JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

COMMUNICATION

pubs.acs.org/JACS

Impact of ^{15}N R_2/R_1 Relaxation Restraints on Molecular Size, Shape, and Bond Vector Orientation for NMR Protein Structure Determination with Sparse Distance Restraints

Yaroslav Ryabov,[†] Charles D. Schwieters,[*,†] and G. Marius Clore[*,‡]

[†]Division of Computational Bioscience, Building 12A, Center for Information Technology, National Institutes of Health, Bethesda, Maryland 20892-5624, United States

[‡]Laboratory of Chemical Physics, Building 5, National Institutes of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-0520, United States

Supporting Information

ABSTRACT: ^{15}N R_2/R_1 relaxation data contain information on molecular shape and size as well as on bond vector orientations relative to the diffusion tensor. Since the diffusion tensor can be directly calculated from the molecular coordinates, direct inclusion of ^{15}N R_2/R_1 restraints in NMR structure calculations without any a priori assumptions is possible. Here we show that ^{15}N R_2/R_1 restraints are particularly valuable when only sparse distance restraints are available. Using three examples of proteins of varying size, namely, GB3 (56 residues), ubiquitin (76 residues), and the N-terminal domain of enzyme I (EIN, 249 residues), we show that incorporation of ^{15}N R_2/R_1 restraints results in large and significant increases in coordinate accuracy that can make the difference between being able or unable to determine an approximate global fold. For GB3 and ubiquitin, good coordinate accuracy was obtained using only backbone hydrogen-bond restraints supplemented by ^{15}N R_2/R_1 relaxation restraints. For EIN, the global fold could be determined using sparse nuclear Overhauser enhancement (NOE) distance restraints involving only NH and methyl groups in conjunction with ^{15}N R_2/R_1 restraints. These results are of practical significance in the study of larger and more complex systems, where the increasing spectral complexity and number of chemical shift degeneracies reduce the number of unambiguous NOE asssignments that can be readily obtained, resulting in progressively reduced NOE coverage as the size of the protein increases.

The mainstay of protein structure determination by NMR spectroscopy resides in short (<6 Å) interproton distance restraints derived from nuclear Overhauser enhancement (NOE) measurements.[1,2] As proteins get larger, the number of NOE restraints that can be unambiguously assigned decreases as the spectral complexity increases.[3] There is therefore considerable interest in developing methods to facilitate NMR structure determination in cases where only sparse NOE restraints are available.[4−10] In optimal circumstances, backbone chemical shift data for selection of protein fragments with similar chemical shifts from a structure database, combined with the use of sophisticated modeling software to assemble the fragments and minimize the resulting models, can potentially generate structures with accuracies comparable to those obtained using conventional NMR structure determination procedures.[11,12] However, methods based purely on chemical shifts are generally limited to proteins containing fewer than ∼120 residues because of combinatorial explosion in the fragment assembly procedure. Further, the tertiary structure information content inherent in backbone chemical shifts is minimal. Residual dipolar couplings (RDCs), which are measured in weakly aligned media and yield orientational restraints on bond vectors relative to an external alignment tensor,[13,14] have been shown to result in large improvements in coordinate accuracy even with minimal NOE restraints.[7] Transverse (R_2) and longitudinal (R_1) relaxation rates, in addition to providing orientational restraints on bond vectors relative to the diffusion tensor,[15] are also dependent on the shape and size of the molecule.[16−18] In previous work, we have shown that refinement against the rotational diffusion tensor is extremely useful in restraining the molecular shape and size of protein−protein complexes[19] and that direct refinement against ^{15}N R_2/R_1 relaxation rates can accurately drive protein−protein docking even in the absence of any other experimental NMR restraints.[20] However, the former work[19] does *not* include N−H bond vector orientational information and does *not* refine directly against the R_2/R_1 ratios, while the latter[20] requires fairly accurate starting structures for the individual proteins for docking and is therefore *not* applicable for de novo structure determination. Here we show how relaxation data (in concert with a few distances) can be used to determine unknown structures, and we also demonstrate that inclusion of ^{15}N R_2/R_1 restraints in a simulated annealing-based structure determination algorithm results in large increases in the coordinate accuracy of structures generated from sparse distance restraints. This is illutrated by application to the proteins GB3 (56 residues), ubiquitin (76 residues), and the N-terminal domain of enzyme I (EIN, 249 residues),

