J Biomol NMR (2007) 39:247–257
DOI 10.1007/s10858-007-9193-3

ARTICLE

Nearest-neighbor effects on backbone alpha and beta carbon chemical shifts in proteins

Liya Wang · Hamid R. Eghbalnia · John L. Markley

Received: 13 April 2007 / Accepted: 4 September 2007 / Published online: 26 September 2007

Abstract We present a method for analyzing the chemical shift database to yield information on nearest-neighbor effects on carbon-13 chemical shift values for alpha and beta carbons of amino acids in proteins. For each amino acid sequence XYZ, we define two correction factors, $\Delta(^{X}Y)_s$ and $\Delta(Y^{Z})_s$, representing the effects on $(\delta^{13}C^{\alpha} - \delta^{13}C^{\beta})$ for residue Y from the preceding residue (X) and the following residue (Z), where X, Y, and Z represent one of the 20 naturally occurring amino acids, Δ designates the change in value or the correction factor (in ppm), and s is an index standing for one of three "pseudo secondary structure states" derived from chemical shift dispersions, which we show represent residues in primarily α-helix, β-strand, and non-$\alpha\beta$ (coil). The correction factors were obtained from maximum likelihood fitting of $(\delta^{13}C^{\alpha} - \delta^{13}C^{\beta})$ values from the chemical shifts of 651 proteins to a mixture of three Gaussians. These correction factors were derived strictly from the analysis of assigned chemical shifts, without regard to the three-dimensional structures of these proteins. The corrections factors were found to differ according to the secondary structural environment of the central residue (deduced from the chemical shift distribution) as well as by different identities of the nearest neighboring residues in the sequence. The areas subsumed by the sequence-dependent chemical shift distributions report on the relative energies of the sequences in different pseudo secondary structural environments, and the positions of the peaks indicate the chemical shifts of lowest energy conformations. As such, these results have potential applications to the determination of dihedral angle restraints from chemical shifts for structure determination and to more accurate predictions of chemical shifts in proteins of known structure. From a database of chemical shifts associated well-defined three-dimensional structures, comparisons were made between DSSP designations derived from three-dimensional structure and pseudo secondary structure designations derived from nearest-neighbor corrected chemical shift analysis. The high level of agreement between the two approaches to classifying secondary structure provides a measure of confidence in this chemical shift-based approach to the analysis of protein structure.

Keywords Proteins · Carbon chemical shifts · Dihedral angle · Nearest-neighbor effects · Pseudo secondary structure · Hypersurface

Electronic supplementary material The online version of this article (doi:10.1007/s10858-007-9193-3) contains supplementary material, which is available to authorized users.

L. Wang · H. R. Eghbalnia (✉) · J. L. Markley (✉)
National Magnetic Resonance Facility at Madison, 433 Babcock Drive, Madison, WI 53706, USA
e-mail: eghbalni@nmrfam.wisc.edu

J. L. Markley
e-mail: markley@nmrfam.wisc.edu

L. Wang · J. L. Markley
Graduate Program in Biophysics and Center for Eukaryotic Structural Genomics, University of Wisconsin-Madison, Madison, USA

H. R. Eghbalnia · J. L. Markley
Biochemistry Department, University of Wisconsin-Madison, 171a Biochemistry Addition, 433 Babcock Dr, Madison, WI 53706, USA

H. R. Eghbalnia
Mathematics Department, University of Wisconsin-Madison, 811 Van Vleck Hall, 480 Lincoln Drive, Madison, WI 53706, USA

Introduction

Although structural and conformational effects on NMR chemical shift have been known for a long time (Markley et al. 1967; McDonald and Phillips 1967; Nakamura and Jardetzky 1967), it is still difficult to precisely decipher the information contained in the chemical shifts of amino acids in proteins. For example, α- and β-carbon chemical shifts are well known to be related to protein dihedral angles (Spera and Bax 1991; Iwadate et al. 1999), but angular predictions from chemical shifts are still not precise enough to support accurate structure determinations. Only in recent years, with the accumulation of large chemical shift databases, has it become possible to begin to use statistical analysis to unravel this information and to attempt to use it for refinement of NMR structures (Wishart et al. 1992; Wishart and Sykes 1994; Kuszewski et al. 1995; Cornilescu et al. 1999).

As one of the largest correction factors for structure effects, neighboring residue effects have been analyzed by both experimental and statistical methods (Braun et al. 1994; Wishart et al. 1995; Iwadate et al. 1999; Schwarzinger et al. 2001; Wang and Jardetzky 2002a). The experimental approaches had three serious drawbacks. First, they were limited to the random coil and thus neglected possible variation of nearest-neighbor effects with secondary structure. Because data from only 20 short model peptides were used to simulate 8,000 tripeptide sequences (Wishart et al. 1995), the results captured nearest-neighbor effects under limited conditions. Finally, the denaturing solvents used likely had selective effects on the chemical shifts. Earlier statistical approaches, on the other hand, suffered from incorrect referencing (Wang et al. 2002a) and from the limited quantity of chemical shift data available at the time. Furthermore, the quality of the data in these statistical analyses was biased by the use of mean values rather than distributions.

An earlier statistical analysis of nearest-neighbor effects on chemical shifts (Wang et al. 2002a) derived its results from a database of protein chemical shifts associated with known three-dimensional structure. Our approach differs from this in that we analyze protein chemical shifts in the absence of structure and examine the distribution of chemical shifts for individual residue types and dipeptides as modeled by fitting to three Gaussian functions that we associate with "pseudo secondary structural states". Our approach focuses on the chemical shift difference ($\delta^{13}C^{\alpha} - \delta^{13}C^{\beta}$), which serves to sharpen the distribution of chemical shifts corresponding to three pseudo secondary structural states: α-helix; β-strand, and non-αβ ("coil") (Wang et al. 2006). This chemical shift difference can be considered to correspond to a "pseudo atom" that carries important information about the protein backbone conformation.

For the present study, we established a large empirical database containing reference-corrected ^{13}C chemical shifts (Wang et al. 2005) from proteins that had associated three-dimensional structures so that we could subsequently compare pseudo structural classifications with those derived from structure by DSSP analysis (Kabsch and Sander 1983). Use of the three-state model greatly improves robustness in cases of sparse data and enabled us to determine separately the nearest-neighbor effects on the chemical shifts of residues in these three major states. The maximum likelihood estimate for the three Gaussian functions was obtained by optimally fitting the corresponding log-likelihood function to the empirical data in the least squared sense. Henceforth, we simply use the term "fitting" to describe the maximum likelihood estimation.

