J Biomol NMR (2009) 44:95–99
DOI 10.1007/s10858-009-9324-0

COMMUNICATION

Empirical correlation between protein backbone ^{15}N and ^{13}C secondary chemical shifts and its application to nitrogen chemical shift re-referencing

Liya Wang · John L. Markley

Received: 13 April 2009 / Accepted: 22 April 2009 / Published online: 13 May 2009

Abstract The linear analysis of chemical shifts (LACS) has provided a robust method for identifying and correcting ^{13}C chemical shift referencing problems in data from protein NMR spectroscopy. Unlike other approaches, LACS does not require prior knowledge of the three-dimensional structure or inference of the secondary structure of the protein. It also does not require extensive assignment of the NMR data. We report here a way of extending the LACS approach to ^{15}N NMR data from proteins, so as to enable the detection and correction of inconsistencies in chemical shift referencing for this nucleus. The approach is based on our finding that the secondary ^{15}N chemical shift of the backbone nitrogen atom of residue *i* is strongly correlated with the secondary chemical shift difference (experimental minus random coil) between the alpha and beta carbons of residue $i - 1$. Thus once alpha and beta ^{13}C chemical shifts are available (their difference is referencing error-free), the ^{15}N referencing can be validated, and an appropriate offset correction can be derived. This approach can be implemented prior to a structure determination and can be used to analyze potential referencing problems in database data not associated with three-dimensional structure. Application of the LACS algorithm to the current BMRB protein chemical shift database, revealed that nearly 35% of the BMRB entries have δ^{15}N values mis-referenced by over 0.7 ppm and over 25% of them have $\delta^{1}H^{N}$ values mis-referenced by over 0.12 ppm. One implication of the findings reported here is that a backbone ^{15}N chemical shift provides a better indicator of the conformation of the preceding residue than of the residue itself.

Keywords Validation of chemical shifts · Carbon-13 chemical shift · Proton chemical shift · Nitrogen-15 chemical shift · LACS

The difference between the chemical shift (δ) of an amino acid and its random coil chemical shift (δ_{coil}) is the secondary chemical shift ($\Delta\delta$), which is widely used in protein secondary structure prediction (Wishart and Sykes 1994; Eghbalnia et al. 2005) and backbone dihedral angle constraint estimation (Cornilescu et al. 1999). Values for $\Delta\delta^{13}C^{\alpha}$, $\Delta\delta^{13}C^{\beta}$, $\Delta\delta^{13}C'$, $\Delta\delta^{1}H^{\alpha}$, $\Delta\delta^{1}H^{N}$ and $\Delta\delta^{15}N$ assigned to a given residue generally are combined and used to estimate the secondary structure propensity of the residue or to derive geometrical constraints on the backbone torsion angles. Because chemical shift values are relative to a standard compound, the accuracy of such predictions depends critically on whether the chemical shift referencing is consistent and follows standard norms. The accuracy of back calculated chemical shifts from high-resolution protein structures is sufficiently accurate that it can be used to assay chemical shift referencing accuracy. By back calculating chemical shifts from proteins with high-resolution structures and comparing them to chemical shifts deposited in the BioMagResBank (BMRB; Seavey et al. 1991; Ulrich et al. 2008), it was found that up to 20%

Electronic supplementary material The online version of this article (doi:10.1007/s10858-009-9324-0) contains supplementary material, which is available to authorized users.

L. Wang
Cold Spring Harbor Laboratory, Williams 5, 1 Bungtown Rd, Cold Spring Harbor, NY 11724, USA

J. L. Markley (✉)
Biochemistry Department, University of Wisconsin, Madison, WI 53705, USA
e-mail: markley@nmrfam.wisc.edu; markley@biochem.wisc.edu

