Journal of Biomolecular NMR (2006) 35: 155–165
DOI 10.1007/s10858-006-9022-0

Article

Probabilistic approach to determining unbiased random-coil carbon-13 chemical shift values from the protein chemical shift database

Liya Wang[a,b], Hamid R. Eghbalnia[a,c,d,*] & John L. Markley[a,b,c]
[a]*National Magnetic Resonance Facility at Madison, 433 Babcock Drive, Madison, WI, 53706, USA;* [b]*Graduate Program in Biophysics; Center for Eukaryotic Structural Genomics, University of Wisconsin-Madison, Madison, USA;* [c]*Biochemistry Department, University of Wisconsin-Madison, 171a Biochemistry Addition, 433 Babcock Dr, Madison, WI, 53706, USA;* [d]*Mathematics Department, University of Wisconsin-Madison, 811 Van Vleck Hall, 480 Lincoln Drive, Madison, WI 53706, USA*

Received 16 January 2006; Accepted 10 April 2006

Key words: adaptive LACS, chemical shifts, conformation dependence, LACS, random coil, RefDB

Abstract

We describe a probabilistic model for deriving, from the database of assigned chemical shifts, a set of random coil chemical shift values that are "unbiased" insofar as contributions from detectable secondary structure have been minimized ($\mathrm{RCCS_u}$). We have used this approach to derive a set of $\mathrm{RCCS_u}$ values for $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ for 17 of the 20 standard amino acid residue types by taking advantage of the known opposite conformational dependence of these parameters. We present a second probabilistic approach that utilizes the maximum entropy principle to analyze the database of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts considered separately; this approach yielded a second set of random coil chemical shifts ($\mathrm{RCCS_{max\text{-}ent}}$). Both new approaches analyze the chemical shift database without reference to known structure. Prior approaches have used either the chemical shifts of small peptides assumed to model the random coil state ($\mathrm{RCCS_{peptide}}$) or statistical analysis of chemical shifts associated with structure not in helical or strand conformation ($\mathrm{RCCS_{struct\text{-}stat}}$). We show that the $\mathrm{RCCS_{max\text{-}ent}}$ values are strikingly similar to published $\mathrm{RCCS_{peptide}}$ and $\mathrm{RCCS_{struct\text{-}stat}}$ values. By contrast, the $\mathrm{RCCS_u}$ values differ significantly from both published types of random coil chemical shift values. The differences ($\mathrm{RCCS_{peptide}}-\mathrm{RCCS_u}$) for individual residue types show a correlation with known intrinsic conformational propensities. These results suggest that random coil chemical shift values from both prior approaches are biased by conformational preferences. $\mathrm{RCCS_u}$ values appear to be consistent with the current concept of the "random coil" as *the state in which the geometry of the polypeptide ensemble samples the allowed region of* (ϕ,ψ)*-space in the absence of any dominant stabilizing interactions* and thus represent an improved basis for the detection of secondary structure. Coupled with the growing database of chemical shifts, this probabilistic approach makes it possible to refine relationships among chemical shifts, their conformational propensities, and their dependence on pH, temperature, or neighboring residue type.

Introduction

The notion of what constitutes a "random coil" configuration in polypeptides has varied over time. Flory's isolated pair hypothesis (Flory, 1969) suggested that the random-coil state of a peptide is *the one in which the* ϕ *and* ψ *angles of each residue are independent of the conformations of neighboring residues*. Although this notion appears suitable to alanine peptides within a restricted region of ϕ,ψ-space, its general applicability is limited (Pappu

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

et al., 2000). An alternative description of the random-coil state as *the well-defined reference state in which no side-chain-side-chain interactions are present* (Shortle, 1996) neglects the intrinsic folding propensities of amino acids. A definition consistent with the current notion of "random coil" would be: *the state in which the geometry of the polypeptide ensemble samples the allowed region of* (ϕ,ψ)*-space in the absence of any dominant stabilizing interactions*. The space sampled depends on the amino acid, the solution conditions (such as pH and temperature), and perhaps on the identity of nearest neighbor residues. The single reported "random coil" state is the energy-weighted distribution of the ensemble of conformational states (Makowska et al., 2006). The above definition captures the experimental observations both in concept and quantitative results. It forms the basis for the model proposed here.

Among available methods for characterizing the random-coil state of peptides and small proteins, the NMR chemical shift stands out as uniquely powerful. In particular, carbon chemical shifts are known to be strongly dependent on backbone torsion angles (Spera and Bax, 1991; Iwadate et al., 1999). Random-coil chemical shift (RCCS) values determined experimentally from spectral analysis of short peptides (Bundi and Wuthrich, 1979; Merutka et al., 1995; Wishart et al., 1995; Schwarzinger et al., 2000) serve as the basis for the determination of a "secondary chemical shift", i.e., the experimental chemical shift minus the random coil chemical shift. Secondary chemical shifts, in turn, have proven useful as measures of protein secondary structure (Wishart and Sykes, 1994).

The influence of the choice of RCCS values on chemical-shift based protein secondary structure identification has been evaluated by reference to five different sets of RCCS standards, including ones derived from experimental data and ones derived from statistical analysis of the chemical shift database (Mielke and Krishnan, 2004). On the basis of their ability to predict known helical and sheet content as a measure of superiority, these authors identified two "best" sets of RCCS standards, one determined from peptide chemical shifts (here denoted by $\text{RCCS}_{\text{peptide}}$) (Schwarzinger et al., 2000) and one derived statistically (here denoted by $\text{RCCS}_{\text{struct-stat}}$) (Lukin et al., 1997). We address here questions of bias in these RCCS standards and present a new way of looking at the problem.

