Journal of Biomolecular NMR **30:** 25–35, 2004.

Backbone assignment of proteins with known structure using residual dipolar couplings

Young-Sang Jung & Markus Zweckstetter*
Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, D-37077 Göttingen, Germany

Received 8 January 2004; Accepted 12 May 2004

Key words: assignment, maltose-binding protein, Mars, NMR, protein-ligand binding, RDC, software, structural genomics

Abstract

A prerequisite for NMR studies of protein-ligand interactions or protein dynamics is the assignment of backbone resonances. Here we demonstrate that protein assignment can significantly be enhanced when experimental dipolar couplings (RDCs) are matched to values back-calculated from a known three-dimensional structure. In case of small proteins, the program MARS allows assignment of more than 90% of backbone resonances without the need for sequential connectivity information. For bigger proteins, we show that the combination of sequential connectivity information with RDC-matching enables more residues to be assigned reliably and backbone assignment to be more robust against missing data. Structural or dynamic deviations from the employed 3D coordinates do not lead to an increased error rate in RDC-supported assignment. RDC-enhanced assignment is particularly useful when chemical shifts and sequential connectivity only provide a few reliable assignments.

Introduction

NMR spectroscopy is a powerful tool to study protein-ligand binding, protein-nucleic acid interactions and protein dynamics. A prerequisite for these studies is assignment of NMR spin resonances (Wüthrich, 2003). In recent years, good progress has been made in automating the assignment process for proteins up to 20 kDa (Moseley and Montelione, 1999) and in the accompanying paper we have introduced the program MARS that allows robust automatic backbone assignment also for unfolded and large proteins (Jung and Zweckstetter, 2003).

Most assignment approaches, such as MARS, rely on methods to connect NMR resonances related to single residues into segments and to map these segments onto the known protein sequence based on the very sensitive relationship between amino acid type and chemical shifts (Grzesiek and Bax, 1993; Moseley and Montelione, 1999; Spera and Bax, 1991). The accuracy of chemical shifts, calculated with current methods from the protein sequence or even from a known 3D structure, is, however, not sufficient, to assign error-free and unambiguously connectivity segments to the protein sequence. This is especially problematic for big proteins and proteins where a significant fraction of data is missing, independent of whether the assignment is performed manually or automatically. To avoid assignment errors in these cases, more conservative approaches have to be taken when connectivity segments are mapped onto the primary sequence. This generally results in a decrease in the number of residues that can be assigned reliably (Jung and Zweckstetter, 2003).

When a three-dimensional (3D) structure of the protein is known already, comparison of NMR parameters back-calculated from this structure with experimental values can potentially be used to improve the assignment process. So far, most studies have focused on incorporation of Nuclear Overhauser Effect (NOE) distance constraints into the assignment process: only assignments that are consistent with distances ob-

*To whom correspondence should be addressed. E-mail: mzwecks@gwdg.de

served in the 3D structure are allowed (Bartels et al., 1996; Bernstein et al., 1993; Pristovsek et al., 2002).

Recently, it was shown that also residual dipolar couplings (RDCs) are very useful for resonance assignment. If no 3D structure is available, RDCs can be used to reduce chemical shift degeneracies in sequential connectivity experiments (Zweckstetter and Bax, 2001). In case of small proteins, they even allow simultaneous resonance assignment and structure determination (Tian et al., 2001). On the other hand, calculation of RDCs from a known 3D structure is straightforward and has been used previously for validation of protein structures (Bax et al., 2001; Prestegard and Kishore, 2001). Therefore, an assignment method for proteins can be envisioned where dipolar couplings calculated from a known 3D structure are compared to experimental values. Initially, such an approach was described for RNA (Al-Hashimi et al., 2002). Hus et al. extended this strategy recently to proteins (Hus et al., 2002). Prestegard and coworkers, on the other hand, employed a manual approach where they assigned five peaks of the human ADP ribosylation factor 1, which could not be assigned using triple-resonance experiments, by matching predicted $^1D_{NH}$ couplings with experimental values (Amor et al., 2002). None of these approaches, however, allows simultaneous use of sequential connectivity information and RDCs, or provides an indication on how reliable an assignment obtained by RDC matching is.

Here we show that RDCs can be routinely included into backbone assignment of proteins with known structure using the program MARS. In case of small proteins, MARS allows RDC-based assignment of more than 90% of backbone resonances without the need for sequential connectivity information. For bigger proteins, we demonstrate that assignment can significantly be enhanced by combining RDC matching with sequential connectivity information and that inaccuracies in the 3D structure do not result in an increased number of assignment errors.

Methods

The assignment algorithm employed in MARS has been described in detail in the accompanying paper (Jung and Zweckstetter, 2003). For RDC-enhanced assignment, this algorithm is extended as described below.

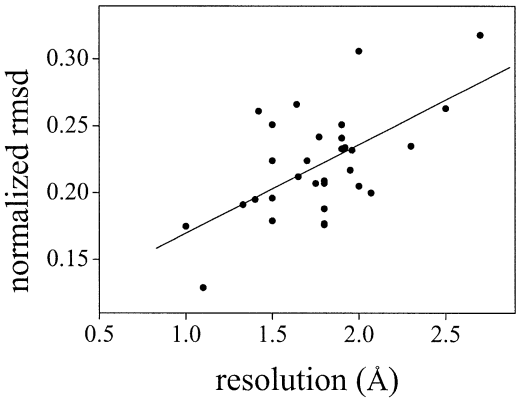

Figure 1. Correlation between resolution of a crystal structure and the fit of dipolar couplings to this structure. 'Normalized rmsd' is the root-mean-square-deviation between experimental and back-calculated RDCs divided by the experimental alignment strength D_a^{HN}. Back-calculation of RDCs was performed by SVD.

Input and output data

When dipolar couplings are to be used for assignment, a PDB file of a known 3D structure or homology model of the protein has to be supplied as input to MARS and the resolution of the structure has to be indicated. In addition, pseudoresidues comprise experimental chemical shifts and RDCs. One-bond RDCs are commonly measured from triple-resonance experiments, such as HNCO, (Bax and Grzesiek, 1993; Bax et al., 2001; Prestegard and Kishore, 2001) and it is therefore straightforward to add these RDCs to pseudoresidues.

