

RDC derived protein backbone resonance assignment using fragment assembly

Xingsheng Wang · Brian Tash · John M. Flanagan · Fang Tian

Received: 23 November 2010 / Accepted: 15 December 2010 / Published online: 30 December 2010
© Springer Science+Business Media B.V. 2010

Abstract Experimental residual dipolar couplings (RDCs) in combination with structural models have the potential for accelerating the protein backbone resonance assignment process because RDCs can be measured accurately and interpreted quantitatively. However, this application has been limited due to the need for very high-resolution structural templates. Here, we introduce a new approach to resonance assignment based on optimal agreement between the experimental and calculated RDCs from a structural template that contains all assignable residues. To overcome the inherent computational complexity of such a global search, we have adopted an efficient two-stage search algorithm and included connectivity data from conventional assignment experiments. In the first stage, a list of strings of resonances (CA-links) is generated via exhaustive searches for short segments of sequentially connected residues in a protein (local templates), and then ranked by the agreement of the experimental $^{13}\text{C}_\alpha$ chemical shifts and ^{15}N - ^1H RDCs to the predicted values for each local template. In the second stage, the top CA-links for different local templates in stage I are combinatorially connected to produce CA-links for all assignable residues. The resulting CA-links are ranked for resonance assignment according to their measured RDCs and predicted values from a tertiary structure. Since the final RDC ranking of CA-links includes all assignable residues and

the assignment is derived from a “global minimum”, our approach is far less reliant on the quality of experimental data and structural templates. The present approach is validated with the assignments of several proteins, including a 42 kDa maltose binding protein (MBP) using RDCs and structural templates of varying quality. Since backbone resonance assignment is an essential first step for most of biomolecular NMR applications and is often a bottleneck for large systems, we expect that this new approach will improve the efficiency of the assignment process for small and medium size proteins and will extend the size limits assignable by current methods for proteins with structural models.

Keywords Backbone resonance assignment · Residual dipolar couplings · Structure-assisted resonance assignment

Introduction

NMR is a well-established tool for the structural analysis of small and medium size proteins in solution and is uniquely capable of studying conformational dynamics, disordered proteins and transient interactions at atomic resolution. The assignment of protein backbone resonances is a prerequisite for structural analysis by NMR. Typically, this is accomplished with $^{13}\text{C}_\alpha$, $^{13}\text{C}_\beta$ and $^{13}\text{C}'$ intrareidue and sequential connectivity data obtained from a set of amide-proton-detected triple resonance experiments on a ^{13}C , ^{15}N uniformly labeled protein (Sattler and Griesinger 1999; Cavanagh et al. 2006). While this protocol has proven to be a most reliable assignment strategy, chemical shift degeneracy and incomplete connectivity data often pose significant barriers to the process as the protein size increases (Wider and Wüthrich 1999). With the growing number of

Electronic supplementary material The online version of this article (doi:10.1007/s10858-010-9467-z) contains supplementary material, which is available to authorized users.

X. Wang · B. Tash · J. M. Flanagan · F. Tian (✉)
Department of Biochemistry and Molecular Biology, College of Medicine, Pennsylvania State University, Hershey, PA 17033, USA
e-mail: ftian@psu.edu

available 3D experimental structures and the maturation of computational methods for structure prediction, there is an awareness that prior structural knowledge can and should be exploited to facilitate protein backbone resonance assignment (Donald and Martin 2009; Bermejo and Llinas 2010). Conceptually, structure assisted resonance assignment is based on a comparison of measured and calculated NMR parameters from a known structure. Current methods fall into two groups: the first mainly relies on NOEs and does not require sequential connectivity from triple resonance experiments (Bartels et al. 1996; Marassi and Opella 2000; Wang et al. 2000; Hus et al. 2002; Pristovsek et al. 2002; Meiler and Baker 2003, 2005; Mesleh and Opella 2003; Langmead and Donald 2004; Apaydin et al. 2008; Stratmann et al. 2008, 2009; Xiong et al. 2008), while the second focuses on improving the conventional assignment process by incorporating data such as RDCs (Tian et al. 2001; Zweckstetter and Bax 2001; Jung et al. 2004; Jung and Zweckstetter 2004). These studies demonstrate the benefit of incorporating high-resolution structural information into the assignment process. Here, we present a new approach that uses RDCs and a structural model to overcome the impediments to resonance assignment due to chemical shift degeneracy and ambiguity in sequential connectivity.

RDCs arise from the anisotropic tumbling of molecules and provide orientational information about internuclear vectors relative a common alignment tensor frame. In the order matrix analysis of RDCs, the order tensor has five independent parameters: the axial and rhombic anisotropies of the alignment tensor, and three angles defining the relative orientation of a rigid unit to the tensor. Given a structural model and the elements of the order tensor, the resulting RDCs can be accurately and quickly predicted. Similarly, starting from a theoretical model and its associated RDCs, the agreement of this model to the actual structure can be evaluated in a straightforward manner. The application of RDCs has had a substantial, and often revolutionary, impact on the study of biomolecules using NMR (Prestegard et al. 2004; Bax and Grishaev 2005). Since RDCs are easily accessible and quantitatively interpretable in the context of a structural model, they provide a unique opportunity to facilitate resonance assignment, especially in conjunction with conventional experimental data (Tian et al. 2001; Jung et al. 2004; Jung and Zweckstetter 2004; Langmead and Donald 2004; Apaydin et al. 2008; Stratmann et al. 2008, 2009). However, when RDCs were included for the assignment in previous studies, it was noted that slight structural and dynamic deviations (“structural noise”) in the template were detrimental to the assignment process. This is due to the sensitivity of RDCs to the exact orientation of the corresponding internuclear vectors (Jung and Zweckstetter 2004; Langmead and

Donald 2004; Stratmann et al. 2009). For example, the RMSD between the back calculated RDCs from a 2 Å resolution crystal structure and the experimental values is in general > 20% of the experimental alignment strength, D_a^{HN} (Jung and Zweckstetter 2004). In the RDC-enhanced backbone resonance assignment program, MARS, these difficulties were partially overcome by reducing the RDC contribution to the scoring function by a factor of 3.3 relative to the chemical shift contribution, and by keeping consistent assignments from multiple cycles of assignments with RDCs perturbed by a large noise, 5σ (Jung and Zweckstetter 2004).

In the present study, we exploit an approach that exhaustively enumerates possible solutions for all residues targeted for assignment in order to identify the optimal global agreement between measured and predicted RDCs. In general, this is a computationally challenging task. For example, for a protein having 50 non-proline residues, there are $\sim 10^{64}$ possible combinations of expected resonances (50!). To overcome the inherent computational complexity of this process, we have adopted a two-stage search algorithm (Bryson et al. 2008) and included intraresidue and sequential connectivity data from conventional experiments to reduce the number of possible assignments. We will illustrate the new strategy with an application to the resonance assignment of ubiquitin using three independent structural templates of varying resolutions. The present approach is further validated on the resonance assignment of several proteins using RDCs and structural templates of differing quality: bovine crystallin, a 21 kDa protein using RDCs collected for murine crystallin; maltose binding protein (MBP), a 42 kDa protein using simulated RDCs perturbed by random errors up to $0.8 * D_a^{HN}$; and the third PDZ3 domain of a human tight junction protein, ZO-1, a 12 kDa protein using predicted models as templates since no experimental structure is currently available. The success of these examples clearly demonstrates that the present approach effectively minimizes the adverse effects of experimental errors and “structural noise” in templates, and thereby overcomes the largest obstacle to RDC-based resonance assignment.

Materials and methods

Sample preparation

Details of the PDZ3 preparation are provided in the supporting information. Briefly, a plasmid expressing the recombinant PDZ3 domain of the human ZO-1 (residues 409–518) protein was introduced into *E. coli* BL21 gold (DE3) cells for expression in M9 medium. Proteins were

purified by Ni–NTA chromatography and then digested with recombinant His-tagged TEV protease. The resulting protein was reapplied to the Ni–NTA column to remove the cleaved His-tag and TEV protease. The flow-through, containing PDZ3, was concentrated and further purified by gel filtration chromatography. For NMR, the protein was concentrated to ~1 mM in a buffer containing 20 mM phosphate, pH 7.0, 100 mM NaCl and 1 mM EDTA.

