Journal of Biomolecular NMR **30:** 143–153, 2004.

An evaluation of chemical shift index-based secondary structure determination in proteins: Influence of random coil chemical shifts*

S.P. Mielke[a,b] & V.V. Krishnan[b,**]
[a]*Biophysics Graduate Group, University of California, Davis, CA 95616 and* [b]*Molecular Biophysics Group, L-448 Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94551*

Received 5 March 2004; Accepted 10 June 2004

Key words: chemical shift, CSI, NMR, random coil, secondary structure

Abstract

Random coil chemical shifts are commonly used to detect protein secondary structural elements in chemical shift index (CSI) calculations. Though this technique is widely used and seems reliable for folded proteins, the choice of reference random coil chemical shift values can significantly alter the outcome of secondary structure estimation. In order to evaluate these effects, we present a comparison of secondary structure content calculated using CSI, based on five different reference random coil chemical shift value sets, to that derived from three-dimensional structures. Our results show that none of the reference random coil data sets chosen for evaluation fully reproduces the actual secondary structures. Among the reference values generally available to date, most tend to be good estimators only of helices. Based on our evaluation, we recommend the experimental values measured by Schwarzinger et al. (2000), and statistical values obtained by Lukin et al. (1997), as good estimators of both helical and sheet content.

Introduction

Determination of secondary structural elements using chemical shift index (CSI) calculations has become a standard procedure in the solution- and solid-state nuclear magnetic resonance (NMR) spectroscopy-based structural characterization of proteins (Gross and Kalbitzer, 1988; Pastore and Saudek, 1990; Spera and Bax, 1991; Szilagyi and Jardetzky, 1989; Williamson, 1990; Wishart et al., 1991a, 1991b, 1992). In this method, characteristic deviations of chemical shifts of certain nuclei in amino acid residues, relative to their random coil values, are used to identify the secondary structures with which those residues are associated. Since structure determination using NMR spectroscopy is strongly dependent on chemical shift assignments, this method is considered to be quick and reliable. In the last few years, several research groups have independently produced reference random coil chemical shifts under a variety of experimental conditions (Bienkiewicz and Lumb, 1999; Braun et al., 1994; Bundi and Wüthrich, 1979; Glushka et al., 1989; Merutka et al., 1995; Plaxco et al., 1997; Richarz and Wüthrich, 1978; Schwarzinger et al., 2000; Thanabal et al., 1994; Wishart et al., 1995a). Though visual inspection of these reference shifts reveals that they differ from each other, in some cases significantly, how these differences influence the quality of estimations of secondary structure in proteins is not known. To address this issue, we present a comparison of secondary structure estimations from the CSI method with those from three-dimensional structures. In our analysis, CSI is invoked using five different random coil reference chemical shift data sets, sampling the range of experimental conditions.

Random coil chemical shifts are the characteristic chemical shifts of the nuclei constituting the amino acid residues of disordered proteins. The effect of a particular secondary structure (helix or strand) on the observed chemical shift is often referred to as the secondary chemical shift, relative to the random coil shift. Secondary chemical shifts are predomin-

**To whom correspondence should be addressed. E-mail: krish@llnl.gov

antly influenced by non-covalent interactions, such as secondary structural changes, hydrogen bonds, and aromatic stackings. Several recent and excellent review articles describe a variety of experimental and computational methods for correlating these secondary chemical shifts with protein three-dimensional structure and dynamics (Ando et al., 2001; Case, 1998, 2000; Case et al., 1994; Oldfield, 1995, 2002; Szilagyi, 1995; Wishart and Case, 2001; Wishart and Nip, 1998).

Several techniques have been developed to characterize and quantify protein and peptide secondary structure using chemical shift data (Cornilescu et al., 1999; Pastore and Saudek, 1990; Schwarzinger et al., 2000; Szilagyi and Jardetzky, 1989; Wang and Jardetzky, 2002b; Wishart and Sykes, 1994b; Wishart et al., 1991a, 1992). These include the $\Delta\delta$ method (Reily et al., 1992), the probability-based method of Wang and Jardetzky (Wang and Jardetzky, 2002b), and the CSI method (Wishart and Sykes, 1994a; Wishart et al., 1992). The common goal of these techniques is to compare the observed chemical shifts of nuclei with a reference list of random coil shifts, or values related to random coil shifts, and use deviations of the observed from the reference values to estimate efficiently and accurately type and location of secondary structure.

The CSI method is a simple, quantitative, and accurate empirical procedure for determining elements of secondary structure. It is predicated on the observation that amino acid nuclei experience an upfield or downfield shift, relative to their random coil values, depending upon whether they are associated with a helical or an extended β-strand configuration (Wishart et al., 1992) (H^{α} and C^{β} nuclei experience an upfield shift in helices and a downfield shift in strands, while C^{α} and C' nuclei experience a downfield shift in helices and an upfield shift in strands). In the first stage of the CSI algorithm, observed chemical shifts are compared with a set of reference shifts, and a chemical shift index (a ternary index having values -1, 0 or 1) is designated for all residues with known chemical shift assignments. A particular residue is assigned -1, 0 or 1 if its observed chemical shift falls lower than, in the empirically determined range of, or higher than the corresponding reference value, respectively. When chemical shift data are available for at least three of the four above-mentioned nuclei, the algorithm permits a 'consensus' prediction based on the structure determined by two out of three, or three out of four, nuclear comparisons. In the original applications of the method, the reference values were empirically optimized with respect to a selected set of proteins. However, it is now common practice to use various unrefined, experimentally or statistically determined random coil reference values as input for CSI calculations (Schwarzinger et al., 2001, 2000; Thanabal et al., 1994), and some of these values have been incorporated, along with the CSI algorithm, into widely used NMR data processing software.