Besides the standard output provided by MARS, an alignment tensor is returned that has been optimized during the assignment process together with RDCs back-calculated from the 3D structure.

Matching of experimental RDCs to back-calculated values

When the 3D structure of the protein is unknown, mapping of single pseudoresidues (PR) or of segments connecting several pseudoresidues relies on comparison of experimental chemical shifts with values calculated from the protein sequence. This could potentially be improved when chemical shifts are calculated from the 3D structure. However, chemical shifts, which were calculated from the protein sequence with the use of correction factors for neighbor residue effects and secondary structure prediction information (Jung and Zweckstetter, 2003), are of comparable quality as values predicted for proteins with known structure us-

ing the program SHIFTS (Xu and Case, 2001, 2002). Therefore, the assignment performance of MARS was not improved when using chemical shifts calculated with SHIFTS (data not shown).

In order to include RDCs into the process of mapping PR segments onto the protein sequence, MARS calculates for all experimentally observed pseudoresidues the deviation of their experimental RDCs and chemical shifts from predicted values according to

$$D(i,j) = w \sum_{k=1}^{N_{CS}} \left\{ \frac{\delta(i)_k^{\exp} - \delta(j)_k}{\sigma_k^{CS}} \right\}^2 + \sum_{l=1}^{N_{RDC}} \left\{ \frac{RDC(i)_l^{\exp} - RDC(j)_l}{\sigma_l^{RDC}} \right\}^2, \quad (1)$$

where $\delta(i)_k^{\exp}$ is the measured chemical shift of type k (e.g., $^{13}C^{\alpha}$ or $^{13}C^{\beta}$) of pseudoresidue i, $\delta(j)_k$ is the predicted chemical shift of type k of residue j, N_{CS} is the number of chemical shift types and σ_k^{CS} is the standard deviation of the statistical chemical shift distribution that is used for calculating $\delta(j)_k$. For $^{1}H^{N}$, ^{15}N, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ $^{13}C'$ and $^{1}H^{\alpha}$ σ_k^{CS} values of 0.82, 4.3, 1.2, 1.1, 1.7 and 0.82 ppm were used, respectively (Jung and Zweckstetter, 2003). Similarly, $RDC(i)_l^{\exp}$ is the experimental RDC of type l (e.g. $^{1}D_{NH}$ or $^{1}D_{CaC'}$) of pseudoresidue i, $RDC(j)_l$ is the back-calculated RDC of type l of residue j, N_{RDC} is the number of RDC types and σ_l^{RDC} is the value used for normalizing RDC deviations. w is a weighting factor that takes into account the different reliability of calculated chemical shifts and RDCs. As back-calculated RDCs are directly influenced by structural and dynamical deviations from the PDB coordinates, empirical optimization resulted in $w = 3.3$, thereby downscaling the contribution of RDCs.

The RDC normalization constant σ_l^{RDC} is adjusted according to the resolution, R_{struc}, of the 3D structure. Figure 1 compares the normalized root-mean-square-deviation between experimental RDCs and values back-calculated from known crystal structures using SVD (based on published assignments) for 31 crystal structures (Table 1). Based on the slope of the linear fit shown in Figure 1, c_{RDC},

$$\sigma_l^{RDC} = c_{RDC} R_{struc} D_a^{HN}, \quad (2)$$

where D_a^{HN} is the magnitude of the alignment tensor required to take into account the overall alignment strength. As the correlation visible in Figure 1 is not very high, σ_l^{RDC} can also be set to a fixed value of 0.21 for R_{struc} ranging from 1.4 to 2.4 Å without strongly affecting the assignment result.

RDCs are back-calculated from user-supplied PDB coordinates according to

$$RDC_{PQ} = -\mu_0 \gamma_P \gamma_Q h / (8\pi^3 < r_{PQ}^3 >) \sum_{i,j} A_{ij} \cos\phi_i^{PQ} \cos\phi_j^{PQ}, \quad (3)$$

where RDC_{PQ} is the dipolar coupling between a pair of spin-1/2 nuclei, P and Q, separated by a distance r_{PQ}, **A** is a second-rank alignment tensor, γ_P and γ_Q are the gyromagnetic ratios, h is Planck's constant, μ_0 is the magnetic permeability of vacuum, and ϕ_i^{PQ} is the angle between the P–Q internuclear vector and the ith molecular axis. As ϕ_i^{PQ} and r_{PQ} can be derived from the 3D structure, the only unknown variable in equation (3) is the alignment tensor **A**.

Alignment tensor determination

The magnitude and rhombicity of a molecular alignment tensor **A** can be obtained accurately without assignment from a histogram of experimental RDCs (Clore et al., 1998; Skrynnikov and Kay, 2000; Warren and Moore, 2001). In order to extract the orientation of the alignment tensor, four different methods are available in MARS: (1) shape and charge/shape-prediction of molecular alignment tensors (Zweckstetter and Bax, 2000; Zweckstetter et al., 2004), (2) singular value decomposition (Losonczi et al., 1999) after an initial assignment step using only chemical shifts, (3) exhaustive back-calculation (Zweckstetter, 2003) and (4) a grid search that optimizes the fit of experimental chemical shifts and RDCs to values predicted from the 3D structure. Shape- and charge/shape prediction is problematic for proteins with long, flexible loops or tails, but has the advantage that the only information necessary is the 3D structure (Zweckstetter and Bax, 2000; Zweckstetter et al., 2004). Exhaustive back-calculation is useful when the amino acid type of some resonances can be identified either by selective labeling or on the basis of the C^{α} and C^{β} chemical shift (Zweckstetter, 2003), as the actual size of the protein is not important, provided that experimental RDCs could be measured accurately. When sufficient chemical shift data are available (for example sequential connectivity information), it is straightforward to obtain the alignment tensor by a two-stage strategy which consists of an initial assignment run using only chemical shifts, followed by a best-fit of experimental

Table 1. Proteins used for evaluation of the correlation between resolution of a crystal structure and the fit of residual dipolar couplings to this structure

