J|A|C|S
ARTICLES
Published on Web 02/26/2002

De Novo Determination of Protein Backbone Structure from Residual Dipolar Couplings Using Rosetta

Carol A. Rohl and David Baker*

Contribution from the Department of Biochemistry, University of Washington, Seattle Washington 98195-7350

Received August 17, 2001. Revised Manuscript Received November 29, 2001

Abstract: As genome-sequencing projects rapidly increase the database of protein sequences, the gap between known sequences and known structures continues to grow exponentially, increasing the demand to accelerate structure determination methods. Residual dipolar couplings (RDCs) are an attractive source of experimental restraints for NMR structure determination, particularly rapid, high-throughput methods, because they yield both local and long-range orientational information and can be easily measured and assigned once the backbone resonances of a protein have been assigned. While very extensive RDC data sets have been used to determine the structure of ubiquitin, it is unclear to what extent such methods will generalize to larger proteins with less complete data sets. Here we incorporate experimental RDC restraints into Rosetta, an ab initio structure prediction method, and demonstrate that the combined algorithm provides a general method for de novo determination of a variety of protein folds from RDC data. Backbone structures for multiple proteins up to ~125 residues in length and spanning a range of topological complexities are rapidly and reproducibly generated using data sets that are insufficient in isolation to uniquely determine the protein fold de novo, although ambiguities and errors are observed for proteins with symmetry about an axis of the alignment tensor. The models generated are not high-resolution structures completely defined by experimental data but are sufficiently accurate to accelerate traditional high-resolution NMR structure determination and provide structure-based functional insights.

Introduction

Protein structure is a critical component of understanding function, both for individual proteins and on a systems level. While significant effort has focused on increasing the speed with which protein structures can be experimentally determined, obtaining a three-dimensional structure is frequently a rate-limiting step in assessing function. Consequently, methods that accelerate structure determination have extensive significance for a wide variety of fields. One area of particular recent focus for rapid NMR structure determination has been the use of orientational restraints provided by residual dipolar couplings (RDCs) measured for molecules partially aligned in a magnetic field.[1] RDCs have been used for multiple fold recognition algorithms,[2] and several approaches have utilized RDCs as a primary source of restraints for de novo structure determination: the high-resolution structure of cytochrome *c*′ was determined using RDC restraints and paramagnetic restraints,[3] the global fold of a three-helix bundle has been determined using RDC restraints supplemented with a limited number of NOE distance restraints,[4] and for the small protein ubiquitin, a high-resolution structure has been determined using only RDC restraints by two independent methods.[5,6]

Despite these promising results, the extent to which orientational restraints obtained from RDCs can be used generally as the sole source of experimental restraints to determine protein backbone structure is unclear. While a RDC restraint restricts the allowable orientations for an internuclear vector, it does not uniquely define this orientation.[1] The degeneracy of RDC restraints implies that the corresponding potential surface is rough, seriously impeding the ability of standard NMR structure determination protocols such as torsion angle dynamics to converge.[7] Methods that rely primarily on RDC restraints for structure determination generally require enough restraints to eliminate or reduce degeneracies so that local geometries can be unambiguously defined.[1,5,6] To generate sufficiently complete data sets, data must be collected for many different internuclear vectors or data must be collected in multiple alignment media. Alternatively, the degrees of freedom must be limited by treating substantial fragments of the protein structure as rigid bodies. Experimental and practical limitations such as exchange broadening, insufficient dispersion, and the absence of amide proton resonances for proline residues restrict the accuracy and completeness of RDC data. Additionally, internal dynamics also

* To whom correspondence should be addressed: (e-mail) dabaker@u.washington.edu; (fax) (206) 685-1792.

(1) Prestegard, J. H.; Al-Hashimi, H. M.; Tolman, J. R. *Q. Rev. Biophys*. **2000**, *33*, 371−424.
(2) (a) Annila, A.; Aitio, H.; Thulin, E.; Drakenberg. T. *J. Biomol. NMR* **1999**, *14*, 223−230. (b) Meiler, J.; Peti, W.; Griesinger, C. *J. Biomol. NMR* **2000**, *17*, 283−294. (c) Andrec, M.; Du, P.; Levy, R. M. *J. Am. Chem. Soc*. **2001**, *123*, 1222−1229.
(3) Hus, J.-C.; Marion, D.; Blackledge, M. *J. Mol. Biol*. **2000**, *298*, 927−936.
(4) Fowler, C. A.; Tian, F.; Al-Hashimi, H. M.; Prestegard, J. H. *J. Mol. Biol*. **2000**, *304*, 447−460.
(5) Delaglio, F.; Kontaxis, G.; Bax, A. *J. Am. Chem. Soc*. **2000**, *122*, 2142−2143.
(6) Hus, J.-C.; Marion, D.; Blackledge, M. *J. Am. Chem. Soc*. **2001**, *123*, 1541−1542.
(7) (a) Clore, G. M.; Starich, M. R.; Bewley, C. A.; Cai, M.; Kuszewski, J. *J. Am. Chem. Soc*. **1999**, *121*, 6513−6514. (b) Meiler, J.; Blomberg, N.; Nilges, M.; Griesinger, C. *J. Biomol. NMR* **2000**, *16*, 245−252.

contribute to observed RDCs, leading to errors when these couplings are interpreted in terms of static structure.