The primary goal of the work presented here is to evaluate the effect on secondary structure prediction of using differing random coil chemical shift reference tables in conjunction with the CSI algorithm. The secondary structure content (the total percentage of helical and sheet content) of a set of 396 folded proteins was calculated using the consensus CSI method. Corresponding structural information was calculated from the three-dimensional structural coordinates of the proteins. A comparison of the results obtained using five different reference tables for CSI calculations to those obtained using a structure-based method allows a critical evaluation of the reliability of the standard protocol for evaluating secondary structure from chemical shift information using CSI.

Materials and methods

Reference random coil chemical shifts

There are several reference random coil chemical shift tables in the literature, and these can be classified into two types: those measured experimentally, and those derived statistically. A complete description of these tables, including the experimental conditions under which they were obtained, can be found in Table 1. Considering the variability of the data represented by Table 1, and following the general recommendations of Wishart and Nip (1998), we have chosen five different sets of random coil chemical shifts for this analysis. In what follows, these five sets are identified by the initials of the first and last authors of the references as *KW, WS, SD, LH,* and *WJ*; i.e., Wüthrich et al. (Braun et al. 1994; Wüthrich, 1986), Wishart et al. (Wishart and Sykes, 1994a; Wishart et al., 1995a, 1995b; Wishart and Case, 2001), Schwarzinger et al. (2000), Lukin et al. (1997), and Wang and Jardetzky (2002a), respectively (bold entries in Table 1). The complete list of random coil values used for the data analysis is available from the authors, upon request. Of the five chosen data sets, three were experimentally derived, while two were obtained using statistics-based approaches. We have re-referenced *KW* and *WS*,

originally referenced to TMS/Dioxane, to DSS. Since reference table *LH* does not derive $^1H^\alpha$ values, the $^1H^\alpha$ reference values of Wang and Jardetzky (2002b) were used for structure estimation using *LH*. Though the experimental values of Plaxco et al. (1997) are relevant for the comparison, these were not considered for the analysis due to lack of heteronuclear chemical shift values.

Chemical shift information

Chemical shift values corresponding to protein atoms were obtained from BMRB NMR-STAR files (Seavey and others, 1991). Only proteins with 50 or more amino acid residues were considered, since these are expected to contain a significant amount of secondary structure. Further, only proteins with at least 70% of their residues assigned chemical shifts were considered. As nearly all recently submitted BMRB chemical shifts are referenced using the widely accepted standard procedure recommended by Wishart et al. (1995a), no re-referencing was performed.

The consensus chemical shift index (CSI) of the proteins was calculated using the procedure outlined by Wishart and Sykes (1994a), using nuclei that are known to be highly sensitive to secondary structural changes ($^1H^\alpha$, $^{13}C^\alpha$, $^{13}C^\beta$ and $^{13}C'$). BMRB NMR-STAR files were converted into the format required by the CSI algorithm. Average values of the two $^1H^\alpha$ resonances for Gly residues were used. Boundaries of the secondary structural regions were determined by the following criteria: (1) a local density of nonzero indices greater than 70% and (2) a minimum of three consecutive negative (-1) indices and four positive ($+1$) indices for helix and strand, respectively. For all three experimental reference sets, and one of the statistical sets (*LH*), the allowed deviation of the random coil values was that suggested by Wishart et al. (1995a, 1995b), while for the other statistically derived set (*WJ*), standard deviations were obtained from the original reference (Wang and Jardetzky, 2002a). Consensus secondary structures were then calculated without any additional filtering or local averaging to determine the total percentages of helical and sheet content; for example, $\%Helix = 100\times(\#\ of\ residues\ identified\ as\ helical)/(total\ \#\ of\ residues)$. The calculations were repeated in an identical fashion for each of the five random coil chemical shift reference tables.

Three-dimensional structural information

Structure files were obtained from the Research Collaboratory for Structural Biology (RCSB) (PDB format, http://www.rcsb.org/pdb/) (Berman et al., 2000; Bernstein et al., 1977). Since most BMRB NMR-STAR files identify several corresponding PDB structures, it was necessary to examine each entry and choose by inspection the most appropriate PDB ID number. When possible, the PDB ID corresponding to the 'best' NMR structure was chosen, though in some cases it was necessary to choose the best X-ray structure (resolution < 2.5 Å). A total of 396 proteins was found to be suitable, and downloaded from the Protein Data Bank. The total percentage of sheet and helix (α and 3_{10}) was determined using the program PROMOTIF (http://www.biochem.ucl.ac.uk/~gail/promotif/promotif.html) (Hutchinson and Thornton, 1996), which uses the atomic coordinate files obtained from the RCSB.

All analyses were performed using scripts written in awk or perl on a Silicon Graphics UNIX workstation. These scripts, as well as a complete list of the 396 proteins studied, with their BMRB accession numbers, PDB codes, and secondary structure contents calculated using both the five different reference tables and three-dimensional coordinates, are available from the authors.

Results

Comparison of secondary structure content estimations from CSI and 3D structures

Figures 1 through 5 show plots of the percentage of helical (left panels) and sheet (right panels) content determined from the random coil chemical shift tables, *KW, WS, SD, LH* and *WJ*, respectively, using CSI, versus the same content calculated from relevant three-dimensional structures. The dashed lines in the figures correspond to an ideal correlation, and the solid lines to an unbiased linear regression analysis of the data. Table 2 lists the coefficients (slope and intercept) of the fit, and the correlation coefficients of the regression analysis. Uncertainties in the former were obtained by a linear model bootstrapping procedure using the R statistical package (www.cran.us.r-project.org) with 512 bootstrap replicates. Based on this analysis, several distinct features are observed.