NMR data

All NMR spectra were recorded at 25°C on a Bruker 600 MHz AVANCE II spectrometer equipped with a cryogenic probe. One bond ^{15}N – ^1H RDCs for PDZ3 in Pf1 phage (Hansen et al. 1998) and 5% PEG (Ruckert and Otting 2000) alignment media were measured with ^{15}N IPAP-HSQC experiments (Ottiger et al. 1998). The deuterium splitting for the protein samples with phage and PEG were 10.7 and 24.6 Hz, respectively. To validate the resonance assignments obtained with our new approach, the backbone resonances of PDZ3 were independently assigned using the conventional approach from HNCACB, CBCA(CO)NH and 3D ^{15}N -edited NOESYHSQC data. Amide protons having fast exchange with water were detected using CLEANEX-PM-FHSQC experiments with mixing times of 10 and 20 ms (Hwang et al. 1998).

$^{13}\text{C}_\alpha$ chemical shifts and two sets of one bond ^{15}N – ^1H RDCs for ubiquitin were obtained from the Protein Data Bank (PDB code: 1D3Z). Two sets of one bond ^{15}N – ^1H RDCs for murine crystallin were obtained from the Biological Magnetic Resonance Bank (BMRB, mrblock_ID: 16460) and its $^{13}\text{C}_\alpha$ chemical shifts were from NRG-CING (nmr.cmbi.ru.nl/cing/NRG-CING.html). $^{13}\text{C}_\alpha$ chemical shifts for the ligand-free MBP were obtained from the BMRB (accession code: 4986). Residues 234 to 236 were not included in the assignment process since their chemical shifts were missing.

Data analysis

Seven theoretical structures of PDZ3 were produced with three modeling programs. Five models were predicted by the I-TASSER online server (<http://zhang.bioinformatics.ku.edu/I-TASSER>). A sixth model was generated with Modeller (version 9.5) using the third PDZ domain of PSD-95 (PDB ID: 1BFE) as a template. Of 1,000 models generated, the lowest energy structure was selected and energy minimized with UCSF Chimera. A seventh model was produced from the SWISS-MODEL (<http://swissmodel.expasy.org/>) online server. To quantitatively evaluate the quality of these structural models, REDCAT (Valafar and Prestegard 2004) was used to perform order matrix analyses with the experimental RDCs. When RDC data from

Table 1 Order matrix analysis of seven predicted structures of PDZ3 with RDCs from phage alignment medium by REDCAT

Models	Residue 16 to 105		39 residues from structured regions	
	Q-factor	Normalized RMSD ^a	Q-factor	Normalized RMSD
Model 1	0.53	0.42	0.45	0.33
Model 2	0.55	0.43	0.36	0.26
Model 3	0.57	0.45	0.34	0.25
Model 4	0.7	0.53	0.5	0.37
Model 5	0.6	0.47	0.36	0.26
Model 6	0.46	0.36	0.32	0.24
Model 7	0.64	0.49	0.4	0.3

^a Normalized RMSD is defined as the RMSD between the experimental and calculated RDCs divided by the experimental alignment strength, D_a^{HN}

phage alignment medium for residues 16 to 105 were used for analysis, the Q-factors and normalized RMSDs ranged from 0.46 to 0.7, and from 0.36 to 0.53, respectively, indicating poor structural quality. However, when only the residues from structured regions (α -helix and β -sheet) were used for analysis, the Q-factors and normalized RMSDs significantly decreased, ranging from 0.32 to 0.5, and 0.24 to 0.37, respectively. Thus the structured regions are better predicted than other regions. The results of this REDCAT analysis are listed in Table 1 and shown in Figure S1.

Searches for the resonance assignment utilizing the present strategy were carried out with in-house scripts written in Python and Perl on a Mac Powerbook with a 2.4 GHz Intel Core 2 Duo processor.

Results

The assignment method

The conventional strategy for backbone resonance assignment relies on the sequential and intraresidue correlations derived from scalar coupling mediated triple resonance experiments, and typically requires that the chemical shifts of $^{13}\text{C}_\alpha$, $^{13}\text{C}_\beta$ or $^{13}\text{C}'$ are resolved for reliable resonance connections. Incomplete connectivity data and chemical shift degeneracy are common problems with this approach. In the present study, we chose to use only $^{13}\text{C}_\alpha(i)$ and $^{13}\text{C}_\alpha(i-1)$ connectivity data to mimic this situation since $^{13}\text{C}_\alpha$ chemical shifts are known to have high degeneracy. These data were supplemented with two sets of ^{15}N – ^1H RDC measurements and a structural template for resonance assignment. Our approach, as illustrated by the flowchart in Fig. 1, operates in two stages. In the first stage, short strings of sequentially linked resonances are constructed

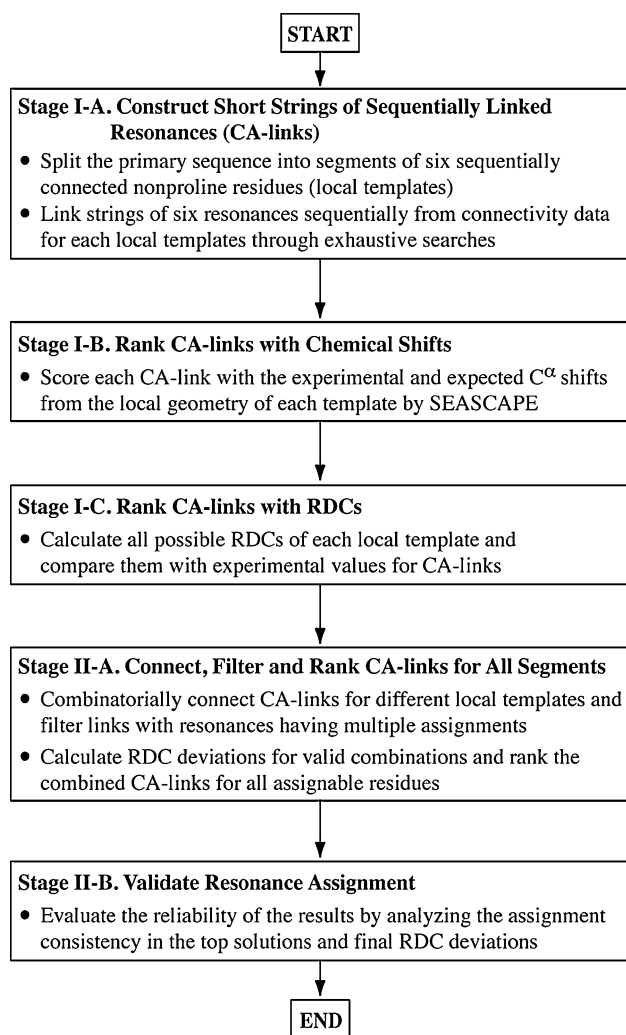


Fig. 1 Flowchart for RDC derived protein backbone resonance assignment using fragment assembly

using $^{13}C_\alpha(i)$ and $^{13}C_\alpha(i-1)$ connectivity data (referred to below as CA-links). The likelihood of their assignments to a segment of residues in a protein sequence is then determined based on their observed and expected $^{13}C_\alpha$ shifts and RDCs. In the second stage, CA-links for different segments of the protein are combinatorially connected to generate CA-links for all assignable residues for final resonance assignment. These long CA-links are ranked according to the agreement between their experimental RDCs and the predicted values from a tertiary structure. Since only top ranking CA-links from stage I are included in the construction of the CA-links for all assignable residues, the global search of stage II becomes computationally feasible. Details of the process are described below using the assignment of ubiquitin as an example.

Stage I-A: Prepare local structural templates by splitting the primary sequence into segments of six sequentially connected nonproline residues, and construct strings of

sequentially linked resonances (CA-links) for each template from connectivity data by conducting exhaustive searches. As with any combinatorial search, the number of CA-links grows rapidly as segment length of local templates increases; thus we have chosen to use a segment length of six residues since the number of viable CA-links, typically, remains reasonable and allows an efficient computational search. Moreover, the SEquential Assignment by the Structure and Chemical shift Assisted Probability Estimation (SEASCAPE) program (Morris et al. 2004), employed for chemical shift scoring in stage I-B, has been explicitly tested with a segment length of six. In addition, when no measurements are missing, the number of RDCs in a CA-link is twice the number of variables used in the RDC ranking of stage I-C. For ubiquitin, the sequence was divided into 10 segments (Table 2). A tolerance of 0.25 ppm of the $^{13}C_\alpha$ chemical shift was used to construct CA-links for each segment, but the quality of these matches did not influence the selection of a viable assignment. Resonances from 6 Gly and 6 residues succeeding Gly were identified from the $^{13}C_\alpha(i)$ and $^{13}C_\alpha(i-1)$ chemical shifts. This data was taken into account in constructing the CA-links. The total number of CA-links for each segment is listed in Table 2. Examples of the constructed CA-links for segments II and IV are displayed in Fig. 2a. In general, there are many potential CA-links for each segment.

