Journal of Biomolecular NMR (2005) 32: 13–22
DOI 10.1007/s10858-005-1717-0

Article

Linear analysis of carbon-13 chemical shift differences and its application to the detection and correction of errors in referencing and spin system identifications

Liya Wang[a,b,c], Hamid R. Eghbalnia[a,d,*], Arash Bahrami[a,b,c] & John L. Markley[a,b,c]
[a]*National Magnetic Resonance Facility at Madison, Biochemistry Department, 433 Babcock Drive, Madison WI 53706, USA;* [b]*Center for Eukaryotic Structural Genomics, Biochemistry Department, 433 Babcock Drive, Madison WI 53706, USA;* [c]*Graduate Program in Biophysics, University of Wisconsin–Madison, Madison WI 53706, USA;* [d]*Mathematics Department, University of Wisconsin-Madison, 811 Van Vleck Hall, 480 Lincoln Drive, Madison WI 53706, USA*

Received 02 December 2004; Accepted 24 January 2005

Key words: carbon-13 chemical shifts, linear analysis of chemical shifts (LACS), protein backbone geometry, proton chemical shifts, RefDB, TALOS

Abstract

Statistical analysis reveals that the set of differences between the secondary shifts of the α- and β-carbons for residues *i* of a protein ($\Delta\delta^{13}C_i^{\alpha} - \Delta\delta^{13}C_i^{\beta}$) provides the means to detect and correct referencing errors for ^{1}H and ^{13}C nuclei within a given dataset. In a correctly referenced protein dataset, linear regression plots of $\Delta\delta^{13}C_i^{\alpha}$, $\Delta\delta^{13}C_i^{\beta}$, or $\Delta\delta^{1}H_i^{\alpha}$ vs. ($\Delta\delta^{13}C_i^{\alpha} - \Delta\delta^{13}C_i^{\beta}$) pass through the origin from two directions, the helix-to-coil and strand-to-coil directions. Thus, linear analysis of chemical shifts (LACS) can be used to detect referencing errors and to recalibrate the ^{1}H and ^{13}C chemical shift scales if needed. The analysis requires only that the signals be identified with distinct residue types (intra-residue spin systems). LACS allows errors in calibration to be detected and corrected in advance of sequence-specific assignments and secondary structure determinations. Signals that do not fit the linear model (outliers) deserve scrutiny since they could represent errors in identifying signals with a particular residue, or interesting features such as a *cis*-peptide bond. LACS provides the basis for the automated detection of such features and for testing reassignment hypotheses. Early detection and correction of errors in referencing and spin system identifications can improve the speed and accuracy of chemical shift assignments and secondary structure determinations. We have used LACS to create a database of offset-corrected chemical shifts corresponding to nearly 1800 BMRB entries: ~300 with and ~1500 without corresponding three-dimensional (3D) structures. This database can serve as a resource for future analysis of the effects of amino acid sequence and protein secondary and tertiary structure on NMR chemical shifts.

Introduction

BioMagResBank (BMRB) (Seavey et al., 1991) currently contains more than 3000 chemical shift datasets. A preliminary survey of these suggested that up to 20% of ^{13}C shifts and 30% of ^{15}N shifts are improperly referenced (Wishart and Case, 2001). Different methods have been proposed to detect and correct for such errors. One approach has been to determine deviations of secondary shifts (experimental shifts minus the corresponding

*To whom correspondence should be addressed. E-mail: eghbalni@nmrfam.wisc.edu

random coil shifts) from those predicted by the chemical shift hypersurface (Spera and Bax, 1991); this method is used to correct input data to the TALOS program (Cornilescu et al., 1999). Wishart et al. (Zhang et al., 2003) recently used statistical notions to produce a database of reference-corrected chemical shifts (RefDB) for proteins with known three-dimensional (3D) structure. Their reference corrections were based on differences between the deposited chemical shifts and those predicted from chemical shift hypersurfaces and classically calculated ring-current (Haigh and Mallion, 1979) and electric field (Osapay and Case, 1991) effects. They incorporated correction factors to account for empirically derived nearest neighbor effects and local side chain effects and used averaged differences between predicted and observed chemical shift values to derive offsets to correct the chemical shifts. This approach indicated that nearly 30% of protein entries in BMRB have significant referencing errors.

The robustness of calculations based on chemical shift hypersurfaces and empirical corrections is limited by factors such as sample bias and the quality of protein structures. For example, whereas dihedral angles can be specified precisely for structures at 1.5 Å resolution, they are considered to be merely "well-constrained" for structures at 2.5 Å. Since most NMR structures correspond to ≥2.5 Å resolution, chemical shift estimates may be imprecise. In the absence of a "fortunate cancellation of errors", such inaccuracies could lead to large uncertainties in the chemical shift correction factor. Therefore, a complementary and independent approach for calibrating NMR resonances should prove useful. If, in addition, the calibration could be performed prior to structure determination, it could greatly improve secondary structure determinations by any algorithm, and in turn lead to more refined dihedral angle restraints.

Relationships between chemical shifts and backbone geometry have been observed experimentally since the 1960s (Markley et al., 1967; Sternlicht and Wilson, 1967). Specifically, it has been demonstrated that the average secondary shifts for α helix are 3.09 ppm for $^{13}C^{\alpha}$ and −0.38 ppm for $^{13}C^{\beta}$ (Spera and Bax, 1991) and −0.14 ppm for $^{1}H^{\alpha}$ (Osapay and Case, 1994). The average secondary shifts for β-sheet are −1.48 ppm for $^{13}C^{\alpha}$ and 2.16 ppm for $^{13}C^{\beta}$ (Spera and Bax, 1991) and 0.45 ppm for $^{1}H^{\alpha}$ (Osapay and Case, 1994). The chemical shifts of other nuclei ($^{13}C'$, $^{1}H^{N}$, and ^{15}N) also have been shown to depend on protein backbone geometry (Wishart et al., 1991; Le and Oldfield, 1994). Several techniques have been developed to identify secondary structure on the basis of chemical shift data: these include the $\Delta\delta$ method (Reiley et al., 1992), the chemical shift index method (Wishart and Sykes, 1994), the probability-based method (Wang and Jardetzky, 2002), and the chemical shift with sequence information combination method (Hung and Samudrala, 2003).