For the sake of comparison, Figure 1 panels (a) and (b) show the *KW* correlation without re-referencing

Table 1. List of available reference random coil chemical shift data

Sample	Nuclei	Solvents	Referenced to	T (°C)	pH	Correction	Reference
Experiment-based random coil shifts							
H-GG-X-A-OH	^{1}H, ^{13}C	D_2O	TMS	35	Varied	None	(Bundi and Wüthrich, 1979; Richarz and Wüthrich, 1978)
Apamin, BPTI	^{1}H, ^{15}N	90% H_2O / 10% D_2O	TSP	50, 65	2.2–4.6	None	(Glushka et al., 1989, 1990)
GG-X-GG	^{13}C	D_2O / 10, 20, or 30% acetonitrile or TFE	TSP	25	2.0–3.5	None	(Thanabal et al., 1994)
H-GG-X-A-OH *(KW)*	^{15}N	90% H_2O / 10% D_2O	TSP	35	2.0 and 5.0	Sequence-corrected	(Braun et al., 1994; Wüthrich, 1986)
H-GG-X-GG-OH	^{1}H	90% H_2O / 10% D_2O and TFE% varied	DSP	278–318	5.0	None	(Merutka et al., 1995)
GG-X-Y-GG, Y=A,P *(WS)*	^{1}H, ^{13}C, ^{15}N	95% H_2O / 5% D_2O	DSS	25	5.0	Nearest neighbor	(Wishart et al., 1995a, 1995b)
Ac-GG-X-GG-NH_2	^{1}H	90% H_2O / 10% D_2O (50 mM Sodium Phosphate) 2, 3, 6, 8 M GuHCl	DSS	20	5.0	None	(Plaxco et al., 1997)
Ac-GG-X-GG-NH2 (X = phosphorylated amino acid)	^{1}H, ^{13}C, ^{15}N	90% H_2O / 10% D_2O	DSS	25	2.0–9.0	None	(Bienkiewicz and Lumb, 1999)
Ac-GG-X-GG-NH_2 *(SD)*	^{1}H, ^{13}C, ^{15}N	90% H_2O / 10% D_2O and 8 M Urea	DSS	20	2.3	None	(Schwarzinger et al., 2000)
Ac-GG-X-GG-NH_2	^{1}H, ^{13}C, ^{15}N	90% H_2O / 10% D_2O	DSS	20	2.3	Sequence-corrected	(Schwarzinger et al., 2000, 2001)
Statistically derived random coil shifts							
Manual	^{1}H, ^{13}C, ^{15}N	Aqueous solution	DSS	–	–	None	(Wishart et al., 1995a, 1995b)
Probability-based *(LH)*	^{13}C, ^{15}N	Aqueous solution	DSS	–	–	None	(Lukin et al., 1997)
Probability-based (BMRB)	^{1}H, ^{13}C, ^{15}N	Aqueous solution	DSS	–	–	None	
Probability-based *(WJ)*	^{1}H, ^{13}C, ^{15}N	Aqueous solution	DSS	–	–	Neighboring residue effect	(Wang and Jardetzky, 2002b)

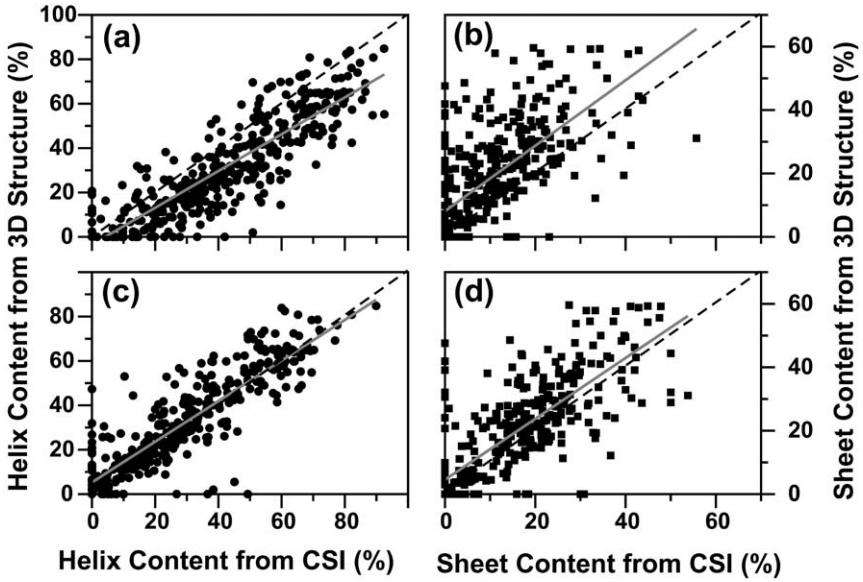

Figure 1. Plots of secondary structure content in percentage determined from chemical shifts and three-dimensional coordinates. Panels (a) helical and (b) sheet content for the original *KW* random coil reference values correspond to (Table 1), while (c) and (d) show the corresponding correlations after the references are corrected relative to DSS. The dashed line corresponds to an ideal correlation, while the solid line represents the linear regression analysis results (Table 2).