Stage I-B: Rank CA-links by the agreement of their experimental $^{13}C_\alpha$ chemical shifts to the expected values derived from the local geometry of each short template. In recent years, significant progress has been made toward constructing protein structures directly from assigned chemical shifts (Cavalli et al. 2007; Shen et al. 2008, 2010; Jensen et al. 2010). In the present application, the chemical shift data are employed to assess the likelihood of the assignment of a CA-link to a particular segment. $^{13}C_\alpha$ shifts are exquisitely sensitive to amino acid type and local backbone structure, but, except for Gly, identification of the amino acid type by its $^{13}C_\alpha$ chemical shift alone is difficult. This obstacle can be overcome if five or more sequentially connected resonances in a fragment of known local backbone geometry are available. Prestegard et al. previously developed the SEASCAPE program to map a string of connected resonances to the most probable primary sequence location for a given protein (Morris et al. 2004). In this approach, the probability of each amino acid type for a given $^{13}C_\alpha$ shift and ϕ , φ torsion angles was constructed from a database of $^{13}C_\alpha$ chemical shifts for proteins with high-resolution structures. SEASCAPE combines the probability of each residue in a short fragment and produces a score that reflects the likelihood of the assignment of this fragment to a particular amino acid sequence. In the present work, a minor modification of the

Table 2 Chemical shift and RDC ranking results of the correct CA-link for each segment of ubiquitin

Segment index	Residues	Total # of CA-links	Chemical shift ranking (IARR)	RDC ranking (RDC deviation, Hz)		
				With 1F9J ^a	With 1ARR ^a	With 1D3Z ^b
1	2–7	30,837	150	849 (7.3)	214 (3.8)	46 (1.7)
2	8–13	1,985	8	7 (2.2)	2 (1.6)	7 (1.4)
3	20–25	30,837	467	75 (4.5)	131 (5.2)	6 (1.5)
4	26–31	30,837	1	15 (3.2)	5 (2.7)	2 (1.2)
5	39–44	30,837	567	727 (4.8)	921 (5.2)	35 (1.4)
6	45–50	1,985	20	14 (5.2)	4 (2.1)	2 (1.1)
7	51–56	1,985	502	– ^c	–	–
8	57–62	30,837	522	905 (6.1)	452 (4.4)	13 (1.2)
9	63–68	30,837	340	570 (5.7)	344 (4.5)	8 (1.1)
10	69–74	30,837	74	–	–	–

^a The estimated magnitudes and rhombicities of the order tensors, 0.00067 and 0.25, and –0.0011 and 0.84 for the first and second alignment media, respectively, were used for the RDC calculations

^b The magnitudes and rhombicities of the order tensors used for RDC calculations were 0.00075 and 0.2, and 0.0012 and 0.74 for the first and second media, respectively, derived from the order matrix analysis of the RDCs for residues consistently assigned in the top 10 solutions of the first run

^c Due to missing measurements, some initial CA-links had fewer than four experimental RDCs. These links were not included in the RDC ranking of stage I–C, but were connected with CA-links for other segments in the calculations of stage II–A

SEASCAPE program allowed the ranking of CA-links according to the agreement of their observed $^{13}\text{C}_\alpha$ shifts to the expected values from the local template. Table 2 lists the SEASCAPE ranking results of the correct CA-link for assignment using a 2.3 Å X-ray structure (1AAR). For all segments, except one, the correct assignments are in the top 2% of the solutions. In this study, CA-links with SEASCAPE scores not in top 20% were eliminated from the assignment pools to expedite subsequent calculations.

Stage I–C: Rank the remaining CA-links by the agreement of their experimental ^{15}N - ^1H RDCs to the expected values from the local geometry of each short template. Following chemical shift ranking by SEASCAPE, ^{15}N - ^1H RDCs for each local template are calculated in all possible orientations. The predicted RDCs are then compared to the experimental value of each residue in a CA-link and the RDC absolute difference (*Absolute_Error*) is calculated and normalized according to (1),

$$Absolute_Error^i = \sum_{j=1}^N \frac{|RDC(j)_{\text{exp}}^i - RDC(j)_{\text{cal}}^i(\alpha, \beta, \gamma)|}{S_{zz}^i * 1000 * N * \sigma_{\text{exp}}^i} \quad (1)$$

where S_{zz}^i is the magnitude of the order tensor for alignment medium i , N is the number of residues in a particular fragment, $RDC(j)_{\text{exp}}^i$ and $RDC(j)_{\text{cal}}^i(\alpha, \beta, \gamma)$ are experimental and calculated RDCs, respectively for residue j , the Euler angle (α, β, γ) defines the relative orientation of the structured fragment to the order tensor, and σ_{exp}^i is the



error of the experimental RDCs. The minimum RDC difference of a CA-link for each alignment medium from a grid search of (α, β, γ) , $Absolute_Error_{\text{min}}^i$, is selected and summed according to (2),

$$Absolute_Error_{\text{total}} = \sum_{i=1}^M Absolute_Error_{\text{min}}^i \quad (2)$$

where M is the number of alignment media for RDC measurements. These summed absolute differences are then used to rank the CA-links. Here, we chose the absolute difference to avoid overemphasizing large deviations.

For RDC calculations, the magnitude and rhombicity of the order tensor are required. Several approaches have been developed to estimate these values prior to assigning resonances (Clare et al. 1998; Warren and Moore 2001; Zweckstetter 2003, 2008; Mukhopadhyay et al. 2009). For ubiquitin, estimates for the magnitudes and rhombicities from the extremes of the experimental RDCs were 0.00067 and 0.25, and –0.0011 and 0.82 for the first and second alignment media, respectively; subsequently these values were refined as described in stage II–A. RDC rankings of the correct CA-link for each local template are listed in Table 2. Examples of the results are provided in Fig. 2b. Initially, the correct CA-links for assignments are not near the top when two X-ray structures at 2.3 (1ARR) and 2.7 (1F9J) Å resolution were used as templates. For all segments except one, the correct CA-links for assignment are ranked lower with the 1ARR than with the 1F9J templates. This was expected since errors in templates and the

(a)

Segment II: LTGKTI (8-13)	Segment IV: VKAKIQ (26-31)	
		
1 2 35 36 38 31 0	1 2 3 4 5 6 0	
1 2 35 36 38 39 1	1 2 3 4 5 8 1	
1 2 35 36 64 34 2	1 2 3 4 5 25 2	
.	.	
6 2 35 36 64 65 17	26 27 28 29 23 67 3696	
8 9 10 11 12 13 18	26 27 28 29 30 31 3697	
8 9 10 11 12 24 19	26 27 28 29 30 39 3698	
.	.	
.	.	

(b)

20 9 10 48 7 8 1.5	12 24 6 49 26 27 2.4
8 9 10 11 12 13 1.6	26 27 28 29 23 24 2.5
32 9 10 54 7 8 1.9	38 39 73 17 57 58 2.7
.	.
.	.
32 9 10 54 55 30 3.0	26 27 28 29 30 31 2.7
58 9 10 54 55 30 3.1	26 27 28 29 23 13 2.8
32 9 10 54 55 23 3.1	12 24 6 55 30 39 2.9
.	.
.	.

Fig. 2 Output for segments II (*left*) and IV (*right*) from assignment procedures described in **(a)** stage I-A and **(b)** stage I-C for ubiquitin using a 2.3 Å X-ray structure (1AAR). The correct CA-links for assignments are colored in blue. The first six columns in **(a)** and **(b)** are resonance numbers. The *last column* in **(a)** is the index for CA-links, and in **(b)** is the RDC deviations sorted in ascending order. The local structure of segments II and IV are shown as ribbons. For simplicity, resonances are numbered according to the ubiquitin residue number

experimental RDCs, as well as uncertainties in the estimates of the magnitudes and rhombicities of order tensors, will affect the RDC ranking. For comparison, when the accurate magnitude and rhombicity of the alignment tensor, determined by order matrix analysis in stage II-A, and a RDC refined structural template (1D3Z) were used, the correct CA-link for each of the segments has smaller RDC deviations and are ranked closer to the top (last column of Table 2). These results indicate that RDC rankings for short segments are sensitive to the quality of input data and emphasize the necessity for subsequent global RDC ranking for CA-links that target all assignable residues in stage II.