Protein	Crystal structure (PDB code)	Resolution (Å)	RDCs (PDB code)	Normalized rmsd[a]
Maltose-binding protein	1DMB	1.8	1EZP	0.24
Maltose-binding protein	1OMP	1.8	1EZP	0.19
Barrier-to-autointegration factor	1CI4	1.9	2EZX	0.23
B1 IgG-binding domain	1IGD	1.1	1P7E	0.13
B1 IgG-binding domain	1PGA	2.1	1P7E	0.20
B1 IgG-binding domain	1PGB	1.9	1P7E	0.23
N-terminal domain of enzyme I	1ZYM	2.5	3EZA	0.26
Histidine-containing phosphocarrier protein	1POH	2.0	3EZA	0.21
Histidine-containing phosphocarrier protein	1OPD	1.5	3EZA	0.20
Ubiquitin	1UBQ	1.8	1D3Z	0.18
Ubiquitin	1UBI	1.8	1D3Z	0.18
Ubiquitin	1F9J	2.7	1D3Z	0.32
Ubiquitin	1AAR	2.3	1D3Z	0.24
Bovine pancreatic trypsin inhibitor	5PTI	1.0	[b]	0.18
Bovine pancreatic trypsin Inhibitor	1QLQ	1.4	[b]	0.26
Cyanovirin-N	1L5B	2.0	2EZM	0.31
Cyanovirin-N	3EZM	1.5	2EZM	0.18
Lysozyme	1FLQ	1.8	1E8L	0.21
Lysozyme	1UIG	1.9	1E8L	0.22
Lysozyme	1UIH	1.8	1E8L	0.21
Lysozyme	1H87	1.7	1E8L	0.22
Lysozyme	1H6M	1.6	1E8L	0.27
Lysozyme	1GWD	1.8	1E8L	0.24
Lysozyme	193L	1.3	1E8L	0.19
Lysozyme	1IEE	1.5	1E8L	0.25
Lysozyme	194L	1.4	1E8L	0.20
Lysozyme	1AKI	1.5	1E8L	0.22
Lysozyme	1AT5	1.8	1E8L	0.21
Lysozyme	1DPX	1.7	1E8L	0.21
Lysozyme	1F0W	1.9	1E8L	0.25
Lysozyme	1KXW	2.0	1E8L	0.23

[a] 'Normalized rmsd' is the root-mean-square-deviation between experimental and back-calculated RDCs divided by the experimental alignment strength D_a^{HN}.

[b] RDCs were kindly provided by Ben Ramirez and Ad Bax.

RDCs to the 3D structure (Losonczi et al., 1999) based on this assignment. Tests show that, even when the percentage of correct assignment is below 50%, the alignment tensor is very close to its correct orientation. As C^{α}/C^{β} chemical shifts depend very much on the type of secondary structure, exchange of assignments mainly takes place between residues located on the same type of regular secondary structure. If these secondary structure elements are close to collinear, such as two β-strands in a β-sheet, residues located in these strands can have similar RDCs and back-calculated alignment tensors are not severely affected by an interchanged assignment (data not shown).

The most general method for extracting the orientation of the alignment tensor is a grid search in which the fit between experimental and predicted RDCs and chemical shifts is optimized. In this gird search 1116 uniformly distributed alignment tensor orientations are systematically sampled (Eisenhaber et al., 1995) and for each orientation the deviation $D(i, j)$ between experimental and back-calculated RDCs is determined (equation 1). All sampled orientations are ranked according to their corresponding $D(i, j)$ values and the

Table 2. RDC-enhanced assignment of ubiquitin for varying amount of data

RDCs[a]	Chemical shifts for linking[b]	Chemical shifts for matching[c]	assignment score (%)[d]					
			1UBQ			1AAR		
			Total[e] correct	Reliable	Wrong reliable[f]	Total[e] correct	Reliable	Wrong reliable[f]
Without sequential connectivity information								
–	–	$C'_{i-1}, C^{\alpha}_{i-1}, C^{\beta}_{i-1}$	36.1	19.5	5.6	36.1	19.5	5.6
$^1D_{NH}$	–	$C'_{i-1}, C^{\alpha}_{i-1}, C^{\beta}_{i-1}$	47.2	16.7	1.4	38.9	9.7	1.4
$^1D_{NH}$, $^1D_{CaC'}$	–	$C'_{i-1}, C^{\alpha}_{i-1}, C^{\beta}_{i-1}$	76.4	44.4	0.0	68.1	33.4	2.8
$^1D_{NH}$, $^1D_{CaC'}$, $^1D_{NC'}$	–	$C'_{i-1}, C^{\alpha}_{i-1}, C^{\beta}_{i-1}$	91.7	55.6	0.0	83.3	50.0	0.0
With sequential connectivity information								
–	C^{α}	$C'_{i-1}, C^{\alpha}_{i-1}, C^{\alpha}_{i}$	80.6	25.0	0.0	80.6	25.0	0.0
$^1D_{NH}$	C^{α}	$C'_{i-1}, C^{\alpha}_{i-1}, C^{\alpha}_{i}$	93.1	51.4	0.0	97.2	37.5	0.0
$^1D_{NH}$, $^1D_{CaC'}$	C^{α}	$C'_{i-1}, C^{\alpha}_{i-1}, C^{\alpha}_{i}$	100.0	90.3	0.0	100.0	73.6	0.0
$^1D_{NH}$, $^1D_{CaC'}$, $^1D_{NC'}$	C^{α}	$C'_{i-1}, C^{\alpha}_{i-1}, C^{\alpha}_{i}$	100.0	100.0	0.0	100.0	100.0	0.0

[a]RDCs observed in nearly neutral bicelles were used (Cornilescu et al., 1998).
[b]Chemical shifts used for establishing sequential connectivity. The connectivity threshold was 0.2 ppm.
[c]Chemical shifts used for mapping pseudoresidue segments to the protein sequence. In addition to the mentioned values, HN and N chemical shifts were used.
[d]Relative to the number of assignable residues, 72 in case of ubiquitin.
[e]# of correct assignments in Ass^{global} ; Ass^{global} was obtained from a MARS run without addition of noise (Jung and Zweckstetter, 2003).
[f]Assignments that were identified as reliable but are incorrect, i.e., the error rate.