$\text{RCCS}_{\text{peptide}}$ values typically are based on chemical shifts of amino acids in short, glycine-flanked peptides. Such peptides are too short to form any structure that can be stabilized by peptide H-bonds, and they do not have side-chain–side-chain interactions. However, for the short peptide GGAGG (pH 4.6, 20 °C), the polyproline II (P_{II}) conformation (a left-handed 3_1 helical conformation occupied by collagen and peptides containing proline with torsion angles $\phi = -75°$ and $\psi = 145°$) is reported to dominate the ensemble (Ding et al., 2003). In the presence of the organic solvent 2,2,2-trifluoroethanol (TFE), the predominant conformation of the pentamer changes from P_{II} to internally H-bonded γ- or β-turns (Liu et al., 2004). The model peptides $GG(A)_nGG$ (n = 1–3) dominantly adopt the P_{II} conformation at lower temperatures, but transition into higher population of extended structure content at higher temperatures (>40 °C) (Chen et al., 2004). Alanine-based (AXA) tripeptides in an aqueous solution have been shown to have different conformational preferences depending on the nature of X (X = A, D, E, G, V, L, M, K, S, H, P, Y, W, or F) (Eker et al., 2004). Remarkably, even a small alanine dipeptide (a single blocked amino acid) has a P_{II} conformational preference (Mehta et al., 2004). Finally, pH effects on RCCS values are an important consideration that has been cumbersome to survey experimentally. The experimental data used in determining the favored experimentally derived RCCS standard (Schwarzinger et al., 2000) were collected at pH 2.3, far from the pH at which most protein NMR data are collected; this introduces significant bias in the RCCS values for aspartate ($pK_a = 3.8$) and glutamate ($pK_a = 4.1$) residues (Richarz and Wuthrich, 1978). Results of this kind suggest that intrinsic conformational preferences and environmental factors have a marked influence on RCCS values derived from model peptides.

The alternative statistical approach uses information from chemical shifts associated with known structure to derive $\text{RCCS}_{\text{struct-stat}}$ values from the chemical shifts of residues in a "coil library" consisting of residues that are neither helix nor sheet (Wishart et al., 1991; Lukin et al., 1997; Wang and Jardetzky, 2002). This approach suffers from a number of problems. The *a priori* classification of chemical shifts by three states (helix, sheet, and "other") may be an unsuitable simplifying assumption; protein secondary structures can be classified into at least seven different, but overlapping, categories besides random coil (Kabsch and Sander, 1983). Furthermore, the assignment of the mean value of the "other" (non-helix, non-sheet) set to the RCCS value assumes that its distribution is unimodal. Yet another difficulty is that some unfolded proteins cannot be characterized by a Gaussian-distributed random coil model (Fitzkee and Rose, 2004). Additional potential problems with the statistical approach arise from the limited quantity of chemical shift data associated with proteins of known structure, referencing problems (Zhang et al., 2003), and possible effects of neighboring residues (Braun et al., 1994; Schwarzinger et al., 2001).

We present here two novel probabilistic approaches for analyzing the database of assigned protein chemical shifts *without* reference to known structure. We show that sets of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts assigned to individual residue types can be analyzed individually by a maximum entropy approach to yield $RCCS_{max\text{-}ent}$ values that are quite similar to the $RCCS_{peptide}$ and $RCCS_{struct\text{-}stat}$ values. An alternative probabilistic model is to make use of additional information to refine the database by removing chemical shifts from residues determined to have significant helical or strand propensity on the basis of chemical shift combinations. We show that unbiased random coil chemical shift values ($RCCS_u$) can be derived for $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ by taking advantage of the known opposite conformational dependence of these parameters. In addition to providing values that lead to more accurate identification of secondary structure from chemical shifts, the latter model suggests an approach for further refining our understanding of relationships between chemical shifts and protein structure.

Methods and results

Our model for the "random coil" is the probabilistic one (described above) *in which the geometry of the polypeptide ensemble samples the allowed region of* (ϕ,ψ)*-space in the absence of any dominant stabilizing interactions*. The experimental "random coil" state is the energy-weighted distribution of the ensemble of such conformational states. If the conformational ensemble has no preferential subset of states, then it is reasonable to describe the ensemble in terms of a maximally entropic symmetric distribution. On the other hand, a maximally entropic symmetric distribution would be unsuitable as a random coil reference if individual amino acids have preferential states arising from intrinsic stabilizing interactions; in such a case, a different approach would be needed to determine the properties of the distribution.

In order to illustrate the probabilistic model and to underscore the influence of these assumptions on the results, we contrast two approaches to computing RCCS values without reference to any independent structural information. We first show that a maximum entropy statistical interpretation returns results ($RCCS_{max\text{-}ent}$ values) very similar to experimentally reported RCCS values. Next, we base the detection of the random coil state on consensus from both $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ values. The algorithm for the detection of the consensus signal uses iterative, piecewise linear statistical regression analysis to account for the structural propensities and hence to yield unbiased RCCS values ($RCCS_u$).

We have applied both models to the analysis of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts, because they are influenced oppositely by helix and sheet and because they are relatively insensitive to sequence effects, provided that the C-terminal residue is not proline (Iwadate et al., 1999; Schwarzinger et al., 2001). The opposite influence of helix and sheet on $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts makes it convenient to examine the $^{13}C^{\alpha}$–$^{13}C^{\beta}$ distributions. We could have worked directly with $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ distributions, but the distributions are better resolved when plotted against the chemical shift difference $^{13}C^{\alpha}$–$^{13}C^{\beta}$ than against chemical shift alone. The better resolved distributions make our estimates more robust, but still allow the separate $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ distributions to be extracted.