Stage II-A: Connect, filter and rank CA-links for all assignable residues. To expedite this step, we adopted an approach that exploits multiple cycles of linking, ranking and filtering (Bryson et al. 2008). At each cycle, two CA-links with RDC deviations below a given a threshold (defined by the quality of structural models and experimental

measurements) are combinatorially connected, and those with resonances having multiple assignments are eliminated from the assignment pool. The remaining links are ranked by RDCs. During the final stage, CA-links for all assignable residues are constructed and ranked with the experimental and expected RDCs from the tertiary structure. For ubiquitin, a set of consistent assignments from the first run were utilized to define the magnitudes and rhombicities of alignment tensors with the order matrix calculations of REDCAT (Valafar and Prestegard 2004). The refined magnitude and rhombicity values of 0.00075 and 0.2, and 0.0012 and 0.74, respectively, for the first and second media were used to repeat the searches described in stages I-C and II-A. The final assignment results are shown in Fig. 3. The correct CA-link for resonance assignment, colored blue, is one of two solutions that have the lowest RDC deviations in all three cases. Another solution has the assignments for residue 75 and 53 exchanged, since experimental RDCs for both residues are not available. Residues inconsistent assigned in the top ten solutions are colored orange. Analysis of the top five and ten solutions show consistent assignments for 66 and 64 out of 72 residues, respectively, in three cases and these consistent assignments are all correct, although the final RDC deviations are dependent on the quality of the structural templates. Moreover, there are no consistent incorrect assignments (false positives), illustrating the reduced sensitivity of the current approach to the overall quality of the structural models used for assignment.

Stage II-B: Validate resonance assignments. The final RDC deviations and the consistency of assignments in the top solutions provide an indication of assignment reliability. As shown in Fig. 3 for ubiquitin, the summed RDC deviations of the top solution from two alignment media are 5.3, 4.8 and 2.0 Hz using the structural templates of 1F9J, 1ARR and 1D3Z, respectively. These are less than 25% of the sum of the normalized maximum RDC, $D_{NH}/1000$ (D_{NH} is the N-H dipolar coupling constant, ~ 23 kHz). In addition, 89% of resonances show consistent assignments in the top ten solutions. This implies that the set of internally consistent assignments are likely correct, while the subset of inconsistent assignments need to be verified with additional experiments or thorough visual inspection of the spectra. For instance, when NOESY data are available, the sequential H^N-H^N NOEs in the α -helix and the across strand H^N-H^N NOEs in the β -sheets can be used for validation.

Resonance assignment of crystallins and MBP

To further explore the utility of this approach, we examined assignments of two larger proteins, bovine crystallin (174 residues) and MBP (370 residues), using data of varying quality. For bovine crystallin, we chose to use

Fig. 3 Assignment results for ubiquitin using the present approach with **a** 2.7 Å X-ray structure (1F9J), **b** 2.3 Å X-ray structure (1ARR), and **c** RDC refined NMR structure (1D3Z) as templates. The *last column* shows final RDC deviations. *Other columns* are resonance numbers. For simplicity, resonances are numbered with the ubiquitin residue number. The correct CA-link for assignments appears in *blue*. *Orange* and *black* numbers are inconsistent and consistent assignments in the top ten solutions, respectively

	I.	II.	III.	IV.	V.	VI.	VII.	VIII.	IX.	X.	
(a)	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 59 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 59 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 59 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 59 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 59 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 59 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 39 40 41 42 43 44 45 46
	47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 5.3	47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 5.3	47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 5.4	47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 5.4	47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 67 63 64 65 66 62 68 69 70 71 72 73 74 75 76 5.8	47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 67 63 64 65 66 62 68 69 70 71 72 73 74 75 76 5.8	47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 67 63 64 65 66 62 68 69 70 71 72 73 74 75 76 5.8	47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 67 63 64 65 66 62 68 69 70 71 72 73 74 75 76 5.8	47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 67 63 64 65 66 62 68 69 70 71 72 73 74 75 76 5.9	47 48 49 50 72 52 75 54 55 56 57 58 59 60 61 67 63 64 65 66 62 68 69 70 71 51 73 74 53 76 5.9	
(b)	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 59 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 59 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 59 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 59 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 59 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 39 40 41 42 43 44 45 46
	47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 4.8	47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 4.8	47 48 49 50 72 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 51 73 74 75 76 5.0	47 48 49 50 72 52 75 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 51 73 74 53 76 5.1	47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 5.1	47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 53 76 5.1	47 48 49 50 72 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 51 73 74 75 76 5.2	47 48 49 50 72 52 75 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 51 73 74 53 76 5.2	47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 67 63 64 65 66 62 68 69 70 71 72 73 74 75 76 5.3	47 48 49 50 51 52 75 54 55 56 57 58 59 60 61 67 63 64 65 66 62 68 69 70 71 72 73 74 53 76 5.3	
(c)	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 59 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 59 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 59 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 59 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 59 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 39 40 41 42 43 44 45 46	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32 59 34 35 36 39 40 41 42 43 44 45 46
	47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 2.0	47 48 49 50 51 52 75 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 53 76 2.0	47 48 49 50 72 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 51 73 74 75 76 2.2	47 48 49 50 72 52 75 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 51 73 74 53 76 2.2	47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 2.3	47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 67 63 64 65 66 62 68 69 70 71 72 73 74 75 76 2.4	47 48 49 50 72 52 53 54 55 56 57 58 59 60 61 67 63 64 65 66 62 68 69 70 71 72 73 74 75 76 2.4	47 48 49 50 72 52 75 54 55 56 57 58 59 60 61 67 63 64 65 66 62 68 69 70 71 51 73 74 53 76 2.4	47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 51 73 74 75 76 2.5	47 48 49 50 72 52 75 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 51 73 74 53 76 2.5	

experimental RDCs measured for murine crystallin and a crystal structure (1AMM) of bovine crystallin as the template to mimic some practical complications of input data. RDCs of murine crystallin were transferred to the corresponding residues of bovine crystallin with the exception of eight residues (82 to 89) in the loop connecting two domains after structural alignment. RDCs for these residues were considered to be missing. In total, experimental

RDCs were available for ~80% of assignable residues. Furthermore, the backbone RMSD between the structures of bovine and murine crystallins, excluding eight loop residues, is ~1.3 Å, effectively mimicking “structural noise” in the assignment template with respect to the measured RDCs. This was clearly reflected in the Q-factor and normalized RMSD of 0.31 and 0.34, respectively, from the order matrix analysis of the transferred RDCs and the

structural template of 1AMM (Table 3 and Figure S2). Since experimental $^{13}\text{C}_\alpha$ chemical shifts were not available for bovine crystallin, values calculated by SHIFTX (Neal et al. 2003) were used to construct CA-links with a tolerance of 0.15 ppm. In this example, we assumed that resonances from Ala, Phe, Lys and Val were known so that each segment of six residues contained at least one residue from a known amino acid type to expedite the calculations during stage I. Practically, these data can be obtained from $^{13}\text{C}_\beta$ chemical shifts or from amino acid specific labeling samples. Table 3 lists the final assignments. Analysis of these results indicates that 96% of the residues are correctly assigned in the top solution, and ~93% of the residues are correctly and consistently assigned in all top 10 solutions. Furthermore, there are no false positive assignments. For comparison, assignments for murine crystallin are also listed in Table 3. While the observed results for murine crystallin is expected (since its assignment template was refined with RDCs), the comparable success for resonance assignment of bovine crystallin is quite remarkable considering that a structural homology was used as the template and ~20% of the residues were missing experimental RDCs.