Table 3. RDC-enhanced assignment of 370-residue maltose-binding protein for varying amount of data

RDCs[a]	Chemical shifts for linking[b]	Chemical shifts for mapping[c]	Missing chemical shifts (%)[d]	Assignment score (%)[e]		
				Total correct[f]	Reliable	Wrong reliable[g]
–	C^{α}, C^{β}	$C'_{i-1}, C^{\alpha}_{i-1}, C^{\alpha}_{i}, C^{\beta}_{i-1}, C^{\beta}_{i}$	4	95.8	87.2	0.0
$^1D_{NH}$	C^{α}, C^{β}	$C'_{i-1}, C^{\alpha}_{i-1}, C^{\alpha}_{i}, C^{\beta}_{i-1}, C^{\beta}_{i}$	4	98.5	94.6	0.0
$^1D_{NH}$, $^1D_{CaC'}$	C^{α}, C^{β}	$C'_{i-1}, C^{\alpha}_{i-1}, C^{\alpha}_{i}, C^{\beta}_{i-1}, C^{\beta}_{i}$	4	98.5	93.4	0.0
$^1D_{NH}$, $^1D_{CaC'}$, $^1D_{NC'}$	C^{α}, C^{β}	$C'_{i-1}, C^{\alpha}_{i-1}, C^{\alpha}_{i}, C^{\beta}_{i-1}, C^{\beta}_{i}$	4	99.1	94.6	0.0
–	C^{α}, C^{β}	$C'_{i-1}, C^{\alpha}_{i-1}, C^{\alpha}_{i}, C^{\beta}_{i-1}, C^{\beta}_{i}$	20	82.7	44.2	0.7
$^1D_{NH}$	C^{α}, C^{β}	$C'_{i-1}, C^{\alpha}_{i-1}, C^{\alpha}_{i}, C^{\beta}_{i-1}, C^{\beta}_{i}$	20	88.8	51.4	0.0
$^1D_{NH}$, $^1D_{CaC'}$	C^{α}, C^{β}	$C'_{i-1}, C^{\alpha}_{i-1}, C^{\alpha}_{i}, C^{\beta}_{i-1}, C^{\beta}_{i}$	20	95.0	58.0	0.4
$^1D_{NH}$, $^1D_{CaC'}$, $^1D_{NC'}$	C^{α}, C^{β}	$C'_{i-1}, C^{\alpha}_{i-1}, C^{\alpha}_{i}, C^{\beta}_{i-1}, C^{\beta}_{i}$	20	94.3	62.6	0.0

[a]RDCs were measured for MBP dissolved in Pf1 bacteriophage (Hansen et al., 1998; Yang et al., 1999).
[b]Chemical shifts used for establishing sequential connectivity. A common connectivity threshold of 0.5 ppm was used for C^{α} and C^{β}.
[c]Chemical shifts used for mapping pseudoresidue segments to the protein sequence. In addition to the mentioned values, HN and N chemical shifts were also used.
[d]Percentage of non-proline residues for which HN and N chemical shifts were not present.
[e]Relative to the number of assignable residues, i.e. those residues with HN and N chemical shifts.
[f]# of correct assignments in Ass^{global} ; Ass^{global} was obtained from a MARS run without addition of noise (Jung and Zweckstetter, 2003).
[g]Assignments that were identified as reliable but are incorrect, i.e., the error rate.

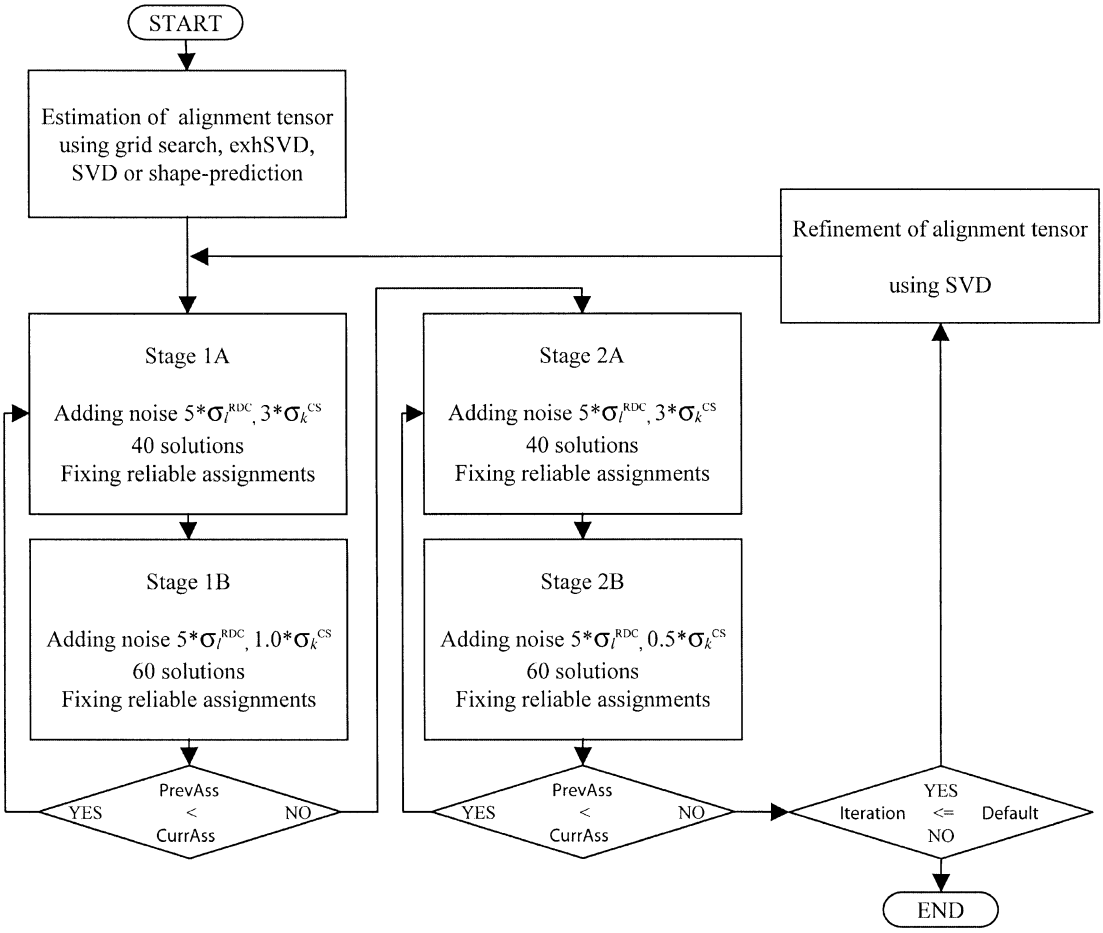

Figure 2. Empirically optimized scheme for avoiding errors due to inaccuracies in calculated RDCs and chemical shifts when mapping pseudoresidue segments to the protein sequence. Opposite to the original scheme (Jung and Zweckstetter, 2003), two full assignment runs are performed and in the second run a refined alignment tensor, which has been obtained by SVD, is used. $5 * \sigma_l^{RDC}$ is the width of the Gaussian distribution function from which RDC noise is drawn. By default two iterations (2 <= Default) are performed. See text for a definition of σ_l^{RDC}.