The structure determination protocol makes use of the molecular structure determination package Xplor-NIH[21] in combination with the E_{relax} potential,[20] which directly minimizes the difference between observed and calculated ^{15}N R_2/R_1 ratios. The latter are computed from the coordinates and the rotational

Received: February 9, 2011
Published: April 04, 2011

dx.doi.org/10.1021/ja201020c | J. Am. Chem. Soc. 2011, 133, 6154–6157

Table 1. Summary of Structural Statistics

	without/with ^{15}N R_2/R_1 restraints		
	GB3	ubiquitin	EIN
accuracy (Å)[a]	3.2/1.1	3.5/1.8	14.7/4.1
precision (Å)[b]	1.4/1.2	1.3/1.7	11.2/8.1
experimental restraints			
R_2/R_1 χ^2 [c]	4.0 ± 0.6/2.0 ± 0.2	5.7 ± 0.7/3.6 ± 0.5	128 ± 30/2.2 ± 0.4
no. of R_2/R_1 data points excluded[c]	4.9 ± 0.6/4.7 ± 0.3	5.0 ± 0/5.0 ± 0	15.2 ± 1.2/4.9 ± 0.7
rms deviation from distance restraints (Å)[d]	0.01 ± 0.00/0.01 ± 0.00	0.01 ± 0.00/0.02 ± 0.01	0.04 ± 0.00/0.05 ± 0.00
rms deviation from ϕ/ψ torsion angle restraints (deg)[d]	0.13 ± 0.07/0.36 ± 0.14	0.48 ± 0.10/0.69 ± 0.10	2.87 ± 0.30/3.56 ± 0.33
R-factors for independent validation against RDCs (%)[e]			
bicelles	24 ± 4/18 ± 5	47 ± 7/39 ± 7	–
phage	47 ± 11/32 ± 5	–	63 ± 5/55 ± 2

[a] Defined as the Cα atomic rms difference between the restrained regularized mean structure and the reference X-ray structure. The PDB codes for the GB3, ubiquitin, and EIN reference X-ray structures are 1IGD,[22] 1UBQ,[23] and 1ZYM,[24] respectively. Residues 72−76 of ubquitin are disordered in solution[25] and therefore were excluded in calculating the accuracy. [b] Defined as the Cα atomic rms difference between the 10 lowest-energy structures and the restrained regularized mean coordinates. [c] The χ^2 values were normalized over the number of experimental ^{15}N R_2/R_1 ratios used in the calculations. Outlier ^{15}N R_2/R_1 data points were automatically excluded during the calculation as described in the text. [d] The number of experimental restraints in each case is provided in the text. It should be noted that for each hydrogen bond, there were two distance restraints: N−O and HN−O were set to 1.8−3.3 and 1.8−2.3 Å, respectively. [e] The RDC R-factor, R_{dip}, is expressed as $R_{dip} = [\langle (D_{obs} - D_{calcd})^2 \rangle / (2\langle D_{obs}{}^2 \rangle)]^{1/2}$ where D_{obs} and D_{calcd} are the observed and calculated RDCs, respectively.[26] The latter were calculated by singular value decomposition using Xplor-NIH.[21] The RDCs for GB3 and ubiquitin were taken from refs 32 and 33, respectively. The RDCs for free EIN were obtained from ref 38.

diffusion tensor, which is itself calculated from the shape and size of the molecule as described previously.[19,20] The effects of increased viscosity at higher protein concentrations, giving rise to increased R_2/R_1 ratios and concomitantly to an increase in the rotational correlation time, are handled by iterative optimization (during the course of simulated annealing) of the apparent diffusion tensor temperature (within a specified range of ±10°), which collects uncertainties in sample temperature, viscosity, and hydration layer description.[19] The protocol starts from a random coil conformation and employs extensive (200 ps) torsion angle dynamics sampling of conformational space[27] at high temperature (3500 K) followed by simulated annealing. Further details of the protocol are provided in the Supporting Information (SI). The target function comprised only experimental NMR restraints, a multidimensional torsion angle database potential of mean force,[28] a quartic van der Waals repulsion potential,[29] and terms to maintain idealized covalent geometries. For each example, we calculated 100 structures and selected for analysis the 10 structures having the lowest total energy.