to DSS, and panels (c) and (d) show the same with re-referencing. Before correction, *KW* tends to overestimate helical content, and, at the same time, underestimate sheet content, for most of the proteins, as seen by significant negative (−3.45%) and positive (8.1%) intercepts, respectively (Table 2). However, correcting the random coil values improves the quality of the fit significantly, in particular for sheet content (the correlation coefficient in this case increases from 0.67 to 0.77), illustrating the sensitivity of secondary structure estimation from CSI to proper referencing. Though the correlation coefficient for uncorrected *KW* is comparable for helical content with the other reference tables, this reference set gives the lowest correlation for sheet proteins. Chemical shift reference set *SD* underestimates helix content (correlation coefficient 0.88), while *WS* overestimates helix content (correlation coefficient 0.86), particularly for predominantly helical proteins. Both of these experimental random coil tables (*WS* and *SD)* estimate sheet content comparably, with the correlation coefficient of *SD* slightly better than that of *WS*. Results from both statistically derived data sets (*LH* and *WJ*) exhibit similar fit parameters for helical content, with a slightly larger offset for *WJ*, while *LH* provides a much more reliable estimation of sheet content than *WJ*. Overall, *WS*, *KW* and *LH* provide comparable correlation coefficients for helical proteins, with *WS* having the smallest intercept, while *SD* and *LH* provide a significantly better estimation of sheet content than the other reference tables considered.

Figures 1–5 contain several significant outliers along both the abscissa and ordinate. Though removing these outliers might have affected the correlations (Table 2), they were left in the data set in order that our results reflect as accurately as possible the quality of available experimental information.

Discussion

Recent progress in semi-empirical correlations between protein structural information and chemical shifts has led to the possibility of the direct refinement of protein structures using chemical shift-based target functions (Cornilescu et al., 1999; Kuszewski et al., 1995a, 1995b, 1996; Laws et al., 1993; Luginbuhl et al., 1995; Williamson et al., 1995). These refinement protocols have been included in several mainstream structure calculation programs, such as AMBER and CNS. In initial approaches to implementing such protocols, chemical shift-based constraints weren't very stringent; for example, secondary $^{13}C^{\alpha}$ chemical shifts of more than 1.5 ppm constrained peptide dihedral angles to $\pm 100^{\circ}$ (Luginbuhl et al., 1995). Chemical shift hy-

Table 2. Linear regression analysis of CSI vs. structure-based helical and sheet content estimates

Random coil reference[a]	Helical content (%)			Sheet content (%)		
	Intercept	Slope	CC[b]	Intercept	Slope	CC[b]
KW[c]	5.33 ± 0.88	0.91 ± 0.02	0.82	4.56 ± 0.59	0.95 ± 0.02	0.77
KW	−3.45 ± 1.02	0.82 ± 0.03	0.77	8.11 ± 0.58	1.03 ± 0.03	0.67
WS[c]	1.21 ± 1.06	0.80 ± 0.03	0.86	5.79 ± 0.69	0.84 ± 0.03	0.73
SD	6.6 ± 0.85	0.93 ± 0.02	0.88	4.27 ± 0.61	0.91 ± 0.03	0.77
LH	5.97 ± 0.89	0.89 ± 0.02	0.90	4.12 ± 0.58	0.98 ± 0.02	0.79
WJ	6.92 ± 0.77	0.94 ± 0.02	0.88	6.6 ± 0.61	0.96 ± 0.03	0.73

[a]Random coil values correspond to the block letter references of Table 1.
[b]Correlation coefficient.
[c]Original values are corrected relative to DSS (Wishart and others, 1995a). Results from the uncorrected *WS* data set are omitted, because differences are negligible.

persurface calculations (Asakura et al., 1999; Oldfield, 1995; Wishart and Case, 2001) and other dihedral angle prediction methods, such TALOS (Cornilescu et al., 1999), have significantly improved the accuracy of the correlation between the range of dihedral angles and observed secondary chemical shifts. For example, for highly helical proteins, in some cases it is now common to use dihedral angle constraint limits in the range of ±10 to 20°. Furthermore, where three bond J-coupling (Wüthrich, 1986) and cross-correlated relaxation data (Reif et al., 1997) can only be used to provide dihedral angle constraints of either ϕ or ψ, respectively, secondary chemical shift effects are utilized to constrain both ϕ and ψ. This study demonstrates that choice of reference random coil chemical shifts reflects upon predicted secondary structures, suggesting that caution should be exercised in the application of secondary chemical shifts to establish structural constraints.

Secondary structure content

We have compared the secondary structures of proteins in terms of helical and sheet content, rather than accounting for the actual secondary structural elements at the residue level. One of the disadvantages of using secondary structure content in this way is that the actual number of residues in a helix or strand, though available during the analysis, gets averaged when the percentage is calculated (see Materials and methods). On the other hand, secondary structure content has proven to be a useful and directly measurable structural parameter in other biophysical characterization methods, such as CD and IR spectroscopies (Sanders et al., 1993; Sreerama and Woody, 1994) as well as NMR (Mielke and Krishnan, 2003; Sibley et al., 2003). In addition, due to its simplicity, secondary structure content estimation allows a straightforward comparison of the ranges of helical and sheet content predicted by different methods of calculation.

CSI for secondary structure determination

Secondary chemical shift effects can be related to secondary structural changes using a variety of methods. These include $\Delta\delta$ plots (Dalgarno et al., 1983), smoothed $^{1}H^{\alpha}$ (Pastore and Saudek, 1990) and $^{13}C^{\beta}$-$^{13}C^{\alpha}$ (Metzler et al., 1993) plots, the CSI method, and the more recent probability-based secondary structure identification (PSSI) method (Wang and Jardetzky, 2002b). Our choice of CSI to calculate secondary structure content is motivated by several factors unique to this method, including: (1) the consensus chemical shift index is defined on the basis of nuclei that are most sensitive to secondary structural changes, and (2) it is easy to automate the calculation for a large number of proteins with different sets of random coil values. Having said this, we expect similar correlation trends from alternative approaches, since all the methods predict secondary structure content with comparable accuracy.