Data set used

About 300 proteins in the RefDB (Zhang et al., 2003) have assigned $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ signals. Because these chemical shifts are considered most reliable (Neal et al., 2003), they were downloaded and used as the basis for the present study. These RefDB values were filtered by LACS analysis

(Wang et al., 2005) to remove outliers. We excluded three residue types: glycine (no C$^{\beta}$), cysteine (not enough data), and proline (coil dominated). For the reason stated above, we removed chemical shifts from all residues followed by proline; by doing so, sequence effects on chemical shifts were comparable to the average ~0.1 ppm effects across all amino acids (Schwarzinger et al., 2001). The resulting number of residues for each amino acid is shown in Table 1.

Maximum entropy analysis of random coil chemical shifts

The customary approach to the analysis of NMR chemical shifts is to categorize them by their secondary structure designation: helix, sheet, and "other" (non-helix, non-sheet). In this context, the mean of chemical shifts in the "other" category is reported as the random coil chemical shift. Our approach was to examine the ^{13}C$^{\alpha}$–^{13}C$^{\beta}$ distribution of all chemical shifts in the database without reference to their structural context. In the absence of preferential states within the "random coil" population, a parsimonious description of the ^{13}C$^{\alpha}$–^{13}C$^{\beta}$ distribution would be the mixture of three distributions. These three distributions would comprise a maximally entropic center distribution that represents the "random coil" ensemble and two other distributions that arise from the opposite movement of ^{13}C$^{\alpha}$ and ^{13}C$^{\beta}$ chemical shifts in response to different structural states. Helices and sheets are the two well-known structural regions with dominant stabilizing interactions and account for most of the content of the two outer modes of the ^{13}C$^{\alpha}$–^{13}C$^{\beta}$ distribution. Note that we have not divided the chemical shifts by their structural categories and that the outer modes of the mixture may contain other structural states. For a fixed variance, the Gaussian distribution is "maximally entropic" (Shannon 1948), and therefore is the least biased selection for a distribution. The focus of our maximum entropy approach is to obtain the properties of the center mode of the Gaussian mixture representing random coils; the additional modes represent states (including helix and sheet states) that cannot be accounted for by a single featureless Gaussian. To find the mean of the random coil state, we optimally fit the empirical distribution of the data to a

Table 1. "Random-coil" chemical shift values for common amino acids

Amino acid	RCCS$_u$ ^{13}C$^{\alpha}$ (ppm)	RCCS$_{peptide}$[a] ^{13}C$^{\alpha}$ (ppm)	RCCS$_u$ ^{13}C$^{\beta}$ (ppm)	RCCS$_{peptide}$[a] ^{13}C$^{\beta}$ (ppm)	# of residues in the database	P$_{II}$ preference [b](%)
Ala	51.9	52.5	19.9	19.1	2284	76.0
Asp	53.7	54.2	41.1	41.1	1969	55.3
Glu	55.9	56.6	30.8	29.9	2548	63.7
Phe	57.1	57.7	40.0	39.6	1236	53.0
His	55.5	55.0(56.3[c])	30.9	29.0(30.8[c])	626	52.7
Ile	60.9	61.1	38.2	38.8	1768	46.8
Lys	55.9	56.2	33.0	33.1	2408	59.1
Leu	54.2	55.1	43.1	42.4	2677	67.6
Met	55.2	55.4	33.0	32.9	693	62.3
Asn	53.1	53.1	38.6	38.9	1434	49.0
Gln	55.3	55.7	29.7	29.4	1314	56.1
Arg	55.8	56.0	30.8	30.9	1578	57.2
Ser	58.0	58.3	64.0	63.8	1865	63.2
Thr	62.4	61.8	69.6	69.8	1812	50.5
Val	62.3	62.2	32.4	32.9	2196	46.6
Trp	57.8	57.5	29.8	29.6	416	58.8
Tyr	58.2	57.9	38.9	38.8	1113	52.1

[a] Chemical shifts of X in GGXAGG measured at pH 5 (Wishart et al., 1995).

[b] Calculated from a non-α-helix, non-β-strand, non-β-turn fragment database extracted from the PDB (Fleming et al., 2005).

[c] Chemical shifts of X in GGXGG measured at pH 9 (Richarz and Wüthrich, 1978); the chemical shift referencing was adjusted (Wishart and Case, 2001) to match those of the other RCCS$_{peptide}$ values.

mixture of three Gaussians, subject to the requirement that the center distribution has maximum variance – i.e., is maximally entropic. The algorithmic details concerning this approach and figures showing the chemical shift distributions are presented in the supporting information.

Consensus signal analysis of random coil chemical shifts

The above considerations suggest that the following two properties could be used as an initial step in further refining the random coil chemical shifts: (1) $\delta^{13}C^{\alpha}$ and $\delta^{13}C^{\beta}$ will shift away, possibly in different directions, from the true or unbiased random coil chemical shift ($RCCS_u$) values whenever the conformation deviates from pure random coil; and (2) $\delta^{13}C^{\alpha}$, $\delta^{13}C^{\beta}$ and [$\delta^{13}C^{\alpha}$–$\delta^{13}C^{\beta}$] all reach their true random coil values under the same conditions. Experimental results for alanine in small peptides support our first assumption and show that $\delta^{13}C^{\alpha}$ can vary from 52.77 to 51.53 ppm, and $\delta^{13}C^{\beta}$ can vary from 19.34 to 21.18 ppm, under conditions that yield different conformational preferences (Mehta et al., 2004). The second assumption appears reasonable and is validated, as discussed below, by the agreement of our results with existing experimental data.