of δ^{13}C and 30% of δ^{15}N were improperly referenced (RefDB; Zhang et al. 2003). It is of importance also to detect and correct possible referencing errors in protein NMR data sets that are not associated with three-dimensional structures (more than 60% of the data sets in BMRB; Wang and Wishart 2005). Approaches to this problem have been based either on secondary structure prediction tools (Wang and Wishart 2005; Marsh et al. 2006; Ginzinger et al. 2007) or on linear relationships between $\Delta\delta^{13}C_i^{\alpha}$, $\Delta\delta^{13}C_i^{\beta}$, $\Delta\delta^{13}C_i'$, or $\Delta\delta^{1}H_i^{\alpha}$ and $(\Delta\delta^{13}C_i^{\alpha} - \Delta\delta^{13}C_i^{\beta})$ (Wang et al. 2005). The latter approach, called LACS (linear analysis of chemical shifts; Wang et al. 2005), utilizes "backbone geometry driven" linear correlations among chemical shifts themselves, instead of relying on secondary structure prediction. The performance of CheckShift (Ginzinger et al. 2007), the most recent approach based on predicted secondary structure, is claimed to equal that of LACS, under conditions of good secondary structure prediction accuracy. Whereas the initial LACS implementation (Wang et al. 2005) provided re-referencing only for δ^{13}C (and $\delta^{1}H^{\alpha}$), CheckShift can be used to determine re-referencing offsets also for δ^{15}N.

We report here the extension of LACS to the re-referencing of δ^{15}N (and $\delta^{1}H^{N}$) chemical shifts. Whereas we earlier found linear relationships between $\Delta\delta^{13}C_i$ (and $\Delta\delta^{1}H_i^{\alpha}$) and $(\Delta\delta^{13}C_i^{\alpha} - \Delta\delta^{13}C_i^{\beta})$, our recent statistical examination shows that linear relationships actually hold between $\Delta\delta^{15}N_i$ (and $\Delta\delta^{1}H_i^{N}$) and $(\Delta\delta^{13}C_{i-1}^{\alpha} - \Delta\delta^{13}C_{i-1}^{\beta})$, where i is the residue whose chemical referencing is examined and $i - 1$ is the index of the preceding residue. Correlations had been reported previously between $\Delta\delta^{15}N_i$ (and $\Delta\delta^{1}H_i^{N}$) and ϕ_i and ψ_{i-1} (Le and Oldfield 1994), and between ϕ_i and ψ_i and $(\Delta\delta^{13}C_i^{\alpha} - \Delta\delta^{13}C_i^{\beta})$ (Spera and Bax 1991; Wang et al. 2007).

The random coil chemical shift difference $(\delta^{13}C_{coil}^{\alpha} - \delta^{13}C_{coil}^{\beta})$ of each residue type, statistically derived from our maximum entropy analysis (Wang et al. 2007) and consistent with experimental observations (Wishart et al. 1995), was used to calculate $(\Delta\delta^{13}C_i^{\alpha} - \Delta\delta^{13}C_i^{\beta})$. $\Delta\delta^{15}N_{coil}$ for each residue type X was taken from the experimental data for the hexapeptide Gly-Gly-X-Ala-Gly-Gly, with the neighboring effect of X on $\Delta\delta^{15}N_{coil}$ of Ala corrected by using data provided in the same report (Wishart et al. 1995). Neighboring effects of X on Gly also have been determined experimentally from data on shorter peptides Gly-Gly-X-Gly-Gly (Schwarzinger et al. 2001), and the corrections derived from the two sets of data are similar. Nearest neighbor corrections also have been derived statistically from database information (Wang and Jardetzky 2002), but these values are less consistent with experiment, probably because they were based on limited data. Ideally, neighboring effects should be measured from all 400 Gly-Gly-X-Y-Gly-Gly hexapeptides. Lacking this information, we made the simplifying assumption that the effect of X on $\Delta\delta^{15}N_{coil}$ of other residue types is the same as it is on Ala.

A different set of random coil chemical shifts, determined at pH of 2.3 for short peptides Gly-Gly-X-Gly-Gly (Schwarzinger et al. 2000), which takes account the significant changes in the $(\delta^{13}C_{coil}^{\alpha} - \delta^{13}C_{coil}^{\beta})$ values of Asp and Glu near and below their side chain p*K*a values (3.8 for Asp, and 4.1 for Glu), has been used for proteins at pH < 4.