One can consider protein backbone geometry as an information bridge that relates the chemical shifts of different nuclei, at least statistically. The desired property of resonances to be used in detecting and correcting referencing errors is reference independence. Protons do not have a statistically significant referencing problem (Wishart and Case, 2001) and could be considered as potential candidates. However, the practical suitability of the proton is limited by the fact that $\delta\ ^{1}H$ is strongly influenced by ring current and sequence effects and is sensitive to temperature and pH (Baxter and Williamson, 1997).

$^{13}C^{\alpha}$ and $^{13}C^{\beta}$ shifts are typically measured in the same experiment; thus their chemical shift differences are reference independent. Except in residues followed by proline, $\delta^{13}C^{\alpha}$ and $\delta^{13}C^{\beta}$ are not affected significantly by protein sequence, including aromatic rings (Blanchard et al., 1997). Empirical analysis has confirmed that protein backbone geometry is the major factor affecting $\delta^{13}C^{\alpha}$ and $\delta^{13}C^{\beta}$ (Iwadate et al., 1999). Most importantly, $\delta^{13}C^{\alpha}$ and $\delta^{13}C^{\beta}$ shift in opposite directions when present in either α-helical or β-strand structures. These observations have motivated our use of the difference, $(\Delta\delta^{13}C^{\alpha} - \Delta\delta^{13}C^{\beta})$, as a reference-independent measure.

The method of re-referencing we introduce is a statistical procedure that relies on average behavior of sample populations and is used to identify observed data that deviate from "expected" values to a statistically significant degree. Therefore, the interpretation of LACS results must be subjected to statistical measures. Re-referencing indicated by LACS does not necessarily indicate a problem. Whereas recalibration to a common reference point may be useful for statistical studies of chemical shift distributions or for assignments of

chemical shifts and secondary structure, the original data sets must remain the primary source for individual experiments.

Methods

We use the following linear equations to describe the relationship between secondary chemical shifts ($\Delta\delta^{13}C^{\alpha}$, $\Delta\delta^{13}C^{\beta}$, $\Delta\delta^{1}H^{\alpha}$, or $\Delta\delta^{13}C'$) and ($\Delta\delta^{13}C^{\alpha}$-$\Delta\delta^{13}C^{\beta}$),

$$Y = \begin{cases} K_{\alpha}X + O_{\alpha} & \text{if} \quad X >= 0 \\ K_{\beta}X + O_{\beta} & \text{if} \quad X < 0 \end{cases} \qquad (1)$$

In this equation, $X = (\Delta\delta^{13}C^{\alpha}\text{-}\Delta\delta^{13}C^{\beta})$ is a reference-independent variable. Y can denote the reference-dependent values of $\Delta\delta^{13}C^{\alpha}$, $\Delta\delta^{13}C^{\beta}$, $\Delta\delta^{1}H^{\alpha}$ or $\Delta\delta^{13}C'$. K_{α} and K_{β} are the slopes for the coil-helical and sheet-coil regions, respectively. O_{α} and O_{β} are Y-intercepts, which report the value of the reference offset; ideally they are zero in the absence of a reference error.

For n data points, standard least squares regression (Equation 1) can be used to recover the offset values (O_{α} and O_{β}). However, errors in identifying the chemical shifts with individual residues can potentially compromise least squares regression results, particularly if the chemical shifts of the partially mis-identified resonances are far from the correct values. To compensate for this, we use a 'robust analysis' (Holland and Welsch, 1977) procedure to ensure stable regression results even in the presence of a small number of errors in spin system identifications, This method iteratively computes a set of weights for data points in the least squares algorithm, with the weights at each iteration calculated by applying a carefully selected function to the residuals from the previous iteration (see Supplementary Materials for details). The weights assign a lower significance to features of the data set, called outliers, which significantly deviate from expected values. The results are consistent with standard least squares regression when no outliers are present and are practically insensitive to outliers when they are present in the input data.

The robust analysis approach serves to automatically detect outliers. Data points given low weights during the robust regression procedure are marked as possible outliers. These are of interest, because they may indicate mis-assignments or irregular structure. A comprehensive exposition on techniques in outlier detection can be found in (Barnett and Lewis, 1994).

Nearly 300 of the proteins in the database of reference-corrected chemical shifts (RefDB) (Zhang et al., 2003) have assigned $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ signals. Chemical shift values corresponding to the backbone atoms ($^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{1}H^{\alpha}$, and $^{13}C'$) of these proteins were downloaded (in a file named RefDB–C.db) and used for generating Figure 1 and Table 1. The corresponding uncorrected values were downloaded from the BMRB.

The determination of secondary chemical shifts relies on knowledge of random coil chemical shifts. In the current study, these were taken from experimentally determined values (Wishart et al., 1995), as it has been suggested (Schwarzinger et al., 2001) that these are a reliable source for random coil chemical shifts. Values for residues C-terminal to proline were corrected (Wishart et al., 1995) for the known effect on $\delta^{13}C^{\alpha}$ and $\delta^{13}C^{\beta}$ (Iwadate et al., 1999). It is important to note that the values used in the analysis were obtained at 25 °C. This leads to the statistical tendency to re-reference chemical shifts to this temperature, a desirable effect particularly for $^{1}H^{\alpha}$, which is sensitive to temperature (Baxter and Williamson, 1997).

