J|A|C|S COMMUNICATIONS

Published on Web 12/03/2010

Sequence-Specific Random Coil Chemical Shifts of Intrinsically Disordered Proteins

Kamil Tamiola, Burçin Acar, and Frans A. A. Mulder*

Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Nijenborgh 4, 9749 AG Groningen, The Netherlands

Received June 28, 2010; E-mail: f.a.a.mulder@rug.nl

Abstract: Although intrinsically disordered proteins (IDPs) are widespread in nature and play diverse and important roles in biology, they have to date been little characterized structurally. Auspiciously, intensified efforts using NMR spectroscopy have started to uncover the breadth of their conformational landscape. In particular, polypeptide backbone chemical shifts are emerging as powerful descriptors of local dynamic deviations from the "random coil" state toward canonical types of secondary structure. These digressions, in turn, can be connected to functional or dysfunctional protein states, for example, in adaptive molecular recognition and protein aggregation. Here we describe a first inventory of IDP backbone ^{15}N, ^{1}H^{N}, ^{1}H$^{\alpha}$, ^{13}C^{O}, ^{13}C$^{\beta}$, and ^{13}C$^{\alpha}$ chemical shifts using data obtained for a set of 14 proteins of unrelated sequence and function. Singular value decomposition was used to parametrize this database of 6903 measured shifts collectively in terms of 20 amino acid-specific random coil chemical shifts and 40 sequence-dependent left- and right-neighbor correction factors, affording the *ncIDP* library. For natively unfolded proteins, random coil backbone chemical shifts computed from the primary sequence displayed root-mean-square deviations of 0.65, 0.14, 0.12, 0.50, 0.36, and 0.41 ppm from the experimentally measured values for the ^{15}N, ^{1}H^{N}, ^{1}H$^{\alpha}$, ^{13}C^{O}, ^{13}C$^{\beta}$, and ^{13}C$^{\alpha}$ chemical shifts, respectively. The *ncIDP* prediction accuracy is significantly higher than that obtained with libraries for small peptides or "coil" regions of folded proteins.

Introduction

In recent years, NMR spectroscopy has proven to be singular in its capacity to study intrinsically disordered proteins (IDPs) with atomic detail.[1-7] Because of the lack of a unique three-dimensional structure, the conformational state of IDPs is described by extensive ensembles derived from a thoroughgoing analysis of various experimental data.[5,6,8-10] As an alternative to comprehensive structure determination, NMR chemical shifts are of significant value, since they reflect the conformational preferences of polypeptide chains with atomic resolution.[11-13] Flexible peptides and unfolded proteins display "random coil" chemical shifts, which in turn can be used as a hallmark of disorder. The deviation of a measured chemical shift from its random coil value indicates the relative tendency of the polypeptide chain to adopt either helical or extended conformations at that point in the primary sequence,[11] thereby offering a sensitive and accurate proxy for changes in protein (dis)order and dynamics.[12,14-16]

Here we describe the first neighbor-corrected random coil chemical shift library for intrinsically disordered proteins, *ncIDP*, which enables the straightforward and accurate prediction of nuclear shielding constants for a polypeptide sequence. To generate this library, we manually compiled a list of the chemical shifts for 14 polypeptides that have been demonstrated in independent studies to be intrinsically disordered. For 12 of these, the resonance assignments were obtained from the BioMagResBank (BMRB) repository,[17] and two further IDPs were assigned in our lab [see Table S1 and the Supporting Information (SI) for details]. Using a total of 6903 experimental nuclear shielding constants, we solved the following equation:

$$\delta^{n}(x,a,y,i) = \delta^{n}_{\mathrm{RC}}(a) + \Delta^{n}_{-1}(x) + \Delta^{n}_{+1}(y) + \varepsilon^{n}(i) \qquad (1)$$