Quality of the data

Algorithms to determine secondary structures

The correlations presented here suggest that secondary chemical shift-based methods for predicting secondary structure content are better indicators of helical regions than sheet regions in proteins. This could be due to insufficient sensitivity of secondary chemical

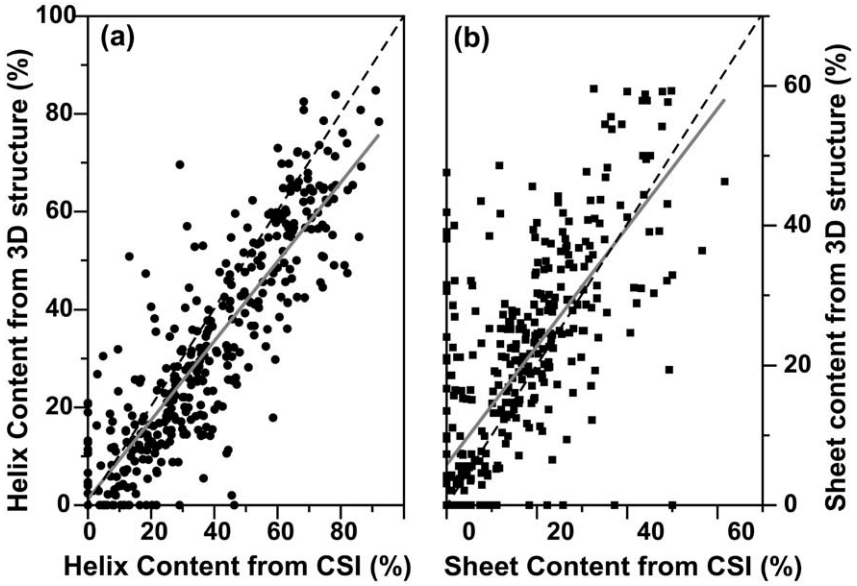

Figure 2. Plots of helical (a) and sheet (b) secondary structure content in percentage determined from chemical shifts and three-dimensional coordinates. The reference random coil values correspond to *WS*. Other features correspond to Figure 1.

shifts for identifying sheets. Ambiguity in the definition of a β-sheet, by contrast with that of an α-helix, may also contribute to this error (Kabsch and Sander, 1983; Richards and Kundrot, 1988). In calculating the secondary structure content from three-dimensional coordinates, we have used the program PROMOTIF, which uses the DSSP (database of secondary structure assignments) algorithm of Kabsch and Sander (1983). Definitions of secondary structure by PROMOTIF (Hutchinson and Thornton, 1996) closely follow IUPAC convention rule 6.3, and have been widely accepted amongst crystallographers. Other commonly used programs for secondary structure determination include STRIDE (secondary structure assignment from atomic coordinates) and DEFINE (determine the secondary and first level supersecondary structure) (Frishman and Argos, 1995). Cuff and Barton (1999) have performed a comprehensive comparison of these three methods (DSSP, STRIDE and DEFINE), and shown that DSSP and STRIDE have an overall and segment-wise agreement of 95%. As the secondary structure definitions are based on the coordinates of a model derived by X-ray crystallography or NMR, any algorithm will be affected by the quality of the underlying data. The best estimation rate varies widely depending on the choice of algorithm (Cuff and Barton, 1999; Figureau et al., 1999, 2003). However, of the many different methods of defining secondary structure proposed, DSSP has most successfully stood the test of time, and is widely used in the field of structural biology. Consequently, using PROMOTIF to perform NMR-based secondary structure calculations seems well justified. Moreover, any variation in the secondary structure content determined from three-dimensional coordinates, though it might alter correlations with secondary structure predicted from CSI using a given reference set of random coil values, will not influence systematic variations arising from the use of different reference sets.

Chemical shifts in the BMRB

In light of the present results, it is necessary to point out possible sources of error in the original data. Though the BMRB is highly useful, it is a relatively new database by comparison with three-dimensional structural databases, such as the PDB, and currently lacks a proper strategy for ensuring the accuracy of the information it provides. According to Wishart and Case (2001), one problem has been the lack of proper chemical shift referencing for some nuclei, such as ^{15}N. Other problems might include the possibility that some fraction of the assignments is missing or incorrect, and the inevitable mismatch between chemical shift data and three-dimensional structures, especially when X-ray determined coordinates are used. We note that Zhang et al. (2003) have recently assembled a secondary chemical shift database called RefDB, in which chemical shift information obtained from the

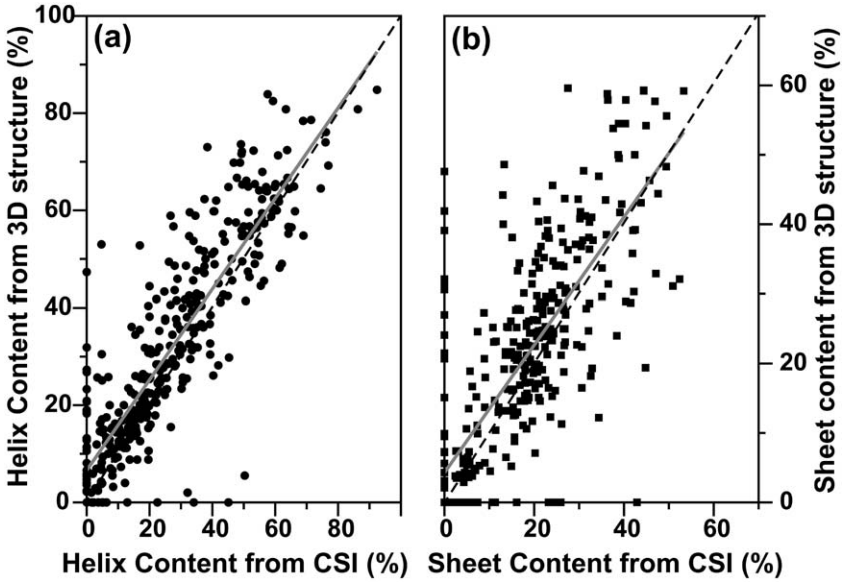

Figure 3. Plots of helical (a) and sheet (b) secondary structure content in percentage determined from chemical shifts and from three-dimensional coordinates. The reference random coil values correspond to *SD*.