In contrast to our previous work[20] on protein docking, in which outliers in the ^{15}N R_2/R_1 data (due to either large amplitude ps−ns motions or conformational exchange line broadening[30]) could easily be excluded because the structures of the individual component proteins of the complex were known, a priori exclusion of outliers was not possible in this instance. We therefore employed, during the course of the structure calculations, a fully automated, iterative data-filtering procedure wherein the mean (m_{diff}) and standard deviation (σ_{diff}) of the differences $d = \rho_{exptl} - \rho_{calcd}$ between the experimental and calculated R_2/R_1 ratios ($\rho = R_2/R_1$) are used to establish a threshold (Δ_{cut}) for excluding outliers. This threshold is given by the expression $\Delta_{cut} = |m_{diff}| + w_{cut}\sigma_{diff}$, where $w_{cut} > 0$ is a constant. E_{relax} is then defined as follows:

$$E_{relax} = k_{relax} \sum_{i=1}^{n} \frac{F(d_i)}{\sigma_i^2} \qquad (1)$$

where

$$F(d_i) = \begin{cases} d_i^2, & |d_i| \le \Delta_{cut} \\ A + B|d_i|^{-\alpha}, & |d_i| > \Delta_{cut} \end{cases} \qquad (2)$$

where i enumerates all of the experimental data points, the σ_i are the errors in the data, and k_{relax} is a force constant. The constants $A = (2 + \alpha)\Delta_{cut}{}^2/\alpha$ and $B = -2\Delta_{cut}{}^{\alpha+2}/\alpha$ are chosen to ensure that E_{relax} and its gradients are continuous functions. The exponent α determines the rate at which E_{relax} reaches its asymptotic behavior for $|d_i| > \Delta_{cut}$. In the current calculations, α was set to 8. Thus, in the region $|\rho_{exptl} - \rho_{calcd}| \le \Delta_{cut}$, E_{relax} has the usual χ^2 form, while outside these boundaries, E_{relax} rapidly becomes independent of the difference between the experimental and calculated R_2/R_1 ratios.

The energy term updates the values of m_{diff} and σ_{diff} during the course of the structure calculation protocol, concomitantly with tessellation of the protein surface (used to compute the diffusion tensor from the molecular shape and size)[19,20] to avoid any numerical discontinuities in the time-dependent behavior of E_{relax}. During the initial stages of the protocol, when the protein conformation is far from the final state, the value of m_{diff} can readily deviate from zero and exceed the value of the standard deviation (i.e., $|m_{diff}| > \sigma_{diff}$). Consequently, having the term $|m_{diff}|$ in the definition of Δ_{cut} ensures that not too many relaxation data points are excluded during the early stages of the calculation. Toward the end of the calculation, where $m_{diff} \approx 0$, only the value of σ_{diff} determines Δ_{cut}. In all of our calculations, we used $w_{cut} = 1.5$, which provides the same average fraction of excluded outliers in the relaxation data as the previously used filtering procedure based on a known structure.[20] In addition, the identities of the excluded residues were very similar (see the SI), indicating that the iterative procedure reliably identifies outliers arising from either local motions or errors in the experimental data. In this regard, the majority of excluded residues were

dx.doi.org/10.1021/ja201020c | J. Am. Chem. Soc. 2011, 133, 6154–6157

located in either tails, loops, or hinge regions at junctions between secondary structure elements.

To assess the impact of ^{15}N R_2/R_1 relaxation restraints on the coordinate accuracy of structures computed on the basis of sparse distance restraints, we made use of three examples. For two small proteins, GB3 (56 residues; diffusion anisotropy of ~1.3)[31,32] and ubiquitin (76 residues; diffusion anisotropy of ~1.2),[25,33] the distance restraints corresponded exclusively to backbone hydrogen bonds that could be easily identified from a qualitative interpretation of the backbone NOE data.[1] For the larger protein EIN (249 residues; diffusion anisotropy 1.7),[34] the backbone hydrogen-bond restraints were supplemented by NH−NH, NH−methyl, and methyl−methyl NOE restraints that could be readily assigned from an analysis of three- or four-dimensional heteronuclear-filtered NOE spectra acquired on [^{13}CH3-ILV]/[^{2}H/^{13}C/^{15}N]-labeled samples.[35,36] (These NOE restraints were selected out of the previously published complete set of NOE restraints[34].) For all three cases, the NOE data were supplemented by backbone ϕ/ψ torsion angle restraints obtained directly from backbone ^{1}H/^{15}N/^{13}C chemical shifts using the program TALOS+.[37] For GB3[32] and ubquitin,[33] there were 51 and 68 ^{15}N R_2/R_1 restraints, respectively, measured at a spectrometer frequency of 600 MHz; 35 and 28 backbone hydrogen bonds (with two distance restraints per hydrogen bond), respectively; and 104 and 130 ϕ/ψ restraints, respectively. For EIN, there were 117 ^{15}N R_2/R_1 restraints measured at 750 MHz,[38] 114 backbone hydrogen bonds, 804 NOE restraints involving only NH and methyl groups, and 484 ϕ/ψ torsion angle restraints. The results of the calculations are summarized in Table 1, and comparisons of the structures calculated with and without ^{15}N R_2/R_1 restraints versus the corresponding reference X-ray structures are shown in Figure 1. In each instance, the parameters of the diffusion tensor calculated from the molecular shape and size of the 10 lowest-energy structures were in excellent agreement with those calculated directly from the N−H bond vector orientations in the reference structure (see the SI).