The three states represented by the three overlapped Gaussians represent a subset of possible secondary structural states. These states are defined solely on the basis of the chemical shift distribution; their definition does not rely on any other information, such as inferred hydrogen bonds. We demonstrate that residues identified as being in the three pseudo secondary structural states on the basis of chemical shift dispersions are largely in agreement with corresponding identifications derived from backbone conformation. One important difference is that a larger number of conformational states can be derived from backbone conformation than the three from chemical shift alone. Turns are one example: residues classified at being in turns on the basis of backbone conformation were found to correspond to residues with bimodal ($\delta^{13}C^{\alpha} - \delta^{13}C^{\beta}$) chemical shift distributions, with one Gaussian component largely in the pseudo coil region and the other in the pseudo helix region.

Methods

Data sets used

The data set used to derive sequence-dependent chemical shift dispersions was extracted from BioMagResBank (Seavey et al. 1991) in April, 2006, for 651 proteins that met the following criteria: sequence length longer than 50 residues, assigned $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ signals, data collected at a pH value >5 and at a temperature within 25 ± 15 °C. The requirement of pH >5 served to filter out data from protonated aspartate and glutamate residues, which have pK_a value of 4.3 and 4.7, respectively. Effects of pH are much less important above pH 5 (Richarz and Wüthrich 1978; Wishart and Case 2001). In addition, within the selected temperature range, thermal effects on chemical shifts are expected to be small. We used the validation software package LACS (Wang et al. 2005) to correct errors in

referencing and to remove outliers in the database that may represent mis-assignments. Although this study made use of the chemical shift difference ($\delta^{13}C^{\alpha} - \delta^{13}C^{\beta}$), which is reference error-free, correct referencing is still necessary for studying neighboring effects on glycine (only $\delta^{13}C^{\alpha}$ available) and for identifying trans/cis proline and oxidized/reduced cysteine (on the basis of $\delta^{13}C^{\beta}$ shifts). Because insufficient data were available for oxidized cysteine and cis prolyl residues, we restricted the study to reduced cysteine residues (as indicated by $\delta^{13}C^{\beta} < 32$ ppm; Sharma and Rajarathnam 2000) and to prolines involved in trans Xaa-Pro linkages ($\delta^{13}C^{\beta} < 33$ ppm; Schubert et al. 2002).

For the comparison of pseudo secondary structure designations with those from DSSP analysis (Kabsch and Sander 1983), this database of ~50,000 residues was filtered to include only residues associated with well-defined three-dimensional structure. This yielded a second database (subset of the first) of ~30,800 residues with chemical shifts and associated DSSP codes.

Chemical shift hypersurfaces related to dihedral angles were built on the TALOS database (Cornilescu et al. 1999) downloaded on May 2005.

The three-state model

The relationship between chemical shifts and protein secondary structure has been known for a long time, and it has been utilized effectively in several sophisticated tools for protein secondary structure prediction (Wishart and Sykes 1994; Wang and Jardetzky 2002b; Hung and Samudrala 2003; Eghbalnia et al. 2005). Although eight different forms of secondary structure can be identified in a protein on the basis of backbone conformation, as defined by DSSP (Kabsch and Sander 1983), all existing methods for NMR chemical shift analysis classify residues into three groups, α-helix, random coil, and β-strand.

Our earlier work showed that ($\delta^{13}C^{\alpha} - \delta^{13}C^{\beta}$) has a distribution that can be well fitted by the sum of three Gaussian functions (Wang et al. 2006). Because of the opposite effects of backbone conformation on the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts, the three states are much better visualized by the distribution of ($\delta^{13}C^{\alpha} - \delta^{13}C^{\beta}$) than by $\delta^{13}C^{\alpha}$ or $\delta^{13}C^{\beta}$ alone. This is shown for the case of alanine chemical shifts in Fig. 1. We proposed that these chemical shift distributions represent three distinct states of individual residues that correspond to the three most common forms of secondary structure. The peak values in this distribution represent the chemical shift values of residues in the most highly populated (lowest energy or most stable) conformation within each form of secondary structure.

To determine nearest-neighbor effects for each amino acid, we fitted the ($\delta^{13}C^{\alpha} - \delta^{13}C^{\beta}$) data for each different preceding and following amino acid residue type with three Gaussian functions to yield peak positions and areas (Table 1). As an example, the data and fitted curves for alanine (as the central Y residue) with different preceding and following residue types are shown in Fig. 2. The bold line represents the fit for the given residue type (alanine) to all of the data, and the dashed lines represent fits to 40 different data subsets (for each of the 20 proceeding residue types and each of the 20 following residue types). Because the three-state model holds for the subsets of chemical shift data with varying neighbors, the nearest-neighbor effects on chemical shifts can be determined from the distance (in ppm) between the corresponding peaks.

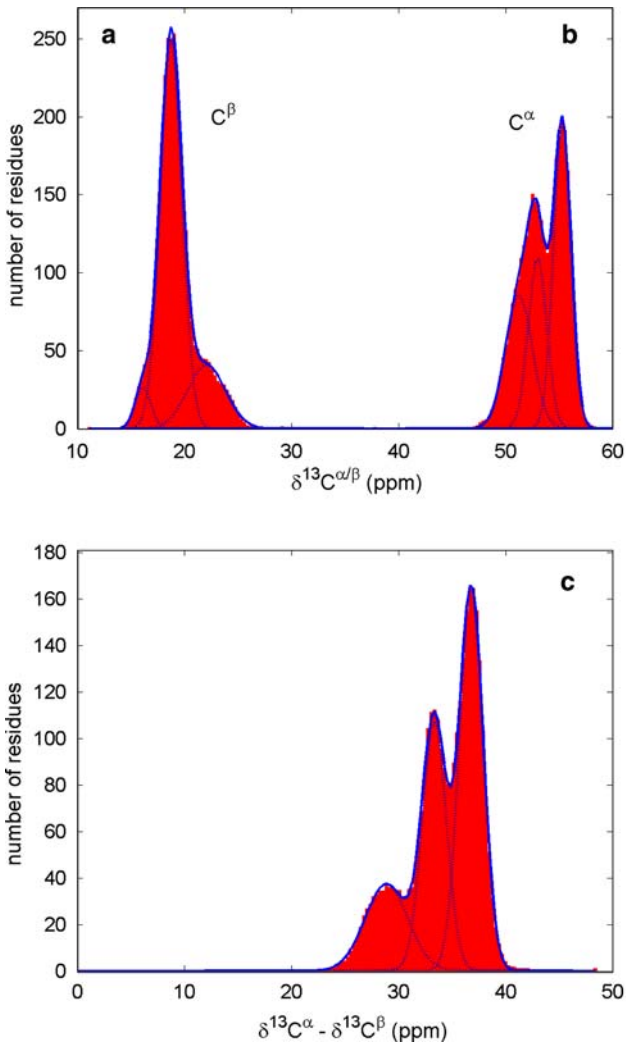

Fig. 1 Gaussian fitting of (a) $\delta^{13}C^{\alpha}$, (b) $\delta^{13}C^{\beta}$, and (c) ($\delta^{13}C^{\alpha} - \delta^{13}C^{\beta}$) chemical shifts from the database for alanine