BMRB is uniformly referenced, and unassigned or missing chemical shifts are predicted using empirical correlations. However, we have chosen to rely upon the BMRB itself, since the accuracy of such predictions is unclear, and the minimal amount of inconsistent chemical shift referencing in the BMRB, particularly for $^{1}H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$, is expected to have a negligible impact upon our results.

Sequence-dependent effects

Sequence-dependent corrections of random coil chemical shifts have recently been noted using experimental (Schwarzinger et al., 2001; Wishart et al., 1995a, 1995b) and statistical (Wang and Jardetzky, 2002a) methods. Schwarzinger et al. (2001) have experimentally measured a subset of petapeptides to investigate the effect of neighboring residues, and elegantly utilized the results to determine the residual secondary structures in partially unfolded proteins. Wang and Jardetzky (2002a) have recently determined a statistical distribution of nearest neighbor effects from chemical shift data obtained from the BMRB. Though the nearest neighbor effects determined by the statistical method bear a trend similar to that of the experimental results in 8 M urea for random coil chemical shifts, the former approach inherently assumes that residues that are neither helical nor sheet must be 'random coil'. In practice, however, it would be necessary to collect experimental data on at least 8000 different tri-peptide samples to determine nearest neighbor effects completely. Since this would require a monumental effort, and none of the available databases provide a complete set of experimental random coil chemical shifts, we did not exclusively account for nearest neighbor effects.

Conclusions

We have presented a complete analysis of the effect of variations in random coil chemical shift reference values upon secondary structure content estimates from the consensus chemical shift index (CSI) method. Correlations between secondary structure content estimates from CSI, and corresponding estimates from three-dimensional coordinates, were used for the evaluation. A considerable amount of effort has gone into determining random coil chemical shifts, but the specific consequences of using a particular data set to determine protein secondary structures have not been investigated in detail. Over a selected set of well-characterized protein structures, it has been suggested that CSI-based secondary structure determination is 93% accurate in comparison to X-ray structure-based determinations (Wishart and Case, 2001). Our analysis of a considerable amount of raw data from the BMRB and PDB shows that CSI estimates helical and sheet structures to an accuracy of only 90% and 79%,

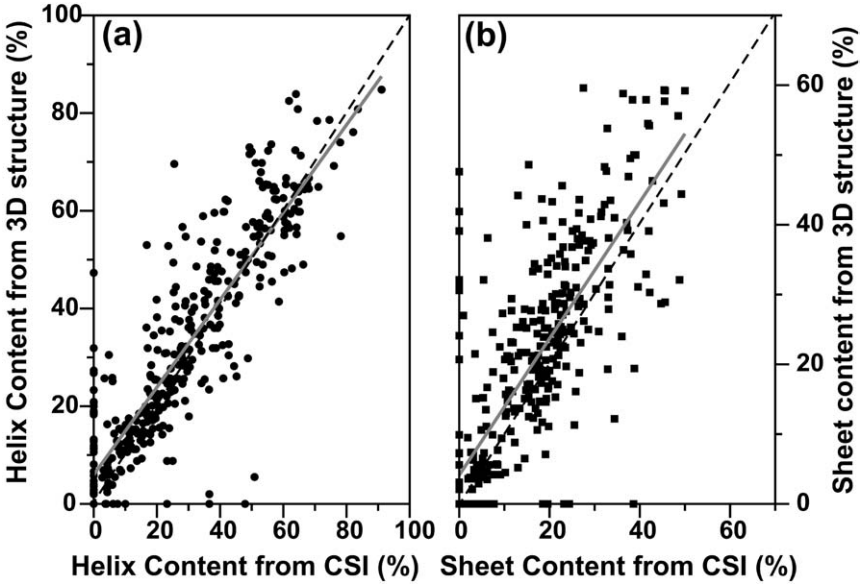

Figure 4. Plots of helical (a) and sheet (b) secondary structure content in percentage determined from chemical shifts and from three-dimensional coordinates. The reference random coil values correspond to *LH*.