To incorporate our desired properties into a computational framework, we could pursue a number of approaches, including the extension of the maximum entropy model to account for more modes and possibly asymmetric distributions. However, a model we had used earlier to demonstrate 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 ($\Delta\delta^{13}C^{\alpha}$ – $\Delta\delta^{13}C^{\beta}$) by using a piecewise linear function (Wang et al., 2005) turned out to be robust and efficient. The model is described by the equations,

$$Y = \begin{cases} k_{\alpha}X + O_{\alpha} & \text{if } X \geq 0 \\ k_{\beta}X + O_{\beta} & \text{if } X \leq 0 \end{cases} \tag{1}$$

where, $X = (\Delta\delta^{13}C^{\alpha} - \Delta\delta^{13}C^{\beta})$ is a reference-independent variable, and Y denotes 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 and report the value of the reference offset; ideally equal to zero in the absence of a reference error. We earlier used this model, in combination with RCCS values from the literature (Wishart et al., 1995; Wishart and Case, 2001), as a means for detecting possible referencing offsets in sets of assigned NMR chemical shifts without the need for structural information (Wang et al., 2005). We named this approach "LACS" for Linear Analysis of Chemical Shifts.

To address the computational aspects of refinement for RCCS values, we have turned the LACS analysis around. We start with a database of chemical shifts, RefDB (Zhang et al., 2003) that have been reference-corrected on the basis of known three-dimensional structure, and treat the RCCS as the unknown. Thus we proceed under the assumption that the input data have been independently verified and properly referenced. This assumption enables us to use the linear analysis of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts according to our model to optimize an initially chosen RCCS value. Note again that, although we maintain the notion of three dominant modes with unknown conformational distributions, we avoid the division of chemical shifts into three discrete subsets representing helix, sheet, and coil.

We restate Eq. (1) with equations that describe the relationship between $\delta^{13}C^{\alpha}$, $\delta^{13}C^{\beta}$ and ($\delta^{13}C^{\alpha}$ – $\delta^{13}C^{\beta}$),

$$\begin{cases} Y_{\alpha}(X) = k_{\alpha}X + O_{\alpha} & \text{if } X \geq R_c \\ Y_{\beta}(X) = k_{\beta}X + O_{\beta} & \text{if } X \leq R_c \\ R_c = Y_{\alpha}(X) - Y_{\beta}(X) & \text{if } X = R_c \end{cases} \tag{2}$$

In the above equations, X=($\delta^{13}C^{\alpha}$ – $\delta^{13}]C^{\beta}$), and Y_{α} and Y_{β} denote $\delta^{13}C^{\alpha}$ and $\delta^{13}C^{\beta}$, respectively; K_{α} and K_{β} are the slopes for the coil-helical region for C^{α} and sheet-coil region for C^{β}, respectively; O_{α} and O_{β} are Y-intercepts; and R_c is the unknown that represents the chemical shift difference between $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ in the random coil state,

$$R_c = \delta^{13}C^{\alpha}_{\text{randomcoil}} - \delta^{13}C^{\beta}_{\text{randomcoil}} \tag{3}$$

If R_c is known, the two lines represented by Eq. (2) can be fitted to the data to obtain parameters K_{α}, K_{β}, O_{α}, and O_{β} (Figure 1). In our case, R_c is unknown so that Eq. (2) is solved by an iterative procedure to yield R_c. Data for the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts of valine (plotted against $\delta^{13}C^{\alpha}$ – $\delta^{13}C^{\beta}$ in Figure 1) illustrate the approach. The upper line, representing the first line in Eq. (2), is the linear fitting of the points with $^{13}C^{\alpha} > R_c$, and the lower line, second line in Eq. (2), is the linear fitting of the points with $^{13}C^{\beta} < R_c$. The above two fittings are based on a given R_c value, which is just a rough estimation and will be evaluated by iterations. With fitted lines and R_c value, the new R_c value can be estimated from the position where the vertical distance ($\delta^{13}C^{\alpha}_{\mathrm{random\ coil}} - \delta^{13}C^{\beta}_{\mathrm{random\ coil}}$) between these two fitted lines is equal to the old R_c value. To make this clearer, the inset of Figure 1 plots the vertical distance between the two fitted lines (solid line) as a function of ($\delta^{13}C^{\alpha}- \delta^{13}C^{\beta}$); the new R_c value is obtained from the point where the solid line crosses the diagonal dashed line. By repeating the above fitting process with new R_c values, a converged R_c value is acquired after several iterations. Once the converged R_c value has been determined, the RCCS$_u$ values for $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ are obtained from the ordinates of its intersection with the two lines (Figure 1).

Computational procedures of the consensus model

For computational purposes, the initial value of R_c (Eq. (3)) can be selected randomly, but for computational convenience, we used the RCCS values for $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ determined experimentally from the chemical shifts of GGXAGG peptides at 25 °C and pH 5 (Wishart et al., 1995). We used a random sampling procedure in order to avoid possible biasing of the resulting RCCS values by misassignments or by incorrectly referenced proteins in RefDB. In each step of the iteration, 80% of the data points were randomly selected, and the mean value of R_c obtained in ten runs was carried to next iteration. In addition, we used a 'robust analysis' procedure (Holland and Welsch, 1977) to ensure stable regression results. This method iteratively computes a set of weights for data points from the least squares algorithm, with the weights at each iteration calculated by applying a carefully selected function to the residuals from the previous iteration. 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.