Results and discussion

For each residue i, we used robust regression (Equation 1) to establish the statistical relationship between the secondary chemical shifts of $^{13}C_i^{\alpha}$, $^{13}C_i^{\beta}$, $^{13}C_i'$, and $^{1}H_i^{\alpha}$ as a function of $(\Delta\delta^{13}C_i^{\alpha} - \Delta\delta^{13}C_i^{\beta})$. Only small deviations from the regression line were observed for deposited chemical shift data for both helical and sheet regions (Figure 1). In our regression analysis, we excluded three amino acids: glycine (no C^{β}), cysteine (limited data available), and proline (coil dominated). From comparisons of the regression coefficients for the 17 remaining amino acids, we determined that the slope of the regression line is relatively independent of the particular amino acid (Figure 2). This result justifies our subsequent use of one or two regression lines for all the amino acids in a protein. Since the data from all amino acids can be combined, the analysis can be carried out prior to sequence-specific assignment.

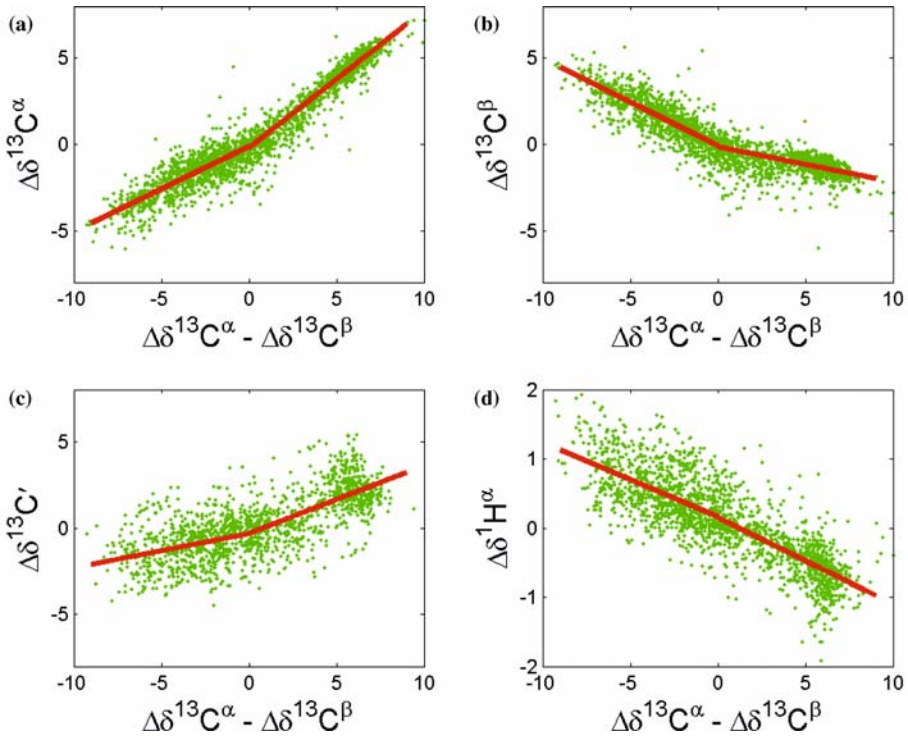

Figure 1. Plot of the secondary chemical shifts (in ppm) for $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, and $^{1}H^{\alpha}$ of valine as a function of ($\Delta\ \delta^{13}C^{\alpha}$-$\delta^{13}\Delta C^{\beta}$) (in ppm). Data (in the file RefDB–C.db) were taken from RefDB (Zhang et al., 2003). The solid lines represent the results of linear regression analyses.

Table 1. Systematic offsets (in ppm) determined by LACS for the data in RefDB[a]

Amino acid	$O(\delta^{13}C^{\alpha})$	$O(\delta^{13}C^{\beta})$	$O(\delta^{1}H^{\alpha})$	$O(\delta^{13}C')$	$O(\Delta\ \delta^{13}C^{\alpha}$-$\Delta\delta^{13}C^{\beta})$
Ala	−0.1647	−0.2017	0.0313	−0.133	0.037
Asp	−0.0397	−0.0891	0.0372	−0.1111	0.0494
Glu	−0.0957	−0.0613	−0.0007	0.0671	−0.0344
Phe	0.0744	0.0252	0.0894	−0.0778	0.0492
His	0.0323	−0.0384	0.096	0.0192	0.0707
Ile	−0.2322	−0.1908	0.1676	−0.1485	−0.0414
Lys	0.0137	−0.0064	0.0151	0.032	0.0201
Leu	−0.0241	−0.0242	0.0194	−0.1155	0.0001
Met	−0.2026	−0.2987	0.0902	0.4391	0.0961
Asn	−0.0595	−0.0562	0.0607	0.0712	−0.0033
Gln	−0.0616	−0.0764	0.0201	−0.0249	0.0148
Arg	−0.1254	−0.1706	0.0517	−0.0614	0.0452
Ser	−0.0328	−0.0266	0.0267	−0.2711	−0.0062
Thr	0.0756	0.0743	0.0651	−0.6508	0.0013
Val	−0.1577	−0.124	0.1543	−0.3374	−0.0337
Trp	0.2158	0.2213	0.0483	0.3361	−0.0055
Tyr	0.172	0.1796	0.0824	−0.0904	−0.0076
Mean (Standard deviation)	−0.036 (0.125)	−0.051 (0.132)	0.062 (0.0472)	−0.062 (0.243)	0.015 (0.038)

[a]Zhang et al. (2003).