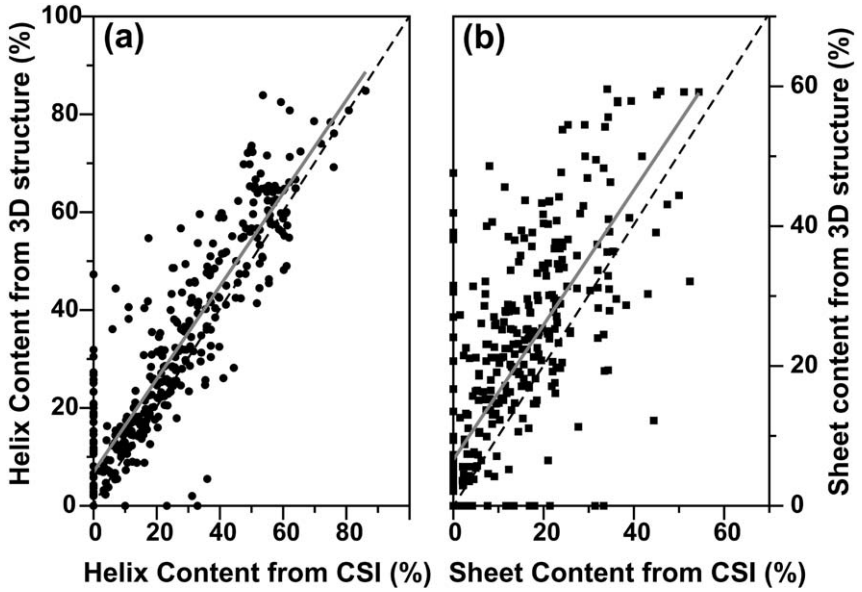

Figure 5. Plots of helical (a) and sheet (b) secondary structure content in percentage determined from chemical shifts and from three-dimensional coordinates. The reference random coil values correspond to *WJ*.

respectively. These results do not reflect the quality of the CSI method itself, but rather the sensitivity of the method to the choice of reference chemical shifts, and the large variation inherent in chemical shift data. Our results further suggest that secondary chemical shifts are more reliable for identifying helical regions of proteins than strand regions. Sharman et al. (2001) have recently proposed that long-range effects from distant amino acids are one of the dominant factors in determining experimental chemical shifts in β-sheets. The absence of a good correlation for β-sheets in the data presented here is perhaps suggestive of this. Though rigorous experimental and statistical methods have been able more accurately to estimate random coil shifts in the last decade, our findings indicate that additional experimental and theoretical

developments are mandatory for an explanation of the observed deviations. The present analysis forms a critical evaluation of the current status of the reliability of secondary chemical shifts as a direct refinement parameter in structure calculations. Though caution must be advised, since this work relies only on secondary chemical shifts, it nevertheless suggests the importance of pursuing a combined experimental, theoretical, and database-driven approach to secondary structure estimation that can provide a better understanding of the factors governing both the chemical shift, and its relationship to protein structure.

Acknowledgements

SPM acknowledges the Student Employee Graduate Research Fellowship (SEGRF). This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory, under contract No. W-7405-Eng-48, and the Laboratory Wide Director's Initiative Grant LW-068.

References

Ando, I., Kuroki, S., Kurosu, H. and Yamanobe, T. (2001) NMR chemical shift calculations and structural characterizations of polymers. *Prog. Nucl. Magn. Reson. Spectrosc.*, **39**: 79–133.

Asakura, T., Iwadate, M., Demura, M. and Williamson, M.P. (1999) Structural analysis of silk with C-13 NMR chemical shift contour plots. *Int. J. Biol. Macromol.*, **24**: 167–171.

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

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, Jr., E.E., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**: 535–542.

Bienkiewicz, E.A. and Lumb, K.J. (1999) Random-coil chemical shifts of phosphorylated amino acids. *J. Biomol. NMR*, **15**: 203–206.

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

Bundi, A. and Wüthrich, K. (1979) ^{1}H-NMR parameters of the common amino acid a\residues measured in aqueous solutions of the linear tetrapeptides H-Gly-Gly-X-LAla-OH. *Biopolymers*, **18**: 285–297.

Case, D.A. (1998) The use of chemical shifts and their anisotropies in biomolecular structure determination. *Curr. Opin. Struct. Biol.*, **8**: 624–630.

Case, D.A. (2000) Interpretation of chemical shifts and coupling constants in macromolecules. *Curr. Opin. Struct. Biol.*, **10**: 197–203.

Case, D.A., Dyson, H.J. and Wright, P.E. (1994) Use of chemical shifts and coupling constants in nuclear magnetic resonance structural studies of peptides and proteins. *Meth. Enzymol.* **239**: 392–416.

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

Cuff, J.A. and Barton, G.J. (1999) Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins*, **34**: 508–519.

Dalgarno, D.C., Levine, B.A. and Williams, R.J. (1983) Structural information from NMR secondary chemical shifts of peptide alpha C-H protons in proteins. *Biosci. Rep.*, **3**: 443–452.

Figureau, A., Soto, M.A. and Toha, J. (1999) Secondary structure of proteins and three-dimensional pattern recognition. *J. Theor. Biol.*, **201**: 103–111.

Figureau, A., Soto, M.A. and Toha, J. (2003) A pentapeptide-based method for protein secondary structure prediction. *Protein Eng.*, **16**: 103–107.

Frishman, D. and Argos, P. (1995) Knowledge-based protein secondary structure assignment. *Proteins*, **23**: 566–579.

Glushka, J., Lee, M., Coffin, S. and Cowburn, D. (1989) ^{15}N chemical shifts of backbone amides in bovine pancreatic trypsin inhibitor and apamin. *J. Amer. Chem. Soc.*, **111**: 7716–7722.

Glushka, J., Lee, M., Coffin, S. and Cowburn, D. (1990) ^{15}N chemical shifts of backbone amides in bovine pancreatic trypsin inhibitor and apamin. (correction). *J. Amer. Chem. Soc.*, **112**: 2843.

Gross, K.-H. and Kalbitzer, H.R. (1988) Distribution of chemical shifts in ^{1}H nuclear magnetic resonance spectra of proteins. *J. Magn. Reson.*, **76**: 87–99.

Hutchinson, E.G. and Thornton, J.M. (1996) PROMOTIF – a Program to Identify and Analyze Structural Motifs in Proteins. *Protein Sci*, **5**: 212–220.

Kabsch, W. and Sander, C. (1983) A dictionary of protein secondary structure. *Biomolymers*, **22**: 2577–2637.

