Journal of Biomolecular NMR (2005) 31: 143–148

A simple method to adjust inconsistently referenced ^{13}C and ^{15}N chemical shift assignments of proteins

Yunjun Wang[a,]* & David S. Wishart[b]
[a]*Mesolight LLC, Fayetteville, AR 7270, U.S.A.;* [b]*Departments of Biological Sciences and Computing Science, University of Alberta, Edmonton, AB, Canada T6G 2E8*

Received 27 July 2004; Accepted 29 November 2004

Key words: chemical shift, chemical shift referencing, NMR, protein, secondary structure identification

Abstract

Inconsistent ^{13}C and ^{15}N chemical shift referencing is a continuing problem associated with protein chemical shift assignments deposited in BioMagResBank (BMRB). Here we describe a simple and robust approach that can quantitatively determine the ^{13}C and ^{15}N referencing offsets solely from chemical shift assignment data and independently of 3D coordinate data. This novel structure-independent approach permitted the assessment and determination of ^{13}C and ^{15}N reference offsets for all protein entries deposited in the BMRB. Tests on 452 proteins with known 3D structures show that this structure-independent approach yields ^{13}C and ^{15}N referencing offsets that exhibit excellent agreement with those calculated on the basis of 3D structures. Furthermore, this protocol appears to improve the accuracy of chemical shift-derived secondary structural identification, and has been formally incorporated into a computer program called PSSI (http://www.pronmr.com).

Introduction

Over the past 15 years, improvements in NMR instrumentation and pulse sequence design coupled with key developments in NMR assignment software have greatly enhanced our ability to rapidly and accurately assign protein chemical shifts. As a result, a growing number of ^{13}C, ^{15}N, and ^{1}H chemical shift assignments are now being deposited into the BMRB (Seavey et al., 1991). This comprehensive collection of chemical shift data is giving biomolecular NMR spectroscopists an unprecedented opportunity to compare, explore and decipher the rich structural and dynamic information encoded protein chemical shifts. Indeed, it has been through the BMRB and related chemical shift databases, that the correlation between protein secondary structure and chemical shifts was first identified (Wishart et al., 1991; de Dios et al., 1993). Likewise, the BMRB permitted better protocols for shift-based secondary structure identification to be developed and tested (Wishart et al., 1992, 1994; Metzler et al., 1993; Gronenborn and Clore, 1994). Similarly, the BMRB has enabled the detailed exploration of nearest-neighbor effects on protein backbone chemical shifts (Wang and Jardetzky, 2002b) as well as the development of new and improved methods to predict ^{1}H, ^{13}C and ^{15}N chemical shifts from 3D structure coordinates (Le and Oldfield, 1994; Xu and Case, 2001; Neal et al., 2003; Wang 2004; Wang, and Jardetzky, 2004).

Despite of the increasing importance of BMRB, the value of its chemical shift data is diminished somewhat by the inconsistency in ^{13}C and ^{15}N chemical referencing for many protein entries. Indeed, this problem has been periodically noted for quite some time (Iwadate et al., 1999; Wishart and Case, 2001). A recent study

*To whom correspondence should be addressed. E-mails: yjwang@mesolight.com, yunjunwang@yahoo.com by Zhang et al. (2003) showed that a large number (~25%) of BMRB entries with ^{13}C and ^{15}N chemical shift assignments deviate significantly from established IUPAC/IUB referencing conventions (Wishart et al., 1995; Markley et al., 1998). Since the structural and dynamic information contained in chemical shifts is exquisitely sensitive to chemical shift referencing, 'inconsistently' referenced chemical shifts can easily mislead NMR spectroscopists in their interpretation or utilization of these data for secondary structure identification or 3D structural refinement. Given the prevalence of this problem, we believe there is a clear need for a simple method or computer program to allow protein NMR spectroscopists to easily identify and correct inconsistently referenced chemical shift assignments. One approach for correcting chemical shift referencing errors has already been described (Zhang et al., 2003). However, this method requires that the 3D structure be known in order to determine if there has been a chemical shift referencing error. Ideally it would be better to identify any ^{13}C or ^{15}N referencing errors prior to determining or refining the 3D structure.