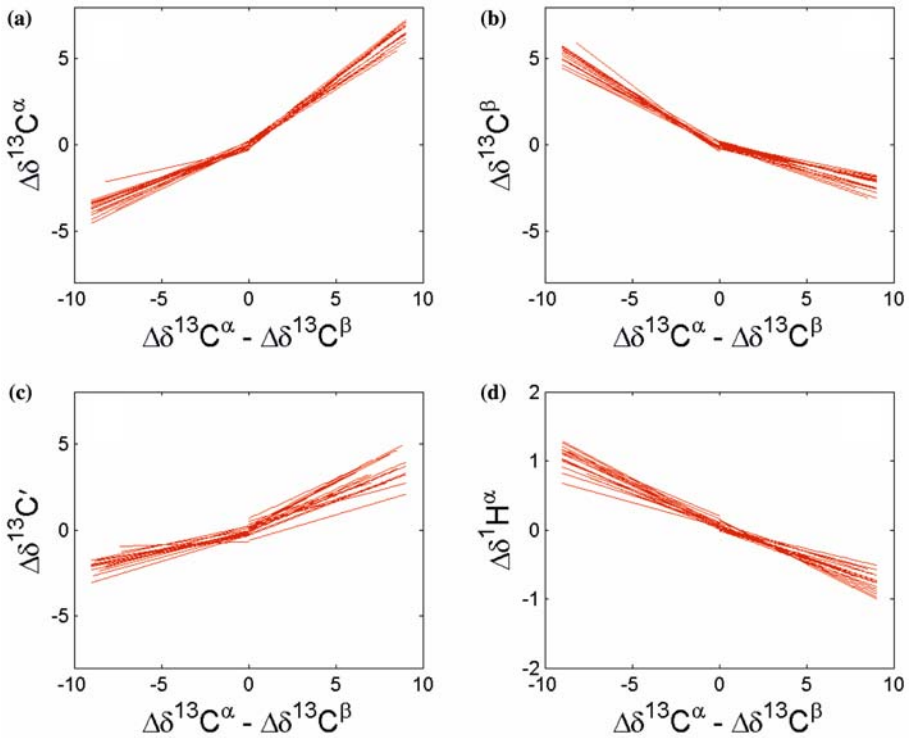

Figure 2. Linear regression results for 17 amino acids (20 standard amino acids with the exception of Cys, Gly, and Pro). Individual lines represent different amino acids. All values are in ppm.

Proteins containing both α-helix and β-sheet require two linear regression lines for $\delta^{13}C^{\alpha}$ and $\delta^{13}C^{\beta}$. This result simply confirms that whereas $\delta^{13}C^{\alpha}$ experiences larger changes between coil and helix than between coil and strand, this trend is reversed for $\delta^{13}C^{\beta}$. The use of two regression lines instead of one for $\delta^{13}C^{\alpha}$ and $\delta^{13}C^{\beta}$ leads to more precise linear relationships.

To gain insight into the statistical significance of the linear analysis of chemical shifts (LACS) method, we compared referencing offsets determined from LACS with those derived from RefDB (the database of reference corrected values). We compared the average values, $O = (O_{\alpha} + O_{\beta})/2$, obtained by LACS for the 17 amino acids with the reference corrections from RefDB (Table 1). Specifically, over a range from –10 to 10 ppm, the mean and variance of $O(\Delta\delta^{13}C^{\alpha}\text{-}\Delta\delta^{13}C^{\beta})$ were close to zero (0.015 and 0.038 ppm, respectively). This confirmed the statistical reliability of our approach for offset analysis. We have made the statistical assumption that, without referencing error, $\Delta\delta^{13}C^{\alpha}$, $\Delta\delta^{13}C^{\beta}$, $\Delta\delta^{1}H^{\alpha}$, and $\Delta\delta^{13}C'$ will approach zero as $(\Delta\delta^{13}C^{\alpha}\text{-}\Delta\delta^{13}C^{\beta})$ approaches zero. Because sampling errors in either the helical or strand regions may cause small deviations in the intercept of the line away from zero, we approximated the intercept by the average of the two intercepts O_{α} and O_{β}.

The majority of the systematic offsets between the LACS and RefDB results (Table 1) are much smaller than the threshold that would influence structure determination and refinement (0.5 ppm for carbon chemical shifts and 0.1 ppm for protons) (Wishart and Nip, 1998). The largest systematic offsets relative to the data from RefDB (Zhang et al., 2003) were for $^{1}H^{\alpha}$ of isoleucine, $^{1}H^{\alpha}$ of valine, and $^{13}C'$ of threonine (Table 1). These discrepancies may stem from the random coil chemical shifts (Wishart et al., 1995) we adopted for these residues, which were derived from experimental values for these residues when followed by alanine and located in regions of coil. However, in the case of other amino acids, this approach for determining random coil chemical shifts yielded excellent agreement with RefDB for $^{13}C^{\alpha}$ and $^{13}C^{\beta}$.

For all residues, the spread of points about the regression line (variance) increased progressively for $\Delta\delta^{13}C^{\alpha}$ or $\Delta\delta^{13}C^{\beta}$, $\Delta\delta^{1}H^{\alpha}$, and $\Delta\delta^{13}C'$. The significantly more dispersed patterns for $\Delta\delta^{13}C'$ and $\Delta\delta^{1}H^{\alpha}$ as compared to $\Delta\delta^{13}C^{\alpha}$ or $\Delta\delta^{13}C^{\beta}$ are consistent with the general finding that adjacent residues have larger effects on $\Delta\delta^{13}C'$ and $\Delta\delta^{1}H^{\alpha}$. The trend demonstrated by the relation between $\delta^{13}C'$ and $\delta^{13}C^{\alpha}$, $\delta^{1}H^{\alpha}$ and $\delta^{13}C^{\beta}$ (Figures 1 and 2) is consistent with empirical $\Delta\delta$ hypersurfaces (Wishart and Nip, 1998).