Equation 1 states that for each protein entry *i*, the observed chemical shift of a nucleus $n \in \{^{1}\mathrm{H}^{\alpha}, {}^{1}\mathrm{H}^{\mathrm{N}}, {}^{13}\mathrm{C}^{\alpha}, {}^{13}\mathrm{C}^{\beta}, {}^{13}\mathrm{C}^{\mathrm{O}}, {}^{15}\mathrm{N}\}$ in an amino acid *a* embedded in the tripeptide sequence $x-a-y$ consists of a random coil reference value $\delta^{n}_{\mathrm{RC}}(a)$, a left-neighbor correction $\Delta^{n}_{-1}(x)$, and a right-neighbor correction $\Delta^{n}_{+1}(y)$. The fourth parameter, $\varepsilon^{n}(i)$, is available to account for chemical shift offsets due to alternative referencing and also subsumes systematic deviations due to variations in pH or temperature. A single offset is included for chemical shifts of type *n* for each entry *i*. In a first round, the linear set of eqs 1 was solved for the 6903 experimental chemical shifts using singular value decomposition (SVD). The SVD algorithm effectively determined the *ncIDP* random coil chemical shift library, which comprises the reference chemical shift values of the 20 amino acids *a* when adjoined by glycine along with the 40 amino acid-specific corrections. The presence of structure results in local changes in the (ensemble distribution of) bond angles, which are manifested through sequence-dependent deviations from the random coil chemical shifts. For example, the ^{13}C$^{\alpha}$ chemical shift increases upon formation of α-helix and decreases in the context of a β-strand. On the basis of various types of experimental data, reports in the literature for the IDPs studied here indicate that these polypeptides do not form stable secondary or tertiary structures but sometimes display small segments that attain weakly populated, transient forms of organization. Thus, if accurate random coil chemical shifts are available, the distribution of secondary chemical shifts would be expected to consist of a sharp peak centered at zero for those nuclei present in random coil regions, augmented with broader features arising from segments that exhibit various levels of digression from the random coil state. Figure 1 demonstrates that this is indeed what was observed when the *ncIDP* library was used as a reference set.

The features observed in Figure 1 are not unique to ^{1}H$^{\alpha}$ chemical shifts but are visible for all chemical shifts that are sensitive to backbone conformation[12] (see Figure S1 in the SI). Since a portion of the data contains conformational bias away from the random coil state, as gauged from the secondary chemical shifts, we devised a self-consistent optimization protocol based on multiple linear regression (described in the SI) to identify outliers in the experimental data and subsequently eliminate them prior to the derivation of a new, curated *ncIDP* library from the remaining data. Through

10.1021/ja105656t

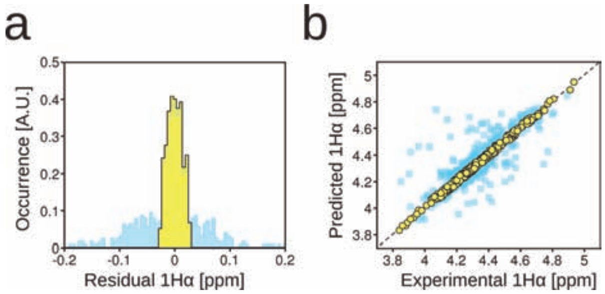

Figure 1. Distribution of ${}^{1}H^{\alpha}$ secondary chemical shifts for 14 IDPs obtained using the *ncIDP* library as a reference set for the random coil chemical shift: (a) histogram of the distribution of secondary chemical shifts; (b) predicted vs experimentally observed chemical shifts. Outliers removed by the self-consistent optimization protocol are shown in blue, and the data retained in the curated set are shown in yellow.

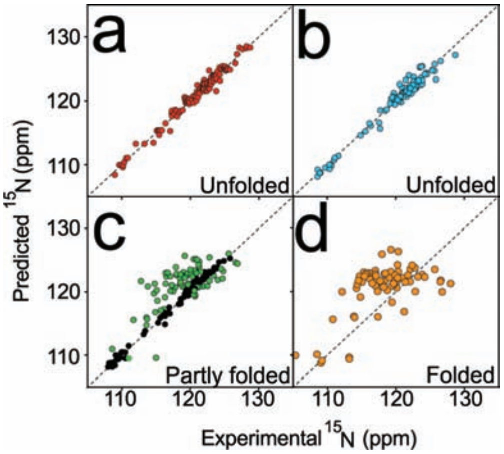

Figure 2. Correlation between experimental and computed ${}^{15}N$ chemical shifts for (a) γ-synuclein, (b) endosulphin-α, (c) human prion protein (hPrP^{C}), and (d) calcium-loaded calbindin D_{9k} P43G.

this iterative procedure the tails of the secondary chemical shift distributions for all nuclei were eliminated (Figure S1). Only the data shown in yellow in Figure 1 and Figure S1 were finally included in the derivation of the definitive *ncIDP* library. The values for $\delta^{n}_{RC}(a)$, $\Delta^{n}_{-1}(x)$, and $\Delta^{n}_{+1}(y)$ are given in Table S2 together with full details of their derivation. With the *ncIDP* library, knowledge of the primary sequence of a query protein alone is sufficient to predict its backbone proton, carbon, and nitrogen random coil chemical shifts.