For each BMRB entry, a robust fitting procedure (Wang et al. 2005) was used to linearly fit $\Delta\delta^{15}N_i$ with $(\Delta\delta^{13}C_i^{\alpha} - \Delta\delta^{13}C_i^{\beta})$ or $(\Delta\delta^{13}C_{i-1}^{\alpha} - \Delta\delta^{13}C_{i-1}^{\beta})$. The distributions of the fitted slopes are shown in Fig. 1. The mean values (slopes) of these two distributions show that $\Delta\delta^{15}N_i$ is statistically four times more sensitive to $(\Delta\delta^{13}C_{i-1}^{\alpha} - \Delta\delta^{13}C_{i-1}^{\beta})$ [Gaussian distribution $N(-0.4, 0.20)$] than to $(\Delta\delta^{13}C_i^{\alpha} - \Delta\delta^{13}C_i^{\beta})$ [Gaussian distribution $N(-0.1, 0.22)$]. The mean slope for the fit to $(\Delta\delta^{13}C_{i-1}^{\alpha} - \Delta\delta^{13}C_{i-1}^{\beta})$ was -0.4, whereas the mean slope for the fit to $(\Delta\delta^{13}C_i^{\alpha} - \Delta\delta^{13}C_i^{\beta})$ was -0.1. The fact that the latter value is close to zero indicates that $\Delta\delta^{15}N_i$ is unrelated statistically to the intra-residue values $(\Delta\delta^{13}C_i^{\alpha} - \Delta\delta^{13}C_i^{\beta})$.

Fitting of $\Delta\delta^{1}H_i^{N}$ with $(\Delta\delta^{13}C_{i-1}^{\alpha} - \Delta\delta^{13}C_{i-1}^{\beta})$ yielded slopes with distribution $N(-0.07, 0.046)$, and fitting of $\Delta\delta^{1}H_i^{N}$ with $(\Delta\delta^{13}C_i^{\alpha} - \Delta\delta^{13}C_i^{\beta})$ yielded $N(-0.02, 0.034)$. These values show the similar trends but are five times smaller, owing to the smaller scale of $\Delta\delta^{1}H^{N}$. However, the root mean square errors of the inter and intra-residue secondary chemical shifts fittings are indistinguishable (data not shown), which indicates that the correlation is very weak and explains why the slopes are so different among proteins.

The variation of the fitted slopes comes from the different amino acid content, the different α, β structure

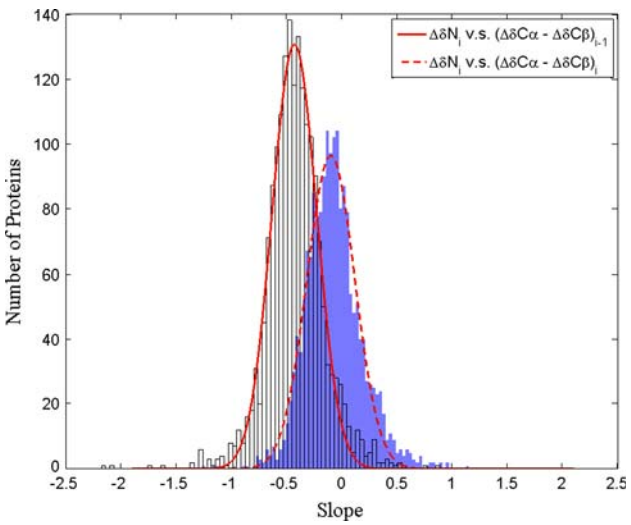

Fig. 1 Comparison of the distributions of the fitted slopes for $\Delta\delta^{15}N_i$ with $(\Delta\delta^{13}C_{i-1}^{\alpha} - \Delta\delta^{13}C_{i-1}^{\beta})$ and $(\Delta\delta^{13}C_i^{\alpha} - \Delta\delta^{13}C_i^{\beta})$

content of these proteins, and variability of the available chemical shift data. In order to deal with such variations, we developed the following procedure for checking the referencing of δ^{15}N for a protein on the basis of the available set of NMR chemical shift data.