Here we wish to describe a simple and effective approach to ^{13}C or ^{15}N chemical shift reference correction, which does not require prior knowledge of the 3D structure. The concept is based on the observation that ^{1}H shifts are almost always correctly referenced and the fact that inconsistent secondary structure assignments generated by ^{1}H, ^{13}C and/or ^{15}N chemical shift methods are strongly indicative of incorrectly referenced ^{13}C or ^{15}N chemical shifts.

Materials and methods

As stated earlier, our main objective was to develop a simple, structure-independent protocol which would determine the necessary chemical shift corrections or chemical shift offset that must be added or subtracted to all ^{13}C and/or ^{15}N shifts to ensure they are consistently referenced to IUPAC/IUBMB standards. In general terms, the method involves iteratively calculating chemical shift derived secondary structures and the corresponding ^{13}C or ^{15}N chemical shift reference offsets using a set of observed and expected (or idealized) shifts. Specifically, the chemical shift reference offset (defined as $\Delta\delta^{\text{offset}}$) is determined using the following three steps (Figure 1).

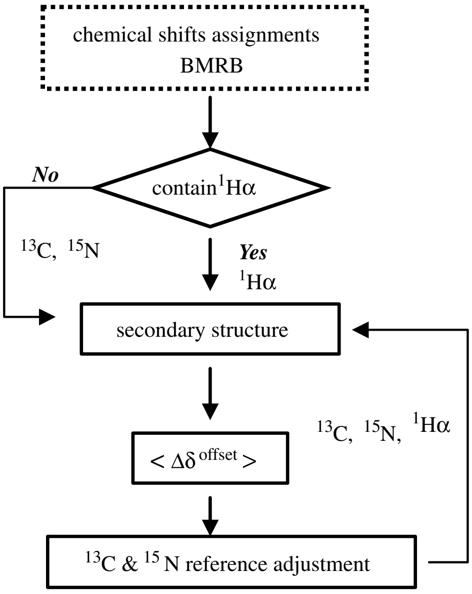

chemical shifts assignments
BMRB
No
contain^{1}Hα
^{13}C, ^{15}N
Yes
^{1}Hα
secondary structure
$< \Delta\delta^{\text{offset}} >$
^{13}C, ^{15}N, ^{1}Hα
^{13}C & ^{15}N reference adjustment

Figure 1. An outline of protocol described in this study for the determination of ^{13}C and ^{15}N referencing offsets. The secondary structures are determined from chemical shifts using the program PSSI (Wang and Jardetzky, 2002a).

Step 1. Secondary structure identification. As has been shown previously (Zhang et al., 2003), protein ^{1}H chemical shift assignments are not generally prone to referencing offset problems. This is likely due to the much better established protocols and inherently simpler methods used for ^{1}H referencing than for ^{13}C or ^{15}N referencing. In light of this fact, the secondary structure for each residue in the protein is first determined via the PSSI program (Wang and Jardetzky, 2002a) using only the ^{1}Hα chemical shifts. PSSI is a chemical shift-based method for secondary structure determination that uses nearest neighbor information to improve the quality of its secondary structure ID. This ^{1}H 'only' approach avoids the inclusion of inconsistently referenced ^{13}C or ^{15}N assignments that might adversely affect the secondary structural identification.

Step 2. Calculation of $\Delta\delta^{offset}$ and $< \Delta\delta^{offset} >$. The ^{13}C and ^{15}N reference offsets are then calculated using the secondary structural information obtained in step 1 using the following equation (for each residue in the protein):

$$\Delta\delta^{\text{offset}} = \delta^{\text{ave}}(\sigma_i) - \delta^{\text{obs}},$$

where, $\delta^{\text{ave}}(\sigma_i)$ is the statistically averaged chemical shift for each nucleus (^{13}CO, ^{13}Cα, ^{13}Cβ, and ^{15}N) for each of the 20 amino acids categorized according to their secondary structural states; σ_i represents the three secondary structural states (β-strand, random coil, and α-helix) and δ^{obs} is the observed chemical shift. In this calculation we used the averaged, secondary structure-dependent ^{13}CO, ^{13}Cα, ^{13}Cβ, and ^{15}N chemical shifts for δ^{ave} as reported by Wang and Jardetzky (2002a).