Table 1 Mean unbiased chemical shift values for residues in different secondary structures; percentages of residues in the database with each secondary structure

Amino acid	$(\delta^{13}C^{\alpha} - \delta^{13}C^{\beta})_{\beta}$ (ppm)	$(\delta^{13}C^{\alpha} - \delta^{13}C^{\beta})_{coil}$ (ppm)	$(\delta^{13}C^{\alpha} - \delta^{13}C^{\beta})_{\alpha}$ (ppm)	β-strand (%)	Coil (%)	α-helix (%)
Ala	29.3	33.6	36.9	18	33	49
Cys[a]	26.6	30.2	36.6	18	57	25
Asp	10.8	13.3	16.7	23	40	37
Glu	22.7	26.6	29.9	19	27	54
Phe	14.5	18.4	22.1	36	27	37
Gly[b]	42.9	45.4	47.5	1	87	11
His	23.7	26.1	29.7	34	31	36
Ile	18.5	22.4	27.4	22	45	33
Lys	19.6	23.4	27.1	19	36	45
Leu	9.7	13.0	16.2	30	22	48
Met	19.6	23.0	26.3	27	29	45
Asn	12.7	14.5	17.6	31	32	37
Pro[c]	29.3	31.1	33.7	10	66	24
Gln	22.9	26.6	30.4	20	32	48
Arg	21.7	25.5	29.3	24	32	44
Ser	–8.0	–5.2	–1.5	32	36	32
Thr	–12.0	–8.0	–2.4	18	59	23
Val	25.2	29.4	35.0	27	43	30
Trp	25.0	29.2	33.4	45	43	12
Tyr	15.3	19.3	23.4	34	38	27

[a] Reduced cysteine only

[b] $\delta^{13}C^{\alpha}$ only

[c] Trans Xaa-Pro peptide bond configuration only

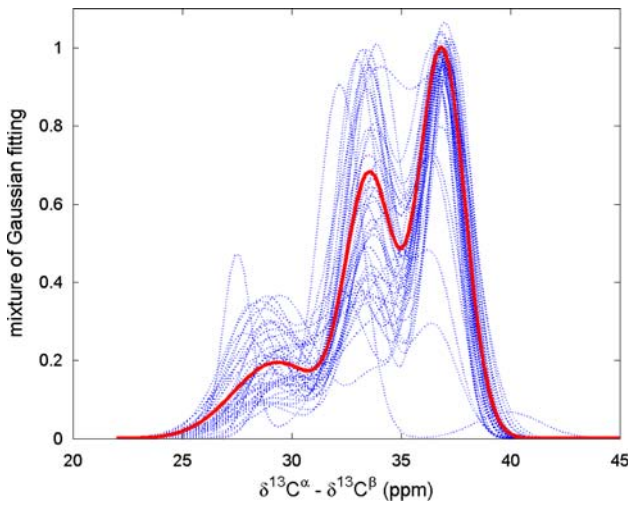

Fig. 2 Gaussian fitting of data for alanine with different neighboring residues. The solid curve is the mean for all neighbors, and the dotted curves are for all 40 cases: Ala with each of 20 preceding and 20 following residues

Determination of $\Delta(^{X}Y)_s$ and $\Delta(Y^{Z})_s$

The preceding effect of residue X on Y in state s for sequence XYZ is defined as

$$\Delta\left(^{X}Y\right)_s = \delta\left(^{X}Y\right)_s - \delta\left(^{A}Y\right)_s \quad (1)$$

And the following effect of residue Z on Y in state s for sequence XYZ is defined as

$$\Delta\left(Y^{Z}\right)_s = \delta\left(Y^{Z}\right)_s - \delta\left(Y^{A}\right)_s \quad (2)$$

Here X, Y, and Z each correspond to one of the 20 amino acids; s represents different secondary structures corresponding to three fitted peaks; A represents alanine; and the δ value is determined from the fitted peak position corresponding to s in the relevant subset of data from the reference corrected chemical shift database.

We have used alanine as the fixed residue (either X = Ala as the preceding residue when examining the effects of different Z, or Z = Ala as the following residue when examining the effects of different X). The use of alanine in this way simplifies the procedure for calculating nearest-neighbor effects from Eqs. 1 and 2. The underlying assumption is that X as alanine before Y does not change the distribution for different Z and that Z as alanine following Y does not change the distribution for different X. Alanine was chosen because it has a high abundance in proteins (better statistics) and because it is less likely than other residues to interact with residues at position $n \pm 2$. Alanine also has been used as a reference residue for measurements of nearest-neighbor effects in short peptides (Wishart et al. 1995). An additional reason for using alanine here is that its chemical shift distribution is well separated into three peaks, making it easy to distinguish nearest-neighbor effects (Figs. 1 and 2).

Neighboring information aids protein secondary structure prediction

The prediction of protein secondary structure from sequence alone has been widely studied since the 1970s (Chou and Fasman 1974; Lim 1974; Garnier et al. 1978). We previously showed that chemical shift dispersions contain information about propensities of amino acid residues to be found in the three pseudo secondary structural states (Wang et al. 2006). Here we examine whether the use of information on nearest-neighbor effects on chemical shift dispersions can improve this approach to secondary structure prediction from sequence.

The percentage of residues Y assigned to pseudo α-helix, random coil or β-strand is defined as,

$$p_s = \frac{S_s}{\sum_s S_s} \tag{3}$$

where s represents one of the three secondary structures and S represents the area under corresponding peaks. We use the following equation to calculate the secondary structure propensity of residue Y given a protein sequence of XYZ,

$$p_s(\mathrm{Y}) = \frac{p_s^{\mathrm{X_Y}}(\mathrm{Y})p_s^{\mathrm{Y}}(\mathrm{Y})p_s^{\mathrm{Y^Z}}(\mathrm{Y})}{\sum_s p_s^{\mathrm{X_Y}}(\mathrm{Y})p_s^{\mathrm{Y}}(\mathrm{Y})p_s^{\mathrm{Y^Z}}(\mathrm{Y})} \tag{4}$$

The definitions of each term in Eq. 4 are given below.