In the case of both GB3 and ubiquitin, hydrogen-bond restraints alone provided an approximate fold. The accuracy of the resulting coordinates was poor, however, with Cα atomic root-mean-square (rms) differences of 3.2 and 3.5 Å, respectively, with respect to the corresponding reference X-ray structures. Inclusion of the ^{15}N R_2/R_1 restraints improved the accuracy by a factor of ~3, resulting in Cα rms differences with respect to the reference structures of 1.1 and 1.8 Å, respectively, for the restrained regularized mean coordinates (Table 1 and Figure 1). Interestingly, inclusion of ^{15}N R_2/R_1 restraints did not increase the precision. This is important because in the absence of ^{15}N R_2/R_1 restraints, the coordinate precision was a factor of 2−3 higher than the coordinate accuracy, whereas the precision and accuracy were comparable when ^{15}N R_2/R_1 restraints were included. In addition, independent validation against N−H residual dipolar couplings (RDC) indicated that inclusion of the ^{15}N R_2/R_1 restraints resulted in relative improvements of 17−25% in the RDC R-factor.

For the larger EIN protein, hydrogen-bond restraints alone were not sufficient to obtain a correct fold irrespective of the inclusion of the ^{15}N R_2/R_1 restraints (Cα rms difference of 17−21 Å with respect to the X-ray coordinates). However, the addition of sparse NOE restraints involving only NH and methyl groups permitted an approximate fold to be obtained in the presence of ^{15}N R_2/R_1 restraints. The accuracy of the Cα positions of the restrained minimized mean coordinates was 4.1 Å, compared with 14.7 Å without ^{15}N R_2/R_1 restraints, and the relative improvement in the RDC R-factor was 10−15%.

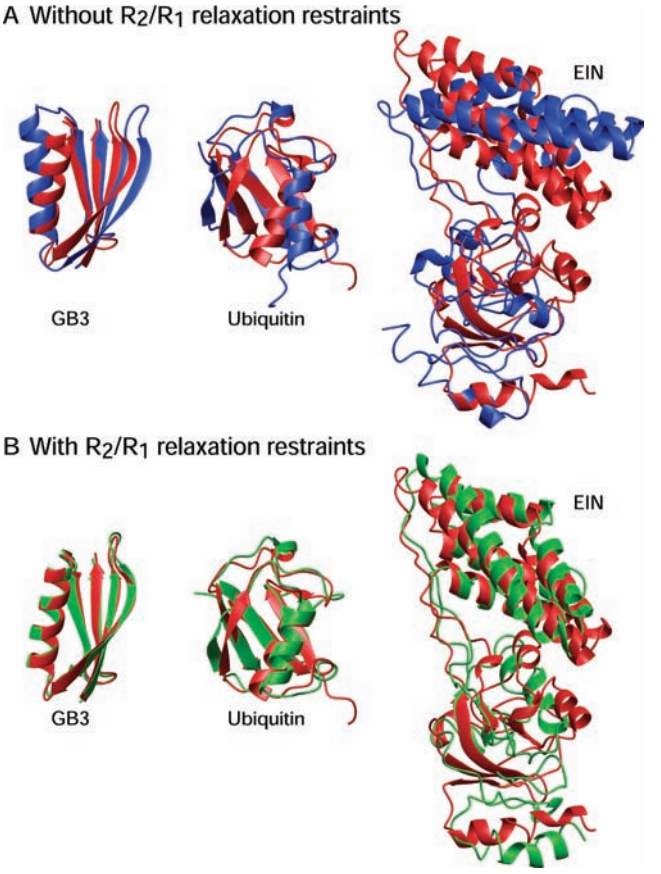

Figure 1. Comparison of structures calculated with sparse distance restraints either (A) without (blue) or (B) with (green) the inclusion of ^{15}N R_2/R_1 relaxation restraints vs the corresponding X-ray structures (red). For GB3 and ubquitin, the sparse restraints consisted exclusively of backbone hydrogen-bond restraints, while for EIN they also included NOE-derived interproton distance restraints involving NH and methyl groups. The PDB codes for the X-ray structures are 1IGD,[22] 1UBQ,[23] and 1ZYM.[24].