The averaged $\Delta\delta^{\text{offset}}$ over all residues in the protein, $< \Delta\delta^{\text{offset}} >$, is then calculated for each backbone nucleus (^{13}Cα, ^{13}Cβ, ^{13}CO and ^{15}N). This value is used as an initial estimate of that nucleus' chemical shift referencing offset. During the calculation of $< \Delta\delta^{\text{offset}} >$, unusually large $\Delta\delta^{\text{offset}}$ values (e.g., greater than three standard deviations away from the running mean) are excluded to avoid the inclusion of typographical or assignment errors that might affect the overall for a given nucleus.

Step 3. ^{13}C and ^{15}N referencing adjustment and back calculation of $< \Delta\delta^{offset} >$. In this step, all observed ^{13}C and ^{15}N chemical shift assignments are initially adjusted using the nucleus-specific reference offsets, $< \Delta\delta^{\text{offset}} >$ obtained in step 2. Subsequently the secondary structure is re-evaluated using the original ^{1}Hα chemical shifts and the reference-adjusted ^{13}C and ^{15}N chemical shifts. A new set of reference offsets, $\Delta\delta^{\text{offset}}$ and their corresponding averaged value $< \Delta\delta^{\text{offset}} >$ for each nucleus are then re-calculated. This step is repeated twice to minimize the effects arising from inconsistent referencing. More specifically, the requirement for a double iteration was determined through a series of tests in which it was found that the reference offsets, $< \Delta\delta^{\text{offset}} >$ always converged to steady state values after two iterations.

Since the reference offsets, $< \Delta\delta^{\text{offset}} >$ are determined on the basis of an iterative statistical procedure, the reliability of this method clearly depends on the number of chemical shifts that are being averaged for each nucleus. As a rule of thumb, the minimum number of chemical shifts required to determine a reference offset for a given nucleus should be greater than 25. This suggests that short peptides (25 residues) are less likely to be correctly re-referenced using this method.

The protocol described here has been implemented into a newly updated version of a freely available computer program called PSSI, a probability-based protein secondary structure identification program that uses a comprehensive set of derived protein chemical shift data (Wang and Jardetzky, 2002a,b; http://www.pronmr.com). To assess the performance of this structure-independent approach to reference offset calculation, we decided to compare it to the structure-dependent approach originally described and validated by Zhang et al. (2003). A total of 452 protein entries were selected from the BMRB for which matching PDB entries could be identified (Wang, 2004). The structure-derived $\Delta\delta^{\text{offset}}$ values were calculated using a protocol similar to that used by Zhang et al. (2003). More specifically, the ^{13}C and ^{15}N chemical shifts were first predicted from the observed backbone and side chain torsion angles (ϕψ and χ_1) using the RSS (Wang, 2004) and PRSI (Wang and Jardetzky, 2003) shielding surfaces. The structure-derived ^{13}C, ^{15}N $\Delta\delta^{\text{offset}}$ values were then determined by averaging the difference between the observed and the RSS-predicted chemical shifts for each nucleus. Offsets calculated using the RSS method or other structure-based shift prediction methods such as SHIFTX (Neal et al., 2003) or SHIFTS (Xu and Case, 2002) were found to be essentially identical. Once the structure-dependent offsets were determined, we then calculated the structure-independent ^{13}C and ^{15}N chemical shift offsets using the protocol described here for the same 452 proteins. Direct comparison between the two sets of predicted reference offsets yielded an excellent agreement (Figure 2). The correlation coefficients between the two sets of data are 0.96, 0.96, 0.98, and 0.93 for ^{13}CO, ^{13}Cα, ^{13}Cβ, and ^{15}N shifts respectively, with the rmsd values being just 0.16, 0.16, 0.17, and 0.31 ppm. The averaged differences between the two sets data are less than 0.15 ppm and for ^{13}C (^{13}CO, ^{13}Cα, and ^{13}Cβ,) and less than 0.50 ppm for ^{15}N shifts. These offset differences are statistically and structurally insignificant for the purposes of interpreting protein chemical shifts.