- $p_s^{\mathrm{X_Y}}(\mathrm{Y})$, the percentage of residue Y (if preceded by X) in state s, is calculated with Eq. 3 by estimating S with the chemical shifts of Y preceded by X.
- $p_s^{\mathrm{Y^Z}}(\mathrm{Y})$ is estimated from the chemical shifts of Y with a following Z.
- $p_s^{\mathrm{Y}}(\mathrm{Y})$, the internal secondary structure propensity of Y averaged over all neighbors, is estimated from all of the chemical shifts contained in the database except those with a following proline.

Although this method for calculating pseudo secondary structure propensities could be extended to sequences of up to five or more residues, two reasons guided our decision to consider only the adjacent residues: (a) the limited size of the data set could result in the inadvertent introduction of large uncertainties if a longer sequence was considered, and (b) neighboring effects are believed to be limited mostly to neighboring residues (or the local environment). By applying Eq. 4 to each residue of the protein, one obtains an estimate of its secondary structure propensity in the given protein sequence. This estimation is based on sequence information only (no chemical shift data are needed). This method avoids the biases that could arise from the presence of unevenly sampled data in the database and corrects biases that could be caused by direct counting from a non-representative chemical shift database.

Protein secondary structure prediction using both sequence and chemical shifts

The combined use of sequence information and chemical shifts has been successful in identifying protein secondary structural elements (Hung and Samudrala 2003; Eghbalnia et al. 2005). On the basis of our three-state model, we use the following equation to predict protein secondary structure of Y (for sequence XYZ) by combining sequence information and chemical shifts:

$$p_s(\mathrm{Y}) = \frac{G_s^{\mathrm{X^Y}}(\delta_\mathrm{X})G_s^{\mathrm{X_Y}}(\delta_\mathrm{Y})G_s^{\mathrm{Y^Z}}(\delta_\mathrm{Y})G_s^{\mathrm{Y_Z}}(\delta_\mathrm{Z})}{\sum_s G_s^{\mathrm{X^Y}}(\delta_\mathrm{X})G_s^{\mathrm{X_Y}}(\delta_\mathrm{Y})G_s^{\mathrm{Y^Z}}(\delta_\mathrm{Y})G_s^{\mathrm{Y_Z}}(\delta_\mathrm{Z})} \tag{5}$$

where $G_s^{\mathrm{X^Y}}(\delta_\mathrm{X})$ is the value of the Gaussian function (corresponding to peak s) at a certain point δ_X. This Gaussian function is acquired by fitting the distribution in the database of each residue X followed by Y. The value of s is 1, 2, or 3, indicating, respectively, α-helix, random coil, or β-strand. The difference between the chemical shifts of $^{13}\mathrm{C}^\alpha$ and $^{13}\mathrm{C}^\beta$ is δ_X.

Equation 5 not only combines sequence information with chemical shifts but also considers secondary structure identification in a size 3 window. As a result, the prediction is naturally "smoothed" with sequence information from the nearest neighbors on the left and right. The sequence-specific information was derived by least squares fitting to the three-state model of the appropriate subset of the chemical shift database. The resulting secondary structure propensity is considered to be statistically more reliable than that achieved by sampling, given the sparsity of protein chemical shift data. As more data become available, the prediction accuracy achievable by this approach will increase.

Chemical shift hypersurface and its improvement by incorporating nearest-neighbor corrections

A chemical shift hypersurface, $\Delta(\varphi, \psi)$, can be used to empirically correlate secondary chemical shifts of a given residue with its dihedral angles (φ, ψ) (Spera and Bax 1991). We have created a hypersurface for the chemical shift difference $(\delta^{13}\mathrm{C}^\alpha - \delta^{13}\mathrm{C}^\beta)$ by convoluting of each of the chemical shift values, $\delta(\varphi_k,\psi_k)$, with a Gaussian function of dihedral angles, prior to addition and normalization, and by extending the summation over all residues k.

$$\Delta(\varphi,\psi)=\frac{\sum\delta(\varphi_k,\psi_k)\exp\left(-\left((\varphi-\varphi_k)^2+(\psi-\psi_k)^2\right)/N\right)}{\sum\exp\left(-\left((\varphi-\varphi_k)^2+(\psi-\psi_k)^2\right)/N\right)} \quad (6)$$

where N is the number of residues in the database and $\delta(\varphi_k, \psi_k)$ is replaced with the chemical shifts difference between alpha and beta carbons to build a new hypersurface.

The neighboring effect corrected hypersurface is constructed in two steps. First the neighboring effects on Y for sequence XYZ are adjusted with the following equation,

$$\delta_{\text{adjusted}}(\mathrm{Y})=\delta_{\text{observed}}(\mathrm{Y})-\Delta\left({}^{\mathrm{X}}\mathrm{Y}\right)-\Delta\left(\mathrm{Y}^{\mathrm{Z}}\right) \quad (7)$$

where s is omitted because only non-helix non-strand residues (identified with Eq. 5 with a cut-off of 0.5) are used. The nearest-neighbor correction factors that we have derived are presented in Supplementary Table s1. The second step is to construct a hypersurface by substituting the nearest-neighbor adjusted chemical shifts into Eq. 6.

One use of a chemical shift hypersurface is in predicting chemical shifts from known structure (Wishart and Nip 1998). Previous applications of this kind have not made use of nearest-neighbor corrections. With our corrected hypersurface, the neighboring effects can be added back to the predicted chemical shifts ($\delta^0_{\text{predicted}}$) with the equation

$$\delta_{\text{predicted}}(\mathrm{Y})=\delta^0_{\text{predicted}}(\mathrm{Y})+\Delta\left({}^{\mathrm{X}}\mathrm{Y}\right)+\Delta\left(\mathrm{Y}^{\mathrm{Z}}\right) \quad (8)$$

Results and discussion

Nearest-neighbor effects on chemical shifts

Derived nearest-neighbor effects for alanine are shown in Fig. 2. The solid line represents the fit with all chemical shift data available for alanine, and the dashed line represents fits with subsets of the chemical shift data for alanine with each of the 20 preceding and each of the 20 following residue types. The curves (Fig. 2) clearly show that the correction factor depends on the nature (residue type) and location (preceding or following) of the neighboring residue. The largest correction factor is for proline as a following residue. The nearest-neighbor effects also differ among the three Gaussians, showing that they depend on the secondary structure in which the residue is located.