A statistical analysis of data shows that reference offsets of absolute value less than or equal to 0.1 ppm for $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ are not statistically significant. Values with reference offset in this range should be considered as correct. In our data analysis, this class forms approximately 30% of the cases. Approximately 25% of values have detected reference shifts larger than 0.5 ppm. Shifts larger than 0.5 ppm are statistically significant, and data within this range should be scrutinized carefully. The remainder of data has shifts in the intermediate range (0.1–0.5 ppm) in absolute value. Examination of a few of these suggests that there may be multiple reasons for the offset, which include temperature, pH, and others. For example, in BMRB entry 5358, BPTI bound to trypsin, the −0.16 ppm shift for $^{13}C^{\alpha}$, and $^{13}C^{\beta}$ could simply be a consequence of protein–protein interactions. This category of values may be a useful and "interesting" class to examine in further detail, although this range of shifts does not generally impact the quality of NMR structures.

Offset calibration. Proteins contain various amino acids, and the slopes for their resonance data computed according to Equation 1 differ (Figure 2). Based on our statistical criteria for error, and considering the linear nature of the relationship for the 17 amino acids, we considered it reasonable to use Equation 1 for a single fit the data from all residue types. This single fit proved to be sufficient for the purpose of determining the reference offset. Offsets calculated using $-(O_{\alpha}+O_{\beta})/2$ showed close agreement with those from RefDB. For example, the offset for $\delta^{13}C^{\alpha}$ in BMRB entry 4998 was determined to be 2.10 ppm from LACS as compared to 2.14 ppm from RefDB (Figure 3).

For recalibration of $\delta^{13}C'$ or $\delta^{1}H^{\alpha}$, which show much greater dispersion than $\delta^{13}C^{\alpha}$ or $\delta^{13}C^{\beta}$, a single regression line proved to be as statistically accurate as a pair of lines. LACS of proteins containing only α-helix or only β-sheet is based on the single available regression line, but this has been shown to be sufficiently robust for estimating offsets for $^{13}C^{\alpha}$ and $^{13}C^{\beta}$.

We have compared LACS offsets with corresponding ones derived from RefDB (Figure 4). The RefDB offsets were obtained from differences between predicted and observed values for the four types of nuclei ($^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, and $^{1}H^{\alpha}$). RefDB predicts "final" offsets for $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ from the average of the two individual offsets. The LACS and RefDB offsets are in closest agreement for $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts (Figure 4) followed by $^{13}C'$ and $^{1}H^{\alpha}$. Although $^{1}H^{\alpha}$ does not have a statistically significant referencing problem (Wishart and Case, 2001), both methods returned similar values in cases where large referencing errors were detected.

The level of agreement between the LACS and RefDB offsets was found to depend on the resolution of the structures used in generating the RefDB offsets. The individual offsets for $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ reported by RefDB tended to be different for structures at lower resolution but became more equivalent for structures at higher resolution. Agreement was considerably higher for structures with 1.8 Å resolution or better (Figure 5d–f); in these cases, almost all the LACS and RefDB offsets agreed to within 0.25 ppm for $^{13}C^{\alpha}/^{13}C^{\beta}$, 0.5 ppm for $^{13}C'$, and 0.1 ppm for $^{1}H^{\alpha}$. Errors with either method can arise from bias in sampling, actual experimental conditions, incorrectly assigned data, or incorrectly reported data. Both approaches gave offset estimates within the expected uncertainty. The data from the 283 entries used to generate Figures 4 and 5 are listed in Table S1 of the Supplementary Materials.

When data from either $\delta^{13}C^{\alpha}$ or $\delta^{13}C^{\beta}$ are unavailable, it is possible to recalibrate on the basis of proton NMR results. In this case, $\Delta\delta^{13}C^{\alpha}$, $\Delta\delta^{13}C^{\beta}$, or $\Delta\delta^{13}C'$ is plotted as a function of $\Delta\delta^{1}H^{\alpha}$, and the offset is taken from the chemical shift at which the regression line crosses $\Delta\delta^{1}H^{\alpha} = 0$. The $\Delta\delta^{1}H^{\alpha}$ approach to the determination of recalibration offsets when compared to $(\Delta\delta^{13}C^{\alpha}–\Delta\delta^{13}C^{\beta})$ LACS or RefDB led to significantly lower accuracy in ~20% of the cases (Figure 6).

Outlier detection. In the automated mode of LACS, the data are examined by reference to a threshold. Residues identified as having large