Interestingly, several proteins were found to have significant differences between the $\Delta\delta^{\text{offset}}$

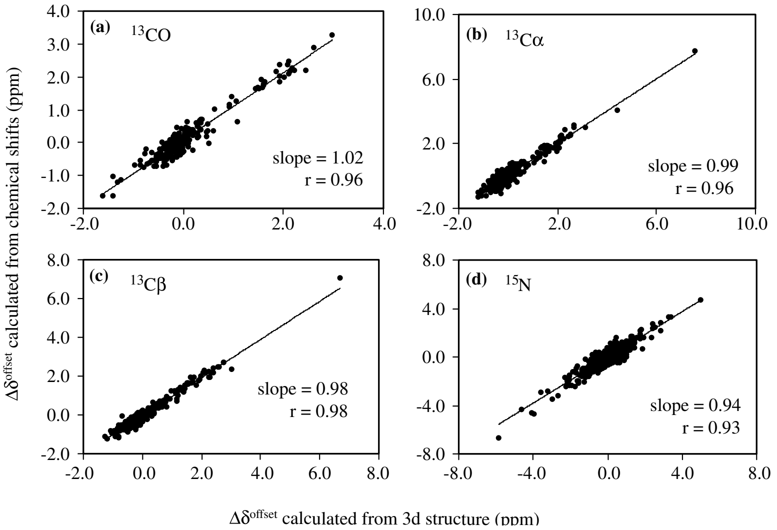


Figure 2. Graph showing correlation between ^{13}CO (a), ^{13}Cα(b), ^{13}Cβ (c) and ^{15}N (d) reference offsets, < $\Delta\delta^{\mathrm{offset}}$ > determined from chemical shifts and those from 3D structures. The slope and correlation coefficients of the linear regression fitting for the two sets of data are shown the bottom right corner of each graph.

calculated using the protocols described in this manuscript and those derived from the corresponding 3D coordinates. BMRB #4272 (Synaptobrevin II) stands out for having the largest discrepancies between the two sets of predicted reference offsets. The structure-dependent reference offsets calculated from 3D structure for this protein are 1.59, 2.52, −0.67, and −1.60 ppm for ^{13}CO, ^{13}Cα, ^{13}Cβ, and ^{15}N shifts respectively; whereas those calculated using our structure-independent approach are −0.13, 0.07, −0.08, and −0.77 ppm. A further investigation revealed that the IUPAC/IUB recommended referencing procedure (i.e., using DSS as a direct reference for ^{1}H, and an indirect reference for ^{15}N and ^{13}C nuclei) was indeed followed during NMR measurement for this protein. We believe the offset differences are likely due to significant solution vs. crystal structure differences. For BMRB #5158 (apoMb), the reference offsets calculated from the 3D structure are 0.89 and 1.19 ppm for ^{13}CO and ^{13}Cα shifts respectively. These values are very close to those reported in RefDB (1.09 and 1.21 ppm), but differ significantly from those calculated using our new structure-independent method (−0.15 ppm for ^{13}CO shifts; and 0.03 ppm for ^{13}Cα shifts). Further investigation revealed that these differences are likely due to the fact that the PDB files used to calculate the structure-dependent reference offsets are quite different from the protein samples used to generate the NMR assignments. Specifically, the PDB file (from a crystal structure) used for this protein exhibits a very high helix content; but the NMR data clearly indicates the protein was partially unfolded (Hazzarda et al., 1999; Cavagnero et al., 2001). These two examples (apoMb and Synaptobrevin II) underline one potential disadvantage of using 3D structures to calculate the chemical shift reference offsets. Namely, the differences between crystal and solution structures can significantly influence calculated reference offsets. Obviously the structure-independent approach introduced here avoids this pitfall. Another advantage to the structure-independent approach to reference offset correction is that it can be applied to proteins where no 3D structure is available. This situation applies to more than 50% of protein entries in the BMRB.

