Journal of Biomolecular NMR (2005) 33: 83–93
DOI 10.1007/s10858-005-2601-7

Article

Estimating the accuracy of protein structures using residual dipolar couplings

Katya Simon, Jun Xu, Chinpal Kim & Nikolai R. Skrynnikov*
Department of Chemistry, Purdue University, West Lafayette, IN, 47907, U.S.A

Received 7 February 2005; Accepted 5 August 2005

Key words: NMR, PALES software, PDZ2 domain from human phosphatase hPTP1E, precision and accuracy of protein structures, residual dipolar couplings, X-ray crystallography

Abstract

It has been commonly recognized that residual dipolar coupling data provide a measure of quality for protein structures. To quantify this observation, a database of 100 single-domain proteins has been compiled where each protein was represented by two independently solved structures. Backbone ^{1}H–^{15}N dipolar couplings were simulated for the target structures and then fitted to the model structures. The fits were characterized by an *R*-factor which was corrected for the effects of non-uniform distribution of dipolar vectors on a unit sphere. The analyses show that favorable $\tilde{R}$ values virtually guarantee high accuracy of the model structure (where accuracy is defined as the backbone coordinate rms deviation). On the other hand, unfavorable $\tilde{R}$ values do not necessarily suggest low accuracy. Based on the simulated data, a simple empirical formula is proposed to estimate the accuracy of protein structures. The method is illustrated with a number of examples, including PDZ2 domain of human phosphatase hPTP1E.

Introduction

Protein structures solved by NMR spectroscopy currently account for 15% of all depositions in the RCSB Protein Data Bank (Berman et al., 2000). In contrast to X-ray crystallography, NMR spectroscopy lacks simple means for assessing the quality of the obtained structures. This issue is of importance for studies that require high level of resolution such as studies of enzyme mechanisms or binding specificities. Overestimating the quality of structures in this context "can be misleading, wasteful, and costly" (Zhao and Jardetzky, 1994).

Protein structures can be characterized in terms of *precision*, *indirectly estimated accuracy*, and *directly estimated accuracy*. Precision, according to a commonly accepted definition, refers to the reproducibility of the calculated structure. Accuracy, on the other hand, describes the agreement between the calculation and the true protein structure.

Precision of an NMR structure is usually expressed through the rmsd of atomic coordinates calculated over the ensemble of low-energy structures, i.e. represents the 'thickness of a bundle'. Other measures include torsion angle rmsd and quantities characterizing the input data, e.g. number of restraints per residue. In this discussion, we use the standard measure of precision expressed in the units of angstroms. Generally, precision provides a lower bound for the accuracy. Aside from that, the correlation between precision and accuracy is poor (Zhao and Jardetzky, 1994; Chalaoux et al., 1999). The main reason for that is the presence of errors in the experimental restraints used for structure calculations. For instance, interpretation of the NOE data suffers from lack of information on internal dynamics (in particular

*To whom correspondence should be addressed. E-mail: nikolai@purdue.edu

with respect to mobile side chains (Fejzo et al., 1991; Smith et al., 1993)) and, consequently, from difficulties in treating spin diffusion. Errors in the input data lead to distortions in the calculated structures. Since NMR restraints tend to be sparse, these distortions do not necessarily cause inconsistencies in structure calculations. As a result, the obtained structures can be well-defined (i.e. precise), yet inaccurate. Finally, it should also be noted that precision is highly sensitive to the details of the refinement protocols and software used (Doreleijers et al., 1998; Spronk et al., 2002, 2003).

Indirectly estimated accuracy is based on a comparison of a given structure with a structural database. Goodness of the structure is judged in this case by using the criteria such as the packing quality or compliance with the Ramachandran map (Vriend, 1990; Laskowski et al., 1993). The disadvantage of this approach is that unusual structures can be unfairly penalized. Furthermore, the reference databases are typically comprised of high-resolution crystallographic structures which may cause a certain bias against NMR structures (e.g. with respect to side-chain disorder which is more pronounced under the conditions of NMR experiments).

Finally, *directly estimated accuracy* is based on experimental data from the protein of interest. The structural model is tested against any independently obtained data that have not been used in the structure calculations. For example, good agreement between independently solved NMR and X-ray coordinates indicates high fidelity of both structures. NMR observables used for structure validation (or cross-validation) include NOEs (Brunger et al., 1993; Bonvin and Brunger, 1995), chemical shifts (Williamson et al., 1995), and residual chemical shifts (Cornilescu et al., 1998).

The validation procedure can be calibrated to quantitatively estimate the accuracy. For example, Williamson and co-workers calculated proton chemical shifts for a number of crystallographic structures. They found that high-resolution structures tend to produce a better agreement between the calculated and experimentally observed shifts. A linear relationship has been proposed to characterize this (tentative) trend (Williamson et al., 1995).

The potential of residual dipolar couplings for structure validation was recognized early on (Tjandra and Bax, 1997; Clore and Garrett, 1999; Pääkkönen et al., 2000; Bewley, 2001; Schwalbe et al., 2001; Spronk et al., 2002; Clore and Kuszewski, 2003; Tugarinov and Kay, 2003). The key advantage of the residual dipolar coupling (RDC) data is the possibility for simple and highly accurate structural interpretation. The consistency between the RDC dataset and the structural model can be assessed following a straightforward fit of the alignment tensor. Furthermore, in the case of well-defined structures, the alignment tensor can be accurately predicted from first principles (Zweckstetter and Bax, 2000). In this situation the potential for quantitative estimation of accuracy is obvious (Bax, 2003).

In this paper we develop the tools for estimating the true accuracy of protein structures based on RDC data. We begin by compiling a database of 100 proteins where each protein is represented by two sets of coordinates: crystallographic and NMR. In each pair we define a 'true' and a 'model' structure; the accuracy of the 'model' is described by the root-mean-square deviation between the backbone coordinates of the two structures. RDC data are simulated for the 'true' structure using the program PALES (Zweckstetter and Bax, 2000; Zweckstetter et al., 2004) and then fitted to the 'model' structure. To characterize the goodness of fit we employ a modified version of an *R*-factor (Clore and Garrett, 1999) which is empirically corrected for the effects of non-uniform distribution of dipolar vectors on a unit sphere, $\tilde{R}$.