Calculate the carbon secondary chemical shift difference $(\Delta\delta^{13}C^{\alpha}_{i-1} - \Delta\delta^{13}C^{\beta}_{i-1})$ and the nearest-neighbor corrected nitrogen secondary chemical shifts, $\Delta\delta^{15}N_i$.

1. Use the robust fitting procedure, which reduces effects of outlier points to fit $\Delta\delta^{15}N_i$ with $(\Delta\delta^{13}C^{\alpha}_{i-1} - \Delta\delta^{13}C^{\beta}_{i-1})$, to derive the slope k and intercept b:

$$\Delta\delta^{15}N_i = k\left(\Delta\delta^{13}C^{\alpha}_{i-1} - \Delta\delta^{13}C^{\beta}_{i-1}\right) + b$$

2. If k is close to the expected mean (0.4 ± 0.1), report $-b$ as the reference offset.
3. If k is outside the range (0.4 ± 0.1), and if the data are insufficient for slope determination or if the distribution of chemical shifts along x-axis suggests that the protein is unfolded or contains only helix or sheet, then use a bounded slope (k within 0.4 ± 0.05) in least square linear fitting $\Delta\delta^{15}N_i$ with $(\Delta\delta^{13}C^{\alpha}_{i-1} - \Delta\delta^{13}C^{\beta}_{i-1})$ to re-estimate b. Report $-b$ as the reference offset. The criterion for insufficient data is when the total number of chemical shift pairs available [N_{total} = number of $(\Delta\delta^{13}C^{\alpha}_{i-1} - \Delta\delta^{13}C^{\beta}_{i-1})$, $\Delta\delta^{15}N_i$) values] is fewer than 66 (arbitrary number). The criterion for poor distribution of chemical shifts is when fewer than 15% of the total number of chemical shift pairs (N_{total}) have $((\Delta\delta^{13}C^{\alpha}_{i-1} - \Delta\delta^{13}C^{\beta}_{i-1}) < 2)$ or $\left(\left(\Delta\delta^{13}C^{\alpha}_{i-1} - \Delta\delta^{13}C^{\beta}_{i-1}\right) > 2\right)$.
4. If k is outside the range (0.4 ± 0.1), and sufficient data of good distribution are available (neither of the two above criteria are met), then report $-b$ determined in step 1 as the reference offset.

In cases where only partial backbone chemical shift values are available, or where the protein is unfolded or contains only α- or β-structure (smaller dispersion of data along the x-dimension), factors other than backbone geometry might dominate the dispersion of δ^{15}N. In these cases, restriction of the slope (as defined in step 3) improves the accuracy of offset estimation (along the y-dimension). The arbitrary numbers are introduced here for lack of a "true" reference offset; otherwise it is possible that the offset values could be optimized by use of a machine learning method.

Figure 2 shows the agreement among LACS, RefDB, and CheckShift for $\delta^{15}N$ offsets. The offsets detected by LACS have a standard deviation of 0.39 ppm with those from CheckShift and a standard deviation of 0.62 ppm with those from RefDB. The same procedure was used to fit $\Delta\delta^{1}H^{N}_{i}$ with $(\Delta\delta^{13}C^{\alpha}_{i-1} - \Delta\delta^{13}C^{\beta}_{i-1})$ (k bounded at -0.07 ± 0.01 for step 3); in this case, the offsets detected by LACS show a standard deviation of 0.11 ppm with those from RefDB. Thus, by using the newly observed linearity between $\delta^{15}N_i$, $\Delta\delta^{1}H^{N}_{i}$ and $(\Delta\delta^{13}C^{\alpha}_{i-1} - \Delta\delta^{13}C^{\beta}_{i-1})$, LACS can give precise offset estimations for both $\delta^{1}H^{N}$ and $\delta^{15}N$.

