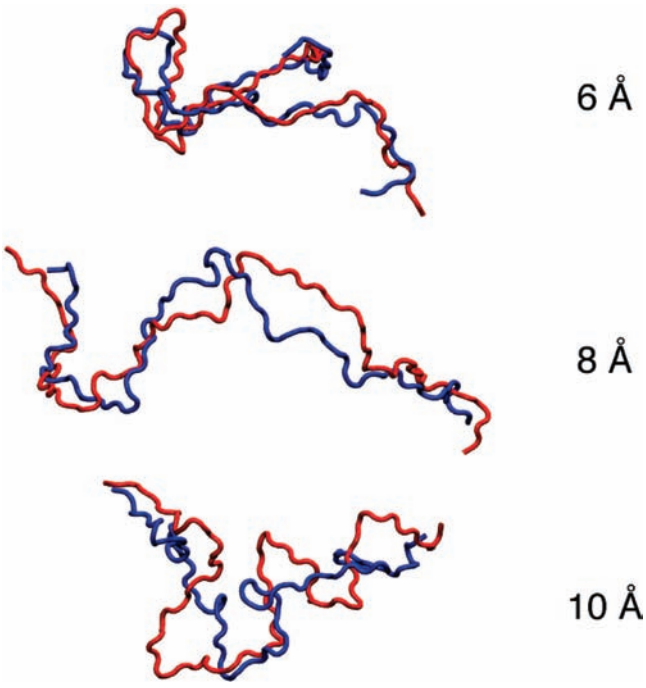

Figure 5. Representative examples of the range of structural corrections induced by refinement, where the SCM ensemble members are shown in red and the refined ensemble members are shown in blue.

where pairs of residues are at a distance shorter than 10 Å. The results that we obtained, shown in Figure 6A, indicated, as expected, that the SCM is devoid of mid- or long-range inter-residue interactions and only presents a tendency to transiently form local (i, i+2 or i, i+3) contacts in the vicinity of residues 9, 23, 34, 47, and 52. It is very interesting to note that all such contacts, except those around position 23, involve Gly residues (G10, G35, G47, G53) and are native (Figure 6B): 9 is part of the turn between strands β1 and β2; 23 and 34 correspond to helical turns at the beginning and end of the α-helix; 47 is part of the turn between strands β3 and β4; 52 is structured as a loop in folded ubiquitin. An analysis of the contact map of the ERIDU ensemble indicated that refinement increased significantly the stability and range of such contacts, while a comparison with the equivalent map of the SCM after MD refinement without restraints did not show such an increase in range but instead suggested local non-specific collapse. This reveals that the changes induced by ERIDU are a direct consequence of the structural information contained in the RDCs and shows, therefore, that when these are used to refine ensembles, both local and non-local interactions can indeed be introduced.[68] We also analyzed whether refinement caused changes in secondary structure by monitoring the Ramachandran plot of the ensembles (Figure S2, Supporting Information) and found that ERIDU does not modify the fraction of dihedrals in the region defined by $\phi < 0°$ and $50° < \varphi < 180°$, which corresponds to extended conformations (71%), contrary to the case when the SCM is refined using MD (63%).

In order to confirm that the non-local interactions displayed in the contact map of the ERIDU ensemble are indeed present, it is necessary to determine whether they are compatible with independent observations based on the measurement of NMR parameters different from RDCs such as trans-hydrogen-bond $^3J_{NC'}$ scalar couplings and PREs. Whereas the former are particularly sensitive to the transient formation of hydrogen bonds in native[69] and partially folded proteins,[37,70] the latter are optimal for the identification of mid- and long-range (up to 20 Å) inter-residue interactions.[56−58,71] An analysis of the $^3J_{NC'}$ and PRE data available for chemically denatured ubiquitin reveals that these are in good qualitative agreement with the results that we have obtained. The $^3J_{NC'}$ data[37] identified a low but detectable population of the native hairpin formed by strands β1 and β2 (labeled β1,2 in Figure 6D) that is also present in the ERIDU ensemble at ca. 10%; the PRE data very recently reported by Huang and Grzesiek[38] indicate mid-range interactions that are without exception present in the ERIDU ensemble. In addition to the contacts between residues corresponding to the first two strands of the native protein, PREs also report on interactions between position 20 and residues that are C-terminal to it, which correspond to the second region of nascent structure of the contact map (Figure 6D), and between residues at positions 33 and 35 and their vicinity, which approximately correspond to the third region of structure. Most importantly, they also reveal that the stretch of sequence between positions 40 and 70 is locally collapsed, again in agreement with the ERIDU map, where a very similar region, that between positions 45 and 65, shows a marked degree of short- and mid-range structure. Although widely used as a model for unstructured proteins, chemically denatured ubiquitin lacks long-range interactions[51] such as those observed in some intrinsically disordered proteins;[57,58,68] this system is therefore not optimal for the assessment of the ability of RDCs to report on long-range interactions in unstructured proteins.