The empirical relationship between $\tilde{R}$ and accuracy can be used to estimate the accuracy of protein structures that have been solved without the use of RDC data. For example, using experimental RDC data for ubiquitin (Cornilescu et al., 1998) we estimated the accuracy of the NOE-based structure 1G6J (Babu et al., 2001) to be better than 1.8 Å. Another recently reported ubiqutin structure, 1XQQ (Lindorff-Larsen et al., 2005), is shown to have the accuracy better than 1.05 Å. A number of other examples are presented, including RDC dataset recorded on the PDZ2 domain of human phosphatase hPTP1E (Kozlov et al., 2000).

Materials and methods

Structural database

A small structural database has been compiled by a random search of the Protein Data Bank

(Berman et al., 2000). Included in the database are 100 proteins and protein domains for which both X-ray and NMR structures are available. Only single-domain entities were selected as verified by the programs 3Dee (Dengler et al., 2001) and DOMAK (Siddiqui and Barton, 1995). In the case of the NMR structures, it was required that they were determined without the use of RDC data.

Secondary structure classification for all coordinate sets was obtained from the Protein Data Bank (determined by the standard method (Kabsch and Sander, 1983)). For each protein (domain) we noted the number of the first and last residues belonging to the regular secondary structure elements, n_f and n_l. Consequently, for each pair of structures we defined the 'consensus' region which extended from $N_f = \max(n_f^{\text{X-ray}}, n_f^{\text{nmr}})$ to $N_l = \min(n_l^{\text{X-ray}}, n_l^{\text{nmr}})$. Residues with numbers less than $N_f - 3$ or greater than $N_l + 3$ were deleted from the structures. In this manner we focused on the well-structured regions and eliminated long unstructured tails (which often differed between the X-ray and NMR structures) as well as, in some cases, extraneous domains. Cases where both the X-ray and NMR structures represented multi-domain entities (e.g. lysozyme) were omitted from consideration because of the uncertainty in domain identification.

Of all X-ray/NMR structure pairs selected in this fashion (see Supplementary Materials), 72 pairs have identical amino acid sequences whereas the remaining 28 differ by up to three point mutations. The backbone coordinates for all structures in the database were complete and contained no gaps. In the situations when a single NMR structure matched several X-ray structures, we selected the X-ray structure with the least number of mutations, highest resolution, and the most recent publication date (in this particular order). In the case of proteins with an obvious structural homology (such as cytochromes) only one NMR/X-ray pair was included in the database.

Single-bond ^{1}H–^{15}N residual dipolar couplings were simulated and analyzed for residues in the 'consensus' region. It was required that the simulated dataset included at least 50 couplings, corresponding to a minimum of 50 non-proline amino acids in the 'consensus' region.

RDC simulations using PALES

X-ray structures described above were protonated using the program MOLMOL (Koradi et al., 1996) and subsequently used to simulate RDC data using the program PALES (Zweckstetter and Bax, 2000; Zweckstetter et al., 2004). For these simulations we selected the commonly used liquid crystalline media composed of *n*-dodecyl-penta(ethylene glycol) (C12E5) and *n*-hexanol (Rückert and Otting, 2000). The calculations were carried out assuming that C12E5/*n*-hexanol mixture forms planar bilayers with the thickness of 28.6 Å (Freyssingeas et al., 1996). The liquid crystal concentration was calculated assuming C12E5/water ratio 5 wt%, C12E5/*n*-hexanol molar ratio 0.96, and 0.3 wt% proportion of free *n*-hexanol in solution (Freyssingeas et al., 1996). It is worth noting that bicelle calculations in PALES have been programmed for DMPC/DHPC mixture so that the program implicitly accounts for the presence of 5 mM of free DHPC in the solvent (Ottiger and Bax, 1998). This was compensated for by introducing a correction into PALES input parameters (specifically, the value 63 mg/ml was used for the liquid crystal concentration). The degree of alignment for C12E5/*n*-hexanol bicelles was assumed to be the same as for DMPC/DHPC (Rückert and Otting, 2000). The magnitude of dipolar couplings simulated in this manner was found to be realistic (typically, up to 20–30 Hz). Note that fine details of the liquid crystalline media are unimportant in our approach so long as the mechanism of steric alignment is correctly modeled by PALES.

RDC fitting and structural parameters

The simulated ^{1}H–^{15}N RDC data were fitted to the NMR coordinates. First, the structures in the NMR ensemble were superimposed with respect to backbone atoms from the 'consensus' region. The entire ensemble was then used to fit the dipolar couplings (Ottiger et al., 1998; Lindorff-Larsen et al., 2005):

$$D_i^{\text{fit}} = (1/N_{\text{nmr}}) \sum_{n=1}^{N_{\text{nmr}}} D_0^{\text{NH}} A_a \{(3\cos^2\theta_{i,n} - 1) + (3/2)\eta \sin^2\theta_{i,n} \cos 2\phi_{i,n}\} \quad (1)$$

where index i refers to the residue number, and the summation is over all members of the ensemble (on average, the size of the ensemble was $N_{nmr} = 21$). It has been assumed that a single alignment tensor is in effect for the entire ensemble. This seems reasonable, since individual structures in the ensemble have a very similar overall shape. Furthermore, it can be argued that individual structures roughly represent the dynamic states associated with fast backbone motion (Bonvin and Brunger, 1996). These states interconvert rapidly on the time scale of molecular alignment, i.e. on the time scale of protein translation (rotation) relative to the surface of liquid crystal. This observation supports the use of the single alignment tensor.

Equation (1) was incorporated in a standard procedure involving optimization of five alignment parameters (A_a, η, plus three Euler angles that define the orientation of the alignment axes in the coordinates of NMR ensemble). Although the majority of the fits were of poor quality (see below) they were nonetheless completely stable and consistent between different numeric algorithms such as simplex search (Tjandra and Bax, 1997) and singular value decomposition (Losonczi et al., 1999). Included in the fitting procedure was the calculation of the R-factor:

$$R = \mathrm{rms}(D - D^{\mathrm{fit}})/\{D_0^{\mathrm{NH}} A_a^{\mathrm{fit}} \sqrt{(4 + 3\eta^{\mathrm{fit}\,2})/5}\}. \quad (2)$$

This expression differs by $\sqrt{2}$ from the original definition (Clore and Garrett, 1999). In a system with ideal isotropic distribution of dipolar vectors, the denominator of Equation (2) corresponds to $\mathrm{rms}(D^{\mathrm{fit}})$. In the case of near-isotropic distribution and a good structural model, R is approximately equal to the quality factor Q, $Q = \mathrm{rms}(D - D^{\mathrm{fit}})/\mathrm{rms}(D)$(Ottiger and Bax, 1999).