In order to validate the statistical robustness of the new reference data set, the *ncIDP* library was derived several times, leaving out one protein entry at a time, and the eliminated chemical shift data were back-predicted. Some results of these predictions are displayed in Figure 2a,b (the predictions for all of the chemical shifts for each of the 14 proteins and peptides are given in Figure S2). Figure 2a,b shows the agreement between the observed and predicted ${}^{15}N$ chemical shifts for the IDPs γ-synuclein[16] (BMRB ID 7244) and endosulfin-α[18] (BMRB ID 15136), respectively. The observed strong correlation between the measured and calculated data (R^2 = 0.99, rmsd = 0.62 ppm and R^2 = 0.96, rmsd = 0.85 ppm, respectively) demonstrates that accurate neighbor-corrected random coil ${}^{15}N$ chemical shifts can be calculated. Figure 2c,d illustrates this point further. The human prion protein (hPrP^{C}; BMRB ID 4402) is known to have an intrinsically disordered N-terminal domain (data shown in black in Figure 2c) while maintaining a well-structured C-terminal part[19] (green data in Figure 2c). Indeed, the predicted chemical shifts are clearly at variance with complete disorder (R^2 = 0.83, rmsd = 2.47 ppm), and the two domains can be readily distinguished. Fully folded proteins display a pattern of even larger deviations, as shown in Figure 2d for the helical protein calbindin D_{9k} P43G (BMRB ID 16340) (R^2 = 0.48, rmsd = 4.35 ppm). Evidently, the *ncIDP* database allows for the accurate detection of protein (dis)order: the root-mean-square difference (rmsd) between the observed and calculated ${}^{15}N$ chemical shifts can be used to define the level of structure in a protein (see Figure S2), with IDPs displaying rmsd values less than 1.0. Above this threshold, persistent structure is to be expected.

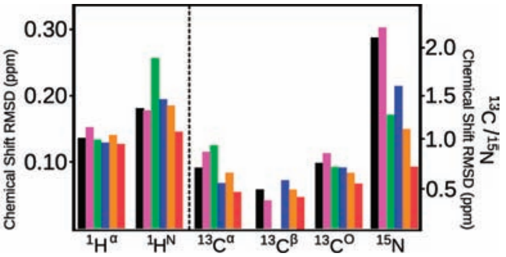

Figure 3. Average chemical shift rmsd's for the following databases and libraries: RefDB[20] (black), Wishart et al.[21] (magenta), Schwarzinger et al.[22] (green), Wang and Jardetzky[23] (blue), De Simone et al.[24] (orange), and this work (red). It should be noted that no rereferencing procedure was applied to the experimental input data.

The performance of *ncIDP* was subsequently benchmarked against alternative random coil chemical shift libraries available in the literature by calculating the rmsd's between the experimental and predicted chemical shifts for 14 unfolded proteins (Table S1), eliminating the query entry prior to building a database from the remaining protein entries. Figure 3 shows the results of these calculations, averaged over the 14 IDPs. (The individual comparisons are given in Figure S3). As a first reference set, we used the RefDB random coil database[20] (black bars), which amends incorrectly referenced chemical shift data submitted to the BMRB.[17] In RefDB, chemical shift averages are reported for regions that do not classify as α-helix or β-sheet, and these are labeled as random coil. Since this database was not developed to predict random coil chemical shifts, it does not utilize the concept of neighbor corrections. Second, we tested the experimental libraries compiled for small synthetic peptides, which are used to mimic polypeptide random coil states. The first library we tested was compiled for Ac-Gly-Gly-*a*-Ala-Gly-Gly-NH_2 in 1 M urea (pH 5) by Wishart et al.[21] (magenta bars). For this reference set, it appears that ${}^{1}H$ and ${}^{13}C$ chemical shifts in particular are different from those for the other databases, which might result from a conformational bias in the case of Ala. For example, the ^{13} C^{α} chemical shifts differ by 0.4 ppm on average from the experimental data for the peptides Ac-Gly-Gly-*a*-Gly-Gly-NH_2 in 8 M urea (pH 2.3) measured by Schwarzinger et al.[22] (Asp and Glu not included). The use of amino acid-specific corrections in the Schwarzinger database improves the correlation between the predicted and experimental data (green bars), but the presence of urea and the low pH cause significant offsets for some nuclei, making this reference state less representative for the chemical shifts obtained under native conditions. As an alternative, neighbor-corrected random coil chemical shift databases have been derived from the nuclear shieldings observed for protein regions found to be outside regular secondary structure elements and turns (i.e., assigned as "coil"), as evaluated from the corresponding Protein Data Bank (PDB) structures. We also compared here the predicted chemical shifts utilizing the databases of Wang and Jardetzky[23] (blue bars) and De Simone et al.[24] (orange bars). The newly derived *ncIDP* library (red bars) clearly demonstrates that the prediction of chemical shifts for natively unfolded