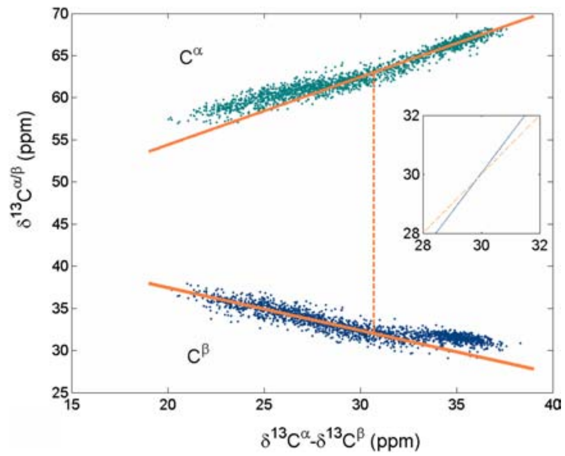

Figure 1. Chemical shifts (ppm) of valine residues from the adjusted RefDB database: for each residue $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ are plotted as a function of ($\delta^{13}C^{\alpha} - \delta^{13}C^{\beta}$). The solid lines represent the results of linear regression analyses for data from the coil-helical region as represented by $^{13}C^{\alpha}$ and the extended-coil region as represented by $^{13}C^{\beta}$. For each ($\delta^{13}C^{\alpha} - \delta^{13}C^{\beta}$) value, we determined the vertical distance between two solid lines (one such line is shown as the dashed line in the figure). The length of the dashed line represents R_c. The inset shows the intersection between the solid line (vertical distance vs. ($\delta^{13}C^{\alpha} - \delta^{13}C^{\beta}$)) and the dashed line (($\delta^{13}C^{\alpha} - \delta^{13}C^{\beta}$) vs. ($\delta^{13}C^{\alpha} - \delta^{13}C^{\beta}$)).

Convergence was rapid; for example, the values for $^{13}C^{\alpha}$ and $^{13}]C^{\beta}$ of alanine and valine converged after a single iteration (Figure 2). The standard deviations for R_c in the ten runs are shown as vertical bars in Figure 2.

To further test the robustness of the method for deriving RCCS$_u$ values, the sampling ratio was varied from 90% to as low as 30%. For each amino acid, 10 runs were carried out at various sampling ratios, and the standard deviations in R_c following convergence (considered to be complete after 5 iterations) were plotted against the sampling ratio (Figure 3). The standard deviations for tryptophan and histidine, for which only limited data are available (416 tryptophan residues and 626 histidine residues), increased rapidly with the decreasing sampling ratio. For the most abundant residues (A, D, E, I, K, L, N, Q, R, S, T and V), the standard deviation was around 0.1 ppm, even when only half of the data were sampled. The remaining small variation can be explained by effects of referencing, mis-assignments, sequence, and pH (discussed below). Other residues exhibited larger standard deviations than expected from their abundance in the database. For example, the deviation for histidine (626 residues) increases much faster than that for methionine (693 residues). As discussed below this anomaly arises from bias caused by pH differences.

Comparison of RCCS$_u$ values with those derived from small peptide chemical shifts

Table 1 compares the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ RCCS$_u$ values derived from "adaptive LACS" analysis (taken as average values from iterations 5–10) with a set of RCCS$_{peptide}$ values (Wishart et al., 1995). Differences between the RCCS$_u$ and RCCS$_{peptide}$ values are shown as the open bars in Figure 5a and b; in this plot, the RCCS$_u$ values are set at zero, and the relative peptide values and corresponding standard deviations of LACS values (from 10 runs at a sampling ratio of 0.8) are shown by the horizontal bars. Also plotted for each residue (Figure 4c) is

$$\begin{aligned}\Delta R_c &= [\Delta\delta^{13}C^{\alpha}] - [\Delta\delta^{13}C^{\beta}] \\ &= [^{13}C^{\alpha}(RCCS_{peptide}) - {}^{13}C^{\alpha}(RCCS_{LACS})] \\ &\quad - [^{13}C^{\beta}(RCCS_{peptide}) - {}^{13}C^{\beta}(RCCS_{LACS})] \end{aligned} \tag{4}$$

These ΔR_c values show a higher sensitivity to the secondary structure than either $\Delta\delta^{13}C^{\alpha}$ or $\Delta\delta^{13}C^{\beta}$ alone (Metzler et al., 1996). Almost half of the 17 amino acids showed ΔR_c values near to or greater than 1 ppm.

Random coil chemical shifts derived from the maximum entropy approach

In our alternative to the "adaptive LACS" approach, the mean of the central Gaussian from the full set of data for each amino acid type (see supplementary material for figures) was used to derive a separate set of values, RCCS$_{max\text{-}ent}$ (Table 2). When compared with RCCS$_{peptide}$ values derived from small peptide chemical shifts (Wishart et al., 1995), the mean absolute deviation was below experimental error. This correspondence suggests that the full set of chemical shift data are biased by the conformational preferences of the amino acid type, just as are experimental data from short peptides.

Relationship with residue intrinsic folding propensity

If we consider the RCCS$_u$ values as true random coil chemical shifts, then they can be used to determine secondary chemical shifts for the GGXAGG peptides. When this was done