In order to account for deviations from isotropic orientation sampling, we calculated for each NMR structure the generalized sampling parameter Ξ (Fushman et al., 2000):

$$\Xi = (1/6) \sum_{\substack{i=\{x,y,z\} \\ j=\{x,y,z\}}} (3\overline{v_i v_j} - \delta_{ij})^2 \quad (3)$$

where v_k are the coordinates of the normalized ^{1}H–^{15}N vectors in the (arbitrary) molecular frame of reference, overbar denotes averaging over all ^{1}H–^{15}N vectors from the 'consensus' region and, on top of it, averaging over all structures in the ensemble, and δ_{ij} is the Kronecker delta symbol. This parameter was used to modify, in empirical fashion, the definition of R:

$$\tilde{R} = R/(1 - \Xi). \quad (4)$$

Atomic coordinate rms deviations were calculated for backbone atoms within the 'consensus' region using the standard set of rules (Zhao and Jardetzky, 1994). In the same vein, rms deviations were determined for orientations of the ^{1}H-^{15}N vectors. First, the mean 'model' was defined by averaging the orientations of ^{1}H-^{15}N bonds over all structures in the NMR ensemble. This 'model' was superimposed on the 'true' structure by minimizing the rms angle between the ^{1}H-^{15}N bonds (within the consensus region). The resulting angular deviation, armsd(model, true), has the units of degrees. No attempt was made to simulate RDC measurement errors since they are usually small and have only minor significance compared to structural differences (Skrynnikov et al., 2000).

Protein production and NMR spectroscopy

Preparation of the sample of PDZ2 domain from human phosphatase hPTP1E and residual dipolar coupling measurements are described in the Supplementary Materials.

Results and discussion

Overview of protein database

In order to quantify the relationship between the quality of the RDC fit and the accuracy of the structure we compiled a small database of 100 single-domain proteins and isolated domains. The selection criteria used in building this database are described in Materials and methods. A special effort has been made to avoid multi-domain proteins. An example of potential complications associated with multi-domain proteins is the situation when a structural model incorrectly predicts the relative position of the domains. While the overall accuracy of such a structural model must be low, it may still produce a good RDC fit.

A realization of such a model can be readily obtained by taking the original structure and performing 'crankshaft' rotations, $\psi_i = \psi_i + \alpha$, $\phi_{i+1} = \phi_{i+1} - \alpha$, in residues i and $i+1$ from the linker region. This transformation achieves the translation of the C-terminal domain and leaves the dipolar couplings invariant. In contrast, single-domain proteins are not expected to suffer from such problems: the structure in this case is held together by a meshwork of NOEs which prevent large translational or rotational displacements. Another reason to avoid multi-domain proteins is that interdomain motion can have a dramatic effect on RDCs (Jacobs et al., 2003).

Initially, we analyzed our 100-protein database in terms of precision (for NMR structures) and agreement between NMR and X-ray structures (which can be viewed, tentatively, as a measure of accuracy). Figure 1 illustrates the correlation between these two parameters. Both quantities are calculated for heavy backbone atoms within the 'consensus' portion of the polypeptide chain (i.e. excluding unstructured tails).

All but three points in the plot Figure 1 fall below the line $y = x$. This is consistent with the well-known observation that the accuracy of an NMR structure cannot be better than its precision (while there are exceptions to this rule, they are statistically rare). More interestingly, for half of the points the accuracy is better than 2.09×precision, in agreement with earlier analyses (Clore and Gronenborn, 1998). Generally, the correlation between precision and accuracy observed in Figure 1 is rather poor. For example, precision of ca. 0.5 Å corresponds to the values of rmsd(*NMR*,*X-ray*) anywhere in the range from 0.6 to 3.3 Å. This is in line with the general observation that precision cannot be used as a reliable predictor of accuracy (Zhao and Jardetzky, 1994; Chalaoux et al., 1999).

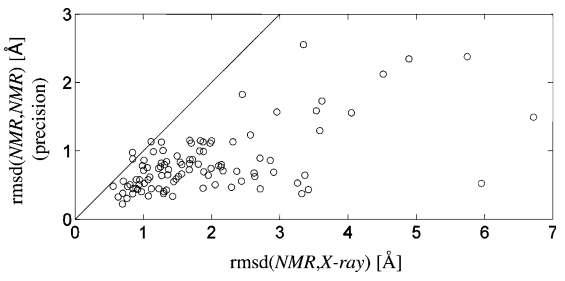

Figure 1. Precision of NMR ensemble vs. deviation between NMR and X-ray structures. Note that rmsd(*NMR*,*X-ray*) is indicative of accuracy but cannot literally be interpreted as accuracy of NMR structure.

We have also analyzed the 100-protein database with an eye on the progress of structure determination methods over the course of years. The results are summarized in Table 1.

The results for the rmsd(*NMR*,*X-ray*) (typeset in bold in Table 1) are somewhat surprising. It appears that the agreement between NMR and X-ray structures has become progressively worse over the course of years. This situation cannot be explained by the increase in the size of studied protein systems. Indeed, the average size of NMR systems shows only modest increase with time (third column in Table 1). In the case of X-ray data, the resolution of the structures has remained approximately constant over the years (right-most column in Table 1), in line with a recent statistical survey (Kleywegt and Jones, 2002).

In some cases large deviations between NMR and X-ray structures can be attributed to vastly different sample conditions (e.g. pH) or the use of different domain constructs. To examine the effect of these outliers, we removed the two proteins with the worst rmsd(*NMR*,*X-ray*) from each of the three groups listed in Table 1. This, however, did not change the trend of increasing divergence between the NMR and X-ray models. We feel that this problem warrants a special study which falls outside the scope of the present work. In this connection note that the results shown in Table 1 are specific to our 100-protein database. In particular, this database does not include recent NMR structures solved with the use of RDC data.

Analyses of simulated RDC data

The database of 100 X-ray/NMR structure pairs was used for the RDC simulation study. The X-ray coordinates were designated as 'true' structures and used to simulate RDC data (Zweckstetter and Bax, 2000). The simulated data were subsequently fitted to the 'model' NMR structures. In this manner two key parameters were correlated: (i) the accuracy of the model expressed in the units of angstroms, rmsd(*true*,*model*), and (ii) *R*-factor characterizing the goodness of the RDC fit, $\tilde{R}$.

