Dipolar Couplings of Proteins

DOI: 10.1002/anie.201206585

Direct Prediction of NMR Residual Dipolar Couplings from the Primary Sequence of Unfolded Proteins**

*Jie-rong Huang, Valéry Ozenne, Malene Ringkjøbing Jensen, and Martin Blackledge**

Over the last decade the accepted paradigm underpinning classical structural biology has been reassessed, with the general realization that a significant fraction of proteins encoded in eukaryotic genomes do not adopt a stable fold in their functional form, but instead are intrinsically disordered either in long contiguous regions, or in many cases throughout their entire length.[1–4] The high degree of flexibility inherent to intrinsically disordered proteins (IDPs) bestows distinct properties allowing them to function differently to folded proteins, but this same attribute complicates characterization of their molecular behavior. To better understand the relationship between primary sequence and biological function in IDPs it is essential that calibrated techniques become available allowing a quantitative determination of their conformational behavior.

Nuclear magnetic resonance (NMR) spectroscopy is one of the most powerful methods for studying disordered proteins, providing atomic resolution, ensemble-averaged information reporting on the conformational energy landscape sampled by each amino acid.[5–9] Among a large variety of NMR methods, residual dipolar couplings (RDCs), reporting on the orientational properties of the internuclear dipole–dipole interaction averaged up to the millisecond time scale, have proven to offer highly sensitive probes of local and long-range conformational sampling.[10–14] To model the vast conformational ensemble that gives rise to the measured NMR signal, statistical coil models have been successfully used to construct protein ensembles by stochastically sampling amino-acid-specific backbone dihedral angle $\{\phi,\psi\}$ energy surfaces.[15–19] This ensemble representation is used to calculate conformationally averaged RDCs, resulting in remarkable reproduction of their distribution along the primary sequences of IDPs. Deviations from coil behavior can then be interpreted in terms of intrinsic propensity to populate the local structure, often in interaction sites of these proteins, or to adopt transient long-range structure.[20–22] Recently RDCs, alone and in combination with chemical shifts, have been used to directly map the conformational energy surface using ensemble approaches.[23–28]

In all of these applications, RDCs are calculated from each member of the ensemble using a steric exclusion model [Eq. (1)].

$$D_{ij}^{n} = K_{ij}\left(A_{a}^{n}\left(3\cos^{2}\Theta^{n}-1\right)+\frac{3}{2}A_{r}^{n}\sin^{2}\Theta^{n}\cos 2\Phi^{n}\right) \quad (1)$$

A_a^n and A_r^n are the axial and rhombic components of the alignment tensor of conformer n. Θ^n and Φ^n are the polar coordinates of the internuclear vector $\{i,j\}$ with respect to the alignment tensor frame, and K_{ij} (Supporting Information) depends on the nuclei and the internuclear distance. The average is then taken over the entire ensemble of N conformers.

Despite the success of this approach, averaging of RDCs over an ensemble of protein conformations presents significant practical difficulties that affect both the accuracy and the portability of the calculation. Most importantly, convergence of the average requires an unmanageably large number of structures (100000 for a protein with 100 amino acids and even larger for longer constructs).[18,24] These convergence characteristics depend on the length of the protein, so that division of the protein into short, uncoupled segments leads to better convergence.[29] It has indeed been demonstrated that a segment length of 15 amino acids is a reasonable length for the so-called local alignment window (LAW) to achieve good accuracy and efficiency for RDC prediction.[24] Nevertheless this approach necessarily sacrifices any long-range information that may be present, so that predicted RDCs from the separate segments need to be corrected with a parametrized baseline to account for any tertiary contacts present in the ensemble.[25]

In this report we demonstrate that RDCs for a given residue are essentially determined by the identity of the amino acid in question and its two neighbors, and sequence-dependent corrections defining local alignment and the polymeric nature of the protein. Combining these effects, we propose a simplified, automatic, and highly accurate method for directly predicting RDCs from the primary sequence of unfolded proteins.