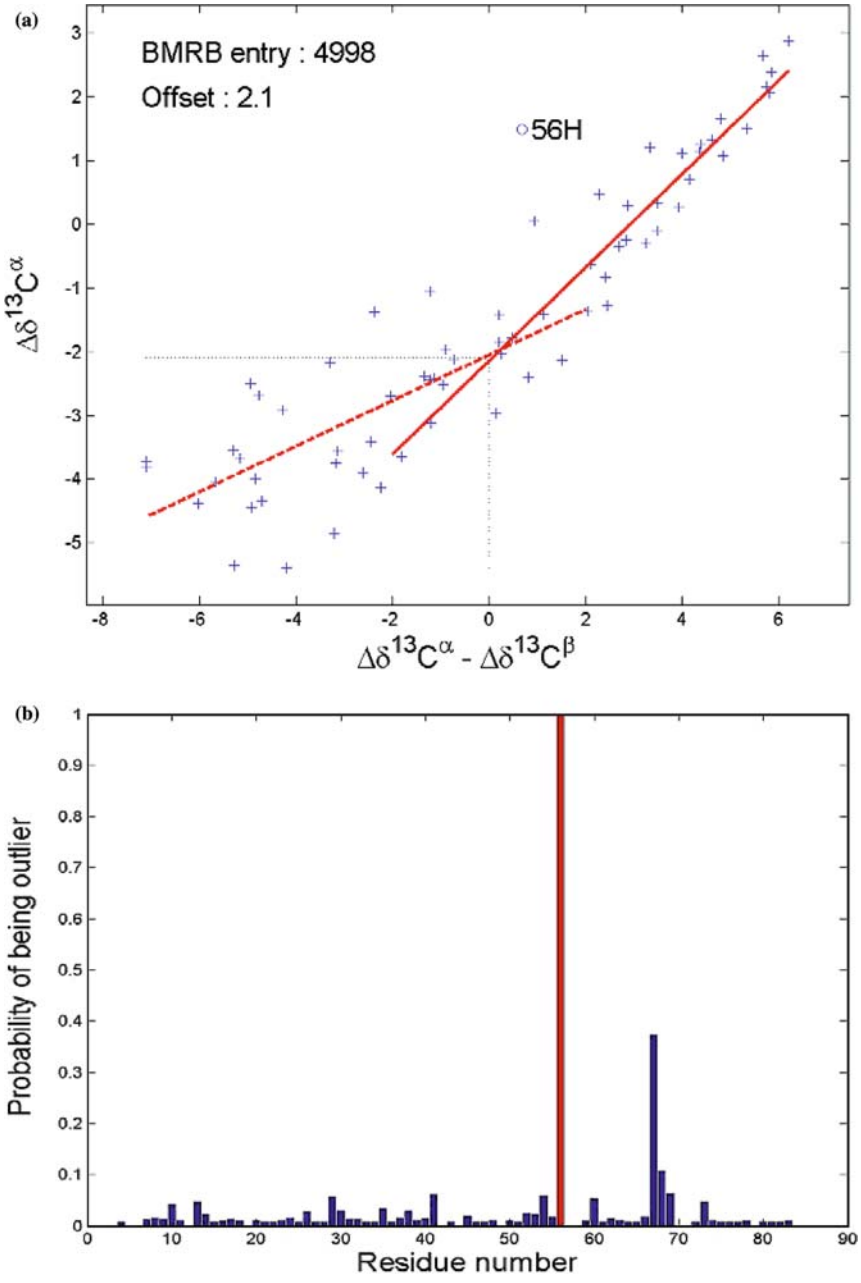

Figure 3. (a) Plot of $\Delta\delta^{13}C^{\alpha}$ as a function of ($\Delta\delta^{13}C^{\alpha}$–$\Delta\delta^{13}C^{\beta}$) for BMRB entry 4998. The solid and dashed lines represent linear regression analysis results for the helical and sheet regions, respectively. The offset (referencing error) is estimated as 2.10 ppm, vs. 2.14 ppm from RefDB (Zhang et al., 2003). The circle identifies an outlier (His56). All values are in ppm. (b) Outlier probability for each of the residues in BMRB entry 4998 with His56 showing a near 100% probability of being an outlier.

(above 90% probability of being an outlier) and moderate (>50% and <90% probability) deviations are reported for further scrutiny.

In RefDB, BMRB entry 4998 is flagged as being ''possibly mis-assigned or deuterate isotope effects'', because the average calculated offsets for $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ were found to differ by more than 0.5 ppm. LACS analysis of BMRB entry 4998 (Figure 3a) identified His56 as a possible outlier (for $^{13}C_i^{\alpha}$ and $^{13}C_i^{\beta}$ chemical shifts), the one with

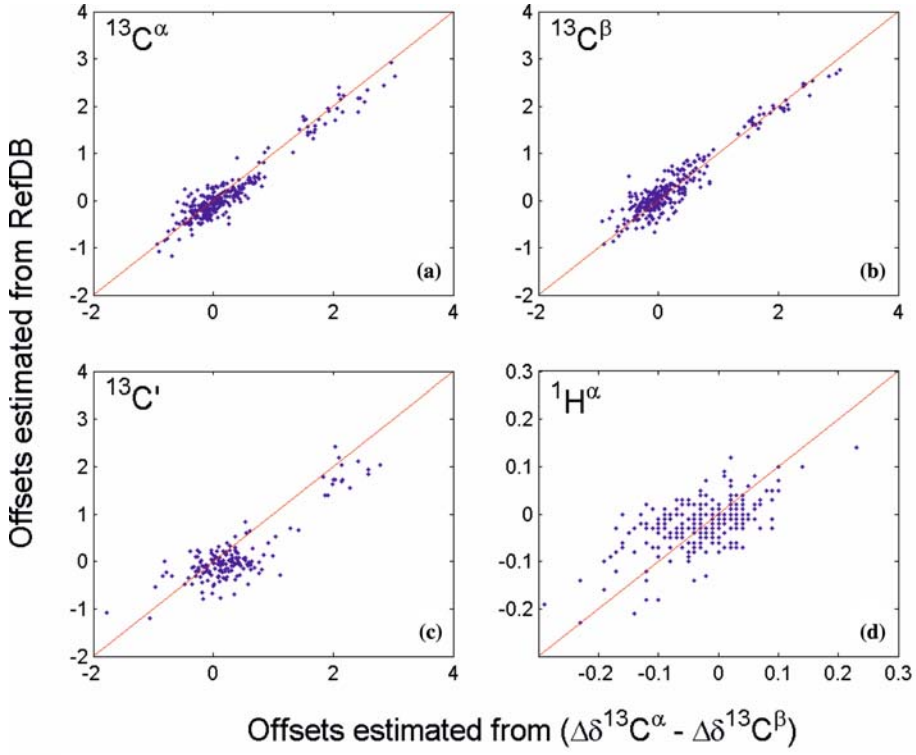

Figure 4. Correlation between offsets (in ppm) for $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, and $^{1}H^{\alpha}$ estimated by LACS (horizontal axes) with those from RefDB (Zhang et al., 2003) (vertical axes). A total of 283 proteins were included. RefDB values are those reported as "predicted – observed".