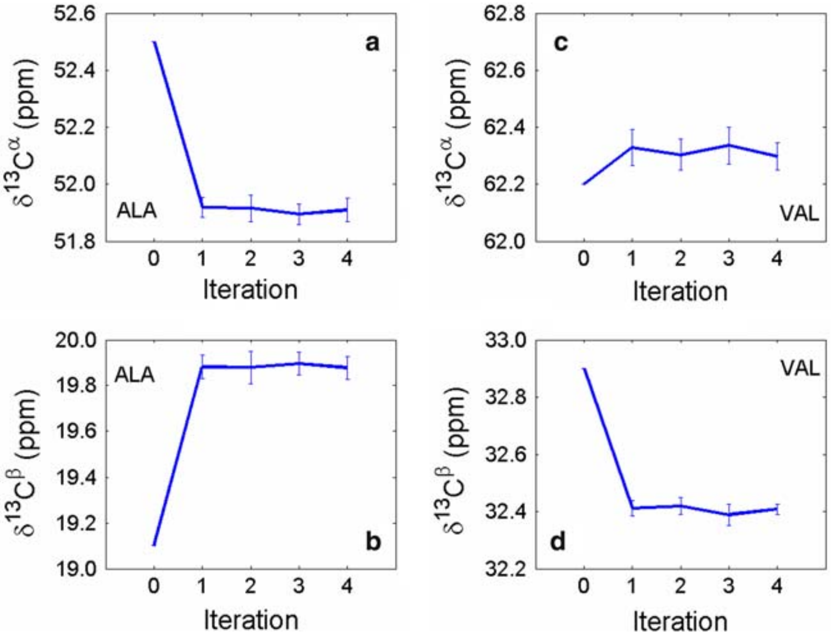

Figure 2. Random coil chemical shifts determined by the adaptive LACS approach for $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ of alanine (a and b) and $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ of valine (c and d) are plotted against the iteration number. A solid line segment connects mean values from each iteration. The standard deviations for each iteration are shown as vertical bars. The converged values are taken as unbiased random chemical shift values (RCCS$_u$).

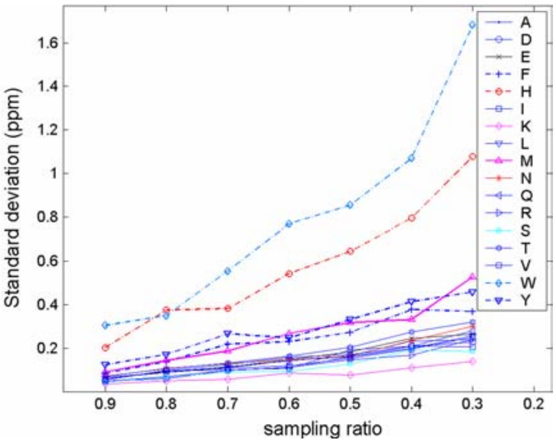

Figure 3. Standard deviations (ppm) of the calculated random coil chemical shifts (expressed as the difference, ($\delta^{13}C^{\alpha} - \delta^{13}C^{\beta}$) determined by the adaptive LACS approach for 17 amino acids (20 standard amino acids with the exclusion of G, C, and P) plotted as a function of the sampling ratio.

(Figure 5), the [$\Delta\delta^{13}C^{\alpha} - \Delta\delta^{13}C^{\beta}$] values were found to correlate with literature values for P_{II}/β-strand preferences of these amino acids (Fleming et al., 2005). The natural P_{II}/β-strand preference was calculated from a coil library consisting of non-α-helix, non-β-strand, and non-β-turn fragments extracted from the Protein Data Bank (Berman et al., 2000). This coil library excluded β-turns, which were present in earlier coil libraries (Swindells et al., 1995; Avbelj and Baldwin, 2003). The goal in building a coil library was to examine the intrinsic residual preferences for ϕ,ψ conformations in the absence of the complicated range of interactions that stabilize secondary structures. Given the large number of proteins in this coil library, the effects of non-local interactions in individual protein structures are expected to be averaged out. Figure 5 clearly shows that a more positive secondary chemical shift (ΔR_c) implies higher P_{II} preference and a more negative secondary chemical shift indicates higher β-strand preference. The fitted line is meant to show the trend rather than to imply a linear relationship. The largest outlier is histidine, which shows a large secondary chemical shift but a low P_{II} preference.

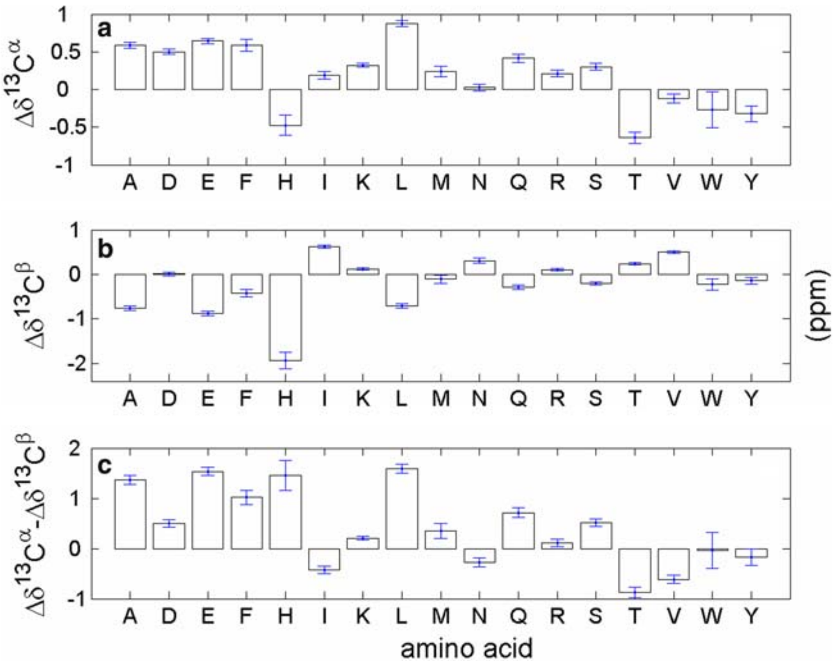

Figure 4. Solid bars indicate differences between random coil chemical shift values measured experimentally from the family of peptides GGXAGG at 25 °C and pH 5.0 (Wishart et al., 1995) ($RCCS_{peptide}$) and the unbiased random coil chemical shifts derived here ($RCCS_u$) from statistical analysis of chemical shifts from a subset of RefDB (Zhang et al., 2003). The horizontal bars indicate the standard deviations from analysis of 20 runs at a sampling ration of 0.8 (see text). (a) Random coil chemical shift differences for $^{13}C^{\alpha}$ for each amino acid. (b) Random coil chemical shift differences for $^{13}C^{\beta}$ for each amino acid. (c) Differences between each of the above values.