lowest $D(i, j)$ value indicates the best estimate for the experimental alignment tensor. All assignment results reported here were obtained using this method.

Assignment schedule

The overall assignment schedule is slightly changed when RDCs are used in addition to chemical shifts. Although the methods described above allow determination of approximate alignment tensors, their accuracy is inferior to singular value decomposition based on a known assignment. Therefore, for RDC-enhanced assignment two complete MARS assignment runs are performed. In the first run, dipolar couplings are back-calculated from the 3D structure using the approximate alignment tensor. After this run a sufficient number of reliable assignments are generally available and based on these assignments MARS can perform a singular value decomposition. This results in an improved tensor that is used in a second assignment run to refine assignment (Figure 2). More assignment runs are generally not required. Due to this two-step procedure the final assignment score obtained by MARS is almost independent from the method that was chosen to get a first estimate of the alignment tensor.

Overcoming structural and dynamic deviations from PDB coordinates

RDCs strongly depend on the exact orientation of their corresponding internuclear vectors (Equation 3) and slight errors in the structure can give rise to significant deviations in back-calculated dipolar couplings. Back-calculated RDCs, however, are used for mapping of

pseudoresidue segments to the protein sequence and incorrect values can lead to wrong assignments. This problem is partially addressed by reducing the weight of RDCs compared to chemical shifts by a factor of 3.3 (Equation 1). To further improve the reliability of RDC-enhanced assignment, a similar approach as for chemical shifts is used (Jung and Zweckstetter, 2003): several assignment phases are performed where back-calculated RDCs are disturbed by addition of noise and only consistent assignments are retained. The addition of noise to RDCs and chemical shifts is done simultaneously and results in an empirically optimized assignment schedule outlined in Figure 2. Opposite to chemical shifts, however, the amount of noise added to back-calculated RDCs is kept fixed at five times σ_l^{RDC}. Such a large amount of variation in back-calculated RDCs is necessary, in order to avoid wrong assignments.

Often parts of proteins, such as flexible termini or loops, are unstructured and are not available in crystal structures or are prone to deviate from their conformation in solution. In order to identify potentially flexible parts of proteins, we estimated NMR S^2 order parameters of N-H^N vectors of the protein backbone from the 3D structure (Zhang and Bruschweiler, 2002). Removal of back-calculated RDCs for residues with estimated S^2 order parameters smaller than 0.75 did, however, not improve MARS assignment results. Therefore, RDCs are back-calculated for all residues that are visible in a crystal structure and are used for enhanced mapping to the protein sequence.

Testing

RDC-enhanced assignment was applied to three proteins for which experimental chemical shifts and dipolar couplings have been reported and a high-resolution crystal structure is available: ubiquitin (76 aa; PDB codes: 1UBQ [1.8 Å] and 1AAR [2.3 Å]; RDCs: PDB code 1D3ZMR; chemical shifts from TALOS) (Cook et al., 1992; Cornilescu et al., 1998, 1999; Vijaykumar et al., 1987), the N-terminal domain of enzyme I of the phosphoenolpyruvate (EIN) (259 aa; PDB code: 1ZYM [2.5 Å]; RDCs: PDB code 3EZAMR; chemical shifts: BMRB code 4106) (Garrett et al., 1997, 1999; Liao et al., 1996) and two-domain maltose-binding protein (MBP) (370 aa; PDB code: 1DMB [1.8 Å]; RDCs: kindly provided by Lewis Kay; chemical shifts: BMRB code 4354) (Gardner et al., 1998; Mueller et al., 2000; Sharff et al., 1993; Yang et al., 1999). Protons were added to crystal structures using MOLMOL (Koradi et al., 1996).

In order to evaluate, how much information is required for successful assignment, we analyze different test cases, such as assignment without sequential connectivity information using only dipolar coupling/chemical shift matching, assignment with only C^α sequential connectivity information and assignment using C^α/C^β chemical shifts. In addition, the effect of including only one, two or three types of RDCs is tested.

Results and discussion

RDC-enhanced assignment without sequential connectivity information

Table 2 shows the results of RDC-enhanced assignment for 76-residue protein ubiquitin. Initially, it was tested how inclusion of RDCs can enhance assignment when no connectivity information at all is available. A situation is assumed where only inter-residual C^α, C^β and C' chemical shifts could be measured, providing information about the amino-acid type of the preceeding residue. When no sequential connectivity information is available, MARS matches single pseudoresidues (comprising $H^N(i)$, $N(i)$, $C'(i-1)$, $C^\alpha(i-1)$ and $C^\beta(i-1)$ chemical shifts) to the primary sequence. Alternatively, one could try to map all possible three-residue fragments (i.e., for a total of 72 pseudoresidues there would be about 360000 possible three-residue fragments that could be matched to each three-residue protein fragment). Tests, however, show that this significantly reduces the assignment quality (data not shown). As calculation of chemical shifts from the protein sequence (or from the 3D structure) gives only approximate values, the total percentage of correct assignment in the absence of RDCs was only 36.1%. Moreover, only 19.5% (out of the total of 72 assignable residues) were labeled as reliable by MARS and about 40% of these were wrong. This highlights that mapping of single pseudoresidues (comprising only chemical shifts) to the protein sequence is not sufficient and that additional information for identification of reliable assignments is required. Comparison of RDCs back-calculated from a known 3D structure with experimental values provides such information. Including only $^1D_{NH}$ couplings into the assignment process increased the overall assignment score to 47.2%, the reliable assignment to 16.7% and

out of these only one assignment was wrong. The situation was further improved when two or three types of RDCs were used. Without trying to distinguish between correct and incorrect assignments, i.e., without applying the MARS criteria for reliability, 66 residues (91.7%) of ubiquitin were assigned correctly. This is in agreement with recent results by Hus et al. (2002). Opposite to their approach, however, MARS allows a clear distinction between reliable assignments and those that are prone to errors. Application of the MARS reliability criteria identifies 55.6% of residues (out of the total of 72 assignable residues) as reliably assigned, with not a single error present.