Our approach immediately raises a question as to whether we believe that X-ray structures are superior to their NMR counterparts. The simple

Table 1. Structural statistics for 100-protein NMR/X-ray database

Year NMR structure solved[a]	Number of proteins	Average system size (a.a.) (NMR)[b]	Precision of NMR structure[c] (Å)	rmsd(*NMR*,*X-ray*)[c] (Å)	Resolution of X-ray structure (Å)
1988–1996	33	110	0.79	**1.44**	1.82
1997–1998	37	134	0.91	**1.97**	2.02
1999–2003	30	129	0.75	**2.37**	1.93

[a]Deposition date as recorded in Protein Databank.
[b]Size of the protein system as investigated in the original NMR study and recorded in Protein Databank. This parameter reflects full size of protein complexes or oligomers as occurred in NMR samples. In our analyses, complex structures were dissected into single-domain units, as described in Materials and methods.
[c]Evaluated for heavy backbone atoms (N, C^{α}, C′, and O) from the 'consensus' region.

answer is that the question is irrelevant. It is only important in our approach to have a pair of reasonable structures which can be characterized by pair-wise rmsd and $\tilde{R}$. The origin of structures does not matter. To demonstrate this fact, we reversed the simulation scheme so that the NMR structures were classified as 'true' and used as input for PALES simulations. The results (not shown) are consistent with those derived from our standard scheme (shown below). Likewise, compatible results were obtained from a small trial set of X-ray/X-ray pairs.

On a more fundamental level, however, we do rely on the fact that X-ray coordinates of (compact and well-structured) protein domains represent a very good approximation to the true solution-state structures. The superb quality of X-ray structures has been confirmed, in particular, by many independent NMR measurements (Smith et al., 1993; Cornilescu et al., 1998; Bax, 2003). It should be stressed that these observations are limited to backbone coordinates of small well-structured proteins (domains) and do not apply to flexible side chains (Ottiger et al., 1998b; Mittermaier and Kay, 2001; Lindorff-Larsen et al., 2005), disordered proteins (Louhivuori et al., 2003; Mohana-Borges et al., 2004), etc.

Although R, Equation 2, is a good measure for quality of the fit, the results may vary depending on how well the ^{1}H-^{15}N vectors sample different orientations on a unit sphere. For example, in a helical bundle protein ^{1}H-^{15}N vectors are predominantly oriented along one direction which roughly coincides, in the case of steric alignment media, with the long axis of the alignment tensor. For this particular orientation RDCs are relatively insensitive to variations in directional angles θ and ϕ. As a result, even a poor model can produce a good (i.e. low) R score (Bax, 2003).

Since the original R-factor is normalized assuming isotropic distribution of dipolar vectors, we used the sampling parameter Ξ to impose an additional 'penalty' on the structures with poor orientation sampling, Equations 2–4. $\Xi = 0$ for the set of vectors uniformly distributed on a unit sphere and $\Xi = 1$ if all vectors are oriented in the same direction (Fushman et al., 2000). For the majority of proteins in our database the difference between the original and modified factors, R and $\tilde{R} = R/(1 - \Xi)$, was found to be less than 5%. As anticipated, however, large differences occurred in the four helical bundle proteins.

Since $\tilde{R}$ was introduced in an *ad hoc* manner, Equation 4, we performed a separate series of simulations to put this parameter on firmer ground. In brief, we examined the response of R to random variations in orientation of ^{1}H-^{15}N vectors. It turned out that the four helical bundles with the largest Ξ values were the least sensitive to structural perturbations (specifically, they were among the 10% of structures with the lowest R). This observation supports the use of Ξ as a correction factor for R. In general, it should be recognized that Ξ provides only a rough measure of orientational sampling. Nonetheless, Ξ (and $\tilde{R}$) are deemed adequate for the purpose of the present analyses.

Figure 2 illustrates the relationship between the $\tilde{R}$ and rmsd(*true*,*model*) found in 100-protein database. The results are obtained on the basis of Equations 1–4 using the PALES-simulated RDC data. The dependence takes a form of a statistical distribution which is similar, for example, to the Ramachandran plot. In the same spirit, the populated region can be delineated in the $\tilde{R}$ vs.

rmsd(*model*,*true*) map. Two empirical curves delimiting this region are shown in the figure.

A salient feature of Figure 2 is a wide spread of points along the *y*-axis and, importantly, a presence of many accurate models with unexpectedly high $\tilde{R}$. For instance, NMR model 1KDE (Sonnichsen et al., 1996) has excellent rmsd(*model*,*true*) = 0.8 Å, but relatively poor $\tilde{R} = 0.81$. This situation does not change when the analysis is restricted to secondary structure elements or when the geometry of peptide planes is regularized. It is important to understand the reasons for this behavior.

The RDC *R*-factor, $\tilde{R}$, directly depends on the orientation of the ^{1}H-^{15}N vectors in the model structure which is in turn correlated with the accuracy of the backbone fold. Using angular rmsd (armsd) to characterize the deviation in the ^{1}H-^{15}N bond orientations, we arrive to the following schematic relationship:

$$\tilde{R} \rightarrow \text{armsd}(model, true) \rightarrow \text{rmsd}(model, true).$$

Let us first consider the correlation between $\tilde{R}$ and armsd(*model*,*true*). As illustrated in Figure 3, there is a substantial uncertainty associated with this relationship. For instance, the models 1KDE and 1JWE (Weigelt et al., 1999) have similar armsd values, 22° and 23°, but widely different $\tilde{R}$, 0.81 vs. 0.39. In fact, the value of $\tilde{R}$ is influenced by individual details of the structure in relation to the alignment tensor. In the case of 1KDE, unusually high $\tilde{R}$ is partly due to a single incorrectly oriented ^{1}H-^{15}N vector that happens to give rise to a large difference in dipolar couplings. In addition, the degree of alignment A_a^{fit} turns out to be significantly underestimated (Zweckstetter and Bax, 2002), which leads to further increase in $\tilde{R}$. In turn, 1JWE also contains a few incorrectly oriented ^{1}H-^{15}N vectors which all occur in one hydrogen-bonded turn. Incidentally, these vectors produce only modest differences in dipolar couplings, although they contribute heavily to the armsd. In this situation, A_a^{fit} is close to its 'true' value and $\tilde{R}$ proves to be lower than could be expected.