To more completely assess this approach to chemical shift reference correction, we decided to calculate the ^{13}C and ^{15}N chemical shift reference offsets for all protein entries deposited in the BMRB. At the time of this writing there were a total of 2982 protein entries (with more than 25 residues) in the BMRB. Of these, 664, 986, 895, and 1141 were found to have ^{13}CO, ^{13}Cα, ^{13}Cβ, and ^{15}N chemical shift assignments respectively. Interestingly, the largest ^{13}CO referencing offset identified was 4.47 ppm for BMRB #5514. The largest offset correction for ^{13}Cα and ^{13}Cβ shifts were 7.82 and 6.91 ppm for BMRB #4431, whereas the largest ^{15}N reference offset correction was 4.70 ppm for BMRB #4127. In total, we found that 160 (^{13}CO), 309 (^{13}Cα), 309 (^{13}Cβ) and 243 (^{15}N) BMRB entries had significant referencing offsets (>0.5 ppm for ^{13}C; >1.0 ppm for ^{15}N). Thus, 24% (^{13}CO), 31% (^{13}Cα), 35% (^{13}Cβ), and 21% (^{15}N) of BMRB entries appear to be inconsistently or incorrectly referenced. The BMRB accession numbers and the corresponding $\Delta\delta^{\text{offset}}$ values for all proteins (646 in total) with significant referencing offsets (>0.5 ppm for ^{13}C; >1.0 ppm for ^{15}N) are listed in the Supplementary Material.

The distribution of the ^{13}CO, ^{13}Cα, ^{13}Cβ, and ^{15}N reference offsets (< $\Delta\delta^{\text{offset}}$ >) for all proteins analyzed in this study are plotted in Figure 3. As can be clearly seen among all ^{13}C (< $\Delta\delta^{\text{offset}}$ >) plots there is a small, but distinct shoulder or secondary peak at ~2 ppm. This appears to be due to the continuing (or inadvertent) use of TMS/dioxane as a ^{13}C reference since this value is consistent with the ~1.7 ppm offset correction for TMS noted by Wishart et al. (1995). A more detailed analysis indicates that 64 (^{13}CO), 104 (^{13}Cα), and 97 (^{13}Cβ) of our BMRB entries have positive reference offsets of above 1.0 ppm. Assuming they all arise from the use of TMS/ dioxane as a ^{13}C reference, we believe that continued use of TMS or dioxane referencing may account for approximately 1/3 of the inconsistently referenced ^{13}C BMRB entries.

These results are not entirely unexpected. Earlier studies by Zhang et al. (2003) demonstrated that heteronuclear chemical shift referencing problems were especially widespread for protein entries made before 1995 (when TMS was widely used). Despite the introduction of detailed chemical shift referencing recommendations (Wishart et al., 1995; Markley et al., 1998) some 8 years earlier, Zhang et al. found that ~20% deposited protein entries are still incorrectly referenced. Even with the introduction of their structure-dependent reference correction software in 2003, the referencing problem still persists. Our latest data indicate that approximately 25% of the protein chemical shift assignments deposited in 2003–2004 are still inconsistently referenced. Hopefully, the introduction of this structure-independent approach to reference offset adjustment will help correct this situation.