It should be noted that the precision of the 10 lowest-energy EIN structures obtained with ^{15}N R_2/R_1 restraints was rather low, and in this instance, there were several local minima with approximately the same overall energy. This is due to several factors: (i) the number of structural restraints in relation to the number of residues in the protein is sparse; (ii) the ^{15}N relaxation data possess intrinsic ambiguity associated with the fourfold symmetry of the ^{15}N R_2/R_1 ratios with regard to the N−H bond vector orientations relative to the diffusion tensor; and (iii) the number of distance restraints between the α and α/β subdomains of EIN (top and bottom in Figure 1) is sparse, causing small rms displacements at the interface of the two subdomains to translate into much larger atomic rms displacements at the outer edges of the molecule. Nevertheless, the inclusion of ^{15}N R_2/R_1 restraints made the difference between obtaining or not obtaining an approximately correct global fold.

In conclusion, we have demonstrated that direct inclusion of ^{15}N R_2/R_1 restraints into NMR structure calculations results in large increases in accuracy when only sparse NOE-derived

dx.doi.org/10.1021/ja201020c | J. Am. Chem. Soc. 2011, 133, 6154–6157

interproton distance restraints are available by providing information on both molecular size and shape and N−H bond vector orientations. The key feature in comparison with earlier work[15] is that the diffusion tensor was calculated at each step of the calculation on the basis of the current molecular surface. This entailed only a relatively modest (∼70%) increase in computational time relative to simulated-annealing calculations without relaxation data restraints. From a practical standpoint, the current results are significant because NOE coverage necessarily becomes sparser with increasing size and complexity of the protein as a result of increasing numbers of chemical shift degeneracies and unresolvable ambiguities in NOE assignments. However, because the method depends on calculation of the diffusion tensor from the shape and size of the molecule, some precautions do have to be taken, as this approach would not be suitable for proteins that aggregate or consist of domains that reorient independently of one another (e.g., proteins such as Ca^{2+}-loaded calmodulin, in which the two domains are connected by a highly flexible linker and there are no stable interdomain contacts). The method, however, is applicable to completely spherical proteins (diffusion anisotropy of 1), since the R_2/R_1 data still provide restraints on shape and size even though information on bond vector orientations is no longer present.

■ ASSOCIATED CONTENT

Supporting Information. Details of the structure determination protocol, including the Xplor-NIH Python script and examples of the relaxation data input files; breakdown of sparse distance restraints; diffusion tensor parameters; statistics of excluded residues; and complete citation for ref 12. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author
charles.schwieters@nih.gov; mariusc@mail.nih.gov

■ ACKNOWLEDGMENT

This work was supported by the NIH Intramural Research Programs of CIT (C.D.S.) and NIDDK (G.M.C.) and by the AIDS Targeted Antiviral Program of the Office of the Director of the NIH (G.M.C.)