To assess the influence of neighboring residues on the RDCs of the central amino acid we used the flexible-meccano[18,19] algorithm to simulate RDCs from pentadecapeptides with different levels of primary sequence identity, and compared these to RDCs predicted from the full-length protein (in this case 76 amino acids in length). The following cases were simulated: A) the content of the pentadecapeptide is identical to the native sequence (from $i-7$ to $i+7$); B) only the middle residue (i) is the same as the native sequence and the other 14 residues are represented by a common amino

[*] Dr. J. R. Huang, V. Ozenne, Dr. M. R. Jensen, Dr. M. Blackledge
Protein Dynamics and Flexibility
Institut de Biologie Structurale Jean-Pierre Ebel
CNRS-CEA-UJF UMR 5075, 41 rue Jules Horowitz
38027 Grenoble Cedex (France)
E-mail: martin.blackledge@ibs.fr

[**] Financial support from the CEA, CNRS, UJF, and the ANR under TAUSTRUCT MALZ 2010 (MB) Protein Disorder JCJC-2010 (MRJ).

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/anie.201206585.

acid, in this case valine; C) the middle and adjacent residues (from $i-1$ to $i+1$) are defined by the native sequence while the remaining 12 amino acids are represented by valines (8000 combinations); D) the middle five ($i-2$ to $i+2$) are defined by the native sequence while the remaining 10 are represented by valines. In each case 10000-strong ensembles were created, and RDCs of the central residues were predicted based on steric exclusion. RDCs from the full-length unfolded sequence were predicted from 50000 conformers (Figure 1).

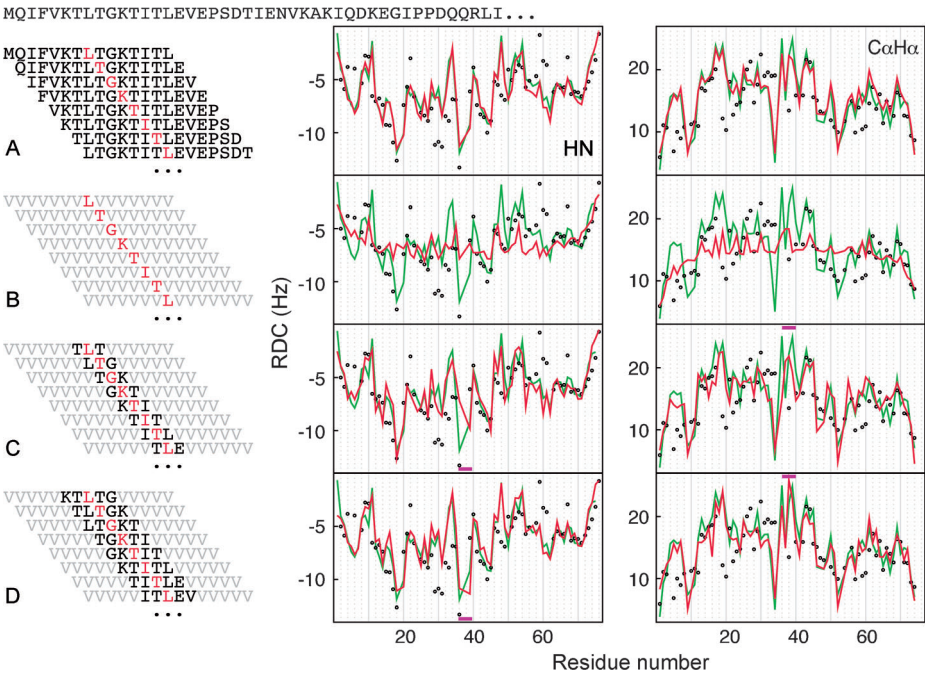

Figure 1. Comparison of RDCs predicted from different levels of identity to primary sequence. Experimental data (D_{HN} and $D_{C\alpha H\alpha}$) from urea-denatured ubiquitin are shown in black. RDCs predicted using global alignment tensor are shown in green. Values using LAWs with 15 amino acids are shown in red. Four levels of identity with respect to the primary sequence were used: A) all amino acids, B) only the central residue, C) the middle three residues, and D) the middle five residues. RDCs of the central residue (red) are shown on the left; bold letters emphasize those residues identical to native sequence, others are shown in gray. In (C) and (D) the red horizontal bars indicate the position of successive proline residues.

In case (A), with the correct sequence, the RDCs reproduce the results from the global tensor calculation very closely, as reported previously for a pentadecapeptide (Figure 1 A).[24] In contrast, with only the central residue retaining its identity, the predicted RDCs bear poor similarity to the results from the global tensor (Figure 1 B). Importantly however, including the identity of the nearest neighbors, already reproduces RDCs predicted from the global tensor very well. In this example of denatured ubiquitin,[30] the most significant discrepancy occurs around residues 35–37 (Figure 1 C). This is due to the presence of successive prolines, where the conformational sampling is locally more restricted than for other amino acids. The comparison improves even further when a quintet of amino acids is considered (Figure 1 D). These simulations indicate that for the general case of an unfolded chain, with a local flexibility described by the overwhelming majority of available combinations, RDCs can be accurately predicted by considering only the identity of neighbors. The effect of the second nearest residues is apparently weak, unless local structure is present.