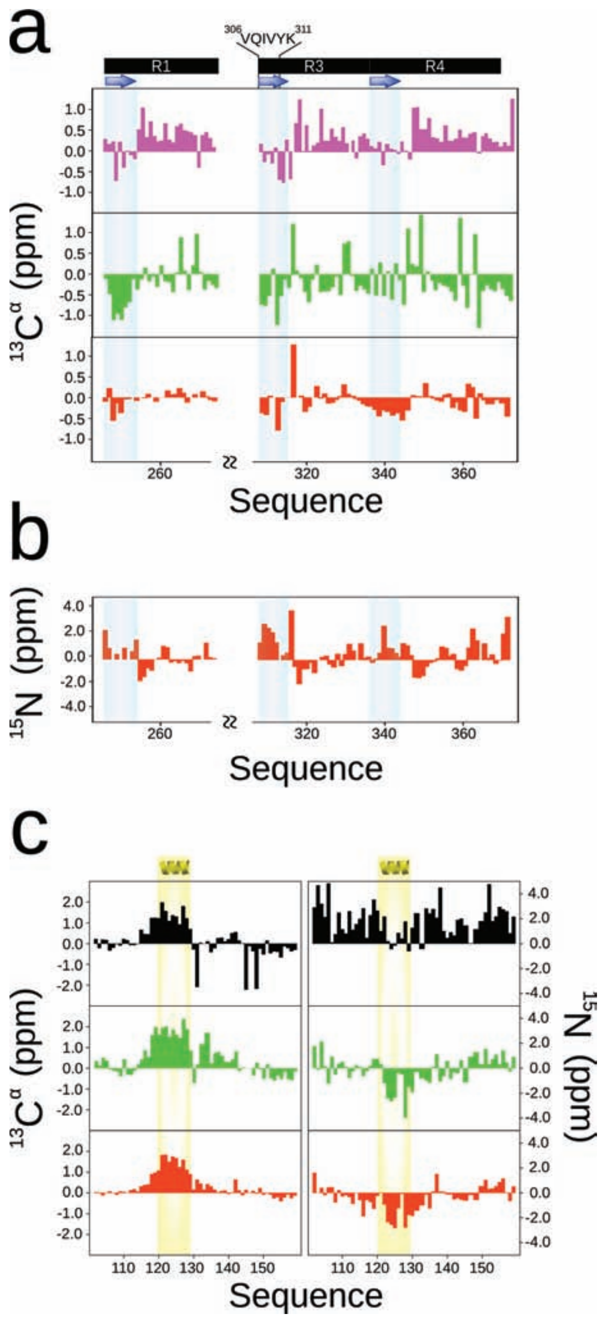

Figure 4. (a) ^{13}C$^{\alpha}$ secondary chemical shifts for the K19 construct of the human tau protein (repeat regions R1, R3, and R4) computed according to the protocols from Eliezer et al.[25] (magenta), Mukrasch et al.[26] (green), and this work (red). A −0.15 ppm referencing offset was computed using routines found in Marsh et al.[16] (b) ^{15}N secondary chemical shifts for human tau protein (repeat regions R1, R3, and R4) computed using *ncIDP*. The position of the hexapeptide essential for aggregation of tau protein is indicated. (c) ^{13}C$^{\alpha}$ and ^{15}N secondary chemical shifts for the kinase inducible domain of rat CREB (residues 101−160) computed using RefDB[20](black), Schwarzinger et al.[22] (green), and this work (red).

proteins benefits most from making use of the experimental data obtained for IDPs as a reference state. It also shows that neighbor corrections are significant for all backbone nuclei.