Next, consider the relationship between armsd and rmsd. As noted before (Doreleijers et al., 1998), the correlation between translational and orientational degrees of freedom in a polypeptide chain is rather loose (see Figure S1). Since torsional angles ψ and ϕ afford considerable freedom, it is possible to build a structural model with an accurate global fold, but bad orientations of the peptide planes. An example of such a structure is 1FRA (Hatanaka et al., 1994), where rmsd(*model*,*true*) = 1.3 Å, armsd(*model*,*true*) = 45°.

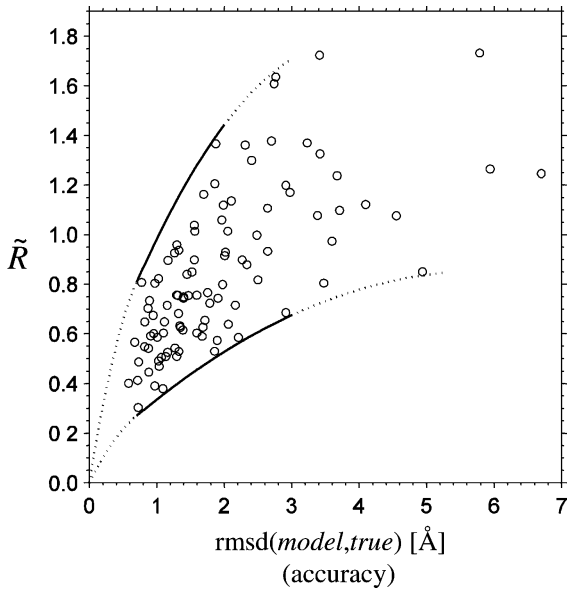

Figure 2. Correlation between the modified *R*-factor $\tilde{R}$ and the accuracy of the structural model as established on the basis of PALES simulations. $\tilde{R}$ is calculated for ^{1}H-^{15}N couplings, rmsd(*model*,*true*) is calculated for backbone atom coordinates (N, C$^{\alpha}$, C′, O, H^{N}, and H$^{\alpha}$); both quantities are restricted to the 'consensus' region in the protein. The populated region is delineated by two empirical curves: rmsd = $3.86 - (18.34 - 10.31\tilde{R})^{1/2}$ (upper branch, defined over the interval 0.7Å < rmsd < 2.0Å) and rmsd = $5.85 - (38.77 - 45.45\tilde{R})^{1/2}$ (lower branch, 0.7 Å < rmsd < 3.0 Å).

In summary, the presence of structures with low rmsd(*model*,*true*) but high $\tilde{R}$ can be explained by the combination of the two factors: (i) individual structural features which make $\tilde{R}$ sensitive to small structural changes and (ii) poor local conformation of the model. Note that the first reason can be prevalent, such as in the case of 1KDE where the quality of local conformation is reasonably good. Thus, a moderately high $\tilde{R}$ value should not be held against the structure.

In addition to $\tilde{R}$ we have investigated a number of other possible figures of merit. In each case we observed the same pattern as in Figure 2 and obtained similar estimates of accuracy for selected protein structures (cf. next section). For instance, we considered an alternative definition of the *R*-factor where RDC data are fitted to the

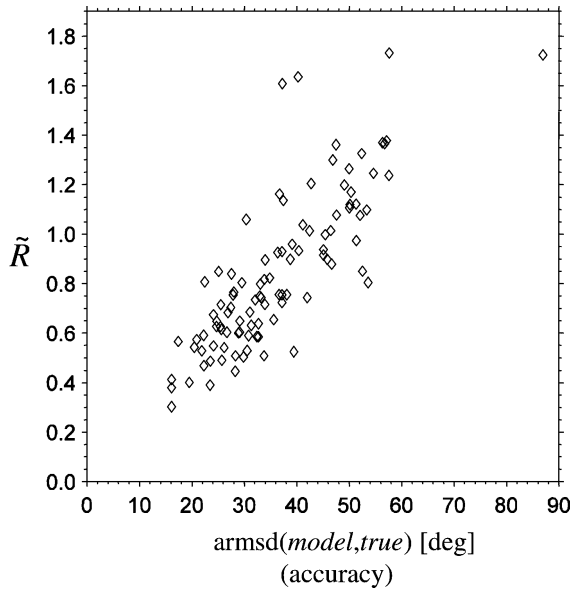

Figure 3. Correlation between $\tilde{R}$ and the accuracy of the structural model as assessed on the basis of the ^{1}H-^{15}N bond orientations. armsd(*model*,*true*) has been obtained by minimizing the angular rms deviation between the two sets of ^{1}H-^{15}N vectors (restricted to the 'consensus' region). Alternative approach, where two structures are first superimposed in a standard fashion and then used to directly calculate armsd, leads to similar values. Note that armsd does not have the same fundamental importance for characterizing the structure quality as rmsd. The correlation coefficient for the data in the plot is $r = 0.81$.

individual structures from the NMR ensemble (this led to the average increase in $\tilde{R}$ of 0.28). In another example, we used mean NMR structures which were generated by overlaying the 'consensus' portions of the backbone (average increase in $\tilde{R}$ of 0.17). While these alternative versions of our treatment are sound, we prefer the definition used in this text, Equations 1–4. Indeed, the best way to judge the quality of the NMR model is to consider the ensemble in its entirety rather than dealing with its individual members or a questionable mean structure.

We also tested the quality factor, Q, as a potential alternative to R. It has been found that the main advantage of R stems from its wide 'dynamic range'. Poor models that bear no resemblance to the 'true' structures give rise to low A_a^{fit} values (Zweckstetter and Bax, 2002) and, consequently, to high R (average value 4.70 was obtained when our database was re-analyzed using random 'models'). This characteristic makes R a more sensitive probe for medium- to low-quality structures. In contrast, Q is restricted to the range from 0.0 to 1.0 and, in fact, can approach the upper limit even for reasonably good models. For instance, the model 1B4M (Lu et al., 1999) with rmsd of 1.9 Å gives rise to $Q = 0.91$.