Calculation of the nearest-neighbor effects required the fitting of a total of 800 subsets of the data. The majority of these sets yielded three, well-isolated Gaussians, e.g. W^P (Fig. 3a). In a minority of cases, however, heavy overlapping of the peaks made it especially difficult to achieve a robust fit, e.g. I^T (Fig. 3b). Although we carefully inspected each fitted data set and adjusted the fit on the basis of prior knowledge about the approximate range for the position of

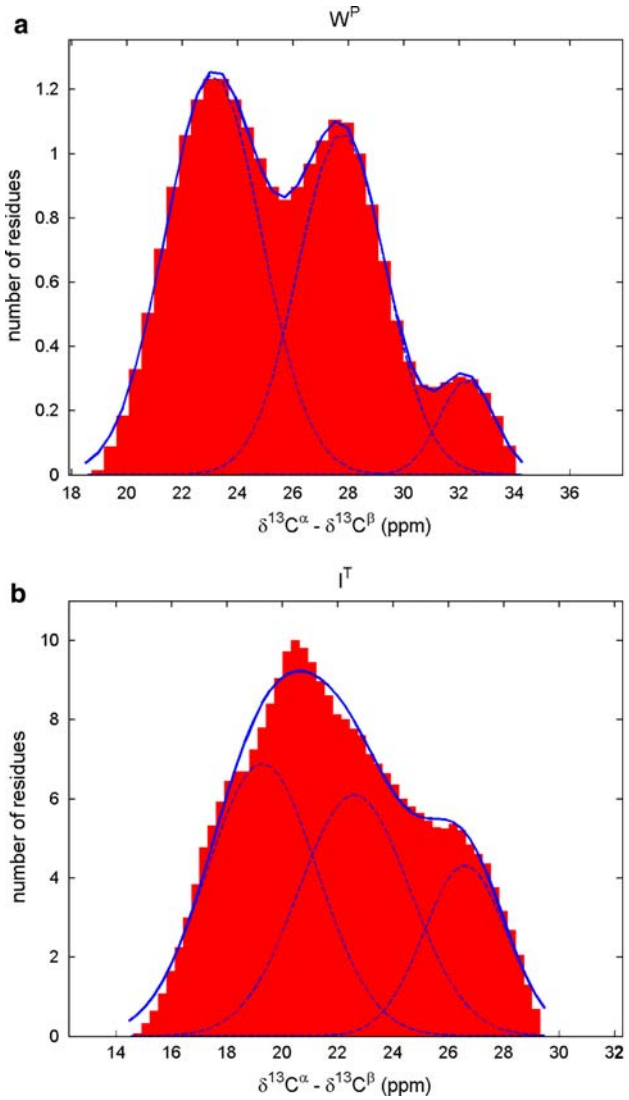

Fig. 3 Gaussian fitting for W^P (Trp followed by Pro) and I^T (Ile followed by Thr)

the peak, some of the results may have been biased by overlap. Figure 3 also shows that in most cases it is the degree of isolation of the peaks (resolution), rather than the number of data points, that determines the fitting robustness. In the majority of cases, excellent peak separation made it possible to calculate nearest-neighbor effects even for low abundance sequences, for example, tryptophan followed by proline (Fig. 3a).

Comparison of computed nearest-neighbor effects with those from experiment

Few experimental data are available for comparison purposes. One experimental data set is that from Wishart et al. (1995) on peptides with a residue X preceded by Gly or followed by Pro. For the available data on 19 residues X, we

compared the quantity $[(\delta^{13}C^{\alpha} - \delta^{13}C^{\beta})_{G-X} - (\delta^{13}C^{\alpha} - \delta^{13}C^{\beta})_{X-P}]$ derived from the experimental peptide data with that derived from our statistical analysis (Fig. 4). The computationally derived data were consistent with the experimental data for 15 of the 19 residues, but cysteine, aspartate, methionine, and tryptophan showed significant differences. We examined the result for one of these outliers to investigate the possible impact of sparse data on the computed result (Fig. 5). Although overlaps in the data for D^P could lead to potential inaccuracy in estimating the position (mean) of the three fitted Gaussians, we determined by cross-validation that the positions of the center peak for D^P as well as for ^GD were determined robustly.

Nearest-neighbor effects on predictions of secondary structure from chemical shifts

The area under each Gaussian curve changes with different neighboring residues (Figs. 2 and 6). Figure 6 shows that, although alanine is believed to prefer an α-helical conformation, it has a very low probability of adopting an α-helical conformation when it has certain neighboring residues, for example when it is followed by asparagine or proline. Thus, our hypothesis is that predictions of secondary structure from chemical shifts will be improved if nearest-neighbor effects are taken into account.

As a test of this idea, we used information from the fitted Gaussians (without and with consideration of nearest-neighbor effects) to predict the position of an α-helix in a particular protein of known three-dimensional structure (PDB ID 1B2F) (Diao 2003) from its sequence alone (Fig. 7). The open bars show the predicted result based solely on the inherent propensity of the amino acid alone

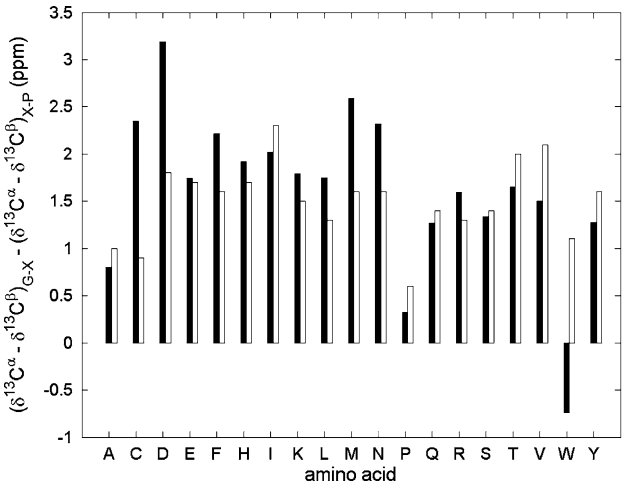

Fig. 4 Comparison between nearest-neighbor effects on chemical shifts derived statistically (solid bars) and measured experimentally (open bars)