Table 2. Similarity of random coil chemical shift values for common amino acids derived from maximum entropy analysis ($RCCS_{max\text{-}ent}$)[a] and from experimental chemical shifts of small peptides ($RCSS_{peptide}$)[b]

Amino acid	$RCCS_{max\text{-}ent}\delta^{13}C^{\alpha}$	$RCSS_{peptide}\delta^{13}C^{\alpha}$	($RCCS_{max\text{-}ent}$ $-RCSS_{peptide})\Delta\delta^{13}C^{\alpha}$	$RCCS_{max-ent}\delta^{13}]C^{\beta}$	$RCSS_{peptide}\delta^{13}C^{\beta}$	($RCCS_{max\text{-}ent}$ $-RCSS_{peptide})\Delta\delta^{13}C^{\beta}$
Ala	52.7	52.5	0.2	19.4	19.1	0.3
Asp	54.3	54.2	0.1	40.9	41.1	−0.2
Glu	56.9	56.6	0.3	30.3	29.9	0.4
Phe	57.9	57.7	0.2	39.8	39.6	0.2
Ile	60.6	61.1	−0.5	39.1	38.8	0.3
Lys	56.5	56.2	0.3	33.1	33.1	0.0
Leu	55.2	55.1	0.1	42.5	42.4	0.1
Met	55.7	55.4	0.3	33.0	32.9	0.1
Asn	53.1	53.1	0.0	38.8	38.9	−0.1
Pro	63.1	63.3	−0.2	32.1	32.1	0.0
Gln	55.8	55.7	0.1	29.5	29.4	0.1
Arg	56.4	56.0	0.4	30.8	30.9	−0.1
Ser	58.7	58.3	0.4	63.9	63.8	0.1
Thr	62.0	61.8	0.2	70.0	69.8	0.2
Val	61.9	62.2	−0.3	33.0	32.9	0.1
Trp	57.2	57.5	−0.3	29.8	29.6	0.2
Tyr	58.3	57.9	0.4	39.3	38.8	0.5

[a] Derived from the center of the central Gaussian distribution (see Figure 1s in Supplementary Materials).
[b] From chemical shift values of peptides (Wishart and Case, 2001).

Dependence of $RCCS_u$ *values on pH*

The anomalous behavior of the histidine $RCCS_u$ values may be explained by known pH effects on histidine $\delta^{13}C^{\alpha}$ and $\delta^{13}C^{\beta}$ values (Richarz and Wuthrich, 1978). Figure 6 shows the distribution of pH values at which the histidine $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts in RefDB were recorded. The majority of the data were collected in the pH range (6–7) in which $\delta^{13}C^{\alpha}$ and $\delta^{13}C^{\beta}$ of histidine (side chain pK_a = ~6.5) exhibit a large pH dependence. By considering the data as a whole, the resulting $RCCS_u$ values for histidine can considered as an average for pH ~6.5. Because the peptide data were collected at a significantly lower pH (pH 5), this difference may explain why the $\Delta\delta^{13}C^{\beta}$ value for histidine is particularly large (Figure 4b) and why it appears as an outlier when plotted as a function of the P_{II}/β-strand preference (Figure 5). The experimental data in RefDB, however, are insufficient to determine accurate pH-dependent $RCCS_u$ values for histidine $^{13}C^{\alpha}$ and $^{13}C^{\beta}$.

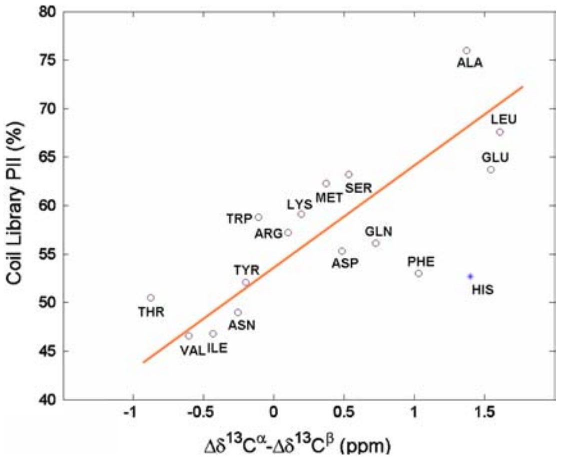

Figure 5. Correlation between intrinsic P_{II} / β-strand preferences for individual amino acids (P_{II}, Fleming et al., 2005) and the secondary chemical shifts ($RCCS_{peptide}$ − $RCCS_u$) determined for 16 GGXAGG peptides (where X is each of the 20 common amino acids except for G, C, and P) on the basis of the set of unbiased random coil chemical shifts derived here. Data are from Table 1. Histidine (denoted by *) appears as an outlier.

Conclusions

We have presented a probabilistic model for random coil chemical shift values of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ nuclei of amino acids in peptides and proteins. Our basic model, which is based on the assumption of

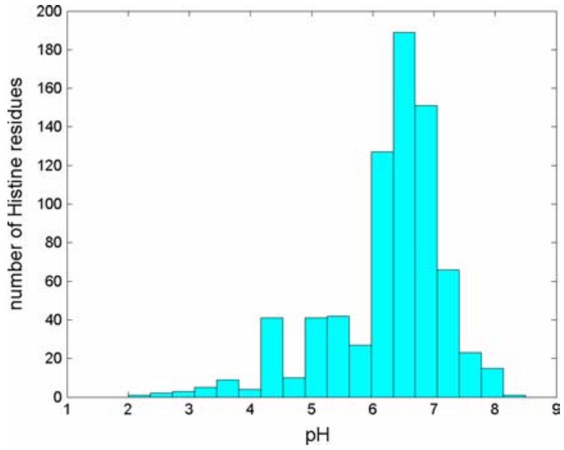

Figure 6. Distribution of pH values at which histidine chemical shift data in the filtered RefDB were collected.