We note that if only the sequence information of either preceding or following neighbors is considered, the predicted values are poorly reproduced (see Figure S1), and that simulations carried out in the absence of side-chain interaction show that explicit steric clashes with neighbors do not contribute to these observations (Figure S2), probably because statistical coil Ramachandran sampling already encodes aspects of side-chain bulkiness. Zweckstetter and co-workers demonstrated that plotting a smoothed distribution of amino acid bulkiness along the sequence also resembles experimental RDCs.[31]

It appears that the main effect of nearest neighbors on the measured RDC originates from a more subtle phenomenon as highlighted in Figure 2. An ensemble consisting of 10^6 conformers of an alanine pentadecapeptide was created using the flexible-meccano algorithm. RDCs were predicted for each conformer, and the averaged RDCs of each amino acid were plotted against the $\{\phi,\psi\}$ sampling of the central amino acid in a 2° × 2° grid over all Ramachandran space. On the right-hand side of the figure the entire sequence samples Ramachandran space uniformly. Perhaps not unexpectedly, the HN RDC depends almost uniquely on the ψ value of the preceding amino acid, and the ϕ of the amino acid of interest (these angles precede and follow the NH bond). Similar observations can be made for other RDCs in the peptide plane, while the CαHα RDC exhibits a weaker dependence. On the left-hand side of the figure the sampling of the central amino acid is still uniform, but the remainder of the sequence samples the conformational potential intrinsic to an alanine amino acid. Constraining the flanking residues to this more physical model significantly modulates the predicted RDCs, showing that the dependence on the $\{\phi,\psi\}$ sampling of the amino acid of interest extends further along the chain. Again, removal of explicit steric interactions does not change this result significantly. The angular sampling available to the peptide of interest clearly depends on the backbone dihedral distribution of the neighbors. These complex dependences, on the sampling of the residue of interest and the neighbor residues, help to rationalize the observation that no individual RDC can provide a "read-out" of Ramachandran sampling, but that

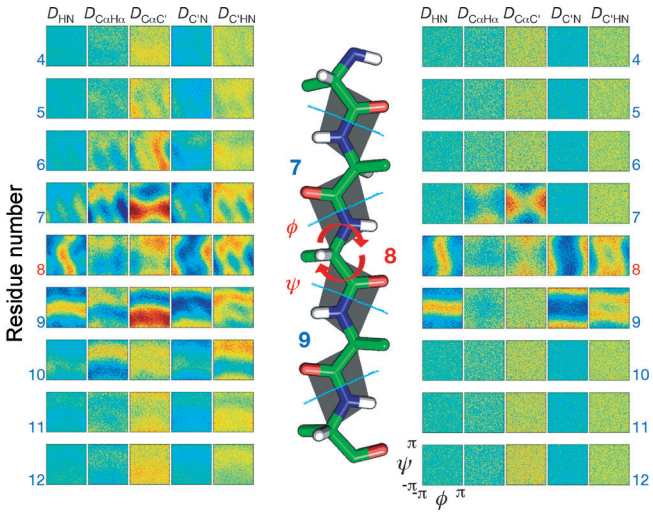

Figure 2. Propagation of the influence of conformational sampling along the peptide chain. Five types of RDCs from a million flexible-meccano structures (pentadeca-alanine) are plotted against the $\{\phi,\psi\}$ distribution of the central residue averaged over $2°\times2°$ grid. Right panel: all residues sample $\{\phi,\psi\}$ space uniformly. Left panel: the central residues sample uniform conformational space, the others sample the $\{\phi,\psi\}$ angles according to the coil library distribution of alanine. Colors from red to blue correspond to higher to lower RDC values over a range chosen to best represent the RDCs of the central residue.