A special series of calculations was carried out to determine the role of the dataset size. Clearly, in the case of small RDC datasets, on the order of ten couplings, $\tilde{R}$ values are loaded with a large statistical uncertainty. In our model calculations, the simulated RDC data (more than 50 couplings per set) were all reduced to the same size (40 couplings per set) by randomly discarding a fraction of data. The fitting procedure using the reduced datasets was used to regenerate the diagram from Figure 2. It was found that the use of smaller samples led to only slight declines in $\tilde{R}$ (on average, by 0.04) and minimal changes in Figure 2. This result suggests that 50 couplings constitute an adequate sample for the problem at hand (Zweckstetter and Bax, 2002).

The result presented in Figure 2 would clearly benefit from the use of a larger database. It should be pointed out, however, that our search for X-ray/NMR pairs appeared to be nearly exhaustive (under the set of rules described in Materials and methods). The region with very low $\tilde{R}$ and rmsd(*model*,*true*) can be, in principle, sampled by considering X-ray/X-ray pairs, Figure S2. This area, however, is of little relevance for analyses of NMR structures since they never seem to achieve very low $\tilde{R}$ values.

Estimating the accuracy of NMR structures

The *simulated* map Figure 2 can be combined with *experimental* RDC data to assess the accuracy of NMR structures. Strictly speaking, this approach is valid only for structures that have been determined without the use of RDCs since the diagram Figure 2 was generated under this assumption. It is clearly futile to attempt structure validation using the same ^{1}H-^{15}N RDC dataset that was previously used in structure refinement (the situation where RDC data are partitioned into two subsets one of which is used for structure calculation and the other for validation (Clore and Garrett, 1999) requires additional investigation).

We have assembled several literature examples to demonstrate applications of the method and, in addition, collected our own RDC data on the PDZ2 domain of human phosphatase hPTP1E.

The *experimental* RDC data were fitted to NMR structures and the resulting $\tilde{R}$ values were directly translated into estimates of accuracy with the help of the curves shown in Figure 2.

The inspection of the results, Table 2, suggests that the upper bound for the accuracy is more useful of the two. For instance, for recently reported ubiquitin structure 1XQQ (Lindorff-Larsen et al., 2005) our analysis claims the accuracy *better than* 1.05 Å.

The lower bound, on the other hand, is not particularly restrictive. For example, the structure 3PDZ (Kozlov et al., 2000) is consistent with the accuracy of 1.6 Å, which is reasonable by the standards of NMR spectroscopy. In general, the following cautious interpretation can be offered: *a low $\tilde{R}$ value guarantees high quality of the structure, whereas a moderately high $\tilde{R}$ value does not reflect on structure quality*.

The obtained estimates of accuracy are based on the boundary curves indicated in Figure 2. These curves are defined over the limited intervals (see caption of Figure 2) so that in some cases, e.g. entry 1MUT (Abeygunawardana et al., 1995) in Table 2, no estimates are available. In other cases, such as 1ACA (Kragelund et al., 1993), the estimates can be trivial as they coincide with the lower bound imposed by the precision of the NMR structure. Nevertheless, in most cases the estimates proved to be informative.

In addition to the proteins listed in Table 2 we came across several other examples that were not included in the table. For instance 1AK8 (Bentrop et al., 1997) appears to have the accuracy better than 2.0 Å. This estimate, however, is based on the RDC data (Chou et al., 2001) that have been recorded in Pf1 phage media with mainly electrostatic alignment mechanism. While we believe that the results of Figure 2 are quite general, additional simulations are recommended to address the case of electrostatic alignment. Another example is zinc-substituted rubredoxin 1ZRP (Blake et al., 1992) with the predicted accuracy better than 1.7 Å. The estimate, however, is based on a limited set of 41 dipolar couplings (Tian et al., 2001). Here we choose to maintain the cutoff level of 50 couplings. One should keep in mind that the notion of accuracy in this context is different from resolution of crystallographic structures. For instance, in the small control group of X-ray/X-ray pairs, Figure S2, the average accuracy of the 'models' is 0.4 Å, whereas the average crystallographic resolution is 2.1 Å.

Table 2. Estimates of accuracy for selected NMR structures based on the quality of RDC fitting

NMR structure PDB code[a]	Precision of NMR structure[b] (Å)	Independent ^{1}H–^{15}N RDC data (source)	No. of RDCs[c]	Alignment media[d]	$\tilde{R}$[e]	Estimated accuracy[f] (Å)
1XQQ	0.9	Cornilescu et al. (1998)	63	DMPC:DHPC	0.34	≤ **1.05**[g]
1G6J	0.4	Cornilescu et al. (1998)	63	DMPC:DHPC	0.49	≤ **1.8**
3CI2	1.0	Tischenko and Boelens (personal communication)	54	C12E5:octanol	0.54	≤ **2.1**
1SYM	1.0	Drohat et al. (1999)	53	DMPC:DHPC	0.55	≤ **2.2**
1MUT	1.6	Massiah et al. (2003)	53	C8E5:octanol	0.77	–
1ACA	0.8	Lerche et al. (2004)	71	DMPC:DHPC	0.89	**0.8** ≤
1D5G	0.8	This work	69	C12E5:hexanol	0.97	**1.0** ≤
3PDZ	0.3	This work	69	C12E5:hexanol	1.30	**1.6** ≤

[a]The structures have been selected using the same principles as previously used in compiling the database. Specifically, only single-domain proteins or separate domains were included, with a minimum of 50 ^{1}H–^{15}N dipolar couplings measured for residues in the $[n_f - n_l]$ region, where n_f and n_l are the first and the last residues displaying secondary structure in the NMR PDB file. Structures 1SYM, 1MUT, and 1D5G/3PDZ do not have crystallographic equivalents; only structure 3CI2 is a part of the 100-protein database.
[b]Defined with regard to all backbone atom coordinates.
[c]Experimental ^{1}H–^{15}N RDC data from residues in the $[n_f - n_l]$ region.
[d]Steric alignment mechanism.
[e]In evaluating $\tilde{R}$, Equation 4, both R and Ξ were calculated over the subset of residues from the $[n_f - n_l]$ region for which experimental ^{1}H–^{15}N couplings were available.
[f]Obtained from the diagram Figure 2 using the empirical curve parameterizations given in the figure legend.
[g]Coordinate set 1XQQ consists of 128 structures calculated with the use of unconventional restraints (relaxation order parameters). When RDC data are re-analyzed against the 20 structures from the top part of the PDB file, the value of $\tilde{R}$ and the estimate of the accuracy remain unchanged. For comparison, fitting the same data to the X-ray structure 1UBQ results in $\tilde{R} = 0.17$.

