Published in final edited form as:

J Am Chem Soc. 2011 April 27; 133(16): 6154–6157. doi:10.1021/ja201020c.

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 Ryabov¹, Charles D. Schwieters¹, and G. Marius Clore²

Charles D. Schwieters: charles.schwieters@nih.gov; G. Marius Clore: mariusc@mail.nih.gov

¹Division of Computational Bioscience, Building 12A, Center for Information Technology, National Institutes of Health, Bethesda, Maryland 20892-5624

²Laboratory of Chemical Physics, Building 5, National Institutes of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-0520

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 not being able to determine an approximate global fold. For GB3 and ubiquitin, good coordinate accuracy is obtained using only backbone hydrogen bond restraints supplemented by ^{15}N - R_2/R_1 relaxation restraints. For EIN, the global fold can be determined using sparse 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 increasing spectral complexity and chemical shift degeneracies reduce the number of unambiguous NOE assignments 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 resides in short ($< 6 \text{ \AA}$) 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 to select protein fragments with similar chemical shifts from a structure database combined, combined with sophisticated modeling software to assemble the fragments and minimize the resulting models, can potentially generate structures of comparable accuracy to those obtained using conventional NMR structure determination procedures.^{11,12} However, methods based purely on chemical shifts are generally limited to proteins less than about

Correspondence to: Charles D. Schwieters, charles.schwieters@nih.gov; G. Marius Clore, mariusc@mail.nih.gov.

Supporting Information **Available:** 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 and statistics of excluded residues. This material is available free of charge via the Internet at <http://pubs.acs.org>.

120 residues owing to combinatorial explosion in the fragment assembly procedure. Further, the tertiary structure information content inherent in backbone chemical shifts is minimal (being primarily limited to proton ring current shifts) and is not made use of in current algorithms.^{10,11} Residual dipolar couplings (RDCs), measured in weakly aligned media, yield orientational restraints on bond vectors relative to an external alignment tensor^{13,14} and 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 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 work¹⁹ does *not* include N-H bond vector orientational information and does *not* refine directly against the R_2/R_1 ratios, while the latter²⁰ 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 can be used (in concert with a few distances) to determine unknown structures and demonstrate that inclusion of ^{15}N R_2/R_1 restraints in a simulated annealing-based structure determination algorithm results in large increases in coordinate accuracy of structures generated from sparse distance restraints. This is illustrated 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-NIH²¹ in combination with the E_{relax} potential²⁰ that 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 diffusion tensor that is itself calculated from the shape and size of the molecule as described previously.^{19,20} 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 taken care of by iterative optimization (during the course of simulated annealing) of the apparent diffusion tensor temperature (within a specified range of $\pm 10^\circ$) that 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 space²⁷ at high temperature (3500 K) followed by simulated annealing. Further details of the protocol are provided in Supplementary. The target function comprises only experimental NMR restraints, a multidimensional torsion angle database potential of mean force,²⁸ a quartic van der Waals repulsion potential²⁹, and terms to maintain idealized covalent geometry. For each example we calculated 100 structures, and selected the 10 lowest total energy structures for analysis.

In contrast to our previous work²⁰ on protein docking where outliers in the ^{15}N R_2/R_1 data (due to either large amplitude ps-ns motions or to conformational exchange line broadening³⁰) could be easily excluded since the structures of the individual component proteins of the complex are known, exclusion of outliers is not possible *a priori* in this instance. We therefore adopted the following fully automated, iterative data filtering procedure during the course of the structure calculations whereby the mean, m_{diff} , and standard deviation, σ_{diff} , of the differences, $d = \rho_{\text{exp}} - \rho_{\text{calc}}$, between experimental and calculated R_2/R_1 ratios ($\rho = R_2/R_1$) are used to establish a threshold Δ_{cut} for excluding outliers. Δ_{cut} is given by $|m_{\text{diff}}| + w_{\text{cut}}\sigma_{\text{diff}}$, where $w_{\text{cut}} > 0$ is a constant. E_{relax} is then defined as follows:

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

where

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

i enumerates all experimental data points, 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 $B = -(2/\alpha)\Delta_{cut}^{\alpha+2}$ 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 once $|d_i| > \Delta_{cut}$. For the current calculations α is set to 8. Thus in the region $|\rho_{exp} - \rho_{calc}| \leq \Delta_{cut}$, E_{relax} has the usual χ^2 form, while outside these boundaries, E_{relax} rapidly becomes independent of the difference between 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 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 for Δ_{cut} ensures that not too many relaxation data points are excluded during the early stages of the calculation. Towards the end of the calculation, $m_{diff} \approx 0$ and it is only the value of σ_{diff} that determines Δ_{cut} . In all our calculations we use $w_{cut} = 1.5$ which provides the same average fraction of excluded outliers in the relaxation data as the filtering procedure based on a known structure used previously.²⁰ In addition, the identity of the excluded residues is very similar (see Supplementary) indicating that the iterative procedure reliably identifies outliers arising either from local motions or errors in the experimental data. In this regard, the majority of excluded residues are located either in 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 make use of three examples. For two small proteins, GB3 (56 residues; diffusion anisotropy ~ 1.3)^{31,32} and ubiquitin (76 residues; diffusion anisotropy ~ 1.2)^{25,33} the distance restraints correspond 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 are supplemented by NH-NH, NH-methyl and methyl-methyl NOE restraints that can be readily assigned from analysis of three or four-dimensional heteronuclear-filtered NOE spectra acquired on [$^{13}\text{CH}_3$ -ILV]/[$^2\text{H}/^{13}\text{C}/^{15}\text{N}$] labeled samples.^{35,36} (These NOE restraints were selected out of the previously published complete NOE restraints³⁴). For all three cases, the NOE data were supplemented by backbone ϕ/ψ torsion angle restraints obtained directly from backbone $^1\text{H}/^{15}\text{N}/^{13}\text{C}$ chemical shifts using the program TALOS+.³⁷ For GB3³² and ubiquitin,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 2 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 (D.S. Garrett & G.M. Clore, unpublished data), 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^{22-24,34} are shown in Fig. 1. In each instance the parameters of the diffusion tensor calculated from the molecular shape and size of the 10 lowest energy structures are in excellent agreement with those calculated directly from the N-H bond vector orientations in the reference structures (see Supplementary).

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

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

It will be noted that the precision of the 10 lowest energy EIN structures obtained with ^{15}N - R_2/R_1 restraints is rather low and in this instance there are several local minima with approximately the same overall energy. This is due to several factors: the number of structural restraints in relation to the number of residues in the protein is sparse; the ^{15}N relaxation data possess intrinsic ambiguity associated with 4-fold 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 the number of distance restraints between the α and α/β subdomains of EIN (top and bottom in Fig. 1) is sparse, such that small rms displacements at the interface of the two subdomains translates to much larger atomic rms displacements at the outer edges of the molecule. Nevertheless, the inclusion of ^{15}N - R_2/R_1 restraints makes 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 interproton distance restraints are available by providing information both on molecular size and shape and N-H bond vector orientations. The key feature, compared to earlier work,¹⁵ is that the diffusion tensor is calculated at each step of the calculation based on the current molecular surface. This only entails 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 since NOE coverage will necessarily become sparser with increasing size and complexity of the protein owing to increasing chemical shift degeneracies and unresolvable ambiguities in NOE assignments. Since the

method depends upon calculating the diffusion tensor from the shape and size of the molecule, some precautions, however, do have to be taken as this approach will 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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. John Wiley & Sons; New York: 1986.
2. Clore GM, Gronenborn AM. Annu Rev Biophys Biophys Chem. 1991; 20:29–63. [PubMed: 1651086]
3. Tugarinov V, Choy WY, Orekhov VY, Kay LE. Proc Natl Acad Sci U S A. 2005; 102:622–627. [PubMed: 15637152]
4. Levitt M. J Mol Biol. 1983; 170:723–764. [PubMed: 6195346]
5. Smith BO, Ito Y, Raine A, Teichmann S, Ben-Tovim L, Nietlispach D, Broadhurst RW, Terada T, Kelly M, Oschkinat H, Shibata T, Yokoyama S, Laue ED. J Biomol NMR. 1996; 8:360–368. [PubMed: 20686886]
6. Gardner KH, Rosen MK, Kay LE. Biochemistry. 1997; 36:1389–1401. [PubMed: 9063887]
7. Clore GM, Starich MR, Bewley CA, Cai M, Kuszewski J. J Am Chem Soc. 1999; 121:6513–6514.
8. Delaglio F, Kontaxis G, Bax A. J Am Chem Soc. 2000; 122:2142–2143.
9. Hus JC, Marion D, Blackledge M. J Mol Biol. 2000; 298:927–936. [PubMed: 10801359]
10. Mueller GA, Choy WY, Yang D, Forman-Kay JD, Venters RA, Kay LE. J Mol Biol. 2000; 300:197–212. [PubMed: 10864509]
11. Cavalli A, Salvatella X, Dobson CM, Vendruscolo M. Proc Natl Acad Sci U S A. 2007; 104:9615–9620. [PubMed: 17535901]
12. Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A. Proc Natl Acad Sci U S A. 2008; 105:4685–4690. [PubMed: 18326625]
13. Prestegard JH, al-Hashimi HM, Tolman JR. Q Rev Biophys. 2000; 33:371–424. [PubMed: 11233409]
14. Bax A, Kontaxis G, Tjandra N. Methods Enzymol. 2001; 339:127–174. [PubMed: 11462810]
15. Tjandra N, Garrett DS, Gronenborn AM, Bax A, Clore GM. Nature Structural Biology. 1997; 4:443–449.
16. Woessner DE. J Chem Phys. 1962; 37:647–654.
17. Ryabov YE, Geraghty C, Varshney A, Fushman D. J Am Chem Soc. 2006; 128:15432–15444. [PubMed: 17132010]
18. Ryabov Y, Fushman D. J Am Chem Soc. 2007; 129:7894–7902. [PubMed: 17550252]
19. Ryabov Y, Suh JY, Grishaev A, Clore GM, Schwieters CD. J Am Chem Soc. 2009; 131:9522–9531. [PubMed: 19537713]
20. Ryabov Y, Clore GM, Schwieters CD. J Am Chem Soc. 2010; 132:5987–5989. [PubMed: 20392103]