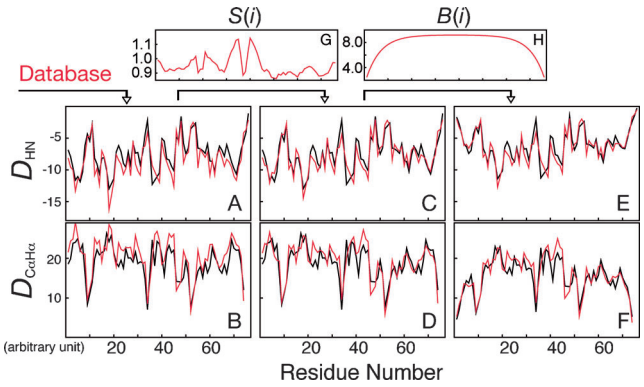

Figure 3. Flowchart of the *seq2rdc* approach. The sequence of ubiquitin and D_{HN} (A, C, E) and $D_{C\alpha H\alpha}$ (B, D, F) are simulated as examples. Black lines represent the simulation from the LAW, and red lines from *seq2rdc*. A and B) database only; C and D) corrected with the scaling factor $S(i)$ depending on the composition of residues $i-7$ to $i-2$ and $i+2$ to $i+7$ (G); E and F) corrected with a length-dependent hyperbolic baseline $B(i)$ (H).

the inclusion of the conformational sampling of the neighbors does allow this.

The characterization of nearest-neighbor effects on conformationally averaged RDCs points directly to the development of a tractable database to predict RDCs from unfolded proteins. Accordingly we have constructed 8000 different combinations of pentadecapeptides in which the middle three residues (numbers 7 to 9) sample all 20 types of amino acid. The remaining 12 sites were modeled as valines. 10000 flexible-meccano structures were generated for each of the 8000 sequences and different types of RDCs of the central residue were deposited in a look-up table (Table S3). By using an efficient internal-coordinate-based algorithm of flexible-meccano[32] and in-house steric alignment prediction,[33] the construction of this look-up table takes 26 h on a single CPU (Intel 2.8 GHz), suggesting that expansion to fourth or fifth amino acids to define a database of quadruplets or quintets is quite feasible. As shown in Figure 3 A and B, RDCs extracted from the database according to this tripeptide information are very similar to the values calculated using the 15 amino acid LAW approach.

Since the conformational sampling for each amino acid is different, we find that there is a slight dependence of the amplitude of the local alignment tensor on the composition of the 12 flanking amino acids comprising the LAW. Thus a polyglycine will be more flexible and therefore exhibit lower average alignment tensor eigenvalues than a polyproline. To account for this variation, 8000 simulations were repeated with N- and C-flanking regions replaced with all 20 types of hexapeptide and scaling factors derived for different amino acids (Table S2). The factor $S(i)$ [Eq. (S8)], is composed of the relative contribution of the 12 flanking amino acids to the local alignment tensor, and is applied to correct for variation as a function of sequence. Applying the amino-acid-dependent scaling factor (Figure 3 C and D), results in better reproduction of predicted RDCs (root mean square, rms, falls by $>60\,\%$) from explicit amino-acid sampling compared to those without correction (Figure 3 A and B). Finally the profile is modified using the baseline $B(i)$ that accounts for the chainlike nature of the disordered protein, as described previously[24] (Figure 3 E and F).

Combining these approaches, we present a simple algorithm (*seq2rdc*), to directly predict RDCs from the primary sequence of IDPs, by extracting the value from the prebuilt database according to the composition of the tripeptide, and modulating this due to the local alignment and baseline factors ($S(i)$ and $B(i)$, respectively). The algorithm, which is around six orders of magnitude faster than existing approaches, provides the user with values from the tripeptide database, $S(i)$ and $B(i)$ (Figures S5 and S6), as well as the final predicted RDCs. Figure 4 compares different types of RDCs predicted from *seq2rdc* with global alignment tensor prediction and published experimental RDCs from denatured apomyoglobin,[11] denatured Δ131Δ construct of staphylococcal nuclease,[10] denatured GB1,[32,34] denatured ubiquitin, α-synuclein,[20] and K18 construct of Tau protein.[21] The prediction accuracy from the two different approaches is very similar. *seq2rdc* is available upon request from the authors.