Finally, it should be noted that the RDC data in Table 2 have been typically recorded without making any special effort to match the conditions used in structure determination studies (cf. third and first columns in the table). This refers to the ligation state, temperature, etc. If RDCs are measured with the explicit purpose of structure validation then these conditions can be usually matched and, as a result, somewhat lower values of $\tilde{R}$ can be expected.

Concluding remarks

In addition to structural information, residual dipolar couplings also carry information on a protein's internal dynamics (Meiler et al., 2001; Wang et al., 2001). Considering the role of dynamics, our approach is expected to produce useful results so long as the amplitudes of internal motions are small and the single (motion-averaged) structure adequately represents the molecule. The credibility of the 'single structure' description for small globular proteins has been proven by the enormous success of crystallographic studies.

While we use the ensembles of NMR structures to fit the RDC data, the main rationale for this approach is that the ensemble in its entirety represents the best approximation to the true structure. Even though NMR ensembles bear certain resemblance to 'snapshots' of fast protein motion (Bonvin and Brunger, 1996), this aspect is relatively unimportant in the analyses of dipolar couplings (Zweckstetter and Bax, 2002; Ulmer et al., 2003; Clore and Schwieters, 2004) and the concept of a single uniquely defined structure remains valid.

As of the end of 2004, the BioMagResBank (Doreleijers et al., 2003) contained 2276 NOE datasets associated with various PDB structures. In contrast, only 116 RDC datasets have been deposited by this time. Thus, the great majority of NMR structures have been solved without the benefit of RDC data. The newly acquired RDC data can serve as a valuable validation tool for these structures. At the same time, RDC data can be used for the purpose of cross-validation in calculating new protein structures (Clore and Garrett, 1999; Drohat et al., 1999). These applications do not necessarily preclude the constructive use of dipolar couplings: after the (cross-validation) procedure is completed, the data can be employed for structure refinement.

It has been demonstrated that the RDC-based validation procedures can be helpful in development of refinement protocols (Spronk et al., 2002). Here we emphasize the application aimed at the accuracy of protein structures. The procedure presented in this work is based on the analysis of a 100-protein database and is, therefore, subject to the usual statistical limitations. The study can be easily extended to include other types of dipolar couplings such as $^{1}H^{\alpha}$-$^{13}C^{\alpha}$, systems with different alignment mechanisms such as Pf1 phage, and datasets of smaller size. The same approach can be used to assess the accuracy of crystallographic coordinates with respect to the solution state of the protein.

In conclusion, we have established a statistical relationship between the accuracy of a structural model and the quality of the RDC fitting that this model provides. While good fits indicate high quality of the structure, poor fits do not necessarily suggest low quality. In fact, structures with very good accuracy, $\sim$1 Å for backbone atoms, can give rise to poor fits. These observations have been put on a quantitative basis, resulting in an empirical recipe for estimating the accuracy of protein structures solved by NMR spectroscopy.

Electronic supplementary material is available at http://dx.doi.org/10.1007/s10858-005-2601-7

Acknowledgements

We thank Drs. Ad Bax, Markus Zweckstetter, and Geoff Barton for valuable discussions, Drs. Wayne Fairbrother, Eugene Tischenko, and Rolf Boelens for sharing their RDC data, and Drs. Christin Gustafson and Lianchun Fan for help with protein expression.

References

Abeygunawardana, C., Weber, D.J., Gittis, A.G., Frick, D.N., Lin, J., Miller, A.F., Bessman, M.J. and Mildvan, A.S. (1995) *Biochemistry*, **34**, 14997–15005.

Babu, C.R., Flynn, P.F. and Wand, A.J. (2001) *J. Am. Chem. Soc.*, **123**, 2691–2692.

Bax, A. (2003) *Protein Sci.*, **12**, 1–16.

Bentrop, D., Bertini, I., Cremonini, M.A., Forsén, S., Luchinat, C. and Malmendal, A. (1997) *Biochemistry*, **36**, 11605–11618.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) *Nucleic Acids Res.*, **28**, 235–242.