To quantitatively investigate the dependence of the current strategy on the quality of input data, we exploited the resonance assignment of ligand-free MBP using synthetic RDCs perturbed by different levels of random errors,

from 0.2 to $0.8 \cdot D_a^{HN}$. The MBP sequence was first split into 47 segments of six sequentially linked nonproline residues. CA-links were constructed for each of these segments using the experimental $^{13}\text{C}_\alpha(i)$ and $^{13}\text{C}_\alpha(i-1)$ shifts with a tolerance of 0.15 ppm and known amino acid resonances from 44 Ala, 21 Lys, 15 Phe and 20 Val residues. The ranking results for every sixth segment are listed in Table 4. When errors are small (0 and $0.2 \cdot D_a^{HN}$), the correct individual CA-links for assignments are in the top solutions, however, as the error increases (0.6 and $0.8 \cdot D_a^{HN}$), the discrimination for the correct solution decreases. This is expected, since increasing perturbations by random errors will reduce the agreement of simulated RDCs with the starting structural model. This is quantitatively reflected in the increased Q-factors and normalized RMSDs from 0.13 and 0.12 to 0.52 and 0.48, respectively, when the error increases from 0.2 to $0.8 \cdot D_a^{HN}$ (Table 5). These results further highlight the difficulty of obtaining resonance assignments by comparing calculated and experimental RDCs for short segments, and emphasize the necessity of a search for the best agreement including all assignable residues. The final assignments, derived from the calculations of stage II using the global optimization criteria, are listed in Table 5. Here, in all cases, more than 96% of residues are correctly and consistently assigned in the top ten solutions, and there are no false positives. The successful assignment with RDCs perturbed by random

Table 3 Assignment results for crystallins using experimental ^{15}N - ^1H RDCs measured for murine crystallin

Structural templates	Q-factor ^a	Normalized RMSD ^a	% of correct assignments in the top solution	% of correct and consistent assignments in the top 10 solutions
Bovine crystallin (1AMM)	0.31	0.34	96	93
Murine crystallin (2A5M)	0.1	0.1	100	96

^a RDCs from gel alignment medium were used for the calculation

Table 4 Initial RDC rankings of the correct CA-links for every sixth segment of MBP using synthetic RDCs perturbed by different levels of random errors

Segment index	Residues	# of CA-links	Rankings with RDCs perturbed by random errors, $\pm m \cdot D_a^{HN}$				
			m = 0	m = 0.2	m = 0.4	m = 0.6	m = 0.8
1	2–7	12,460	1	3	143	265	215
7	41–46	11,436	1	1	197	582	123
13	92–97	935	1	2	6	16	6
19	140–145	78	1	1	4	1	37
25	184–189	166	1	1	2	3	3
31	220–225	4,370	1	1	11	95	271
37	284–289	51,575	1	2	200	617	364
43	335–340	6,122	1	1	23	24	705

RDCs were simulated: $S_{zz}(1) = 0.001$, $\text{Eta}(1) = 0.65$ and $S_{zz}(2) = -0.0006$, $\text{Eta}(2) = 0.3$

Table 5 Final assignment results for MBP using synthetic RDCs perturbed by different levels of random errors

RDC errors ($\pm m * D_a^{HN}$, Hz)	Q-factor	Normalized RMSD	% of correct assignments in the top solution	% of correct and consistent assignments in the top 10 solutions
$m = 0.2$	0.13	0.12	100	96
$m = 0.4$	0.26	0.24	99	97
$m = 0.6$	0.39	0.36	99	97
$m = 0.8$	0.52	0.48	99	97

errors as large as $0.8 * D_a^{HN}$ demonstrates the robustness of the present approach.

Resonance assignment of PDZ3

As a more challenging test for our approach, we exploited theoretical structures as templates for resonance assignment of the third PDZ3 domain of human ZO-1 because predicted models are typically of low quality, with some regions more accurately represented than others. The human PDZ3 domain shares 33% sequence identity and 48% sequence similarity with the PDZ3 domain of rat PSD-95. We generated seven models using three different modeling programs: models 1 to 5 with I-TASSER, model 6 with MODELLER 9 and model 7 with SWISS-MODEL. Figure 4 overlays these models with each residue colored according to its Q_{res} value. Q_{res} is a measurement of structural similarity of each residue in a set of aligned

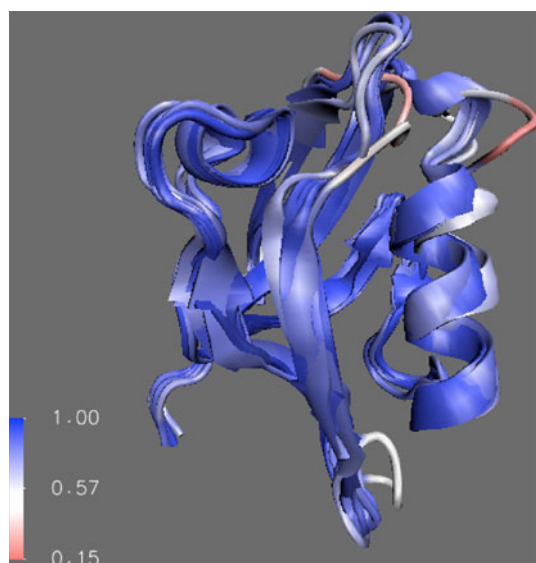


Fig. 4 Overlay of seven predicted PDZ3 structures from residues 16 to 100. Residues from 1 to 15 and 101 to 113 are not shown for clarity. Each residue is colored according to its Q_{res} value. Q_{res} is a measurement of the structural similarity of each residue in a set of overlaid structures. In these seven models, more variations are seen in loops than in structured regions and the packing of an α -helix in one model (model 6) appears to be different from others

structures in the program VMD (Roberts et al. 2006). From consensus in these seven models six structured segments were identified, including five β -sheets and one α -helix (each segment had at least five residues). The average pairwise backbone RMSD for these six segments is ~ 0.6 Å. This is significantly smaller than the RMSD value of ~ 1.1 Å for residues 16 to 104. Furthermore, while residues of five β -sheets in these three models display overall structural similarity, the packing of an α -helix in one model (model 6) appears to be distinct from others and residues from loop regions show low Q_{res} values. This is consistent with the assumption that current modeling methods are more accurate at predicting structured regions than loops, and at predicting local structures than tertiary packings (Baker and Sali 2001). Therefore, we chose to partition these six structured segments into two independent rigid units: five β -sheets in one unit (unit I) and the helix in another (unit II). In other words, the predicted tertiary structure information among β -sheets was used for resonance assignment, while the relative orientation of the α -helix to β -sheets was not used due to variations observed in seven predicted models.

Fifteen and thirteen residues at the N- and C-termini, respectively, of PDZ3 were predicted to be unstructured. We performed CLEANEX-PM-FHSQC experiments (Hwang et al. 1998) with mixing times of 10 and 20 ms and identified amide protons having very fast exchanges with water, including residues 5 to 16, 27, 36 to 38, 47, 72, 87, 99, 100, 103, 104 and 107 to 113. These resonances were excluded from the assignment process for the structured regions, leaving a total of 73 peaks. Of these, 39 resonances were from the structured regions and they were targeted for assignment using the current approach. CA-links were constructed using experimental $^{13}C_{\alpha}$ shifts with a tolerance of 0.25 ppm and known resonances from Val. Table 6 lists the SEASCAPE and RDC ranking results of the correct CA-link for each of the structured fragments in PDZ3 with models 1 and 6 (stage I). For most segments, the correct CA-links for assignment are in the top solutions. Subsequently, segments I to IV were combinatorially linked, filtered and RDC ranked with the predicted structural templates of unit I (stage II-A). A final step was added to connect RDC ranked CA-links for structural units I and II

Table 6 Initial RDC and SEASCAPE rankings of correct CA-links for six structured segments of PDZ3 using predicted models 1 and 6

Fragments	I MKLVK (17–21)	II LRLAG (30–34)	III IFVAGV (40–45)	IV DQILRV (60–65)	V REEAVLFLLD (75–84)	VI VTILAQK (91–97)
Total CA-links	1,247	92	7	562	458,543	8,207
Local structural templates from Model 1 predicted by I-TASSER						
SEASCAPE ranking	29	1	1	12	369	1
RDC ranking	7	1	2	10	3	5
RDC error (Hz)	2.7	1.7	6.6	3.7	2.5	3.8
Local structural templates from Model 6 predicted by MODELLER						
SEASCAPE ranking	29	1	1	12	499	1
RDC ranking	19	1	1	1	1	1
RDC error (Hz)	3.2	2.5	2.2	1.0	3.0	2.2