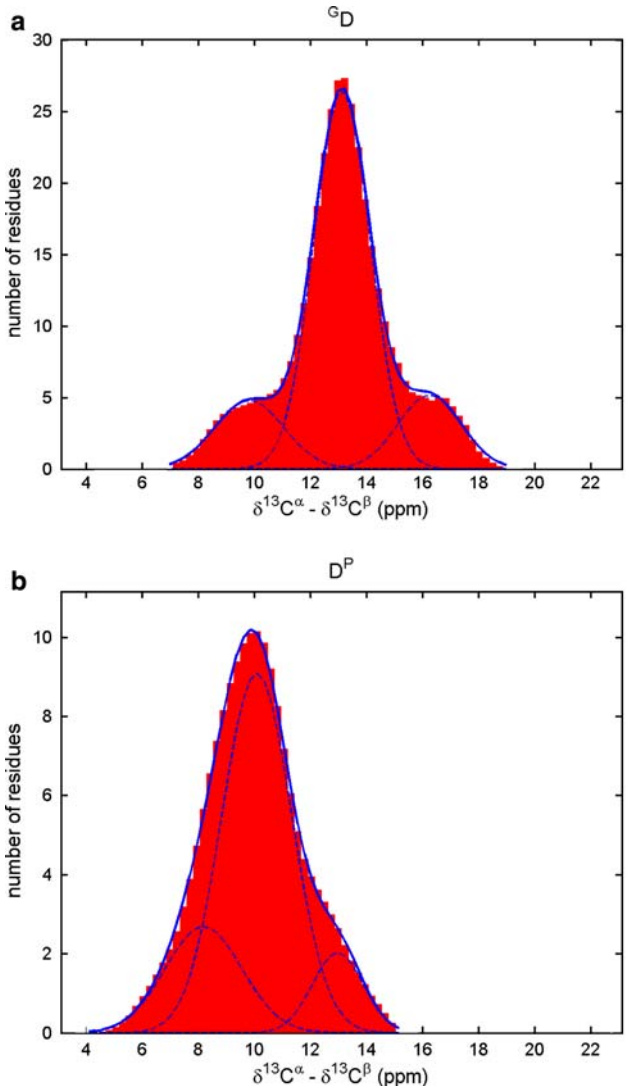

Fig. 5 Gaussian fitting for ^GD (Asp preceded by Gly) and D^P (Asp followed by Pro)

$(P_s^Y(Y))$, and the solid bars show the prediction achieved by adding nearest-neighbor effects $(P_s^{XY}(Y)$ and $P_s^{YZ}(Y))$ to the inherent propensity with Eq. 4. The results show that the position of the α-helix was better determined when neighboring effects were taken into consideration. This is consistent with the large body of experimental data on neighboring effects on α-helix propensities of residues (Rohl and Baldwin 1998).

Comparison of pseudo secondary structure predictions from chemical shift dispersions with secondary structure categories determined from three-dimensional structure

By using the more limited set of chemical shift data associated with well-defined backbone conformation, we

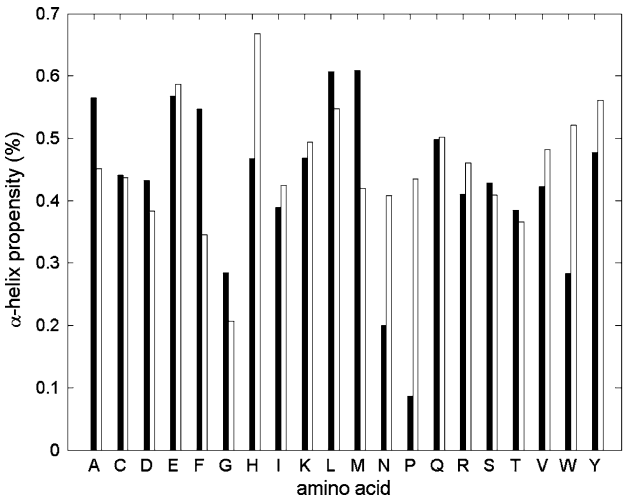

Fig. 6 Nearest-neighbor effects on the α-helix propensity of alanine from the identity of the preceding residue (open bars) and following residue (solid bars)

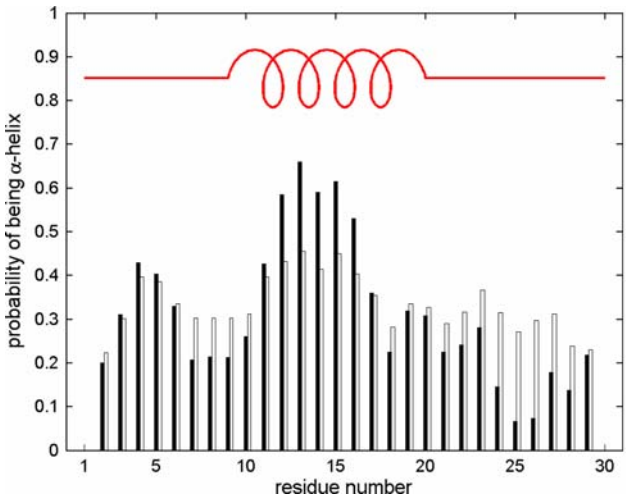

Fig. 7 Predictions of α-helix in a protein (Protein Data Bank 1B2F) from sequence information using Eq. 4 without (open bars) and with (solid bars) nearest-neighbor corrections

compared secondary structure categories predicted from nearest-neighbor corrected chemical shift data (Eq. 5) with secondary structure assignments from DSSP (Kabsch and Sander 1983). The comparison did not involve any tuning of the data set, nor was it necessary to perform any training. The mapping was carried out in two different ways. In the first approach, residues from the database of shifts with three-dimensional structure were sorted into their pseudo secondary structure categories; then these were mapped into the eight DSSP code categories (Table 2). For the three common categories, the level of correspondence was (Table 2): 87.3% (α-helix/α-helix), 61.2% (β-strand/extended strand), and 75.83% (coil/unassigned). In the second (reciprocal) approach, residues from the database of shifts with three-dimensional structure were sorted into their eight DSSP code categories, and then these were mapped into the three pseudo secondary structure categories (Table 3). For the three common categories, the level of correspondence analyzed in this fashion was (Table 3): 90.1% (α-helix/α-helix), 65.1% (β-strand/extended strand), and 82.5% (coil/unassigned).

We acknowledge that other approaches to predicting secondary structure from chemical shifts (with and without sequence information) yield better agreement with DSSP results for these three categories of secondary structure (Eghbalnia et al. 2005; Wang and Jardetzky 2002b). The point we wish to make here is that the high degree of correlation between these completely independent approaches (Tables 2 and 3) suggests that the three-peak model suitably represents the secondary structure in the majority of cases. It is interesting to note that the additional DSSP categories map into mixed pseudo secondary structural states (Tables 2 and 3): β-bridge (to β-strand and coil), 3_{10}-helix (to coil and α-helix), π-helix (to coil and α-helix), hydrogen bonded turn to (coil and α-helix), and bend (primarily to coil but in equal minor measure to α-helix and β-strand).