Bewley, C.A. (2001) *J. Am. Chem. Soc.*, **123**, 1014–1015.
Blake, P.R., Park, J.B., Zhou, Z.H., Hare, D.R., Adams, M.W.W. and Summers, M.F. (1992) *Protein Sci.*, **1**, 1508–1521.
Bonvin, A.M.J.J. and Brunger, A.T. (1995) *J. Mol. Biol.*, **250**, 80–93.
Bonvin, A.M.J.J. and Brunger, A.T. (1996) *J. Biomol. NMR*, **7**, 72–76.
Brunger, A.T., Clore, G.M., Gronenborn, A.M., Saffrich, R. and Nilges, M. (1993) *Science*, **261**, 328–331.
Chalaoux, F.R., O'Donoghue, S.I. and Nilges, M. (1999) *Proteins*, **34**, 453–463.
Chou, J.J., Li, S.P., Klee, C.B. and Bax, A. (2001) *Nat. Struct. Biol.*, **8**, 990–997.
Clore, G.M. and Garrett, D.S. (1999) *J. Am. Chem. Soc.*, **121**, 9008–9012.
Clore, G.M. and Gronenborn, A.M. (1998) *Proc. Natl. Acad. Sci. USA*, **95**, 5891–5898.
Clore, G.M. and Kuszewski, J. (2003) *J. Am. Chem. Soc.*, **125**, 1518–1525.
Clore, G.M. and Schwieters, C.D. (2004) *J. Am. Chem. Soc.*, **126**, 2923–2938.
Cornilescu, G., Marquardt, J.L., Ottiger, M. and Bax, A. (1998) *J. Am. Chem. Soc.*, **120**, 6836–6837.
Dengler, U., Siddiqui, A.S. and Barton, G.J. (2001) *Proteins*, **42**, 332–344.
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.
Doreleijers, J.F., Rullmann, J.A.C. and Kaptein, R. (1998) *J. Mol. Biol.*, **281**, 149–164.
Drohat, A.C., Tjandra, N., Baldisseri, D.M. and Weber, D.J. (1999) *Protein Sci.*, **8**, 800–809.
Fejzo, J., Krezel, A.M., Westler, W.M., Macura, S. and Markley, J.L. (1991) *Biochemistry*, **30**, 3807–3811.
Freyssingeas, É., Nallet, F. and Roux, D. (1996) *Langmuir*, **12**, 6028–6035.
Fushman, D., Ghose, R. and Cowburn, D. (2000) *J. Am. Chem. Soc.*, **122**, 10640–10649.
Hatanaka, H., Oka, M., Kohda, D., Tate, S., Suda, A., Tamiya, N. and Inagaki, F. (1994) *J. Mol. Biol.*, **240**, 155–166.
Jacobs, D.M., Saxena, K., Vogtherr, M., Bernado, P., Pons, M. and Fiebig, K.M. (2003) *J. Biol. Chem.*, **278**, 26174–26182.
Kabsch, W. and Sander, C. (1983) *Biopolymers*, **22**, 2577–2637.
Kleywegt, G.J. and Jones, T.A. (2002) *Struct. Fold. Des.*, **10**, 465–472.
Koradi, R., Billeter, M. and Wüthrich, K. (1996) *J. Mol. Graph.*, **14**, 51–55.
Kozlov, G., Gehring, K. and Ekiel, I. (2000) *Biochemistry*, **39**, 2572–2580.
Kragelund, B.B., Andersen, K.V., Madsen, J.C., Knudsen, J. and Poulsen, F.M. (1993) *J. Mol. Biol.*, **230**, 1260–1277.
Laskowski, R.A., Macarthur, M.W., Moss, D.S. and Thornton, J.M. (1993) *J. Appl. Cryst.*, **26**, 283–291.
Lerche, M.H., Kragelund, B.B., Redfield, C. and Poulsen, F.M. (2004). To be published; PDB deposition 1NTI.
Lindorff-Larsen, K., Best, R.B., DePristo, M.A., Dobson, C.M. and Vendruscolo, M. (2005) *Nature*, **433**, 128–132.
Losonczi, J.A., Andrec, M., Fischer, M.W.F. and Prestegard, J.H. (1999) *J. Magn. Reson.*, **138**, 334–342.
Louhivuori, M., Pääkkönen, K., Fredriksson, K., Permi, P., Lounila, J. and Annila, A. (2003) *J. Am. Chem. Soc.*, **125**, 15647–15650.
Lu, J.Y., Lin, C.L., Tang, C.G., Ponder, J.W., Kao, J.L.F., Cistola, D.P. and Li, E. (1999) *J. Mol. Biol.*, **286**, 1179–1195.
Massiah, M.A., Saraswat, V., Azurmendi, H.F. and Mildvan, A.S. (2003) *Biochemistry*, **42**, 10140–10154.
Meiler, J., Prompers, J.J., Peti, W., Griesinger, C. and Bruschweiler, R. (2001) *J. Am. Chem. Soc.*, **123**, 6098–6107.
Mittermaier, A. and Kay, L.E. (2001) *J. Am. Chem. Soc.*, **123**, 6892–6903.
Mohana-Borges, R., Goto, N.K., Kroon, G.J.A., Dyson, H.J. and Wright, P.E. (2004) *J. Mol. Biol.*, **340**, 1131–1142.
Ottiger, M. and Bax, A. (1998) *J. Biomol. NMR*, **12**, 361–372.
Ottiger, M. and Bax, A. (1999) *J. Biomol. NMR*, **13**, 187–191.
Ottiger, M., Delaglio, F., Marquardt, J.L., Tjandra, N. and Bax, A. (1998) *J. Magn. Reson.*, **134**, 365–369.
Pääkkönen, K., Sorsa, T., Drakenberg, T., Pollesello, P., Tilgmann, C., Permi, P., Heikkinen, S., Kilpeläinen, I. and Annila, A. (2000) *Eur. J. Biochem.*, **267**, 6665–6672.
Rückert, M. and Otting, G. (2000) *J. Am. Chem. Soc.*, **122**, 7793–7797.
Schwalbe, H., Grimshaw, S.B., Spencer, A., Buck, M., Boyd, J., Dobson, C.M., Redfield, C. and Smith, L.J. (2001) *Protein Sci.*, **10**, 677–688.
Siddiqui, A.S. and Barton, G.J. (1995) *Protein Sci.*, **4**, 872–884.
Skrynnikov, N.R., Goto, N.K., Yang, D.W., Choy, W.Y., Tolman, J.R., Mueller, G.A. and Kay, L.E. (2000) *J. Mol. Biol.*, **295**, 1265–1273.
Smith, L.J., Sutcliffe, M.J., Redfield, C. and Dobson, C.M. (1993) *J. Mol. Biol.*, **229**, 930–944.
Sonnichsen, F.D., DeLuca, C.I., Davies, P.L. and Sykes, B.D. (1996) *Struct. Fold. Des.*, **4**, 1325–1337.
Spronk, C.A.E.M., Linge, J.P., Hilbers, C.W. and Vuister, G.W. (2002) *J. Biomol. NMR*, **22**, 281–289.
Spronk, C.A.E.M., Nabuurs, S.B., Bonvin, A.M.J.J., Krieger, E., Vuister, G.W. and Vriend, G. (2003) *J. Biomol. NMR*, **25**, 225–234.
Tian, F., Valafar, H. and Prestegard, J.H. (2001) *J. Am. Chem. Soc.*, **123**, 11791–11796.
Tjandra, N. and Bax, A. (1997) *Science*, **278**, 1111–1114.
Tugarinov, V. and Kay, L.E. (2003) *J. Mol. Biol.*, **327**, 1121–1133.
Ulmer, T.S., Ramirez, B.E., Delaglio, F. and Bax, A. (2003) *J. Am. Chem. Soc.*, **125**, 9179–9191.
Vriend, G. (1990) *J. Mol. Graph.*, **8**, 52–56.
Wang, L.C., Pang, Y.X., Holder, T., Brender, J.R., Kurochkin, A.V. and Zuiderweg, E.R.P. (2001) *Proc. Natl. Acad. Sci. USA*, **98**, 7684–7689.
Weigelt, J., Brown, S.E., Miles, C.S., Dixon, N.E. and Otting, G. (1999) *Struct. Fold. Des.*, **7**, 681–690.
Williamson, M.P., Kikuchi, J. and Asakura, T. (1995) *J. Mol. Biol.*, **247**, 541–546.
Zhao, D.Q. and Jardetzky, O. (1994) *J. Mol. Biol.*, **239**, 601–607.
Zweckstetter, M. and Bax, A. (2000) *J. Am. Chem. Soc.*, **122**, 3791–3792.
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.