An alternate strategy to overcome the inherent degeneracy and incompleteness of RDC data is to supplement the experimental restraints with an alternate source of information, such as the empirical statistics employed by protein prediction methods. Rosetta is an algorithm for ab initio structure prediction that attempts to mimic the interplay of local and global interactions in determining protein structure.[8] The method is based on the experimental observation that local sequence preferences bias but do not uniquely define the local structure of a protein chain. The final native conformation is obtained when these fluctuating local structures come together to yield a compact conformation with favorable nonlocal interactions such as buried hydrophobic residues, paired β-strands, and specific side-chain interactions. In the Rosetta algorithm, the structures sampled by local sequences are approximated by the distribution of structures seen for those short sequences and related sequences in known protein structures: a library of fragments that represent the range of accessible local structures for all short segments of the protein chain are selected from the Protein Data Bank (PDB). Compact structures are then assembled by randomly combining these fragments using a Monte Carlo simulated annealing search. The fitness of individual conformations with respect to nonlocal interactions is evaluated using a scoring function derived from the observed residue distributions in known protein structures. A coarse Monte Carlo fragment insertion-based strategy enables the method to very effectively sample even rough energy landscapes, restricting the search to conformations that are consistent with both the local sequence preferences and the properties of native proteins.

Using only primary sequence information, successful de novo Rosetta predictions yield models on the order of 3−7 Å C^{α} RMSD to native for substantial fragments (>60 residues) of the query sequence. In the recent CASP 4 experiment, fragments of this size were correctly predicted for 16 of 21 attempted domains.[9] Here we describe the addition of residual dipolar coupling restraints to the Rosetta method and use the combined RosettaNMR algorithm to determine backbone structures for a variety of proteins. The results demonstrate that the combination of the experimental data with Rosetta overcomes many of the limitations in using RDC data to define protein folds. For the first time, backbone structures are obtained for multiple proteins other than ubiquitin using RDCs in the absence of distance restraints. These models provide rapid access to a moderate-resolution view of protein structure. In addition to accelerating high-resolution structure determination, when combined with sequence information, moderate-resolution backbone structures are likely to be useful for genome-scale methods for functional annotation, active site detection, or identification of functional specificity determinants.

Theory and Methods

Given a set a molecular coordinates, the residual dipolar coupling between atoms *m* and *n*, D^{mn}, can be calculated according to eq 1, where

$$D^{mn} = D^{mn}_{\max} \sum_{ij=\{x,y,z\}} S_{ij} \cos \phi_i^{mn} \cos \phi_j^{mn} \qquad (1)$$

S is the Saupe order matrix and ϕ_i^{mn} is the angle between the *mn* internuclear vector and the *i*th axis of the molecular frame. $D^{mn}_{\max}$ is given by eq 2, where γ_m is the gyromagnetic ratio of nucleus *m*, and

$$D^{mn}_{\max} = -\left(\frac{\mu_0}{4\pi}\right)\frac{\gamma_m \gamma_n h}{2\pi r_{mn}^{\ 3}} \qquad (2)$$

γ_{mn} is the internuclear distance. Here, each molecular conformation for which RDCs are evaluated is treated as a rigid body, and the Saupe order matrix yielding the least-squares fit to the reduced residual dipolar couplings is determined using singular value decomposition.[11] The normalized χ^2 between the experimental and calculated reduced RDCs is evaluated according to eq 3. The principal component of the

$$\chi^2 = \sum \left(\frac{D^{mn}_{\text{obs}} - D^{mn}_{\text{calc}}}{A^{mn}_{zz} D^{mn}_{\max}} \right)^2 \qquad (3)$$

diagonalized order tensor (A^{mn}_{zz}) is included to normalize data collected in different alignment media, and $D^{mn}_{\max}$ is included to normalize for differences in bond lengths and gyromagnetic ratios for different types of couplings.

Fragment libraries are composed of 200 nine- and three-residue fragments for every overlapping window in the protein sequence, selected from a nonredundant database of protein crystal structures of resolution better than 2.0 Å. For each query protein, all sequence homologues (BLAST *E*-value $<10^{-3}$) are removed from the database prior to fragment selection to avoid biasing the results. Each fragment in the database is scored according to its agreement with a multiple sequence alignment for the query sequence and agreement with experimental chemical shifts using a modification of the TALOS algorithm as previously described.[10,12] The RDC χ^2 is added to the scoring function in cases where sufficient data are available within a sequence window to determine the Saupe order matrix. The top 150 fragments for each sequence window are retained in the final fragment libraries. To ameliorate errors and uncertainties in the experimental data, 50 additional fragments for each window are selected solely on the basis of agreement with the multiple sequence alignment and the sequence-based predicted secondary structure as described previously.[8,10]