The detection of local structural propensity in a protein chain is crucially dependent on the availability of a reliable set of reference chemical shifts for the random coil state.[5,9−12,16,25−27] Figure 4 buttresses this point by demonstrating the effect of different random coil libraries on the secondary chemical shifts for two intrinsically disordered protein domains. The first example, shown in Figure 4a, focuses on the aggregation-prone repeat regions of the human tau protein. Earlier studies presented conflicting results: one investigation[25] indicated significant α-helical propensity on the basis of the Wishart[21] random coil values, whereas another study[26] suggested that weakly populated β-sheet-like regions are predominant on the basis of a comparison with the random coil chemical shifts of Schwarzinger et al.[22] The present results do not identify α-helical regions in the K19 portion of human tau but are in good agreement with the presence of β-sheet propensity at the beginning of each of the repeat regions.[26,27] It is important to emphasize that this difference is not due to the experimental data obtained in the two studies but results entirely from the use of a different reference state for the calculation of secondary chemical shifts. The agreement with the analysis of Mukrasch et al.[26] holds true at a macroscopic level, but there are also differences observed at the residue level. Much of the residue-to-residue variation seen in the middle panel of Figure 4 is strongly suppressed when using *ncIDP* (Figure 4 bottom). In addition, the ^{15}N secondary chemical shifts shown in Figure 4b demonstrate good agreement with those obtained from ^{13}C$^{\alpha}$ (it should be noted that these secondary chemical shifts are strongly anticorrelated) and support the notion that the imperfect hexapeptide repeats contain a strong signal for aggregation into a cross-β structure. The possibility of also detecting transitory helical structure is illustrated with a second example involving the kinase inducible domain (KID) of the rat CREB transcription factor. For CREB-KID, it was previously established that the two helices interacting with the KIX domain of the coactivator CBP are already but differentially populated in the free form.[28] We present in Figure 4c the sequence-dependent secondary chemical shifts for the ^{13}C$^{\alpha}$ and ^{15}N nuclei of CREB-KID (101−160) calculated using RefDB,[20] the database of Schwarzinger et al.,[22] and *ncIDP*. The predictions made with *ncIDP* confirm that the first helix (hα:120−129) is significantly populated in the absence of KIX, whereas the second helix (hβ:134−144) in the complex is better described as a random coil in isolation. Predictions with the program AGADIR[29] yield levels of 30 and 1% for the two helices, respectively. The stark difference in helical propensity of the two regions is also consistent with the observation of "helical" amide (i, $i + 3$) NOE connectivities for hα but not for hβ.[28]

The above examples show that the determination of a representative set of neighbor-corrected random coil chemical shifts for IDPs results in a more continuous pattern of secondary shifts, supporting the reliability of the method. Moreover, although the conformational sensitivity of ^{15}N chemical shifts is well-documented,[12] deviations from random coil values appear not to be useful[16] in the absence of neighbor corrections. The improvement obtained in the prediction of random coil ^{15}N chemical shifts obtained here (Figure 3) clearly facilitates the use of secondary chemical shifts as a gauge for the conformational state of a protein (Figure 2) and the sequence-specific detection of weak signals of transient structure (Figure 4).

Finally, any method that utilizes a comparative analysis of protein chemical shifts with respect to a reference random coil state will greatly benefit from the accuracy and reliability offered by *ncIDP*. This new library can be readily interfaced with available protein chemical shift analysis tools, such as chemical shift index[11] (CSI), structural propensity score assessment for intrinsically disordered proteins,[16] and protein structure modeling from chemical shift information[30] (CS-Rosetta).

Acknowledgment. Prof. Dr. G. Lippens (Lille, France) is acknowledged for bringing the problem of random coil reference data to our attention. We thank Dr. X. Periole and Prof. Dr. D. B. Janssen for their continued support of this project. This work was supported by a VIDI Grant to F.A.A.M. from The Netherlands Organization for Scientific Research (NWO).

Supporting Information Available: Materials and methods, complete ref 17, Tables S1 and S2, and Figures S1−S3. This material is available free of charge via the Internet at http://pubs.acs.org. The neighbor-corrected IDP random coil chemical shift library *ncIDP* and a sequence-based calculation interface are available at http://www.protein-nmr.org/.