three maximally entropic modes, shows very good agreement with experimental observations. We have developed an ''adaptive LACS'' procedure for refining random coil chemical shifts to remove bias from the conformational preferences of amino acid residue types. The resulting unbiased random coil chemical shift ($RCCS_u$) values support a more refined notion of random coil chemical shifts and their propensities. The $RCCS_u$ values should provide a useful basis for predicting peptide and protein secondary structure, including conformational preferences of residues in dynamically disordered regions. The model can also be used to further probe the relationship of chemical shifts and structural state. In this direction, we can use additional chemical shift signals from other nuclei and perform our analysis along extra chemical shift dimensions in order to find consistent and refined structural classifications.

In the analysis presented here, as a proof of principle, we used RefDB (Zhang et al., 2003), an independently derived database of chemical shifts reference corrected on the basis of known structures. Having established the principle of unbiased random coil chemical shifts, we have used the $RCCS_u$ values described here as input to LACS (Wang et al., 2005) to create a larger database of offset-corrected chemical shifts corresponding to ~ 1800 BMRB entries: ~300 with and ~1500 without corresponding 3D structures. This database of chemical shifts is available from the NMRFAM website at http://www.bija.nmrfam.wisc.edu/MANI-LACS/uLACS.db.

Acknowledgements

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, by the National Institute of General Medical Science's Protein Structure Initiative through NIH grants P50 GM64598, and 1U54 GM074901, which support the Center for Eukaryotic Structural Genomics, and the NIH grant 1K22LM8992. We thank Franc Avbelj for helpful discussions and Eldon L. Ulrich and William M. Westler for advice and encouragement.

Electronic supplementary material is available at http://dx.doi.org/10.1007/s10858-006-9022-0

References

Avbelj, F. and Baldwin, R.L (2003) *Proc. Natl. Acad. Sci. USA*, **100**, 5742–5747.

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

Braun, D., Wider, G. and Wuthrich, K. (1994) *J. Am. Chem. Soc*., **116**, 8466–8469.

Bundi, A. and Wuthrich, K. (1979) *Biopolymers*, **18**, 285–297.

Chen, K., Liu, Z. and Kallenbach, N.R. (2004) *Proc. Natl. Acad. Sci. USA*, **101**, 15352–15357.

Ding, L., Chen, K., Santini, P.A., Shi, Z. and Kallenbach, N.R. (2003) *J. Am. Chem. Soc*., **125**, 8092–8093.

Eker, F., Griebenow, K., Cao, X., Nafie, L.A. and Schweitzer-Stenner, R. (2004) *Proc. Natl. Acad. Sci. USA*, **101**, 10054–10059.

Fitzkee, N.C. and Rose, G.D. (2004) *Proc. Natl. Acad. Sci. USA*, **101**, 12497–12502.

Fleming, P.J., Fitzkee, N.C., Mezei, M., Srinivasan, R. and Rose, G.D. (2005) *Protein Sci*., **14**, 111–118.

Flory, P.J. (1969) *Statistical Mechanics of Chain Molecules* Interscience Publishers, New York.

Holland, P.W. and Welsch, R.E. (1977) *Commun. Stat.: Theory Methods*, **A6**, 813–827.

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.

Liu, Z., Chen, K., Ng, A., Shi, Z., Woody, R.W. and Kallenbach, N.R. (2004) *J. Am. Chem. Soc*., **126**, 15141–15150.

Lukin, J.A, Gove, A.P, Talukdar, S.N and Ho, C (1997) *J. Biomol. NMR*, **9**, 151–166.

Makowska, J., Rodziewicz-Motowidlo, S., Baginska, K., Vila, J.A., Liwo, A., Chmurzynski, L. and Scheraga, H.A. (2006) *Proc. Natl. Acad. Sci. USA*, **103**, 1744–1749.

Mehta, M.A., Fry, E.A., Eddy, M.T., Dedeo, M.T., Anagnost, A.E. and Long, J.R. (2004) *J. Phys. Chem. B*, **108**, 2777–2780.

Merutka, G., Dyson, H.J. and Wright, P.E. (1995) *J. Biomol. NMR*, **5**, 14–24.

Metzler, W.J., Leiting, B., Pryor, K., Mueller, L. and Farmer, B.T. 2nd (1996) *Biochemistry*, **35**, 6201–11.

Mielke, S.P. and Krishnan, V.V. (2004) *J. Biomol. NMR*, **30**, 143–153.

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

Pappu, R.V., Srinivasan, R. and Rose, G.D. (2000) *Proc. Natl. Acad. Sci. USA*, **97**, 12565–12570.

Richarz, R. and Wuthrich, K. (1978) *Biopolymers*, **17**, 2133–2141.

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.

Schwarzinger, S., Kroon, G.J., Foss, T.R., Wright, P.E. and Dyson, H.J. (2000) *J. Biomol. NMR*, **18**, 43–48.

Shannon, C.E. (1948) *Bell System Technical Journal*, **27**, 379–423 and 623–656.

Shortle, D. (1996) *Faseb. J.*, **10**, 27–34.

Spera, S. and Bax, A. (1991) *J. Am. Chem. Soc.*, **113**, 5490–5492.

Swindells, M.B., MacArthur, M.W. and Thornton, J.M. (1995) *Nat. Struct. Biol.*, **2**, 596–603.

Wang, L., Eghbalnia, H.R., Bahrami, A. and Markley, J.L. (2005) *J. Biomol. NMR*, **32**, 13–22.

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 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.