Models are generated using the Rosetta Monte Carlo simulated annealing protocol. All backbone atoms in the protein including H^N and H^{α} are explicitly included while each amino acid side chain is represented by a single centroid. Simulations start with the protein chain in an extended conformation and then contiguous sets of backbone torsion angles are replaced with those of fragments chosen randomly from the library. Protein conformations are evaluated according to the Rosetta potential function that favors hydrophobic burial, specific side-chain−side-chain interactions, pairing of β-strands, and overall compactness. This scoring function is derived from the observed residue distributions in known protein structures and has been extensively described elsewhere.[8] The scoring function is modified here to include the normalized RDC χ^2. Following the fragment assembly protocol, the lowest energy structures are subjected to a short Monte Carlo optimization protocol in which dihedral angles of single residues are randomly perturbed. Because Rosetta uses very short Monte Carlo simulations (~1 minute on a 1-GHz Pentium processor), most trajectories are expected to result in the incorrect structure. Consequently, for each protein or data set, multiple simulations are carried out from independent random seeds until the 10 lowest energy structures

(8) (a) Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. *J. Mol. Biol.* **1997**, *268*, 209−225. (b) Simons, K. T.; Ruczinski, I.; Kooperberg, C.; Fox, B. A.; Bystroff, C.; Baker, D. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 82−95.
(9) Bonneau, R.; Tsai, J.; Ruczinski, I.; Chivian, D.; Rohl, C.; Strauss, C. E. M. S.; Baker, D. *Proteins: Struct. Funct. Genet. Suppl.* **2001**, *5*, 119−126.
(10) Bowers, P. M.; Strauss, C. E. M.; Baker, D. *J. Biomol. NMR* **2000**, *18*, 311−318.
(11) Losonczi, J. A.; Andrec, M.; Fischer, M. W. F.; Prestegard, J. H. *J. Magn. Reson.* **1999**, *138*, 334−342.
(12) Cornilescu, G.; Delaglio, F.; Bax, A. *J. Biomol. NMR* **1999**, *12*, 239−302.

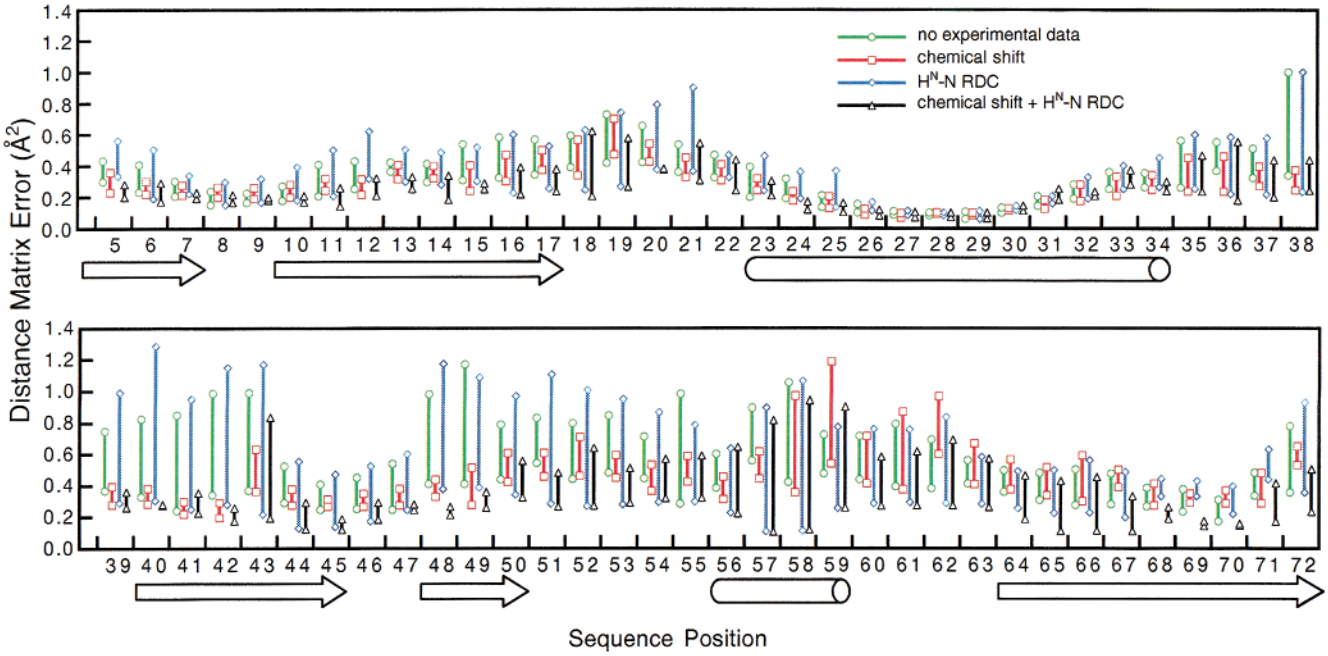

Figure 1. Accuracy of ubiquitin fragment libraries generated with different data sets. The range of distance matrix errors for the 10 fragments with the best match to the native structure is plotted for each 9-residue window in the ubiquitin sequence. Four fragment libraries generated with different experimental data are shown: no experimental data (green circles); backbone chemical shifts (red squares); H^N-N RDCs in one alignment medium (blue diamonds); and the combination of backbone chemical shift and H^N-N RDCs in one alignment medium (black triangles). The locations of regular secondary structure elements are indicated by arrows (strands) and cylinders (helices). Distance matrix error is the root-mean-square difference in distances between all atom pairs in the fragment and the corresponding pairs in the native structure.

cluster to the same global fold. Generally, 1000 simulations are sufficient to meet this criterion.