As a further check of the robustness of our method we also evaluated the internal consistency of our calculated reference offsets. That is, the calculated ^{13}CO, ^{13}Cα, and ^{13}Cβ(especially the ^{13}Cα, and ^{13}Cβ) reference offsets should be approximately equal to each other for any given protein or BMRB entry. For the most part we find these ^{13}C offsets actually agree with each other quite well. However, ~3% of the BMRB entries we tested show significant (>1.0 ppm) differences between their calculated ^{13}Cα and ^{13}Cβ reference offsets (these are marked in the supplementary material and will be flagged in our reference offset

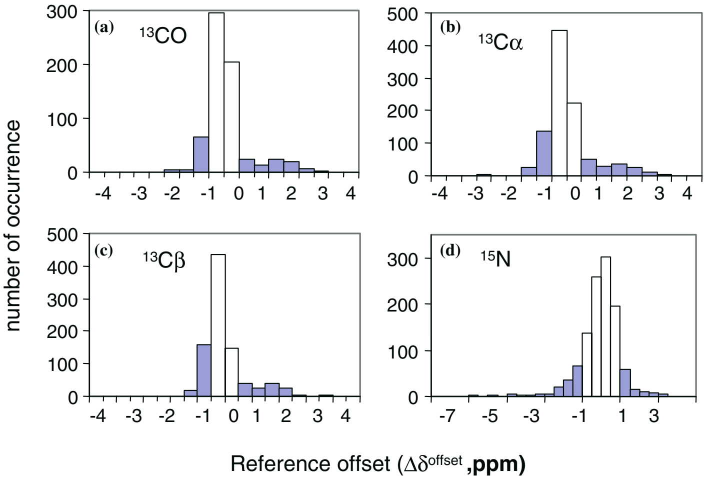
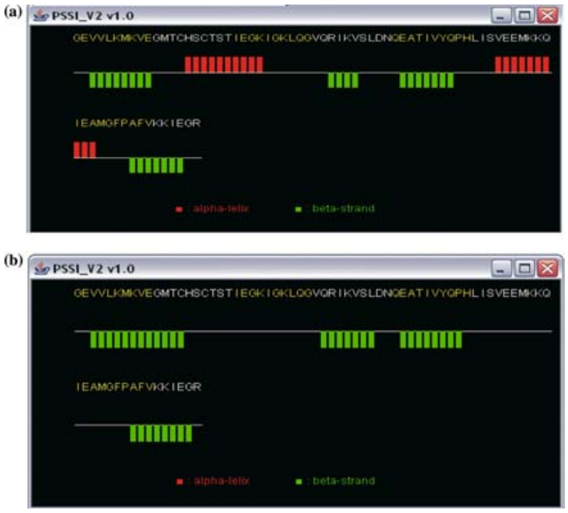


Figure 3. Distribution of the ^{13}CO (a), ^{13}Cα (b), ^{13}Cβ (c), and ^{15}N (d) reference offsets, $\Delta\delta^{\text{offset}}$, Those within 0.5 ppm (for ^{13}C) and 1.0 ppm (for ^{15}N) are expressed using the unshaded bars.



```
GEVVLKMKVEGMTCHSCTSTIEGKIGKLQGVQRIKVSLDNQEATIVYQPHLISVEEMKKQ
IEAMGFPAFVKKIEGR
```

```
GEVVLKMKVEGMTCHSCTSTIEGKIGKLQGVQRIKVSLDNQEATIVYQPHLISVEEMKKQ
IEAMGFPAFVKKIEGR
```

Figure 4. Graphic outputs from PSSI showing the secondary structure derived from NMR chemical shifts for Copper-transporting ATPase 1 protein (BMRB #6130) with (a) and without (b) ^{13}C and ^{15}N reference self-calibration.

program – PSSI). These discordant ^{13}Cα/^{13}Cβ shifts may reflect differential shifting due to deuteration, limited sampling of one nucleus relative to the other, incorrect instrumental calibration, shift biasing (due to TROSY effects, for example) or possibly assignment errors. Because it is almost impossible to know the cause of these discrepancies or to consistently correct for them, we have chosen to treat the calculated reference offsets for ^{13}CO, ^{13}Cα, and ^{13}Cβ nuclei independently. Users are obviously free to average the calculated ^{13}Cα, and ^{13}Cβ (and even the ^{13}CO) offsets to produce a consensus ^{13}C offset, if they wish.