References

(1) Tompa, P. *Trends Biochem. Sci.* **2002**, *27*, 527.
(2) Bartlett, A. I.; Radford, S. E. *Nat. Struct. Mol. Biol.* **2009**, *16*, 582.
(3) Uversky, V. N.; Oldfield, C. J.; Dunker, A. K. *Annu. Rev. Biophys.* **2008**, *37*, 215.
(4) Meier, S.; Blackledge, M.; Grzesiek, S. *J. Chem. Phys.* **2008**, *128*, 052204.
(5) Mittag, T.; Forman-Kay, J. D. *Curr. Opin. Struct. Biol.* **2007**, *17*, 3.
(6) Mulder, F. A. A.; Lundqvist, M.; Scheek, R. M. In *Instrumental Analysis of Intrinsically Disordered Proteins: Assessing Structure and Conformation*, 1st ed.; Uversky, V., Longhi, S., Eds.; Wiley, Hoboken, NJ, 2010; Chapter 3, p 61.
(7) Dyson, H.; Wright, P. *Nat. Rev. Struct. Mol. Biol.* **2005**, *6*, 197.
(8) Shortle, D. R. *Curr. Opin. Struct. Biol.* **1996**, *6*, 24.
(9) Eliezer, D. *Methods Mol. Biol.* **2007**, *350*, 49.
(10) Eliezer, D. *Curr. Opin. Struct. Biol.* **2009**, *19*, 23.
(11) Wishart, D. S.; Sykes, B. D.; Richards, F. M. *Biochemistry* **1992**, *31*, 1647.
(12) Wishart, D. S.; Case, D. A. *Methods Enzymol.* **2001**, *338*, 3.
(13) Mulder, F. A. A.; Filatov, M. *Chem. Soc. Rev.* **2010**, *39*, 578.
(14) Berjanskii, M.; Wishart, D. S. *Nat. Protoc.* **2006**, *1*, 683.
(15) Berjanskii, M. V.; Wishart, D. S. *J. Biomol. NMR* **2008**, *40*, 31.
(16) Marsh, J. A.; Singh, V. K.; Jia, Z.; Forman-Kay, J. D. *Protein Sci.* **2006**, *15*, 2795.
(17) Ulrich, E. L.; et al. *Nucleic Acids Res.* **2008**, *36*, 402.
(18) Boettcher, J. M.; Hartman, K. L.; Ladror, D. T.; Qi, Z.; Woods, W. S.; George, J. M.; Rienstra, C. M. *Biomol. NMR Assignments* **2007**, *1*, 167.
(19) Zahn, R.; Liu, A.; Lührs, T.; Riek, R.; von Schroetter, C.; García, F. L.; Billeter, M.; Calzolai, L.; Wider, G.; Wüthrich, K. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 145.
(20) Zhang, H.; Neal, S.; Wishart, D. S. *J. Biomol. NMR* **2003**, *25*, 173.
(21) Wishart, D. S.; Bigam, C. G.; Holm, A.; Hodges, R. S.; Sykes, B. D. *J. Biomol. NMR* **1995**, *5*, 67.
(22) Schwarzinger, S.; Kroon, G. J.; Foss, T. R.; Chung, J.; Wright, P. E.; Dyson, H. J. *J. Am. Chem. Soc.* **2001**, *123*, 2970.
(23) Wang, Y.; Jardetzky, O. *J. Am. Chem. Soc.* **2002**, *124*, 14075.
(24) De Simone, A.; Cavalli, A.; Hsu, S. D.; Vranken, W.; Vendruscolo, M. *J. Am. Chem. Soc.* **2009**, *131*, 16332.
(25) Eliezer, D.; Barré, P.; Kobaslija, M.; Chan, D.; Li, X.; Heend, L. *Biochemistry* **2005**, *44*, 1026.
(26) Mukrasch, M. D.; Biernat, J.; von Bergen, M.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M. *J. Biol. Chem.* **2005**, *280*, 24978.
(27) Mukrasch, M. D.; Bibow, S.; Korukottu, J.; Jeganathan, S.; Biernat, J.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M. *PLoS Biol.* **2009**, *7*, e1000034.
(28) Radhakrishnan, I.; Pérez-Alvarado, G. C.; Dyson, H. J.; Wright, P. E. *FEBS Lett.* **1998**, *430*, 317.
(29) Lacroix, E.; Viguera, A. R.; Serrano, L. *J. Mol. Biol.* **1998**, *284*, 173.
(30) Shen, Y.; Vernon, R.; Baker, D.; Bax, A. *J. Biomol. NMR* **2009**, *43*, 63.

JA105656T