The method was tested on a set of proteins varying in size and secondary structure content using both simulated and real experimental RDC constraint sets. Data were simulated for the homeodomain of insulin gene enhancer protein (ISL-1),[13] ubiquitin,[14] histidine-containing phosphocarrier protein (HPr),[15] colicin E9 immunity protein (Im9),[16] ribosomal protein L30,[17] profilin I,[18] and G-α interacting protein (GAIP).[19] To construct the simulated RDC data sets, an alignment tensor was calculated from the molecular coordinates of the native structure using the PALES program,[20] and RDCs were calculated according to eq 1. A random 20% variation was introduced into the calculated couplings to simulate experimental error. The simulated data sets are available as Supporting Information. Simulations were also carried out using experimental restraints taken from the literature or PDB depositions for ubiquitin,[21] GAIP,[19] barrier-to-autointegration factor (BAF),[22] and cyanovirin-N.[23] Chemical shift assignments for all proteins were taken from the literature, BioMagResBank (www.bmrb.wisc.edu), or restraint files deposited in the PDB (www.rcsb.org/pdb) or generously provided by NMR laboratories. Chemical shift assignments for C′, C^α, C^β, H^α, and N atoms were used for proteins when available (ubiquitin, profilin I, BAF). For ISL-1, Im9, HPr, and GAIP, carbonyl carbon assignments were not included. For cyanovirin-N, only C^α, C^β, and H^α assignments were used.

Results

The quality of the fragment libraries and the effect of incorporating chemical shift and RDC data into the fragment selection phase of the algorithm are assessed in Figure 1. For ubiquitin, the range of distance matrix errors for the 10 fragments with the closest match to the native structure (out of 200 total fragments) is shown for each nine-residue window along the protein sequence. In the absence of any experimental data (green circles), the accuracy of the fragment library varies substantially with sequence position. In some regions, all of the 10 best fragments are close matches to the native structure (residues 24−33). In other regions, the accuracy of the best fragment is poorer, and the range of accuracies seen in the 10 best fragments is much larger (residues 40−45). With the addition of chemical shift data (red squares), the accuracy of the best fragments selected is increased for some regions of the sequence, but a more substantial difference is seen in the range of accuracies among the top 10 fragments: the best fragments frequently cluster more tightly at higher accuracies. With the addition of a single H^N-N RDC per residue (blue diamonds), better matches to the native fragment can frequently be obtained (residues 50−62) as expected from the sensitive orientational dependence of RDCs. In addition, however, substantially worse fragments are often found within the 10 best matches because the orientational restraints obtained from RDCs are not uniquely defined. Incorporation of both chemical shift and RDC data (black triangles) generally allows the best properties of both of these libraries to be combined, improving both the accuracy of

(13) Ippel, H.; Larsson, G.; Behravan, G.; Zdunek, J.; Lundqvist, M.; Schleucher, J.; Lycksell, P. O.; Wijmenga, S. *J. Mol. Biol.* **1999**, *288*, 689−703.
(14) Cornilescu, G.; Marquardt, J. L.; Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 6836−6837.
(15) Jia, Z.; Quail, J. W.; Waygood, E. B.; Delbaere, L. T. *J. Biol. Chem.* **1993**, *268*, 22490−224501.
(16) Osborne, M. J.; Breeze, A. L.; Lian, L. Y.; Reilly, A.; James, R.; Kleanthous, C.; Moore, G. R. *Biochemistry* **1996**, *35*, 9505−9512.
(17) Mao, H.; Williamson, J. R. *J. Mol. Biol.* **1999**, *292*, 345−359.
(18) Fedorov, A. A.; Magnus, K. A.; Graupe, M. H.; Lattman, E. E.; Pollard, T. D.; Almo, S. C. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 8636−8640.
(19) de Alba, E.; De Vries, L.; Farquhar, M. G.; Tjandra, N. *J. Mol. Biol.* **1999**, *291*, 927−939.
(20) Zweckstetter, M.; Bax, A. *J. Am. Chem. Soc.* **2000**, *122*, 3791−3792.
(21) Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 12334−12341.
(22) (a) Cai, M.; Huang, Y.; Zheng, R.; Wei, S. Q.; Ghirlando, R.; Lee, M. S.; Craigie, R.; Gronenborn, A. M.; Clore, G. M. *Nat. Struct. Biol.* **1998**, *5*, 903−909. (b) Umland, T. C.; Wei, S.-Q.; Craigie, R.; Davies, D. R. *Biochemistry* **2000**, *39*, 9130−9138.
(23) Bewley, C. A.; Gustafson, K. R.; Boyd, M. R.; Covell, D. G.; Bax, A.; Clore, G. M.; Gronenborn A. M. *Nat. Struct. Biol.* **1998**, *5*, 571−578.