the lowest weight in the robust regression (Figure 3b). Whereas $\Delta\delta^{13}C^{\alpha}$ of His56 shows a strong helical structure propensity, ($\Delta\delta^{13}C^{\alpha}$–$\Delta\delta^{13}C^{\beta}$) reveals no such propensity. This inconsistency accounts for the observed deviation from the regression line. SHIFTX (Neal et al., 2003), predicts chemical shifts for $\delta^{13}C^{\alpha}$ and $\delta^{13}C^{\beta}$ of His56 as 57.78 ppm and 29.53 ppm respectively. After correcting these experimental values by the reference offset reported by RefDB for this dataset (2.14 ppm), these chemical shifts become $56.49+2.14=58.63$ ppm and $29.81+2.14=31.95$ ppm. These values differ from those calculated from the structure by SHIFTX by 0.85 ppm for $\delta^{13}C^{\alpha}$ and 2.42 ppm for $\delta^{13}C^{\beta}$. Since SHIFTX yields more precise estimates for $\delta^{13}C^{\beta}$ than for $\delta^{13}C^{\alpha}$, there appears to be a real discrepancy between the structure and chemical shift of the His56 β-carbon. Secondary structure analysis of BMRB entry 4998 by DSSP (Kabsch and Sander, 1983) indicates that His56 is actually located in a bend rather than in a helix as might be indicated by $\Delta\delta^{13}C^{\alpha}$ (Figure 3a). Whereas RefDB simply flagged the entire BMRB entry 4998 as suspicious, the LACS approach identified an anomaly independent of the structure and pinpointed it to a particular atom in the assigned NMR data.

Conclusion

Linear analysis of chemical shift differences (LACS) can serve a variety of purposes. As a data consistency check, LACS can be used to detect possible referencing problems in datasets in advance of sequence-specific assignments and determinations of secondary structure. The method can be used to apply automated offset corrections. LACS can be used as a diagnostic for errors in spin system identifications. Removal of offset bias can result in improved specification of dihedral angle constraints prior to structure determination. LACS also can serve to complement existing validation tools (Moseley et al., 2004).

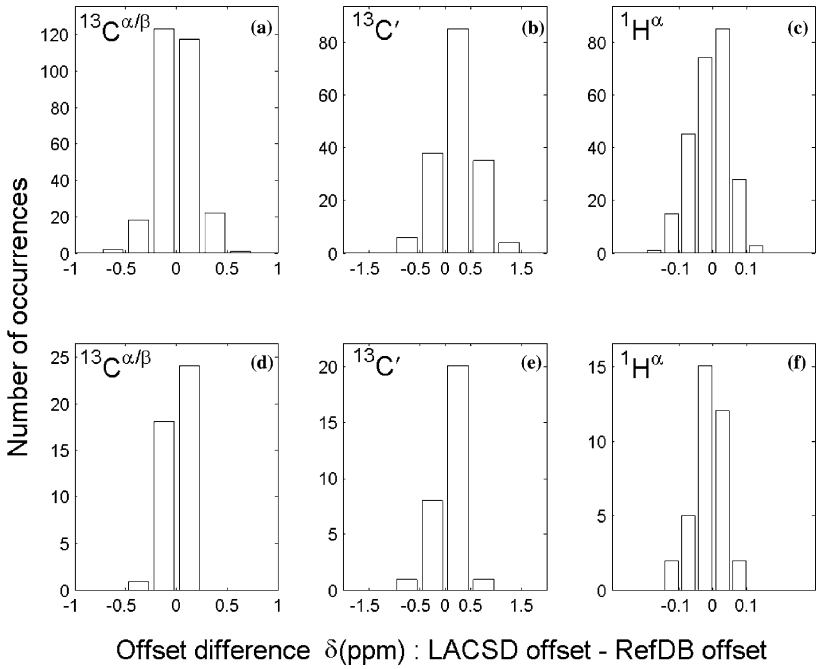

Figure 5. Histogram of offset differences (in ppm) obtained by subtracting values reported by RefDB (Zhang et al., 2003) from those estimated by LACS. (a) $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ for all proteins, (b) $^{13}C'$ for all proteins, (c) $^{1}H^{\alpha}$ for all proteins, (d) $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ for proteins with resolution better than 1.8 Å, (e) $^{13}C'$ for proteins with resolution better than 1.8 Å, (f) $^{1}H^{\alpha}$ for proteins with resolution better than 1.8 Å.

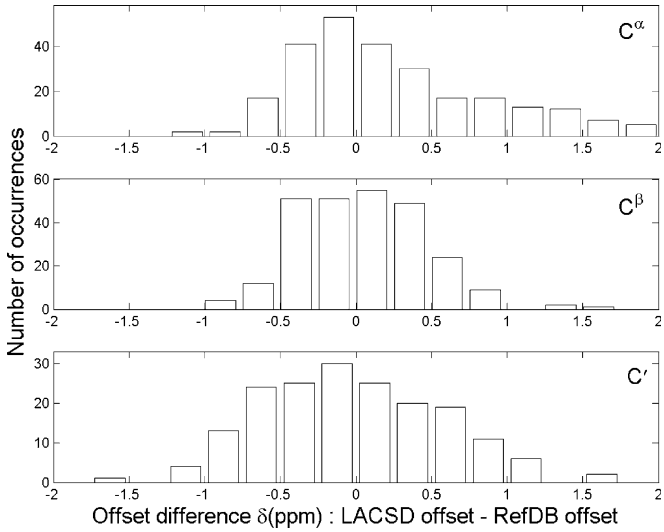

Figure 6. Histogram of offset differences (in ppm) for $\delta^{13}C^{\alpha}$, $\delta^{13}C^{\beta}$, and $\delta^{13}C'$ obtained by subtracting values reported by RefDB (Zhang et al., 2003) from values estimated from the use of $\Delta\ \delta^{1}H^{\alpha}$ as the reference-independent standard.