The estimated magnitudes and rhombicities of the order tensors, 0.00065 and 0.45, and 0.00031 and 0.36 for the phage and PEG alignment media, respectively, were used for RDC calculations

Table 7 Final assignment results for PDZ3 residues of structured regions using predicted structures

Structural models	% (number) of correct assignments in the top solution	% (number) of correct and consistent assignments in the top 10 solutions
Model 1	100 (39)	85 (33)
Model 2	95 (37)	85 (33)
Model 3	95 (37)	85 (33)
Model 4	95 (37)	85 (33)
Model 5	90 (35)	85 (33)
Model 6	92 (36)	82 (32)
Model 7	92 (34)	76 (28)

Thirty-nine residues in structured regions are targeted for assignment for models 1 to 6. For model 7, thirty-seven residues are targeted since segment V of an α -helix is two residues shorter than its counterparts in other models

and to eliminate CA-links with resonances that had multiple assignments. The surviving CA-links were then ranked with the combined RDC deviations for the final assignment (Table 7). More than 90% of the residues from structured regions are correctly assigned in the top solutions for all seven models. More than 80% of residues are consistently and correctly assigned in the top ten solutions with no false positives. Even though the percent of correct assignments is not as high as in previous examples, our approach performed reasonably well considering the poor quality of these models and the demanding task of assigning 39 out of total 73 resonances. Subsequently, by excluding the “assigned” resonances from structured regions we were able to assign $\sim 60\%$ of the remaining residues of PDZ3 using conventional approach. For comparison, less than 10% of resonances were assigned using the same $^{13}\text{C}_\alpha$ chemical shift data.

Discussion

In this study, we have developed a structure-assisted, RDC-based approach for protein backbone resonance assignment. Our approach derives assignments based on the best global agreement between experimental and calculated RDCs from a structural template including all assignable residues. This required the implementation of a two-stage search algorithm and the inclusion of limited connectivity data to overcome the computationally expensive task. We first described the application of this approach to the backbone resonance assignment of ubiquitin and crystallin using experimental structures and RDCs. While RDC rankings of short segments were dependent on the quality of structural templates, the final assignments obtained from the ranking of CA-links for all assignable residues were quite insensitive to the quality of employed models. The reliability of the current approach was further illustrated with the resonance assignment of MBP using synthetic RDCs perturbed by random errors up to $0.8 * D_a^{HN}$. In all cases, more than 95% of the residues were correctly and consistently assigned in the top ten solutions. Finally, we evaluated our approach using theoretical structural templates. The seven predicted PDZ3 structures used in the current work were not high quality, as evidenced from their order matrix analyses (Table 1 and Figure S1). Even for the structured region, the predicted models represented, at best, medium resolution structures based on their Q-factors and normalized RMSDs (ranging from 0.32 to 0.5, and 0.24 to 0.37). Nevertheless, we were able to reliably assign $\sim 80\%$ of residues in the structured regions. For comparison, only a few residues could be assigned using the $^{13}\text{C}_\alpha$ chemical shift data in the examples described above. Therefore, the present approach provides an alternate route to overcoming the chemical shift degeneracy for backbone resonance assignment in the presence of a structural model.

RDCs, in principal, are ideal constraints for structure-based resonance assignment. They can be directly predicted from a structural model and measured with an accuracy of a few tenths of Hz, even for systems of considerable sizes (Yang et al. 1999; Hu et al. 2009; Arbogast et al. 2010; Bhattacharya et al. 2010; Fitzkee and Bax 2010; Mantylahti et al. 2010), and they are routinely collected during NMR studies. However, due to the sensitivity of RDCs to the exact orientation of their corresponding internuclear vector, deviations in the structure or dynamics of the template and errors in the experimental measurement had limited their application for resonance assignment. The current strategy largely circumvents these problems, since assignments are optimized simultaneously for all assignable residues, as opposed to one or several residues at a time. In particular, the success of the current assignment approach can be attributed to several key features. First, in the ranking algorithm of stage I, a string of sequentially linked RDCs are evaluated for their overall agreement to a local structural model, reducing susceptibility to errors in individual RDC measurements and/or deviations of an internuclear bond vector in its orientation (Stratmann et al. 2009). RDC and chemical shift rankings in this stage eliminate most of the incorrect CA-links, making the construction of complete CA-links in stage II computationally feasible. Second, multiple top CA-links for each segment in stage I are kept and combinatorially connected with CA-links from other segments to generate final CA-links for all residues targeted for assignment. The assignment is determined by the “global minimum” of RDC deviations of the final CA-links. This diminishes the potential for incorrect assignments due to “local minimums”. Although this process does not guarantee that the correct solution will be found at the top of the final list, assignments for most resonances are correct in the best solutions of our examples. Third, using ^{15}N - ^1H RDCs from two alignment media for RDC ranking addresses some limitations of RDC data, such as its inherent insensitivity to 180° rotations and varying sensitivity to “structural noise” (since the RDC error for a given structural deviation is a function of the relative orientation of the internuclear vector to the principal alignment frame) (Al-Hashimi et al. 2000; Zweckstetter and Bax 2002; Langmead and Donald 2004). For future application, when the relative orientation between two order tensors is accurately defined, it can be included to reduce the number of variables for RDC ranking (Miao et al. 2008).

The conventional assignment strategy relies on the intraresidue and sequentially connectivity data. While it has proven to be a reliable approach for resonance assignments, resonances of $^{13}\text{C}_\alpha$, $^{13}\text{C}_\beta$ or ^{13}C have to be resolved. In the present study, we exploit the synergy between various types of experimental data as well as prior structural

knowledge for assignment in the presence of chemical shift ambiguities. Sequential connectivity, while potentially ambiguous and incomplete, is very effective at reducing the assignment “space”. For instance, for a protein having 100 non-proline residues, a small segment of six residues could have close to 10^{12} possible assignments without sequential connectivity information ($100 \times 99 \times 98 \times 97 \times 96 \times 95$). However, if each resonance has 10 possible sequential connections instead, the number of viable combinations is dramatically reduced to 10^6 . Thus, a computational search for all possible assignments of a segment of six residues becomes feasible to exploit the primary sequence as an additional constraint to resolve chemical shift ambiguities during the assignment process (Andrec et al. 2001; Coggins and Zhou 2003; Xu et al. 2006b; Crippen et al. 2010). Furthermore, since the resolution of experimental structures varies and the quality of predicted structures is unknown in advance and generally poor, inclusion of the connectivity data into the present assignment strategy offers a reliable route to overcome the adverse effects of “structural noise” in the templates (Stratmann et al. 2009). In practice, $^{13}\text{C}_\alpha(i)$ and $^{13}\text{C}_\alpha(i-1)$ connectivities are collected with most sensitive triple resonance experiments (Nietlispach et al. 2002). For example, the TROSY-HNCA experiment yielded complete resonances for the membrane protein Smr in bicelles (~ 150 kDa), and provided intraresidue and sequential correlation information for a nondeuterated 65 kDa protein (Xu et al. 2006a; Poget and Girvin 2007). On the other hand, RDCs depend on the orientations of their corresponding internuclear vectors instead of the chemical environment. In a sense, long-range, orientational dependent RDCs are “orthogonal” to chemical shifts, and thereby provide an effective route to overcome resonance degeneracy.

The computational algorithm used in stage I–C shares some similarity with the one used by Bax et al. for constructing protein backbone structures from RDCs (Delaglio et al. 2000), but the objectives are entirely different. In previous applications, RDCs from a string of sequentially assigned resonances were screened against the predicted values from a library of structures for peptide fragments, and the best matches were selected for structure determination. In the present application, a “correct” geometry model for a fragment is assumed, and RDCs predicted from this model are used to score and rank a pool of strings of sequentially linked RDCs, or connected resonances for assignment. Compared to the conceptually related RDC-enhanced MARS approach to assignment (Jung and Zweckstetter 2004), the present method is truly “RDC-based”. Chemical shift data contribute to creating and filtering viable CA-links, but final assignments come from an optimum global agreement between the experimental

and calculated RDCs. Adverse effects due to local “structural noise” and random errors in experimental measurements are minimized. In contrast, RDC-enhanced MARS only compares measured and expected RDCs for short fragments, and it is potentially more susceptible to these detrimental effects. Consequently, MARS has to reduce the RDC contribution to the scoring function and final assignments rely mainly on the agreement between observed and calculated chemical shifts.