Table 1. Simulations for Proteins Varying in Secondary Structure Content, Size, and Topological Complexity

protein	reference structure	length	contact order[a]	secondary structure	no. of RDCs	backbone RMSD (Å) lowest score	lowest RMSD	average pairwise
ISL-1 homeodomain	1bw5	52[b]	12.8	α	249	1.91	1.04	1.42 ± 0.49
ubiquitin	1d3z	76	21.2	α/β	364	1.66	1.61	1.09 ± 0.20
HPr	1poh	85	27.0	α/β	414	2.33	1.30	1.97 ± 0.62
IM9	1imq	86	21.9	α	411[c]	6.93	6.93	6.52 ± 1.27
					276[d]	5.86	5.86	7.36 ± 1.26
BAF	1ci4	89	19.4	α	246[e]	2.88[f]	2.45[f]	2.48 ± 0.53[d]
cyanovirin-N	2ezm	101	30.7	β	327[e]	3.24[g]	3.24[g]	6.00 ± 1.71
						2.84[h]	2.84[h]	
ribosomal L30	1ck2	104	32.1	α/β	503[i]	3.00	3.00	5.37 ± 0.74
profilin I	1acf	125	25.4	α/β	600	2.08	1.26	1.54 ± 0.31
					436	2.42	2.13	2.27 ± 0.49
					156[j]	2.99	2.99	4.04 ± 0.55
GAIP	1cmz	128	24.5	α	622	3.02	3.02	6.88 ± 2.47
					291[k]	4.55	4.41	2.69 ± 0.41

[a] Average sequence separation between contacting residues. [b] Residues 8−59. [c] Five of the 10 lowest energy structures represented the same global fold for this data set. [d] Seven of the 10 lowest energy structures represented the same global fold for this data set. [e] Experimental data; normalizations applied by the original authors to the published values were removed prior to use. [f] Structures consistent with a dimer (see text). [g] Residues 5−50. [h] Residues 55−101. [i] A total of 6000 simulations were required before 8 of the 10 lowest energy structures represented the same global fold. [j] $H^{\alpha}-C^{\alpha}$ and $H^{N}-N$ couplings only. [k] Experimental data; signs of couplings involving nitrogen atoms were reversed relative to the published values to account for the negative gyromagnetic ratio of nitrogen.

Table 2. Effect of Data Set Completeness on Model Accuracy for Ubiquitin

	experimental data: RDC	experimental data: chemical shift	backbone RMSD (Å)[a]: lowest score	lowest RMSD	average pairwise
I	137 $H^{N}-N$ 132 $C^{\alpha}-H^{\alpha}$ 136 C−N 134 $C-H^{N}$	real; two alignment media	+	1.17	1.03	0.77 ± 0.17
II	68 $H^{N}-N$ 66 $C^{\alpha}-H^{\alpha}$ 67 C−N 67C−H^{N}	real; one alignment medium	+	1.29	1.29	1.58 ± 0.33
III	72 $H^{N}-N$ 76 $C^{\alpha}-C$ 70 $C^{\alpha}-H^{\alpha}$ 75 C−N 71 $C-H^{N}$	simulated	+	1.17	0.95	1.03 ± 0.20
IV	68 $H^{N}-N$ 67 $C-H^{N}$ 67 C−N	real	+	1.86	1.65	1.23 ± 0.20
V	72 $H^{\alpha}-H^{N}$ 72 $H^{N}-N$	simulated	−	2.04	1.58	1.93 ± 0.29
VI	68 $H^{N}-N$	real	−	2.75	2.75	1.86 ± 0.46

[a] Residues 1−71.

the best fragment and the number of fragments approaching this accuracy.

Tables 1 and 2 summarize the results of all the simulations carried out. In general, the algorithm converges to the correct global fold for a variety of proteins: the accuracy of the lowest energy structure obtained by the method is ∼3 Å or better in most cases, and this fold is reproducibly obtained (Table 1). In the absence of experimental data, structure prediction with the Rosetta algorithm is generally limited to proteins with contact order (average sequence separation between contacting residues) of less that ∼20. In CASP 4, the most complex domain for which a successful prediction was made had a contact order of 15.[9] Here, the combination of Rosetta with RDC data reliably determines folds for proteins with complexities significantly beyond this limit, including the highest contact order protein in the test set, ribosomal L30 (contact order, 32.1). Additional simulations are required to obtain convergence for this protein, suggesting that the combined algorithm will also experience difficulty with extremely complex topologies. The largest errors in the low-energy structures are seen for all helical proteins such as Im9 and GAIP, where translational shifts of helices, as well as the variation of the N- and C-termini, give rise to the relatively high RMSD, despite the correctness of the global fold (Figure 2). Because most nonlocal close contacts in helical proteins occur between side-chain atoms and the RosettaNMR method uses a simplified centroid representation of side chains, incorporation of full-atom side-chain representations and an atomistic Lennard-Jones potential could yield models of increased accuracy. Additionally, it is likely that improvements in the sampling of the RosettaNMR method could produce structures of increased accuracy because the final models frequently showed poorer agreement with the experimental data than do the native structures.