Overall, LACS values showed better agreement to CheckShift than to RefDB, whose offsets are based on chemical shifts back-calculated from high-resolution structures. This result may reflect the fact that δ^{15}N values are difficult to estimate accurately from structure. However, analysis by LACS and CheckShift of the chemical shift data from a few BMRB entries yielded very poor agreement. These outliers, which were excluded from Fig. 2, are listed in Table 1, along with available reference offsets predicted by RefDB. In cases where all three values were available, the LACS values were closer to the RefDB values than to the CheckShift values. Because the RefDB values are associated with three-dimensional structures, this result suggests that LACS may lead to fewer large re-referencing errors than CheckShift. Considering the deviations among LACS, CheckShift, and RefDB, we suggest that experimental δ^{15}N or $\delta^{1}H^{N}$ values should be

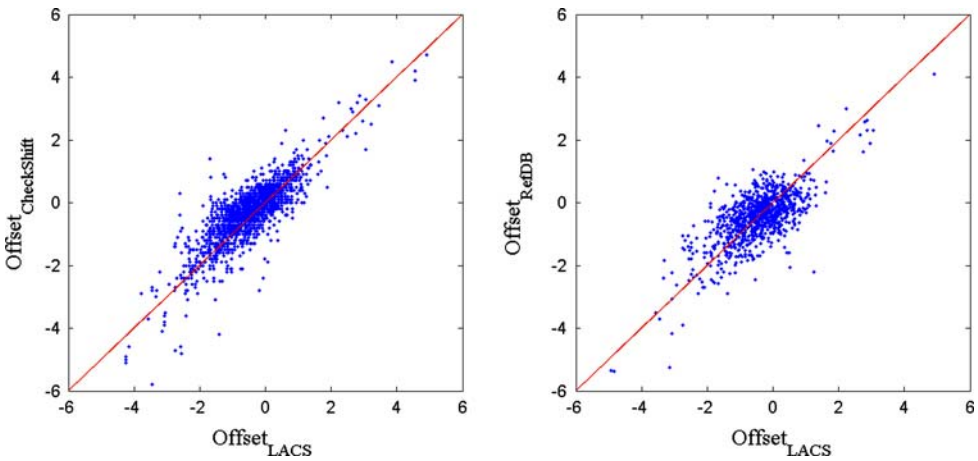

Fig. 2 Comparison of the offsets (ppm) estimated by LACS, CheckShift and RefDB

Table 1 List of BMRB entries that CheckShift and LACS predict significantly different δ^{15}N reference offsets

BMRB entry	δ^{15}N Reference offsets detected (ppm)		
	CheckShift	LACS	RefDB
5,629	11.6	0.6	−0.45
4,094	10.0	0.3	0.00
5,212	6.6	−0.8	–
4,130	0.0	−5.6	−5.38
4,131	0.0	−5.7	−5.34
6,549	−6.7	−1.1	−1.01
6,790	5.9	−0.2	–
6,840	−5.9	−0.3	−0.16
5,753	4.4	−1.2	−2.26

re-referenced if and only if the offsets predicted by LACS for δ^{15}N are >0.7 ppm or for $\delta^{1}H^{N}$ are >0.12 ppm. This approach should reduce the chance of introducing systematic errors when re-referencing a large set of proteins. However, for a single protein, users should always examine the LACS plot to check for possible mis-assignments, uneven distribution of chemical shifts along the *x* axis, and/or insufficient data.

An advantage of LACS is that it can be applied even in cases where the backbone chemical shifts are partially assigned; this makes LACS more widely applicable than other approaches. We used the algorithm to examine possible referencing problems in the current BMRB protein chemical shift database. The results suggest that nearly 35% of the BMRB entries have δ^{15}N values mis-referenced by over 0.7 ppm and over 25% of them have $\delta^{1}H^{N}$ values mis-referenced by over 0.12 ppm.

Previous studies showed that $\Delta\delta^{15}N_i$ and $\Delta\delta^{1}H_i^{N}$ are correlated with ϕ_i and ψ_{i-1} (Le and Oldfield 1994). The results reported here imply that ψ_{i-1} plays a more important role than ϕ_i on $\Delta\delta^{15}N_i$ and $\Delta\delta^{1}H_i^{N}$. Therefore, this study not only extends the LACS approach for re-referencing to δ^{15}N and $\delta^{1}H^{N}$ but also suggests that $\Delta\delta^{15}N_i$ and $\Delta\delta^{1}H_i^{N}$ values should be used as indicators of conformation the $(i-1)$-th residue rather than the i-th. Furthermore, it has also been shown that $\Delta\delta^{15}N_i$ can be predicted from ϕ_i, ψ_{i-1} and χ^1 (Wang and Jardetzky 2004). Omission of χ^1 and ϕ_i in our linear regression analysis might be responsible for the dispersion of $\Delta\delta^{15}N_i$ around the fitted line. Conversely, it might be possible to derive χ^1 and ϕ_i constraints after extracting the backbone effect (where the linearity holds). However, studies of this kind currently are greatly hindered by the limited amount of protein chemical shift data associated with three-dimensional structures.