The results discussed so far were obtained with a 1.8 Å crystal structure of ubiquitin (PDB code: 1UBQ). Such high-resolution structures might not always be available. At the same time, structural noise is a major factor influencing the accuracy of back-calculated alignment tensors and RDCs, whereas the experimental accuracy of RDC data measured with current methods is usually sufficient (Zweckstetter and Bax, 2002). In order to test the robustness of RDC-enhanced assignment against structural deviations from the PDB coordinates, assignment of ubiquitin was also performed using a 2.3 Å crystal structure (PDB code: 1AAR). When only $^1D_{NH}$ couplings were used, the number of reliable assignments was reduced compared to assignment based solely on chemical shifts (Table 2). This is due to the fact that reliability is now tested using both chemical shifts and RDCs, thereby removing some (previously reliable and correct) assignments that do not have very characteristic dipolar couplings. More important, however, is that the error rate was reduced to a single wrong assignment when $^1D_{NH}$ RDCs were introduced. Using two or three types of RDCs, assignment of ubiquitin was significantly enhanced, similar to the results obtained for the 1.8 Å structure. Due to the lower quality of the 1AAR structure, however, the improvement achieved by inclusion of RDCs was not as strong. In particular, a lower number of assignments were identified as reliable, whereas the total number of correct assignments was only slightly affected. Nevertheless, 100% of ubiquitin could be assigned reliably, when C^α connectivity information was combined with RDC-matching of $^1D_{NH}$, $^1D_{CaC'}$ and $^1D_{NC'}$ couplings (see below).

RDC-enhanced assignment with sequential connectivity information

For some applications, such as titration studies, reliable assignment scores of 50% might be sufficient or some wrong assignments are not problematic. Complete and error-free assignment, however, will often still be the major aim. In addition, assignment of small proteins such as ubiquitin is straightforward using C^α/C^β connectivity information obtained from triple-resonance experiments, even without usage of RDCs. For bigger proteins, on the other hand, mapping of pseudoresidues to the protein sequence using only chemical shifts is usually not sufficient to reliably assign 100% of the protein. Especially, when a substantial amount of data is missing due to chemical exchange or incomplete back-exchange of amide protons in deuterated proteins, the number of residues, which can be assigned reliably, significantly decreases (Jung and Zweckstetter, 2003). Therefore, the area where RDC-enhanced assignment has its largest potential is for big, deuterated proteins in combination with standard sequential connectivity information.

Combination of a limited amount of connectivity information with RDC-matching was first tested on ubiquitin (Table 2). Using only C^α connectivity information with a threshold of 0.2 ppm for establishing sequential connectivity together with $^1D_{NH}$, $^1D_{CaC'}$ and $^1D_{NC'}$ couplings, 100% of residues were assigned reliably by MARS without any assignment error (for both the 1UBQ and 1AAR structure). On the other hand, without RDCs, i.e., using just chemical shifts for mapping pseudoresidue segments to the protein sequence, only 25% of residues could be assigned reliably. This indicates the great potential of combining RDCs back-calculated from a known structure with sequential connectivity information.

Table 3 shows results obtained from RDC-enhanced assignment for 370-residue maltose-binding protein (MBP). Using the complete set of C^α/C^β chemical shifts deposited in the BMRB (Doreleijers et al., 2003) but not using any RDC-matching, 87.2% of assignable residues of MBP were assigned reliably. This number was increased to about 94% when at least one RDC type was included. No errors were introduced into assignment by RDC-matching. As the assignment score was already very high, inclusion of more than one RDC type did not further improve assignment significantly. More pronounced was the effect when a substantial amount of data was missing. When 20% of MBP's pseudoresidues were removed

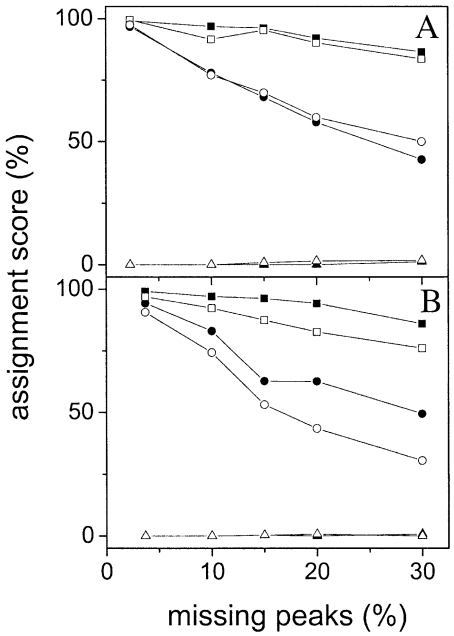

Figure 3. Dependence of RDC-enhanced assignment on the percentage of missing pseudoresidues. Pseudoresidues were randomly deleted. ■ indicate the percentage of all assignments that were correct (not tested for reliability). ● show the percentage of residues that could be assigned reliably (relative to the total number of assignable residues) and ▲ indicate assignments that were identified as reliable but are wrong, i.e., the error rate of MARS. Only C^{α} and C^{β} chemical shifts with a common threshold of 0.5 ppm for establishing sequential connectivity were used. Open symbols indicate the results without RDCs (see accompanying paper). (A) Results for the 259-residue N-terminal domain of enzyme I using RDC-matching of $^{1}D_{NH}$ couplings. (B) Results for the 370-residue maltose-binding protein using RDC-matching of $^{1}D_{NH}$, $^{1}D_{CaC'}$ and $^{1}D_{NC'}$ couplings.

randomly and no RDC-matching was employed, the reliable assignment was reduced to 44.2% and two assignment errors were present (Jung and Zweckstetter, 2003). Enhancing the mapping process by comparison of $^{1}D_{NH}$, $^{1}D_{CaC'}$ and $^{1}D_{NC'}$ couplings back-calculated from MBP's 1.8 Å structure with experimental values, increased the reliable assignment to 62.6% (total correct assignment of 94.3%). In addition, no assignment errors were present any more.