Nearest-neighbor effects on the back-prediction of chemical shifts from structure

Because nearest neighbor effects modify chemical shifts, we hypothesized that the back-prediction of chemical shifts from structure could be improved by incorporation of nearest-neighbor correction factors. A well-characterized database of chemical shifts associated with ϕ, ψ angles is the TALOS database. We used the TALOS database to build an initial set of chemical shift hypersurfaces for individual residue types. We then corrected the ($\delta^{13}C^{\alpha}$ – $\delta^{13}C^{\beta}$) values defining these hypersurfaces for neighboring effects by incorporating factors based on Eq. 7 and the protein sequences. We finally used Eq. 6 to construct the corrected hypersurfaces from the corrected ($\delta^{13}C^{\alpha}$ – $\delta^{13}C^{\beta}$) values. Figure 8a shows the chemical shift hypersurface for alanine corrected for nearest-neighbor effects, with adjacent lines having a chemical shift difference of 0.5 ppm. The centers of α-helix and β-strand are indicated, respectively, by red and blue contours. The results for α-helix or β-strand regions of the final hypersurface for alanine (Fig. 8a) show that large changes in dihedral angles are correlated with chemical shift changes as small as 0.5 ppm. Thus the assignment of a deterministic chemical shift value in these regions may be highly error-prone. Instead, it should be more accurate and useful to make use of a probability distribution over the range of values. For the above reason, accurate back prediction of chemical shifts from structure is

Table 2 Mapping of residues from the database of shifts with three-dimensional structure corresponding to each of the three pseudo secondary structure categories into the eight DSSP code categories

	DSSP codes for residues							
Prediction from sequence-dependent chemical shift dispersion	α-helix	β-bridge	extended strand	3_{10}-helix	π-helix	Hydrogen-bonded turn	Bend	Unassigned
Number of residues in the various DSSP classifications	10760	279	6919	645	11	2512	3086	6592
Pseudo secondary structure category	Mapping into the eight DSSP code categories (% of total residues in the category)							
α-helix	*87.3*	1.8	0.6	45.3	45.5	32	12.3	3.5
β-strand	0.3	43	*61.2*	1.6	0	2.8	11	20.3
Coil	12.4	54.8	37.9	53.2	54.5	64.9	76.3	*75.8*

Italic numbers indicate correspondences between like categories

Table 3 Mapping of residues from the database of shifts with three-dimensional structure corresponding to each of the eight DSSP code categories into the three categories of pseudo secondary structure

DSSP code category	Number of residues in the various DSSP code categories	Mapping to categories of pseudo secondary structure (% of total residues in the category)		
		α-helix	β-strand	Coil
α-helix	9633	*90.9*	0.1	9
β-bridge	185	1.6	38.4	60
Extended strand	4923	0.4	*65.1*	34.5
3_{10} helix	458	44.3	0.7	55
π helix	10	40	0	60
Hydrogen bonded turn	1779	28.7	1.3	70
Bend	2299	10	6.9	83.1
Unassigned	5052	2.3	15.1	*82.5*

Italic numbers indicate correspondences between like categories

limited to residues in regions that are neither α-helix nor β-strand (Fig. 8b, Eq. 5). Residues within 1.645 standard deviation (90% confidence) of the central peak of the three-peak model were designated as neither α-helix nor β-strand and were selected for back prediction.

To illustrate the improvement in performance of a nearest-neighbor corrected hypersurface, we compared structure-based chemical shift predictions derived from a hypersurface for this amino acid residue without and with nearest-neighbor corrections. We chose glutamate because it experiences large nearest-neighbor effects on chemical shifts as shown in Supplementary Tables s3 and s6. Inclusion of the nearest-neighbor corrections in the hypersurface led to and increase in prediction accuracy (Fig. 9) and an increase in the correlation coefficient between experimental and calculated chemical shifts from 0.66 to 0.70.

Conclusions

The approach described here takes advantage of the higher sensitivity of the chemical shift difference ($\delta^{13}C^{\alpha} - \delta^{13}C^{\beta}$) to conformational effects (over chemical shifts of the individual nuclei) and the fact that the dispersion of this parameter ($\delta^{13}C^{\alpha} - \delta^{13}C^{\beta}$) can be fitted by three Gaussians that represent residues in three states of "pseudo secondary structure" (Wang et al. 2006). From a large protein chemical shift database, we have extracted the nearest-neighbor effects on this parameter in each of the pseudo secondary structural states: those corresponding to α-helix, β-strand, and coil (neither helix nor strand). The results of this analysis are nearest-neighbor correction factors to unbiased random-coil chemical shifts for residues in each of these states and factors that indicate the relative energies of dipeptides in the three states. We have shown how these factors can be used to improve the prediction of secondary structure from chemical shifts, improve the prediction of secondary structure from sequence alone, and improve the prediction of protein chemical shifts from a known three-dimensional structure. Although the applications of this approach are constrained severely by the limited protein chemical shift data currently available, the results presented here point to the potential value of nearest-neighbor corrected chemical shift hyperspaces.

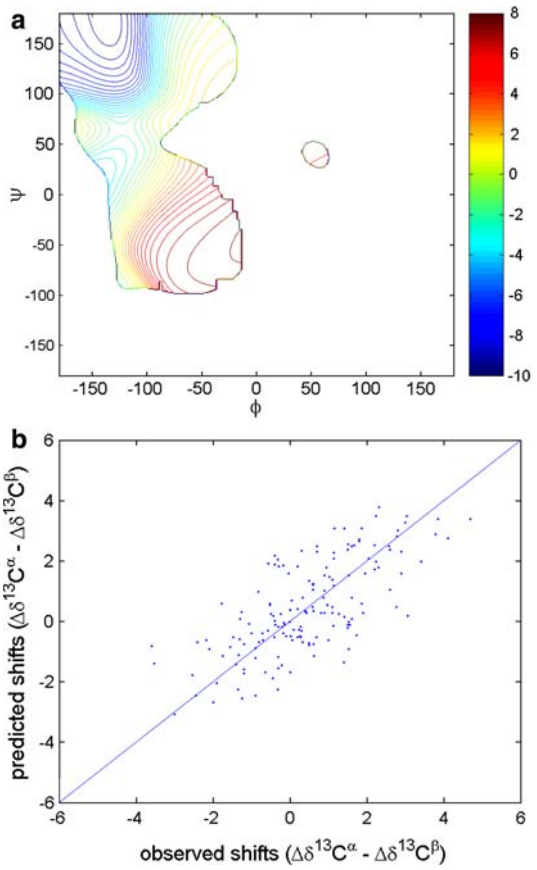

Fig. 8 (**a**) Example of a chemical shift hypersurface with nearest-neighbor corrections. The hypersurface shown is for the alanine carbon secondary chemical shift difference ($\delta^{13}C^{\alpha} - \delta^{13}C^{\beta}$). (**b**) Example of the use of nearest-neighbor corrected hypersurfaces in estimating chemical shifts from local protein structure (ψ, ϕ angles). Shown are experimental and back-calculated data using the TALOS database (Cornilescu et al. 1999)