■ REFERENCES

(1) Wüthrich, K. *NMR of Proteins and Nucleic Acids*; Wiley: New York, 1986.
(2) Clore, G. M.; Gronenborn, A. M. *Annu. Rev. Biophys. Biophys. Chem.* **1991**, *20*, 29.
(3) Tugarinov, V.; Choy, W. Y.; Orekhov, V. Y.; Kay, L. E. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 622.
(4) Levitt, M. *J. Mol. Biol.* **1983**, *170*, 723.
(5) Smith, B. O.; Ito, Y.; Raine, A.; Teichmann, S.; Ben-Tovim, L.; Nietlispach, D.; Broadhurst, R. W.; Terada, T.; Kelly, M.; Oschkinat, H.; Shibata, T.; Yokoyama, S.; Laue, E. D. *J. Biomol. NMR* **1996**, *8*, 360.
(6) Gardner, K. H.; Rosen, M. K.; Kay, L. E. *Biochemistry* **1997**, *36*, 1389.
(7) Clore, G. M.; Starich, M. R.; Bewley, C. A.; Cai, M.; Kuszewski, J. *J. Am. Chem. Soc.* **1999**, *121*, 6513.
(8) Delaglio, F.; Kontaxis, G.; Bax, A. *J. Am. Chem. Soc.* **2000**, *122*, 2142.
(9) Hus, J. C.; Marion, D.; Blackledge, M. *J. Mol. Biol.* **2000**, *298*, 927.
(10) Mueller, G. A.; Choy, W. Y.; Yang, D.; Forman-Kay, J. D.; Venters, R. A.; Kay, L. E. *J. Mol. Biol.* **2000**, *300*, 197.
(11) Cavalli, A.; Salvatella, X.; Dobson, C. M.; Vendruscolo, M. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 9615.
(12) Shen, Y.; et al. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 4685.
(13) Prestegard, J. H.; al-Hashimi, H. M.; Tolman, J. R. *Q. Rev. Biophys.* **2000**, *33*, 371.
(14) Bax, A.; Kontaxis, G.; Tjandra, N. *Methods Enzymol.* **2001**, *339*, 127.
(15) Tjandra, N.; Garrett, D. S.; Gronenborn, A. M.; Bax, A.; Clore, G. M. *Nat. Struct. Biol.* **1997**, *4*, 443.
(16) Woessner, D. E. *J. Chem. Phys.* **1962**, *37*, 647.
(17) Ryabov, Y. E.; Geraghty, C.; Varshney, A.; Fushman, D. *J. Am. Chem. Soc.* **2006**, *128*, 15432.
(18) Ryabov, Y.; Fushman, D. *J. Am. Chem. Soc.* **2007**, *129*, 7894.
(19) Ryabov, Y.; Suh, J. Y.; Grishaev, A.; Clore, G. M.; Schwieters, C. D. *J. Am. Chem. Soc.* **2009**, *131*, 9522.
(20) Ryabov, Y.; Clore, G. M.; Schwieters, C. D. *J. Am. Chem. Soc.* **2010**, *132*, 5987.
(21) Schwieters, C. D.; Kuszewski, J. J.; Clore, G. M. *Prog. Nucl. Magn. Reson. Spectrosc.* **2006**, *48*, 47.
(22) Derrick, J. P.; Wigley, D. B. *J. Mol. Biol.* **1994**, *243*, 906.
(23) Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J. *J. Mol. Biol.* **1987**, *194*, 531.
(24) Liao, D. I.; Silverton, E.; Seok, Y. J.; Lee, B. R.; Peterkofsky, A.; Davies, D. R. *Structure* **1996**, *4*, 861.
(25) Cornilescu, G.; Marquardt, J. L.; Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 6836.
(26) Clore, G. M.; Garrett, D. S. *J. Am. Chem. Soc.* **1999**, *121*, 9008.
(27) Schwieters, C. D.; Clore, G. M. *J. Magn. Reson.* **2001**, *152*, 288.
(28) Clore, G. M.; Kuszewski, J. *J. Am. Chem. Soc.* **2002**, *124*, 2866.
(29) Nilges, M.; Clore, G. M.; Gronenborn, A. M. *FEBS Lett.* **1988**, *229*, 317.
(30) Clore, G. M.; Driscoll, P. C.; Wingfield, P. T.; Gronenborn, A. M. *Biochemistry* **1990**, *29*, 7387.
(31) Ulmer, T. S.; Ramirez, B. E.; Delaglio, F.; Bax, A. *J. Am. Chem. Soc.* **2003**, *125*, 9179.
(32) Hall, J. B.; Fushman, D. *J. Am. Chem. Soc.* **2006**, *128*, 7855.
(33) Tjandra, N.; Feller, S. E.; Pastor, R. W.; Bax, A. *J. Am. Chem. Soc.* **1995**, *117*, 12562.
(34) Garrett, D. S.; Seok, Y. J.; Liao, D. I.; Peterkofsky, A.; Gronenborn, A. M.; Clore, G. M. *Biochemistry* **1997**, *36*, 2517.
(35) Gardner, K. H.; Kay, L. E. *Annu. Rev. Biophys. Biomol. Struct.* **1998**, *27*, 357.
(36) Tugarinov, V.; Kanelis, V.; Kay, L. E. *Nat. Protoc.* **2006**, *1*, 749.
(37) Shen, Y.; Delaglio, F.; Cornilescu, G.; Bax, A. *J. Biomol. NMR* **2009**, *44*, 213.
(38) Garrett, D. S.; Clore, G. M. Unpublished data.