The *seq2rdc* approach therefore provides an immediate probe of protein unfoldedness once RDCs have been measured, allowing for direct comparison to experimental data. For example in the case of Tau protein, positive D_{NH} RDCs in the experimental data deviate from prediction, indicating nonrandom behavior of regions that have been shown to populate β-turns.[21] Any distortion of the underlying baseline, due to long-range contacts or fluctuating tertiary structure, can be readily combined with the local conformational sampling read from the triplet analysis, in analogy to previously developed explicit ensemble approaches.[25]

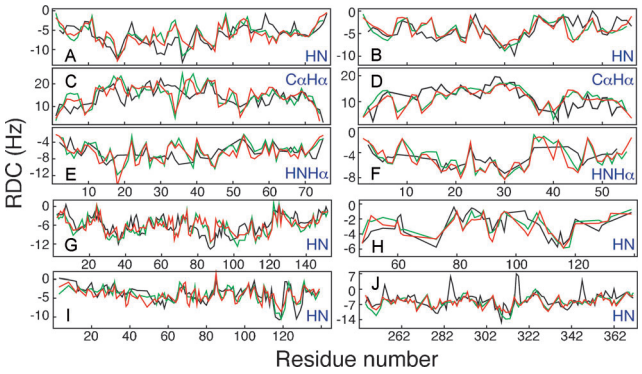

Figure 4. Examples of *seq2rdc* prediction (red) compared to experimentally measured RDCs (black). A,C,E) Denatured ubiquitin, B,D,F) denatured protein GB1, G) denatured apo-myoglobin, H) denatured Δ131Δ staphylococcal nuclease, I) α-synuclein, and J) K18 construct of Tau. Values are compared to prediction from global alignment tensor (green). RDC type is indicated in each panel (HNHα RDC refers to H^{N}_{i}-H^{α}_{i-1}).

In summary, analysis of local and long-range effects, corresponding to the conformational sampling of the region of interest and the chain-like nature of the unfolded protein, respectively, reveals that theoretical RDCs can be determined by consideration of these factors alone. Using insights gained from these studies, we find that RDC prediction can in general be deconvoluted to four components: the sampling of the amino acid of interest, nearest-neighbor-dependent effects, sequence-dependent scaling factors to correct the local alignment tensor, and a length-dependent baseline to incorporate the polymeric nature of the unfolded protein. We demonstrate that a database of combinations of triplets of amino acids, combined with corrections for the presence of the triplet in a chain of known composition, defines to a very good approximation expected random coil values of RDCs in unfolded states. This obviates the need to calculate explicit and extensive ensembles of atomic resolution structures, resulting in a significant improvement in the efficiency of calculating RDCs from unfolded sequences.

Received: August 15, 2012
Revised: October 23, 2012
Published online: November 27, 2012

Keywords: conformational sampling · intrinsic disorder · NMR spectroscopy · proteins · residual dipolar couplings