This study demonstrates how the use of RDCs and prior structural knowledge can facilitate resonance assignment when chemical shift data are insufficient. One of the potential applications of the present method is the study of large proteins with known structures and macromolecular complexes with the structure(s) of individual component(s) in the complex available. In recent years, long-range structural constraints such as RDCs and PREs in combination with data from small angle X-ray scattering experiments have greatly expanded these applications (Pintacuda et al. 2006; Takeuchi and Wagner 2006; Sprangers et al. 2007; Grishaev et al. 2008a, b; Bertelsen et al. 2009; Clore and Iwaha 2009), but backbone resonance assignments (or at least part of them) are required. For high molecular weight systems, chemical shifts and RDCs are likely to be more degenerate and less accurately measured, which increases the number of CA-links for chemical shift and RDC rankings. However, we anticipate that these limitations can be overcome by using computer clusters and by the addition of other data from traditional and recently developed experiments (Frueh et al. 2009; Takeuchi et al. 2010). With perdeuteration some C_β chemical shifts remain accessible, especially for residues from flexible regions, providing amino acid type information and additional sequential connectivity information to reduce the number of CA-links. In addition, the C_β shifts likely will be useful for the assignment of flexible residues, since RDCs of these residues are potentially contaminated by molecular motions. Similarly, ^{15}N -edited NOESY experiments are effective and commonly used to assign resonances for large systems. The application of sequential $\text{H}^{\text{N}}\text{-H}^{\text{N}}$ NOEs in α -helices and across strand $\text{H}^{\text{N}}\text{-H}^{\text{N}}$ NOEs in β -sheets to reduce the number of viable CA-links is straightforward. These are short-range NOEs and usually experimentally observable. These additional data can be easily incorporated into the present strategy. Finally, applications of this new assignment approach will benefit directly from the increasing number of experimental structures and maturation in computational modeling. Over the last decade, significant progress has been made in high throughput protein structure determination, especially by X-ray crystallography. Currently, the PDB contains more than 69,000 structures. With the increasing number of structural templates and coverage of the protein structural space by

experimental data, there are now representative structures in the PDB for many proteins and/or domains. For proteins that have >35% sequence identity to a template, models obtained from the “best” prediction programs often have a $\text{RMSD} < 2 \text{ \AA}$ in the core region when compared with their experimental counterparts (Schueler-Furman et al. 2005; Bujnicki 2006; Zhang 2009). Leverage of predicted structures for resonance assignment is particularly appealing and warrants further investigations.

Acknowledgments We are grateful for financial supports from the National Institutes of Health (5R01GM081793-03), the American Diabetes Association (7-07-RA-34) and the Penn State University College of Medicine. We would like to thank Dr. James H. Prestegard at the University of Georgia, Dr. Homa Valafar at the University of South Carolina and Dr. Maria C. Bewley at Penn State University for helpful discussions. The work was performed on a 600 MHz NMR spectrometer at the NMR Core Facility at the Penn State College of Medicine, which was purchased with funds from NIH 1S10RRO21172 and the Tobacco Settlement Funds awarded by the Pennsylvania Department of Health. Several figures were prepared with the program VMD. VMD was developed with NIH support by the Theoretical and Computational Biophysics group at the Beckman Institute, University of Illinois at Urbana-Champaign.

References

- Al-Hashimi HM, Valafar H, Terrell M, Zartler ER, Eidsness MK, Prestegard JH (2000) Variation of molecular alignment as a means of resolving orientational ambiguities in protein structures from dipolar couplings. *J Magn Reson* 143:402–406
- Andrec M, Du PC, Levy RM (2001) Protein backbone structure determination using only residual dipolar couplings from one ordering medium. *J Biomol NMR* 21:335–347
- Apaydin MS, Conitzer V, Donald BR (2008) Structure-based protein NMR assignments using native structural ensembles. *J Biomol NMR* 40:263–276
- Arbogast L, Majumdar A, Tolman JR (2010) HNC0-based measurement of one-bond amide $^{15}\text{N}\text{-}^1\text{H}$ couplings with optimized precision. *J Biomol NMR* 46:175–189
- Baker D, Sali A (2001) Protein structure prediction and structural genomics. *Science* 294:93–96
- Bartels C, Billeter M, Guntert P, Wuthrich K (1996) Automated sequence-specific NMR assignment of homologous proteins using the program GARANT. *J Biomol NMR* 7:207–213
- Bax A, Grishaev A (2005) Weak alignment NMR: a hawk-eyed view of biomolecular structure. *Curr Opin Struct Biol* 15:563–570
- Bermejo GA, Llinas M (2010) Structure-oriented methods for protein NMR data analysis. *Prog Nucl Magn Reson Spectrosc* 56: 311–328
- Bertelsen EB, Chang L, Gestwicki JE, Zuiderweg ERP (2009) Solution conformation of wild-type *E. coli* Hsp70 (DnaK) chaperone complexed with ADP and substrate. *Proc Natl Acad Sci USA* 106:8471–8476
- Bhattacharya A, Revington M, Zuiderweg ERP (2010) Measurement and interpretation of $^{15}\text{N}\text{-}^1\text{H}$ residual dipolar couplings in larger proteins. *J Magn Reson* 203:11–28
- Bryson M, Tian F, Prestegard JH, Valafar H (2008) REDCRAFT: a tool for simultaneous characterization of protein backbone structure and motion from RDC data. *J Magn Reson* 191:322–334