21. Schwieters CD, Kuszewski JJ, Clore GM. *Prog Nucl Mag Res Sp.* 2006; 48:47–62.
22. Derrick JP, Wigley DB. *J Mol Biol.* 1994; 243:906–918. [PubMed: 7966308]
23. Vijay-Kumar S, Bugg CE, Cook WJ. *J Mol Biol.* 1987; 194:531–544. [PubMed: 3041007]
24. Liao DI, Silverton E, Seok YJ, Lee BR, Peterkofsky A, Davies DR. *Structure.* 1996; 4:861–872. [PubMed: 8805571]
25. Cornilescu G, Marquardt JL, Ottiger M, Bax A. *J Am Chem Soc.* 1998; 120:6836–6837.
26. Clore GM, Garrett DS. *J Am Chem Soc.* 1999; 121:9008–9012.
27. Schwieters CD, Clore GM. *J Magn Reson.* 2001; 152:288–302. [PubMed: 11567582]
28. Clore GM, Kuszewski J. *J Am Chem Soc.* 2002; 124:2866–2867. [PubMed: 11902865]
29. Nilges M, Clore GM, Gronenborn AM. *FEBS Lett.* 1988; 229:317–324. [PubMed: 3345845]
30. Clore GM, Driscoll PC, Wingfield PT, Gronenborn AM. *Biochemistry.* 1990; 29:7387–7401. [PubMed: 2223770]
31. Ulmer TS, Ramirez BE, Delaglio F, Bax A. *Journal of the American Chemical Society.* 2003; 125:9179–9191. [PubMed: 15369375]
32. Hall JB, Fushman D. *J Am Chem Soc.* 2006; 128:7855–7870. [PubMed: 16771499]
33. Tjandra N, Feller SE, Pastor RW, Bax A. *J Am Chem Soc.* 1995; 117:12562–12566.
34. Garrett DS, Seok YJ, Liao DI, Peterkofsky A, Gronenborn AM, Clore GM. *Biochemistry.* 1997; 36:2517–2530. [PubMed: 9054557]
35. Gardner KH, Kay LE. *Annu Rev Biophys Biomol Struct.* 1998; 27:357–406. [PubMed: 9646872]
36. Tugarinov V, Kanelis V, Kay LE. *Nat Protoc.* 2006; 1:749–754. [PubMed: 17406304]
37. Shen Y, Delaglio F, Cornilescu G, Bax A. *J Biomol NMR.* 2009; 44:213–223. [PubMed: 19548092]

A Without R_2/R_1 relaxation restraints **B** With R_2/R_1 relaxation restraints **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 versus the corresponding X-ray structures (red). For GB3 and ubiquitin the sparse restraints consist exclusively of backbone hydrogen bond restraints, while for EIN they also include NOE-derived interproton distance restraints involving NH and methyl groups. The PDB codes for the X-ray structures are 1IGD,²² 1UBQ²³ and 1ZYM.²⁴

Table 1
Summary of structural statistics

	Without/With ^{15}N - R_2/R_I restraints		
	GB3	Ubiquitin	EIN
Accuracy ($\text{\AA}$) ^a	3.2 / 1.1	3.5 / 1.8	14.7 / 4.1
Precision ($\text{\AA}$) ^b	1.4 / 1.2	1.3 / 1.7	11.2 / 8.1
Experimental restraints			
R_2/R_I 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
Number of R_2/R_I excluded data points ^c	4.9±0.6 / 4.7±0.3	5.0±0 / 5.0±0	15.2±1.2 / 4.9±0.7
R.m.s. deviation from distance restraints ($\text{\AA}$) ^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
R.m.s. deviation from ϕ/ψ torsion angle restraints ($^\circ$) ^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-factor 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 Accuracy is defined as the Ca atomic r.m.s. 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 1UBQ²³ and 1ZYM,²⁴ respectively. (Residues 72-76 of ubiquitin are disordered in solution²⁵ and therefore excluded in calculating accuracy).

^b Precision is defined as the Ca atomic rms difference between the 10 lowest energy structures and the restrained regularized mean coordinates.

^c The χ^2 values are normalized over the number of experimental ^{15}N - R_2/R_I ratios used in the calculations. Outlier ^{15}N - R_2/R_I data points are automatically excluded during the calculation as described in the text.

^d The number of experimental restraints in each case is provided in the text. Note that for each hydrogen bond, there are two distance restraints, N-O and HN-O set to 1.8-3.3 and 1.8-2.3 $\text{\AA}$, respectively.

^e The RDC R-factor, R_{dip} , is expressed as $R_{dip} = \{ < (D_{obs} - D_{calc})^2 / (2 < D_{obs}^2) > \}^{1/2}$ where D_{obs} and D_{calc} are the observed and calculated RDCs.²⁶ The latter are 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 are from D.S. Garrett & GM. Clore (unpublished data).