The standalone executable application for using LACS to determine all backbone chemical shift reference offsets can be downloaded from http://brie.cshl.edu/~liyawang/LACS/ or from BMRB (http://bmrb.wisc.edu).

Acknowledgments This work is a continuation of the LACS re-referencing tool development supported by NIH grants P41RR02301 and 1U54 GM074901 to J.L.M.

References

Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 13:289–302

Eghbalnia HR, Wang L, Bahrami A, Assadi A, Markley JL (2005) Protein energetic conformational analysis from NMR chemical shifts (PECAN) and its use in determining secondary structural elements. J Biomol NMR 32:71–81

Ginzinger SW, Gerick F, Coles M, Heun V (2007) CheckShift: automatic correction of inconsistent chemical shift referencing. J Biomol NMR 39:223–227

Le H, Oldfield E (1994) Correlation between nitrogen-15 nuclear magnetic resonance chemical shifts in proteins and secondary structure. J Biomol NMR 4:341–348

Marsh JA, Singh VK, Jia Z, Forman-Kay JD (2006) Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein: implications for fibrillation. Protein Sci 15:2795–2804

Schwarzinger S, Kroon GJ, Foss TR, Wright PE, Dyson HJ (2000) Random coil chemical shifts in acidic 8 M urea: implementation of random coil shift data in NMRView. J Biomol NMR 18:43–48

Schwarzinger S, Kroon GJ, Foss TR, Chung J, Wright PE, Dyson HJ (2001) Sequence-dependent correction of random coil NMR chemical shifts. J Am Chem Soc 123:2970–2978

Seavey BR, Farr EA, Westler W, Markley JL (1991) A relational database for sequence-specific protein NMR data. J Biomol NMR 1:217–236

Spera S, Bax A (1991) Empirical correlation between protein backbone conformation and C and C ^{13}C nuclear magnetic resonance chemical shifts. J Am Chem Soc 113:5490–5492

Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Wenger RK, Yao H, Markley JL (2008) BioMagResBank. Nucleic Acids Res 36:D402–D408

Wang Y, Jardetzky O (2002) Investigation of the neighboring residue effects on protein chemical shifts. J Am Chem Soc 124:14075–14084

Wang Y, Jardetzky O (2004) Predicting ^{15}N chemical shifts in proteins using the residue-specific individual shielding surfaces from phi, psi(i-1), and chi1 torsion angles. J Biomol NMR 28:327–340

Wang Y, Wishart DS (2005) A simple method to adjust inconsistently referenced ^{13}C and ^{15}N chemical shift assignments of proteins. J Biomol NMR 31:143–148

Wang L, Eghbalnia HR, Bahrami A, Markley JL (2005) Linear analysis of carbon-13 chemical shift differences and its application to the

detection and correction of errors in referencing and spin system identifications. J Biomol NMR 32:13–22

Wang L, Eghbalnia HR, Markley JL (2007) Nearest-neighbor effects on backbone alpha and beta carbon chemical shifts in proteins. J Biomol NMR 39:247–257

Wishart DS, Sykes BD (1994) The ^{13}C chemical-shift index: a simple method for the identification of protein secondary structure using ^{13}C chemical-shift data. J Biomol NMR 4:171–180

Wishart DS, Bigam CG, Holm A, Hodges RS, Sykes BD (1995) ^{1}H, ^{13}C and ^{15}N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects. J Biomol NMR 5:67–81

Zhang HY, Neal S, Wishart DS (2003) RefDB: a database of uniformly referenced protein chemical shifts. J Biomol NMR 25:173–195