- Bujnicki JM (2006) Protein-structure prediction by recombination of fragments. *ChemBioChem* 7:19–27
- Cavalli A, Salvatella X, Dobson CM, Vendruscolo M (2007) Protein structure determination from NMR chemical shifts. *Proc Natl Acad Sci USA* 104:9615–9620
- Cavanagh J, Fairbrother WJ, Palmer AG III, Skelton NJ, Rance M (2006) *Protein NMR spectroscopy*, second edition: principles and practice. Elsevier Academic Press, San Diego
- Clore GM, Iwaha J (2009) Theory, practice, and applications of paramagnetic relaxation enhancement for the characterization of transient low-population states of biological macromolecules and their complexes. *Chem Rev* 109:4108–4139
- Clore GM, Gronenborn AM, Bax A (1998) A robust method for determining the magnitude of the fully asymmetric alignment tensor of oriented macromolecules in the absence of structural information. *J Magn Reson* 133:216–221
- Coggins BE, Zhou P (2003) PACES: protein sequential assignment by computer-assisted exhaustive search. *J Biomol NMR* 26:93–111
- Crippen GM, Rousaki A, Revington M, Zhang YB, Zuiderweg ERP (2010) SAGA: rapid automatic mainchain NMR assignment for large proteins. *J Biomol NMR* 46:281–298
- Delaglio F, Kontaxis G, Bax A (2000) Protein structure determination using molecular fragment replacement and NMR dipolar couplings. *J Am Chem Soc* 122:2142–2143
- Donald BR, Martin J (2009) Automated NMR assignment and protein structure determination using sparse dipolar coupling constraints. *Prog Nucl Magn Reson Spectrosc* 55:101–127
- Fitzkee NC, Bax A (2010) Facile measurement of ^1H - ^{15}N residual dipolar couplings in larger perdeuterated proteins. *J Biomol NMR* 48:65–70
- Frueh DP, Arthanari H, Koglin A, Walsh CT, Wagner G (2009) A double TROSY hNCAnH experiment for efficient assignment of large and challenging proteins. *J Am Chem Soc* 131:12880–12881
- Grishaev A, Tugarinov V, Kay LE, Trewheella J, Bax A (2008a) Refined solution structures of the 82-kDa enzyme malate synthase G from joint NMR and synchrotron SAXS restraints. *J Biomol NMR* 40:95–106
- Grishaev A, Ying JF, Canny MD, Pardi A, Bax A (2008b) Solution structure of tRNA^{val} from refinement of homology model against residual dipolar coupling and SAXS data. *J Biomol NMR* 42:99–109
- Hansen MR, Mueller L, Pardi A (1998) Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions. *Nat Struct Biol* 5:1065–1074
- Hu K, Doucleff M, Clore GM (2009) Using multiple quantum coherence to increase the ^{15}N resolution in a three-dimensional TROSY HNCO experiment for accurate PRE and RDC measurements. *J Magn Reson* 200:173–177
- Hus J, Prompers JJ, Bruschweiler R (2002) Assignment strategy for proteins with known structure. *J Magn Reson* 157:119–123
- Hwang TL, van Zijl PCM, Mori S (1998) Accurate quantitation of water-amide proton exchange rates using the phase-modulated CLEAN chemical EXchange (CLEANEX-PM) approach with a Fast-HSQC (FHSQC) detection scheme. *J Biomol NMR* 11:221–226
- Jensen MR, Salmon L, Nodet G, Blackledge M (2010) Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts. *J Am Chem Soc* 132:1270–1271
- Jung YS, Zweckstetter M (2004) Backbone assignment of proteins with known structure using residual dipolar couplings. *J Biomol NMR* 30:25–35
- Jung Y, Sharma M, Zweckstetter M (2004) Simultaneous assignment and structure determination of protein backbones by using NMR dipolar couplings. *Angew Chem Int Edit* 43:3479–3481
- Langmead CJ, Donald B (2004) An expectation/maximization nuclear vector replacement algorithm for automated NMR resonance assignments. *J Biomol NMR* 29:111–138
- Marassi FM, Opella SJ (2000) A solid-state NMR index of helical membrane protein structure and topology. *J Magn Reson* 144:150–155
- Meiler J, Baker D (2003) Rapid protein fold determination using unassigned NMR data. *Proc Natl Acad Sci USA* 100:15404–15409
- Meiler J, Baker D (2005) The fumarate sensor DcuS: progress in rapid protein fold elucidation by combining protein structure prediction methods with NMR spectroscopy. *J Magn Reson* 173:310–316
- Mesleh MF, Opella SJ (2003) Dipolar waves as NMR maps of helices in proteins. *J Magn Reson* 163:288–299
- Miao XJ, Mukhopadhyay R, Valafar H (2008) Estimation of relative order tensors, and reconstruction of vectors in space using unassigned RDC data and its application. *J Magn Reson* 194:202–211
- Morris LC, Valafar H, Prestegard JH (2004) Assignment of protein backbone resonances using connectivity, torsion angles and $^{13}\text{C}_\alpha$ chemical shifts. *J Biomol NMR* 29:1–9
- Mukhopadhyay R, Miao XJ, Shealy P, Valafar H (2009) Efficient and accurate estimation of relative order tensors from lambda-maps. *J Magn Reson* 198:236–247
- Neal S, Nip AM, Zhang HY, Wishart DS (2003) Rapid and accurate calculation of protein ^1H , ^{13}C and ^{15}N chemical shifts. *J Biomol NMR* 26:215–240
- Nietlispach D, Ito Y, Laue ED (2002) A novel approach for the sequential backbone assignment of large proteins: selective intra-HNCA and DQ-HNCA. *J Am Chem Soc* 124:11199–11207
- Ottiger M, Delaglio F, Bax A (1998) Measurement of J and dipolar couplings from simplified two-dimensional NMR spectra. *J Magn Reson* 131:373–378
- Pintacuda G, Park AY, Keniry MA, Dixon NE, Otting G (2006) Lanthanide labeling offers fast NMR approach to 3D structure determinations of protein-protein complex. *J Am Chem Soc* 128:3696–3702
- Poget SF, Girvin ME (2007) Solution NMR of membrane proteins in bilayer mimics: small is beautiful, but sometimes bigger is better. *Biochim Biophys Acta* 1768:3098–3106
- Prestegard JH, Bougault CM, Kishore AI (2004) Residual dipolar couplings in structure determination of biomolecules. *Chem Rev* 104:3519–3540
- Pristovsek P, Ruterjans H, Jerala R (2002) Semiautomatic sequence-specific assignment of proteins based on the tertiary structure—The program stnmr. *J Comput Chem* 23:335–340
- Roberts E, Eargle J, Wright D, Luthey-Schulten Z (2006) MultiSeq: unifying sequence and structure data for evolutionary analysis. *BMC Bioinformatics* 7:382. doi:10.1186/1471-2105-7-382
- Ruckert M, Otting G (2000) Alignment of biological macromolecules in novel nonionic liquid crystalline media for NMR experiments. *J Am Chem Soc* 122:7793–7797
- Sattler M, Griesinger C (1999) Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog Nucl Magn Reson Spectrosc* 34:93–158
- Schueler-Furman O, Wang C, Bradley P, Misura K, Baker D (2005) Progress on modeling of protein structures and interactions. *Science* 310:638–642
- Shen Y, Lange O et al (2008) Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci USA* 105:4685–4690
- Shen Y, Bryan PN, He YN, Orban J, Baker D, Bax A (2010) De novo structure generation using chemical shifts for proteins with high-sequence identity but different folds. *Protein Sci* 19:349–356

- Sprangers R, Velyvis A, Kay LE (2007) Solution NMR of supramolecular complexes: providing new insights into function. *Nat Methods* 4:697–703
- Stratmann D, Heijenoort C, Guittet E (2008) NOEnet—use of NOE networks for NMR resonance assignment of proteins with known 3D structure. *Bioinformatics* 25:474–481
- Stratmann D, Guittet E, van Heijenoort C (2009) Robust structure-based resonance assignment for functional protein studies by NMR. *J Biomol NMR* 46:157–173
- Takeuchi K, Wagner G (2006) NMR studies of protein interactions. *Curr Opin Struct Biol* 16:109–117
- Takeuchi K, Frueh DP, Hyberts SG, Sun ZYJ, Wagner G (2010) High-resolution 3D CANCA NMR experiments for complete mainchain assignments using C^α direct detection. *J Am Chem Soc* 132:2945–2951
- Tian F, Valafar H, Prestegard JH (2001) A dipolar coupling based strategy for simultaneous resonance assignment and structure determination of protein backbones. *J Am Chem Soc* 123:11791–11796
- Valafar H, Prestegard JH (2004) REDCAT: a residual dipolar coupling analysis tool. *J Magn Reson* 167:228–241
- Wang JF, Denny JK et al (2000) Imaging membrane protein helical wheels. *J Magn Reson* 144:162–167
- Warren JJ, Moore PB (2001) A maximum likelihood method for determining D_a and R for sets of dipolar coupling data. *J Magn Reson* 149:271–275
- Wider G, Wüthrich K (1999) NMR spectroscopy of large molecules and multimolecular assemblies in solution. *Curr Opin Struct Biol* 9:594–601
- Xiong F, Pandurangan G, Bailey-Kellogg C (2008) Contact replacement for NMR resonance assignment. *Bioinformatics* 24:I205–I213
- Xu YQ, Zheng Y, Fan JS, Yang DW (2006a) A new strategy for structure determination of large proteins in solution without deuteration. *Nat Methods* 3:931–937
- Xu YZ, Wang XX, Yang J, Vaynberg J, Qin J (2006b) PASA—A program for automated protein NMR backbone signal assignment by pattern-filtering approach. *J Biomol NMR* 34:41–56
- Yang DW, Venters RA, Mueller G, Choy WY, Kay LE (1999) TROSY-based HNCQ pulse sequences for the measurement of $^1\text{HN}-^{15}\text{N}$, $^{15}\text{N}-^{13}\text{CO}$, $^1\text{HN}-^{13}\text{CO}$, $^{13}\text{CO}-^{13}\text{Ca}$ and $^1\text{HN}-^{13}\text{C}^\alpha$ dipolar couplings in ^{15}N , ^{13}C , ^2H -labeled proteins. *J Biomol NMR* 14:333–343
- Zhang Y (2009) Protein structure prediction: when is it useful? *Curr Opin Struct Biol* 19:145–155
- Zweckstetter M (2003) Determination of molecular alignment tensors without backbone resonance assignment: aid to rapid analysis of protein-protein interactions. *J Biomol NMR* 27:41–56
- Zweckstetter M (2008) NMR: prediction of molecular alignment from structure using the PALES software. *Nat Protoc* 3:679–690
- Zweckstetter M, Bax A (2001) Single-step determination of protein substructures using dipolar couplings: aid to structural genomics. *J Am Chem Soc* 123:9490–9491
- Zweckstetter M, Bax A (2002) Evaluation of uncertainty in alignment tensors obtained from dipolar couplings. *J Biomol NMR* 23:127–137