Robustness against missing data

The robustness of RDC-enhanced MARS assignment was further tested by continuously increasing the randomly deleted fraction of observed pseudoresidues from 5 to 30% for MBP and the N-terminal domain of enzyme I (EIN). Similar to the situation when RDCs were not used, the assignment decreased with decreasing number of pseudoresidues and the reliable assignment was most strongly affected (Figure 3B). Whereas, however, without RDCs the percentage of reliable assignments dropped to 31% when 30% of MBP's pseudoresidues were randomly deleted (Jung and Zweckstetter, 2003), it remained at 49% upon inclusion of $^{1}D_{NH}$, $^{1}D_{CaC'}$ and $^{1}D_{NC'}$ couplings. In addition, the total number of correct assignments was increased from 76% to 86%.

For MBP a very extensive set of RDCs was measured by optimized triple-resonance experiments (Yang et al., 1999). For EIN, on the other hand, only $^{1}D_{NH}$ RDCs for 60% of residues were available from two-dimensional HSQC spectra (Garrett et al., 1999). In addition, with a resolution of 2.5 Å and 10 residues not present in the PDB coordinates, the crystal structure available for EIN (PDB code: 1ZYM) is of much lower quality than that of MBP. In this case, RDC-enhanced and RDC-free assignment were virtually identical (Figure 3A). A slight improvement upon inclusion of RDCs, however, is obtained with respect to the error-rate. Whereas for RDC-free assignment two, three and three residues were assigned wrongly at 15, 20 and 30% deleted pseudoresidues, respectively, this was reduced to zero, zero and two residues for RDC-enhanced assignment. Such a small effect is actually not unexpected as the major use of RDCs is improved matching of PR-segments to the primary sequence. When C^{α} and C^{β} chemical shift information is close to complete, as it is the case for EIN, segment placement is already quite robust and incorporation of just $^{1}D_{NH}$ couplings for 60% of pseudoresidues does not have a major impact.

Very often, however, not entire pseudoresidues are missing, but certain chemical shifts are not observable. This situation was simulated by randomly removing chemical shifts within pseudoresidues of EIN. When C^{α} and C^{β} chemical shifts are removed from pseudoresidues this strongly affects the ability to correctly place PR-segments onto the primary sequence. In such situations even a small number of $^{1}D_{NH}$ RDCs can be useful as demonstrated in Figure 4. Although, the number of reliable assignments was only increased by 6% on average, the total number of correct assignments was raised by 22% when 28% of C^{α}/C^{β} chemical shifts were missing. This means that with the help of $^{1}D_{NH}$ RDCs the correct assignment was proposed for 55 additional residues of EIN, providing a significantly improved starting point for manual refinement of the assignment (using for example the analysis software SPARKY). Therefore, even for sparse data

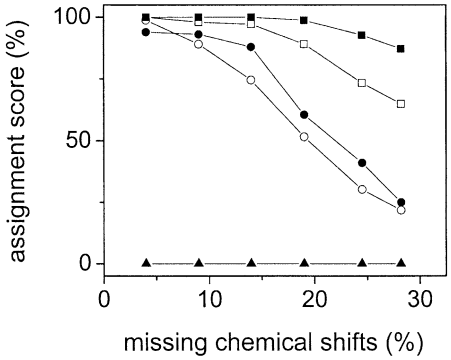

Figure 4. Dependence of RDC-enhanced assignment on the percentage of missing chemical shifts within pseudoresidues for the 259-residue N-terminal domain of enzyme I. Chemical shifts were deleted randomly. ■, ● and ▲ indicate correct, reliable and wrong reliable assignments, respectively. C^{α} and C^{β} chemical shifts with thresholds of 0.2 and 0.4 ppm for establishing sequential connectivity were used. Open symbols indicate the results without RDCs (see Jung and Zweckstetter, 2004). There are zero errors for both RDC-enhanced and RDC-free assignment.

comparison of RDCs back-calculated from a known 3D structure with experimental values is useful for assignment.

Concluding remarks

We have introduced a reliable method for enhancing backbone resonance assignment of proteins with known structure using residual dipolar couplings. This method has been implemented into the automatic assignment program MARS. It is equally applicable to small or big proteins, when only $^{1}D_{NH}$ couplings could be measured for a limited number of residues or when a complete set of dipolar couplings for five different inter-nuclear vectors is available. RDC-enhanced assignment will be especially useful for large proteins where chemical shift data are often missing for a substantial portion of residues and chemical shift degeneracy is too high to allow unambiguous assignment. Similarly, if only a few reliable assignments could be obtained based on chemical shifts and sequential connectivity, RDC matching allows evaluation of remaining assignment possibilities. Safe assignments or connectivities (as established, for example, from manual inspection of assignment strips on the screen) can thereby be fixed.

Structure-enhanced assignment becomes increasingly important due to the rapid increase in the number of high-resolution 3D structures that are determined as part of the worldwide structural genomics effort. Residual dipolar couplings are in this respect particularly interesting, as they can be measured efficiently from two-dimensional ^{1}H-^{15}N HSQC or three-dimensional triple-resonance spectra. Moreover, triple-resonance experiments can be used simultaneously for RDC measurement and to establish sequential connectivity (Tian et al., 2001; Zweckstetter and Bax, 2001), thereby saving spectrometer time and money. At the same time, dipolar couplings will often be measured, in order to validate a crystal structure prior to its usage, for example, in binding studies (Evenas et al., 2001; Jain et al., 2003). In these cases, inclusion of RDCs into the assignment process and therefore improved assignment will not require additional NMR samples or extra measurement time.

Acknowledgements

We thank Christian Griesinger for useful discussions. RDCs in MBP were kindly provided by Lewis Kay. M.Z. is funded by the DFG Emmy Noether-research programme (ZW 71/1-3).

References

Al-Hashimi, H.M., Gorin, A., Majumdar, A., Gosser, Y. and Patel, D.J. (2002) *J. Mol. Biol.*, **318**, 637–649.