That the ERIDU ensemble presents structural β features compatible with independent studies of the same system is an encouraging result but one that needs additional validation by assessing the robustness of the method to the removal of a fraction of the restraints. To this aim we carried out three different series of simulations: a first series of 21 runs where 5% of the RDCs, randomly selected from the list of 406 restraints, were unrestrained; a second series of 7 runs, carried out in triplicate, where each of the sets of RDCs was unrestrained; and a third, very stringent series of 7 runs, again carried out in triplicate, where only one of the sets was restrained whereas the other six where not. Similarly to the case for NOEs in structure determination, where most of the structural information is contained in a small fraction of the restraints,[73] we find that the agreement with experiment of the unrestrained RDCs, expressed as Q_{free}, also critically depends on the identity of the unrestrained set. Therefore, in order to obtain a global view of the robustness of ERIDU, we present in Table 2 the average values of Q_{free} of the SCM and ERIDU ensembles and the average improvement in Q_{free} upon refinement, as well as the worst and the best cross-validations obtained. The results show that, in all test cases, ERIDU decreases, on average, the RDC violation of the unrestrained sets, thus providing additional evidence that the approach is successful at capturing the essential structural features of unstructured proteins.

(68) Bernadó, P.; Bertoncini, C. W.; Griesinger, C.; Zweckstetter, M.; Blackledge, M. *J. Am. Chem. Soc.* **2005**, *127*, 17968–17969.
(69) Grzesiek, S.; Cordier, F.; Dingley, A. J. *Methods Enzymol.* **2001**, *338*, 111–133.
(70) Cordier, F.; Grzesiek, S. *Biochemistry* **2004**, *43*, 11295–11301.
(71) Lietzow, M. A.; Jamin, M.; Jane Dyson, H. J.; Wright, P. E. *J. Mol. Biol.* **2002**, *322*, 655–662.
(72) Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J. *J. Mol. Biol.* **1987**, *194*, 531–544.
(73) Nabuurs, S. B.; Spronk, C. A.; Krieger, E.; Maassen, H.; Vriend, G.; Vuister, G. W. *J. Am. Chem. Soc.* **2003**, *125*, 12026–12034.