To investigate the effect of data set completeness on the algorithm performance, simulations were repeated for profilin and Im9 using less complete data sets. Small decreases in accuracy and precision are observed when ∼30% of the data are randomly removed (Table 1). The correct fold is still reliably

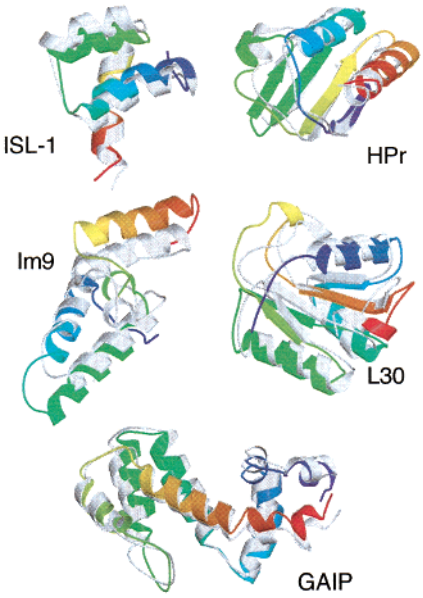

Figure 2. Model structures generated using Rosetta and RDC restraints. For each protein, the native structure is show in gray and the top-ranking model structure is colored from blue to red along the primary sequence. For Im9, the structure shown corresponds to Table 1, line 4. All protein structure diagrams were generated using Molscript and Raster3D.[27]

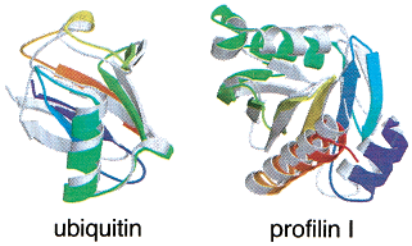

Figure 3. Structures generated using Rosetta with very sparse RDC data sets. Native structures are shown in gray. The ubiquitin structure is the top-ranking model obtained using 68 H^N-N RDCs (data set VI, Table 2). The profilin structure is the top-ranking model obtained using chemical shifts along with 85 H^N-N and 71 $H^\alpha-C^\alpha$ RDCs (Table 1, line 9).

obtained for profilin even when the data set includes only $H^\alpha-C^\alpha$ and H^N-N couplings for roughly two-thirds of the residues (Figure 3). The effect of completeness of the data set on algorithm performance was investigated more thoroughly using ubiquitin as a test case (Table 2). As expected, higher resolution structures are obtained in the presence of increased amounts of experimental data, but all of the data sets reproducibly yield the correct fold. Even with a very limited data set of only 68 H^N-N RDCs, the RosettaNMR algorithm correctly and reproducibly identifies the ubiquitin fold (Figure 3). This data set is significantly less than that required to define the global fold; structures of >20 Å from the native structure that satisfy the experimental restraints can easily be found (data not shown). In addition, the structure obtained from the combined algorithm with this sparse data set is also substantially better (2.8 Å) than the Rosetta predictions obtained in the absence of experimental data that range from 4 to 10 Å to the native structure (data not shown).

Similar accuracies are obtained for both ubiquitin and GAIP using either real or simulated data, and comparable results are obtained for Im9 using simulated data and BAF using experimental data, suggesting that the simulated data are reasonably representative of real experimental data. Interestingly, for GAIP,

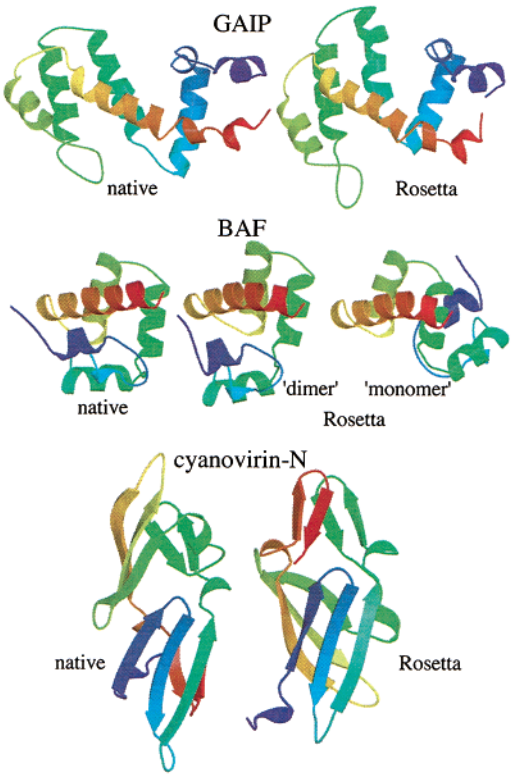

Figure 4. Best scoring protein structures generated for GAIP, BAF, and cyanovirin-N using Rosetta and experimental RDC restraints. For BAF, the "monomer" structure corresponds to the fold determined using only the RDC restraints while the "dimer" structure is that obtained by utilizing the additional information that the protein is a dimer and that residues with amides showing intersubunit NOEs (16, 43, 47, 48, 50, 51, 53) must be accessible to form the dimer interface. This information was not used during the simulations, but was used to remove structures inconsistent with dimer formation from the final population of models

the low-energy structures obtained from the real data set converge to a significantly tighter cluster than those observed with the simulated data set, despite the fact that the simulated data set contains roughly twice as many restraints (Table 1). While the low-energy structure obtained using the experimental data is ~4.5 Å from the reported NMR structure, the correct global fold is obtained reliably by the algorithm (Figure 4), and the structures obtained show better agreement with the experimental RDCs than does the high-resolution NMR structure.