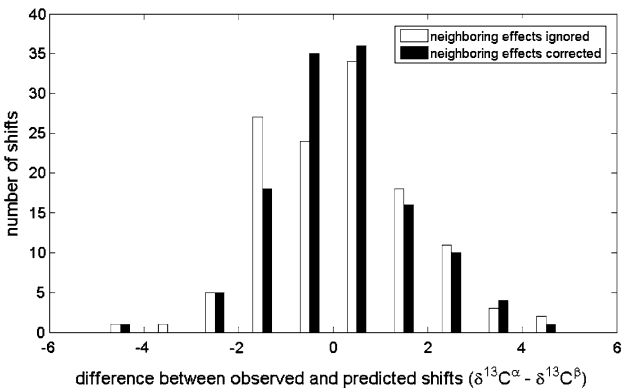

Fig. 9 Comparison of back-predicted results from chemical shift hypersurfaces for glutamate residues, without and with nearest-neighbor corrections. The data shown were extracted from the TALOS database (Cornilescu et al. 1999). Note that the back-predicted results became more accurate when the nearest-neighbor corrections were applied to the input chemical shifts

Detailed examination of the computed neighboring effects (see Supplementary Tables s1–s10) revealed that, for the majority of amino acids, the largest effects arise from the nature of either the preceding or following residue, but not both, irrespective of the secondary structure state. Exceptions are certain amino acids in the β-strand conformation, which experience large chemical shift effects from both the preceding and following amino acids, e.g. methionine and arginine (see Supplementary Tables s2 and s5).

The approach presented here is part of ongoing research aimed at improving the classification of protein secondary structure from chemical shift and sequence information. A feature of chemical shift dispersions not necessarily associated with defined backbone conformation is that they include information from residues that are dynamic or disordered. Because of this fact, the results of this kind of analysis, although more representative for random coil residues (neither helix nor strand), may not provide a simple one-to-one correspondence with secondary structure designations provided by DSSP analysis.

A website has been constructed (http://miranda.nmrfam.wisc.edu/MANI-NACS/) that provides the results of this study in electronically accessible format and includes a tool that accepts a protein sequence as input and/or a file containing rows of assigned $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts (in the same order as the given sequence). The output is an analysis of the secondary structure, and for each residue, with the exception of Gly and Pro, the neighborhood adjusted mean values $\Delta(^{X}Y)_s + \Delta(Y^{Z})_s$ representing the overall effects on the chemical shifts of residue Y from the preceding residue (X) and the following residue (Z). Also available on this website is a copy of the TALOS database used in this study.

Acknowledgements This research was supported by Biomedical Research Technology Program, National Center for Research Resources, through NIH Grant P41 RR02301 (to JLM), which supports the National Magnetic Resonance Facility at Madison, by the National Institute of General Medical Science's Protein Structure Initiative through NIH Grants P50 GM64598 and 1U54 GM074901 (to JLM), which support the Center for Eukaryotic Structural Genomics, and NIH Grant 5K22LM8992 (to HRE).

References

Braun D, Wider G, Wüthrich K (1994) Sequence-corrected N-15 "random coil" chemical shifts. J Am Chem Soc 116:8466–8469

Chou PY, Fasman G (1974) Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. Biochemistry 13:211–222

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

Diao J (2003) Crystallographic titration of cubic insulin crystals: pH affects GluB13 switching and sulfate binding. Acta Crystallogr D 59:670–676

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

Garnier J, Osguthorpe DJ, Robson B (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 120:97–120

Hung LH, Samudrala R (2003) Accurate and automated classification of protein secondary structure with PsiCSI. Protein Sci 12:288–295

Iwadate M, Asakura T, Williamson MP (1999) C alpha and C beta carbon-13 chemical shifts in proteins from an empirical database. J Biomol NMR 13:199–211

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637

Kuszewski J, Qin J, Gronenborn AM, Clore GM (1995) The impact of direct refinement against 13C alpha and 13C beta chemical shifts on protein structure determination by NMR. J Magn Reson B 106:92–96

Lim VI (1974) Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. J Mol Biol 88:857–872

Markley JL, Meadows DH, Jardetzky O (1967) Nuclear magnetic resonance studies of helix-coil transitions in polyamino acids. J Mol Biol 27:25–35

McDonald CC, Phillips WD (1967) Manifestations of the tertiary structures of proteins in high-frequency nuclear magnetic resonance. J Am Chem Soc 89:6332–6341

Nakamura A, Jardetzky O (1967) Systematic analysis of chemical shifts in the nuclear magnetic resonance spectra of peptide chains, I. Glycine-containing dipeptides. Proc Natl Acad Sci USA 58:2212–2219

Richarz R, Wüthrich K (1978) Carbon-13 NMR chemical shifts of the common amino acid residues measured in aqueous solution of the linear tetrapeptides H-Gly-Gly-X-L-Ala-OH. Biopolymers 17:2133–2141

Rohl CA, Baldwin RL (1998) Deciphering rules of helix stability in peptides. Methods Enzymol 295:1–26

Schubert M, Labudde D, Oschkinat H, Schmieder P (2002) A software tool for the prediction of Xaa-Pro peptide bond conformations in proteins based on 13C chemical shift statistics. J Biomol NMR 24:149–154

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 WM, Markley JL (1991) A relational database for sequence-specific protein NMR data. J Biomol NMR 1:217–236

Sharma D, Rajarathnam K (2000) ^{13}C NMR chemical shifts can predict disulfide bond formation. J Biomol NMR 18:165–171

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

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

Wang Y, Jardetzky O (2002b) Probability-based protein secondary structure identification using combined NMR chemical-shift data. Protein Sci 11:852–861

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 (2006) Probabilistic approach to determining unbiased random-coil carbon-13 chemical shift values from the protein chemical shift database. J Biomol NMR 35:155–165

Wishart DS, Case DA (2001) Use of chemical shifts in macromolecular structure determination. Methods Enzymol 338:3–34

Wishart DS, Nip AM (1998) Protein chemical shift analysis: a practical guide. Biochem Cell Biol 76:153–163

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, Sykes BD, Richards FM (1992) The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. Biochemistry 31:1647–1651

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