Amide protons and nitrogens are not considered in our analysis. Two-dimensional scatter plots between the secondary chemical shifts of these nuclei and ($\Delta\delta^{13}C^{\alpha}$–$\Delta\delta^{13}C^{\beta}$) show no discernable relationship (data not shown). These values usually are excluded in secondary structure determinations (Wishart and Sykes, 1994; Hung and Samudrala, 2003), although they have been reported to be helpful in distinguishing a β-strand from a random coil by PSSI (Wang and Jardetzky, 2002). The

chemical shifts of ^{15}N and $^{1}H^{N}$ have been shown to correlate with ϕ_i and ψ_{i-1} (Le and Oldfield, 1994) rather than with ϕ_i and ψ_i. This suggests that referencing problems with these nuclei may be best detected after the structure has been determined.

Our results confirm earlier findings that the major determinants of $\delta^{13}C^{\alpha}$ and $\delta^{13}C^{\beta}$ are local backbone geometry. For $^{13}C'$ and $^{1}H^{\alpha}$, the additional effects of sequence, ring current, and other influences need to be considered. As the database of assigned chemical shifts expands, it may become possible to model these effects in advance of a structure determination.

LACS has been used to create a database of offset-corrected chemical shifts corresponding to nearly 1800 BMRB entries: ~300 with and ~1500 without corresponding 3D structures. This database is available from the NMRFAM website at http://bija.nmrfam.wisc.edu/MANI-LACS/LACS.db. This database can serve as a resource for future analysis of the effects on chemical shifts of amino acid sequence and protein secondary and tertiary structure. This software is part of the MANI tools which can be found at http://bija.nmrfam.wisc.edu/MANI-LACS.

Supporting Information Available: *Table S1*. Correction offsets (in ppm) estimated by two approaches: LACS ($\Delta\delta^{13}C^{\alpha}$–$\Delta\delta^{13}C^{\beta}$) and RefDB, and details on the robust fitting procedure: at http://dx.doi.org/10.1007/s10858-005-1717-0.

Acknowledgments

This research was supported by Biomedical Research Technology Program, National Center for Research Resources, through NIH grant P41 RR02301, which supports the National Magnetic Resonance Facility at Madison, and by the National Institute of General Medical Science's Protein Structure Initiative through NIH grant 1 P50 GM64598, which supports the Center for Eukaryotic Structural Genomics. During part of this work H.E. was supported as a postdoctoral trainee by the National Library of Medicine under grant 5T15LM005359. We thank Eldon L. Ulrich and William M. Westler for advice and encouragement. This work made extensive use of the BioMagResBank.

References

Barnett, V. and Lewis, T. (1994) *Outliers in Statistical Data*, Wiley & Sons, Chichester New York.
Baxter, N.J. and Williamson, M.P. (1997) *J. Biomol. NMR*, **9**, 359–369.
Blanchard, L. Hunter, C.N. and Williamson, M.P. (1997) *J. Biomol. NMR*, **9**, 389–395.
Cornilescu, G. Delaglio, F. and Bax, A. (1999) *J. Biomol. NMR*, **13**, 289–302.
Haigh, C.W. and Mallion, R.B. (1979) *Prog. Nucl. Mag. Res. Sp.*, **13**, 303 .
Holland, P.W. and Welsch, R.E. (1977) *Commun. Stat. Theor. Methods*, **A6**, 813–827.
Hung, L.H. and Samudrala, R. (2003) *Protein Sci.*, **12**, 288–295.
Iwadate, M. Asakura, T. and Williamson, M.P. (1999) *J. Biomol. NMR*, **13**, 199–211.
Kabsch, W. and Sander, C. (1983) *Biopolymers*, **22**, 2577–2637.
Le, H. and Oldfield, E. (1994) *J. Biomol. NMR*, **4**, 341–348.
Mahalanobis, P.C. (1936) *Proc. Natl. Inst. Sci.*, **12**, 49–55 .
Markley, J.L. Meadows, D.H. and Jardetzky, O. (1967) *J. Mol. Biol.*, **27**, 25–35.
Moseley, H.N. Sahota, G. and Montelione, G.T. (2004) *J. Biomol. NMR*, **28**, 341–355.
Neal, S. Nip, A.M. Zhang, H. and Wishart, D.S. (2003) *J. Biomol. NMR*, **26**, 215–240.
Osapay, K. and Case, D.A. (1991) *J. Am. Chem. Soc.*, **113**, 9436–9444.
Osapay, K. and Case, D.A. (1994) *J. Biomol. NMR*, **4**, 215–230.
Reiley, M.D. Thanabal, V. and Omecinsky, D.O. (1992) *J. Am. Chem. Soc.*, **114**, 6251–6252.
Schwarzinger, S. Kroon, G.J. Foss, T.R. Chung, J. Wright, P.E. and Dyson, H.J. (2001) *J. Am. Chem. Soc.*, **123**, 2970–2978.
Seavey, B.R. Farr, E.A. Westler, W.M. and Markley, J.L. (1991) *J. Biomol. NMR*, **1**, 217–236.
Spera, S. and Bax, A. (1991) *J. Am. Chem. Soc.*, **113**, 5490–5492.
Sternlicht, H. and Wilson, D. (1967) *Biochemistry*, **6**, 2881–2892.
Wang, Y. and Jardetzky, O. (2002) *Protein Sci.*, **11**, 852–861.
Wishart, D.S. Bigam, C.G. Holm, A. Hodges, R.S. and Sykes, B.D. (1995) *J. Biomol. NMR*, **5**, 67–81.
Wishart, D.S. and Case, D.A. (2001) *Methods Enzymol.*, **338**, 3–34.
Wishart, D.S. and Nip, A.M. (1998) *Biochem. Cell Biol.*, **76**, 153–163.
Wishart, D.S. and Sykes, B.D. (1994) *J. Biomol. NMR*, **4**, 171–180.
Wishart, D.S. Sykes, B.D. and Richards, F.M. (1991) *J. Mol. Biol.*, **222**, 311–333.
Zhang, H. Neal, S. and Wishart, D.S. (2003) *J. Biomol. NMR*, **25**, 173–195.