The results obtained with BAF and cyanovirin-N illustrate one of the inherent limitations of RDC data: the observed couplings are insensitive to the inversion of any axis of the alignment tensor. BAF is a dimer in solution, but because RDC data cannot determine either the oligomerization state or the relative positions of the two subunits, the folding simulations treated the protein as a monomer. In the absence of any additional information, the RosettaNMR algorithm produces a fold in which the first three helices are rotated en masse by 180° relative to native structure (Figure 4). This fold can be easily identified as inconsistent with the dimer structure because residues with amide resonances that show intersubunit NOEs are buried in the protein core. If the additional information that the protein is a dimer and that helix 4 (green helix in Figure 4) is involved in the dimer interface is used to remove conformations in which this surface is not exposed from the population of models, then the remaining low-energy structures cluster around the correct fold (Figure 4). The agreement of both the

"monomer" and the "dimer" models with the RDC data is equivalent because the axis of rotation coincides with an axis of the alignment tensor, but the "monomer" structure is a lower energy structure (when evaluated as an isolated subunit) because the hydrophobic dimer interface is buried.

A similar phenomenon is observed with cyanovirin-N, a 101-residue β-protein with an unusual topology consisting of two symmetric halves. In the global fold obtained in the Rosetta simulations, the relative orientations of these two halves are reversed (Figure 4). As in BAF, this 180° rotation occurs about an axis of the alignment tensor. Since the RDC data cannot discriminate between these two alternate subdomain orientations, symmetry-related structures clearly must be considered, and additional information is required to resolve this ambiguity. Despite this uncertainty, the model does preserve the correct subdomain folds and would still be useful in aiding the assignment of additional restraints that could easily identify the correct the subdomain orientation.

Discussion

Here we demonstrate the enhancement in de novo fold determination that can be obtained by combining experimental RDC data with the Rosetta method: correct backbone structures can be uniquely determined by the combined algorithm even when the experimental data or Rosetta in isolation are insufficient to define the protein fold. For all the proteins examined, additional information is clearly needed to define the high-resolution structure, but the models provided by the RosettaNMR method are an excellent starting point for such an effort. The backbone structures determined here are of comparable or better accuracy than those obtained by comparative modeling or fold-recognition methods, without the requirement for sequence or structural homology. Models of this quality can and have been used previously to provide a variety of structure-based insights into protein function.[24] For example, residues implicated in DNA binding can be mapped equally well onto either the BAF model, the high-resolution NMR structure, or the crystal structure.[22] Such models will likely be useful for genome-scale functional annotation and characterization of active sites, particularly when combined with sequence information. The RosettaNMR models are also of sufficient accuracy to identify candidates for structural genomics efforts. Recent work has estimated that structures of ~16 000 carefully selected proteins will be required in order to provide reasonable coverage of structure space, but three times this many structures may be required if optimal target-selection methods are not used.[25] The models generated here would clearly be useful in rapidly identifying structures that are already well represented in the database.

The Rosetta method is a fundamentally different approach to utilizing RDC data for structure determination than has been previously described. Molecular dynamics-based NMR structure determination methods have traditionally relied on distance restraints to define global topology and usually incorporate RDCs later in the protocol to refine structures to high resolution.[7] The Monte Carlo fragment-insertion strategy used by Rosetta is capable of effectively searching the complex potential surface that results from the orientational degeneracy of the experimental restraints, allowing these restraints to be used early in the simulations to define overall topology. Previously described methods that utilize RDCs as the primary source of experimental restraints to define the backbone structure rely on unambiguously determining local geometry to initially define the protein fold at low resolution.[5,6] In contrast to such determininstic local buildup methods, both the local and long-range information in RDCs is used to define the backbone topology in Rosetta. Selection of fragments to include in the library does, of necessity, use only the local information, but the long-range information inherent in the RDCs is utilized throughout the fragment assembly process.

Defining protein topology solely from local geometry is problematic because small errors in local structure accrue and propagate. Global topology is defined primarily by the local geometry of loop regions, but RDC restraints are generally sparser and less accurate for loops than for regular secondary structure. Additionally, internal dynamics likely contribute to the observed RDCs to a greater extent in flexible loops making the interpretation of RDCs solely on structural terms even less reliable. Even for the small protein ubiquitin and a very complete and accurate data set, an initial model built from local geometry considerations is ~7 Å from the native structure.[5] For larger proteins with more complex topologies and less complete data, initial models built by satisfying only local geometry restraints are likely to be of very poor quality, and backbone structures for proteins other than ubiquitin have not been previously determined from RDCs in the absence of distance constraints. The Rosetta strategy tolerates errors in the local geometry, as illustrated by the fact that correct global folds are obtained for a variety of proteins even though the fragment libraries contain poor matches to the native structure at most positions and, for some sequence windows, may not contain any good matches (Figure 1). Furthermore, because the Rosetta method is probabilistic rather than deterministic, data sets that are insufficient in isolation to determine the backbone structure can be used to uniquely determine the fold in the combined algorithm.