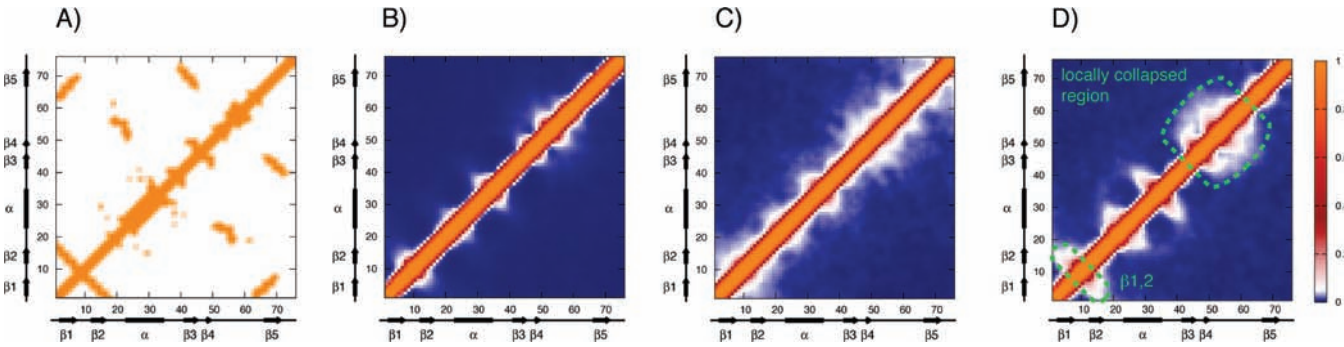

Figure 6. Contact maps for SCM and refined RDC ensembles using 10^3 conformers: (A) native ubiquitin (pdb 1UBQ[72]), (B) SCM reported by Jha et al.,[28] (C) an ensemble obtained by unrestrained MD using the SCM as starting configuration, and (D) SCM refined by MD restrained with steric RDCs using the ERIDU algorithm, where β1,2 identifies contacts between native strands 1 and 2. The maps are color-coded according to the fractional number of contacts between residues (Cα−Cα distance lower than 10 Å) among the ensemble; contact maps obtained with 10^2 conformers are shown in Figure S3 of the Supporting Information.

Table 2. Cross-Validation of the ERIDU Ensemble[a]

RDCs used	$\overline{Q_{free}^{SCM}}$	$\overline{Q_{free}^{ERIDU}}$	ΔQ_{free}^{max}	$\overline{\Delta Q_{free}}$	ΔQ_{free}^{min}
95%[b]	0.64	0.53	−0.48	−0.11	+0.31
6 sets[c,e]	0.73	0.62	−0.32	−0.11	+0.12
1 set[d,e]	0.68	0.63	−0.14	−0.05	+0.02

[a] Simulations were run using 10^2 conformers, and the ability of refinement to improve agreement with experiment of the free, i.e. unrestrained, RDCs (ΔQ_{free}) was assessed using three different approaches: (i) by not restraining 5% of the RDCs, chosen randomly; (ii) by not restraining a complete set of RDCs corresponding to a given bond vector type; and (iii) by restraining only one set. The result of the cross-validation depends on the actual set of randomly selected restraints; in order to illustrate the range of results obtained, we present the average values of Q_{free} before and after refinement as well as the best (ΔQ_{free}^{max}), worst (ΔQ_{free}^{min}) and average changes in Q_{free}. [b] Results obtained after 21 independent runs. [c] Results obtained after 3 independent simulations in which each of the 7 sets was unrestrained. [d] Results obtained after 3 independent runs in which each of the 7 sets was the only one restrained. [e] Details of the results obtained for each set are provided as Supporting Information (Table S1).

4. Conclusions

We have introduced a new method for the structural and dynamical characterization of flexible molecules that uses RDCs measured in steric alignment as restraints in ensemble molecular dynamics simulations and applied it to the refinement of ensembles of conformations that describe the structure and dynamics of chemically denatured ubiquitin. Key features of the ERIDU approach are that, similarly to related methods,[38,44,63] it directly computes the steric alignment tensor of the different ensemble members while, by using SCMs[28,29] as starting configurations, it maximizes its coverage of the vast conformational space available to unstructured proteins. The ubiquitin ensemble thus obtained is in good agreement with experiment and can be cross-validated by predicting RDCs not used as restraints; in addition, it recaptures features also identified in studies of other NMR parameters, such as trans-hydrogen-bond $^3J_{NC'}$ scalar couplings[37] and PREs,[38] that reveal that chemically denatured ubiquitin presents mid-range interactions that are, in certain cases, present in the native structure. That the ERIDU ensemble so comprehensively reproduces the structural and dynamical features of chemically denatured ubiquitin certifies that RDCs are an extremely powerful tool for the characterization of unstructured and intrinsically disordered proteins at high resolution. In addition, that RDCs induced by steric alignment can now be used to address flexible systems suggests that this NMR parameter will be key to understanding how large-scale fluctuations such as hinge motions in interdomain proteins relate to their biological function.

Acknowledgment. This work was supported by grants from IRB (S.E.-M., R.B.F., X.S.) and ICREA (X.S.). The authors acknowledge the use of computational resources of the Red Española de Supercomputación (RES).

Supporting Information Available: Complete ref 45, correlation plots for all restrained RDC sets, 2D (ϕ,φ) plots, table with details of the cross-validation, and additional contacts maps. This material is available free of charge via the Internet at http://pubs.acs.org.

JA906995X