[1] P. Tompa, *Trends Biochem. Sci.* **2002**, *27*, 527 – 533.
[2] H. J. Dyson, P. E. Wright, *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 197 – 208.
[3] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, Z. Obradović, *Biochemistry* **2002**, *41*, 6573 – 6582.
[4] V. N. Uversky, *Protein Sci.* **2002**, *11*, 739 – 756.
[5] H. J. Dyson, P. E. Wright, *Curr. Opin. Struct. Biol.* **2002**, *12*, 54 – 60.
[6] D. Eliezer, *Curr. Opin. Struct. Biol.* **2009**, *19*, 23 – 30.
[7] S. Meier, M. Blackledge, S. Grzesiek, *J. Chem. Phys.* **2008**, *128*, 052204.
[8] R. Schneider, J. Huang, M. Yao, G. Communie, V. Ozenne, L. Mollica, L. Salmon, M. R. Jensen, M. Blackledge, *Mol. Biosyst.* **2012**, *8*, 58 – 68.
[9] J. A. Marsh, C. Neale, F. E. Jack, W.-Y. Choy, A. Y. Lee, K. A. Crowhurst, J. D. Forman-Kay, *J. Mol. Biol.* **2007**, *367*, 1494 – 1510.
[10] D. Shortle, M. S. Ackerman, *Science* **2001**, *293*, 487 – 489.
[11] R. Mohana-Borges, N. K. Goto, G. J. A. Kroon, H. J. Dyson, P. E. Wright, *J. Mol. Biol.* **2004**, *340*, 1131 – 1142.
[12] W. Fieber, S. Kristjansdottir, F. M. Poulsen, *J. Mol. Biol.* **2004**, *339*, 1191 – 1199.
[13] M. Louhivuori, K. Pääkkönen, K. Fredriksson, P. Permi, J. Lounila, A. Annila, *J. Am. Chem. Soc.* **2003**, *125*, 15647 – 15650.
[14] M. R. Jensen, P. R. L. Markwick, S. Meier, C. Griesinger, M. Zweckstetter, S. Grzesiek, P. Bernadó, M. Blackledge, *Structure* **2009**, *17*, 1169 – 1185.
[15] H. Schwalbe, K. M. Fiebig, M. Buck, J. A. Jones, S. B. Grimshaw, A. Spencer, S. J. Glaser, L. J. Smith, C. M. Dobson, *Biochemistry* **1997**, *36*, 8977 – 8991.
[16] L. J. Smith, K. A. Bolin, H. Schwalbe, M. W. MacArthur, J. M. Thornton, C. M. Dobson, *J. Mol. Biol.* **1996**, *255*, 494 – 506.
[17] A. K. Jha, A. Colubri, K. F. Freed, T. R. Sosnick, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 13099 – 13104.
[18] P. Bernadó, L. Blanchard, P. Timmins, D. Marion, R. W. H. Ruigrok, M. Blackledge, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 17002 – 17007.
[19] V. Ozenne, F. Bauer, L. Salmon, J.-R. Huang, M. R. Jensen, S. Segard, P. Bernadó, C. Charavay, M. Blackledge, *Bioinformatics* **2012**, *28*, 1463 – 1470.
[20] P. Bernadó, C. W. Bertoncini, C. Griesinger, M. Zweckstetter, M. Blackledge, *J. Am. Chem. Soc.* **2005**, *127*, 17968 – 17969.
[21] M. D. Mukrasch, P. Markwick, J. Biernat, M. von Bergen, P. Bernadó, C. Griesinger, E. Mandelkow, M. Zweckstetter, M. Blackledge, *J. Am. Chem. Soc.* **2007**, *129*, 5235 – 5243.
[22] M. Wells, H. Tidow, T. J. Rutherford, P. Markwick, M. R. Jensen, E. Mylonas, D. I. Svergun, M. Blackledge, A. R. Fersht, *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 5762 – 5767.
[23] M. R. Jensen, K. Houben, E. Lescop, L. Blanchard, R. W. H. Ruigrok, M. Blackledge, *J. Am. Chem. Soc.* **2008**, *130*, 8055 – 8061.
[24] G. Nodet, L. Salmon, V. Ozenne, S. Meier, M. R. Jensen, M. Blackledge, *J. Am. Chem. Soc.* **2009**, *131*, 17908 – 17918.
[25] L. Salmon, G. Nodet, V. Ozenne, G. Yin, M. R. Jensen, M. Zweckstetter, M. Blackledge, *J. Am. Chem. Soc.* **2010**, *132*, 8407 – 8418.
[26] M. R. Jensen, G. Communie, E. A. Ribeiro, Jr., N. Martinez, A. Desfosses, L. Salmon, L. Mollica, F. Gabel, M. Jamin, S. Longhi et al., *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 9839 – 9844.
[27] J. Huang, F. Gabel, M. R. Jensen, S. Grzesiek, M. Blackledge, *J. Am. Chem. Soc.* **2012**, *134*, 4429 – 4436.
[28] V. Ozenne, R. Schneider, M. Yao, J.-R. Huang, L. Salmon, M. Zweckstetter, M. R. Jensen, M. Blackledge, *J. Am. Chem. Soc.* **2012**, *134*, 15138 – 15148.
[29] J. A. Marsh, J. M. R. Baker, M. Tollinger, J. D. Forman-Kay, *J. Am. Chem. Soc.* **2008**, *130*, 7804 – 7805.
[30] S. Meier, S. Grzesiek, M. Blackledge, *J. Am. Chem. Soc.* **2007**, *129*, 9799 – 9807.
[31] M.-K. Cho, H.-Y. Kim, P. Bernado, C. O. Fernandez, M. Blackledge, M. Zweckstetter, *J. Am. Chem. Soc.* **2007**, *129*, 3032 – 3033.
[32] J.-R. Huang, M. Gentner, N. Vajpai, S. Grzesiek, M. Blackledge, *Biochem. Soc. Trans.* **2012**, *40*, 989 – 994.
[33] J. Huang, S. Grzesiek, *J. Am. Chem. Soc.* **2010**, *132*, 694 – 705.
[34] N. Vajpai, M. Gentner, J.-R. Huang, M. Blackledge, S. Grzesiek, *J. Am. Chem. Soc.* **2010**, *132*, 3196 – 3203.