Amor, J.C., Seidel, R.D., Tian, F., Kahn, R.A. and Prestegard, J.H. (2002) *J. Biomol. NMR*, **23**, 253–254.

Bartels, C., Billeter, M., Güntert, P. and Wüthrich, K. (1996) *J. Biomol. NMR*, **7**, 207–213.

Bax, A. and Grzesiek, S. (1993) *Accounts Chem. Res.*, **26**, 131–138.

Bax, A., Kontaxis, G. and Tjandra, N. (2001) *Meth. Enzymol.*, **339**, 127–174.

Bernstein, R., Cieslar, C., Ross, A., Oschkinat, H., Freund, J. and Holak, T.A. (1993) *J. Biomol. NMR*, **3**, 245–251.

Clore, G.M., Gronenborn, A.M. and Bax, A. (1998) *J. Magn. Reson.*, **133**, 216–221.

Cook, W.J., Jeffrey, L.C., Carson, M., Chen, Z.J. and Pickart, C.M. (1992) *J. Biol. Chem.*, **267**, 16467–16471.

Cornilescu, G., Delaglio, F. and Bax, A. (1999) *J. Biomol. NMR*, **13**, 289–302.

Cornilescu, G., Marquardt, J.L., Ottiger, M. and Bax, A. (1998) *J. Am. Chem. Soc.*, **120**, 6836–6837.

Doreleijers, J.F., Mading, S., Maziuk, D., Sojourner, K., Yin, L., Zhu, J., Markley, J.L. and Ulrich, E.L. (2003) *J. Biomol. NMR*, **26**, 139–146.

Eisenhaber, F., Lijnzaad, P., Argos, P., Sander, C. and Scharf, M. (1995) *J. Comput. Chem.*, **16**, 273–284.

Evenas, J., Tugarinov, V., Skrynnikov, N.R., Goto, N.K., Muhandiram, R. and Kay, L.E. (2001) *J. Mol. Biol.*, **309**, 961–974.

Gardner, K.H., Zhang, X.C., Gehring, K. and Kay, L.E. (1998) *J. Am. Chem. Soc.*, **120**, 11738–11748.

Garrett, D.S., Seok, Y.J., Liao, D.I., Peterkofsky, A., Gronenborn, A.M. and Clore, G.M. (1997) *Biochemistry*, **36**, 2517–2530.

Garrett, D.S., Seok, Y.J., Peterkofsky, A., Gronenborn, A.M. and Clore, G.M. (1999) *Nat. Struct. Biol.*, **6**, 166–173.
Grzesiek, S. and Bax, A. (1993) *J. Biomol. NMR*, **3**, 185–204.
Hansen, M.R., Mueller, L. and Pardi, A. (1998) *Nat. Struct. Biol.*, **5**, 1065–1074.
Hus, J.C., Prompers, J.J. and Bruschweiler, R. (2002) *J. Magn. Reson.*, **157**, 119–123.
Jain, N.U., Noble, S. and Prestegard, J.H. (2003) *J. Mol. Biol.*, **328**, 451–462.
Jung, Y.S. and Zweckstetter, M. (2004) *J. Biomol. NMR*, **30**, 11–23.
Koradi, R., Billeter, M. and Wuthrich, K. (1996) *J. Mol. Graph.*, **14**, 51–&.
Liao, D.I., Silverton, E., Seok, Y.J., Lee, B.R., Peterkofsky, A. and Davies, D.R. (1996) *Structure*, **4**, 861–872.
Losonczi, J.A., Andrec, M., Fischer, M.W.F. and Prestegard, J.H. (1999) *J. Magn. Reson.*, **138**, 334–342.
Moseley, H.N.B. and Montelione, G.T. (1999) *Curr. Opin. Struct. Biol.*, **9**, 635–642.
Mueller, G.A., Choy, W.Y., Yang, D.W., Forman-Kay, J.D., Venters, R.A. and Kay, L.E. (2000) *J. Mol. Biol.*, **300**, 197–212.
Prestegard, J.H. and Kishore, A.I. (2001) *Curr. Opin. Chem. Biol.*, **5**, 584–590.
Pristovsek, P., Ruterjans, H. and Jerala, R. (2002) *J. Comput. Chem.*, **23**, 335–340.
Sharff, A.J., Rodseth, L.E. and Quiocho, F.A. (1993) *Biochemistry*, **32**, 10553–10559.
Skrynnikov, N.R. and Kay, L.E. (2000) *J. Biomol. NMR*, **18**, 239–252.
Spera, S. and Bax, A. (1991) *J. Am. Chem. Soc.*, **113**, 5490–5492.
Tian, F., Valafar, H. and Prestegard, J.H. (2001) *J. Am. Chem. Soc.*, **123**, 11791–11796.
Vijaykumar, S., Bugg, C.E. and Cook, W.J. (1987) *J. Mol. Biol.*, **194**, 531–544.
Warren, J.J. and Moore, P.B. (2001) *J. Magn. Reson.*, **149**, 271–275.
Wüthrich, K. (2003) *Angew. Chem.-Int. Edit.*, **42**, 3340–3363.
Xu, X.P. and Case, D.A. (2001) *J. Biomol. NMR*, **21**, 321–333.
Xu, X.P. and Case, D.A. (2002) *Biopolymers*, **65**, 408–423.
Yang, D.W., Venters, R.A., Mueller, G.A., Choy, W.Y. and Kay, L.E. (1999) *J. Biomol. NMR*, **14**, 333–343.
Zhang, F.L. and Bruschweiler, R. (2002) *J. Am. Chem. Soc.*, **124**, 12654–12655.
Zweckstetter, M. (2003) *J. Biomol. NMR*, **27**, 41–56.
Zweckstetter, M. and Bax, A. (2000) *J. Am. Chem. Soc.*, **122**, 3791–3792.
Zweckstetter, M. and Bax, A. (2001) *J. Am. Chem. Soc.*, **123**, 9490–9491.
Zweckstetter, M. and Bax, A. (2002) *J. Biomol. NMR*, **23**, 127–137.
Zweckstetter, M., Hummer, G. and Bax, A. (2004) *Biophys. J.*, **86**, 3444–3460.