One goal of the results presented here is to define the extent to which RDC data can be used to determine protein folds. Previous work has integrated distance restraints into the Rosetta algorithm and tested the efficacy of sparse backbone NOE restraints for determining proteins folds using a set of proteins similar to those used here.[10] Addition of either RDC data or NOE data to the Rosetta algorithm yields comparable results for most of the proteins tested. Not surprisingly, the relative usefulness of RDC and NOE data for defining folds depends on both the protein and the data involved. For example, the lowest energy GAIP structures obtained using distance constraints were on the order of 9.5−12 Å to the native structure, but the data sets utilized contained at most one long-range NOE (in addition to short-range NOEs). In contrast, structures obtained for Im9 with distance restraints including five to seven long-range NOEs are of higher accuracy (2−3 Å) than those obtained here with RDC data. Realistic sparse data sets are likely to contain a mixture of different data types, and the RosettaNMR algorithm allows both NOE and RDC data to be used simultaneously. The power of combining distance restraints with RDC data is obvious as even extremely limited distance restraints

(24) Vitkup D.; Melamud, E.; Moult, J.; Sander, C. *Nat. Struct. Biol.* **2001**, *8*, 559−566.
(25) Baker, D.; Sali, A. *Science* **2001**, *294*, 93−96.

can easily distinguish between symmetry-related structures that are identical with respect to RDC data.

Complete determination of a high-resolution structure, well defined by experimental data, is still the most common goal of NMR structural characterization, at least in part because of the time investment in obtaining a well-behaved sample and assigning spectra. The utility of rapid fold determination methods, such as the one described here, lies in their ability to generate models very early in the structure determination procedure from a limited number of restraints to expedite assignment and data analysis. From a genomics perspective, the ability to utilize incomplete data sets is particularly important, because high-throughput methods, such as automated backbone assignment protocols, are likely to generate incomplete data sets.[26] Sparse data methods also have particular relevance to systems that are not amenable to conventional structure determination methods for technical reasons. Using limited amounts of data, the Rosetta algorithm can converge on correct backbone structures for a variety of proteins. Even in cases where the data are insufficient to distinguish between alternative structures, the models obtained would still be extremely useful in advancing assignment, and newly assigned restraints can be used to distinguish between alternate conformations. Experimental structure determination is, in general, an iterative procedure, and the RosettaNMR algorithm has the potential to be a general tool to accelerate this process.

Conclusions

While RDC couplings are comparatively rapid to obtain and sensitive to both local and long-range structure, they are likely to be insufficient in isolation to determine protein structure because of both inherent degeneracies and practical limitations. Such underdetermined data sets are expected to be even more common in high-throughput, automated methods for structural genomics. Here we demonstrate that correct backbone folds for multiple proteins can be reproducibly generated by supplementing experimental RDC data with empirical information about protein structure in the form of the de novo structure prediction algorithm Rosetta. For the test set of nine proteins examined here, the most substantial uncertainties are obtained for all α-helical proteins, where translational shifts of helices can give rise to large RMSDs despite the correctness of the chain topology, and in symmetrical proteins, for which RDC restraints cannot distinguish rotations about axes of the alignment tensor. Rosetta is a novel approach to using RDCs for structure determination, combining a Monte Carlo method to search rough energy landscapes with a fragment assembly strategy that utilizes both local and long-range information to define the backbone fold. Additionally, the probabilistic nature of the method enables underdetermined data sets to be utilized. The combined RosettaNMR algorithm is a general method for de novo fold determination from RDC restraints. The models obtained are useful both for accelerating high-resolution NMR structure determination and, in cases where technical or practical limitations preclude determination of a high-resolution structure, for providing structure-based insights into protein function.

The RosettaNMR program is available from the authors. Requests should be directed to rosettaNMR@rosetta.bakerlab.org

Acknowledgment. The authors thank Ad Bax, Lewis Kay, and Lawrence McIntosh for chemical shift assignments and Jens Meiler, William Wedemeyer, Geoffrey Mueller, and James Choy for thoughtful discussion. This work was supported by the Howard Hughes Medical Institute. C.A.R. is a fellow of the Interdisciplinary Training in Genomic Sciences program, T32 HG00035-06.

Note Added in Proof. We thank a reviewer for directing us to the experimental data set of 300 RDCs for protein G (PDB accession code 3gb1). With these RDCs in the absence of chemical shift data, the top five structures obtained with the RosettaNMR algorithm have a pairwise RMSD of 1.49 ± 0.3 Å. Relative to the reference NMR structure, the low-scoring RosettaNMR model has an RMSD of 1.51 Å, and the low RMSD structure has an RMSD of 0.8 Å.

Supporting Information Available: Simulated RDC data sets used for the proteins in Table 1. This material is available free of charge via the Internet at http://pubs.acs.org. See any current masthead page for ordering information and Web access instructions.

JA016880E

(26) Moseley, H. N. B.; Montelione, G. T. *Curr. Opin. Struct. Biol.* **1999**, *9*, 635−642.
(27) (a) Kraulis, P. *J. Appl. Crystallogr.* **1991**, *24*, 946−950. (b) Merritt, E. A.; Bacon, D. J. *Methods Enzymol.* **1997**, *277*, 505−524.