Consistent chemical shift referencing clearly has an impact on how much structural and dynamic information an NMR researcher can recover from their chemical shift assignments. To illustrate this point we used the newly updated version of PSSI to show how this program can automatically identify and adjust inconsistently referenced ^{13}CO, ^{13}Cα, ^{13}Cβ, and ^{15}N chemical shift assignments (Figure 3a). Using data from the Copper-transporting ATPase 1 (BMRB #6130) the new version of PSSI was able to automatically determine the necessary reference offsets for this protein – 2.43 ppm for ^{13}CO, 2.0 ppm for ^{13}Cα, 1.99 ppm for ^{13}Cβ, and 0.91 ppm for ^{15}N (Figure 4). As expected, such large reference offsets completely alter results of the secondary structures that would be derived from the chemical shifts. As shown in Figures 3b and c, this mixed α/β protein was initially mis-identified as a pure β-sheet protein using the original incorrectly referenced chemical shift assignments.

Supplementary material to this paper is available in electronic form at http://dx.doi.org/10.1007/s10858-004-7441-3.

References

Cavagnero, S., Nishimura, C., Schwarzinger, S., Dyson, H.J. and Wright, P.E. (2001) *Biochemistry*, **40**, 14459–14467.

de Dios, A.C., Pearson, J.G. and Oldfield, E. (1993) *Science*, **260**, 1491–1496.

Gronenborn, A.M. and Clore, G.M. (1994) *J. Biomol. NMR*, **4**, 455–458.

Hazzarda, J., Südhofb, T.C. and Rizoa, J. (1999) *J. Biomol. NMR*, **14**, 203–207.

Iwadate, M., Asakura, T. and Williamson, M.P. (1999) *J. Biomol. NMR*, **13**, 199–211.

Le, H. and Oldfield, E. (1994) *J. Biomol. NMR*, **4**, 341–348.

Markley, J.L., Bax, A., Arata, Y., Hilbers, C.W., Kaptein, R., Sykes, B.D., Wright, P.E. and Wüthrich, K. (1998) *J. Biomol. NMR*, **12**, 1–23.

Metzler, W.J., Constantine, K.L., Friedrichs, M.S., Bell, A.J., Ernst, E.G., Lavoie, T.B. and Mueller, L. (1993) *Biochemistry*, **32**, 13818–13829.

Neal, S., Nip, A.M., Zhang, H. and Wishart, D.S. (2003) *J. Biomol. NMR*, **26**, 215–240.

Seavey, B.R., Farr, E.A., Westler, W.M. and Markley, J.L. (1991) *J. Biomol. NMR*, **1**, 217–236.

Wang, Y. (2004) *J. Biomol. NMR*, **30**, 233–244.

Wang, Y. and Jardetzky, O. (2002a) *Protein Sci.*, **11**, 852–861.

Wang, Y. and Jardetzky, O. (2002b) *J. Am. Chem. Soc.*, **124**, 14075–14084.

Wang, Y. and Jardetzky, O. (2004) *J. Biomol. NMR*, **28**, 327–340.

Wishart, D.S. and Case, D.A. (2001) *Meth. Enzymol.*, **338**, 3–34.

Wishart, D.S. and Sykes, B.D. (1994) *J. Biomol. NMR*, **4**, 171–180.

Wishart, D.S., Bigam, C.G., Yao, J., Abildgaard, F., Dyson, H.J., Oldfield, E., Markley, J.L. and Sykes, B.D. (1995) *J. Biomol. NMR*, **6**, 135–140.

Wishart, D.S., Sykes, B.D. and Richards, F.M. (1991) *J. Mol. Biol.* **222**, 311–333.

Wishart, D.S., Sykes, B.D. and Richards, F.M. (1992) *Biochemistry*, **31**, 1647–1651.

Xu, X.P. and Case, D.A. (2001) *J. Biomol. NMR*, **21**, 321–333.

Zhang, H., Neal, S. and Wishart, D.S. (2003) *J. Biomol. NMR*, **25**, 173–195.