Kuszewski, J., Gronenborn, A.M. and Clore, G.M. (1995a) The impact of direct refinement against proton chemical shifts on protein structure determination by NMR. *J. Magn. Reson. Ser. B*, **107**: 293–297.

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

Kuszewski, J., Gronenborn, A.M. and Clore, G.M. (1996) A potential involving multiple proton chemical-shift restraints for nonstereospecifically assigned methyl and methylene protons. *J. Magn. Reson. Ser. B*, **112**: 79–81.

Laws, D.D., Dedios, A.C. and Oldfield, E. (1993) NMR chemical shifts and structure refinement in proteins. *J. Biomol. NMR*, **3**: 607–612.

Luginbuhl, P., Szyperski, T. and Wüthrich, K. (1995) Statistical basis for the use of C-13-alpha chemical shifts in protein structure determination. *J. Magn. Reson. Ser. B*, **109**: 229–233.

Lukin, J.A., Gove, A.P., Talukdar, S.N. and Ho, C. (1997) Automated probabilistic method for assigning backbone resonances of (C-13,N-15)-labeled proteins. *J. Biomol. NMR*, **9**: 151–166.

Merutka, G., Dyson, H.J. and Wright, P.E. (1995) Random coil H-1 chemical shifts obtained as a function of temperature and trifluoroethanol concentration for the peptide series Ggxgg. *J. Biomol. NMR*, **5**: 14–24.

Metzler, W.J., Constantine, K.L., Friedrichs, M.S., Bell, A.J., Ernst, E.G., Lavoie, T.B. and Mueller, L. (1993) Characterization of the 3-dimensional solution structure of human profilin – H-1,

C-13, and N-15 NMR assignments and global folding pattern. *Biochemistry*, **32**: 13818–13829.

Mielke, S.P. and Krishnan, V.V. (2003) Protein structural class identification directly from NMR spectra using averaged chemical shifts. *Bioinformatics*, **19**: 2054–2064.

Oldfield, E. (1995) Chemical shifts and three-dimensional protein structures. *J. Biomol. NMR*, **5**: 217–225.

Oldfield, E. (2002) Chemical shifts in amino acids, peptides, and proteins: from quantum chemistry to drug design. *Annu. Rev. Phys. Chem.*, **53**: 349–378.

Pastore, A. and Saudek, V. (1990) The relationship between chemical shift and secondary structure in proteins. *J. Magn. Reson.*, **90**: 165–176.

Plaxco, K.W., Morton, C.J., Grimshaw, S.B., Jones, J.A., Pitkeathly, M., Campbell, I.D. and Dobson, C.M. (1997) The effects of guanidine hydrochloride on the 'random coil' conformations and NMR chemical shifts of the peptide series GGXGG. *J. Biomol. NMR*, **10**: 221–230.

Reif, B., Hennig, M. and Griesinger, C. (1997) Direct measurement of angles between bond vectors in high-resolution NMR. *Science*, **276**(5316): 1230–1233.

Reily, M.D., Thanabal, V. and Omecinsky, D.O. (1992) Structure-induced carbon-13 chemical shifts: A sensitive measure of transient localized secondary structure in peptides. *J. Amer. Chem. Soc.*, **114**: 6251–6252.

Richards, F.M. and Kundrot, C.E. (1988) Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins*, **3**: 71–84.

Richarz, R. and Wüthrich, K. (1978) Carbon-13 NMR chemical shifts of the common amino acid residues measured in aqueous solutions of the linear tetrapeptides H-Gly-Gly-X-LAla-OH. *Biomolymers*, **17**: 2133–2141.

Sanders, J.C., Haris, P.I., Chapman, D., Otto, C. and Hemminga, M.A. (1993) Secondary structure of M13 coat protein in phospholipids studied by circular dichroism, Raman, and Fourier transform infrared spectroscopy. *Biochemistry*, **32**: 12446–12454.

Schwarzinger, S., Kroon, G.J.A., Foss, T.R., Chung, J., Wright, P.E. and Dyson, H.J. (2001) Sequence-dependent correction of random coil NMR chemical shifts. *J. Amer. Chem. Soc.*, **123**: 2970–2978.

Schwarzinger, S., Kroon, G.J.A., Foss, T.R., Wright, P.E. and Dyson, H.J. (2000) Random coil chemical shifts in acidic 8 M urea: Implementation of random coil shift data in NMRView. *J. Biomol. NMR*, **18**: 43–48.

Seavey, B.R., Farr, E.A., Westler, W.M. and Markley, J.L. (1991) A relational database for sequence-specific protein NMR data. *J. Biomol. NMR*, **1**: 217–236.

Sharman, G.J., Griffiths-Jones, S.R., Jourdan, M. and Searle, M.S. (2001) Effects of amino acid phi,psi propensities and secondary structure interactions in modulating H alpha chemical shifts in peptide and protein beta-sheet. *J. Amer. Chem. Soc.*, **123**: 12318–12324.

Sibley, A.B., Cosman, M. and Krishnan, V.V. (2003) An empirical correlation between secondary structure content and averaged chemical shifts in proteins. *Biophys. J.*, **84**: 1223–1237.

Spera, S. and Bax, A. (1991) Empirical correlation between protein backbone conformation and C-alpha and C-beta C-13 nuclear magnetic resonance chemical shifts. *J. Amer. Chem. Soc.*, **113**: 5490–5492.

Sreerama, N. and Woody, R.W. (1994) Protein secondary structure from circular dichroism spectroscopy. Combining variable selection principle and cluster analysis with neural network, ridge regression and self-consistent methods. *J. Mol. Biol.*, **242**: 497–507.

Szilagyi, L. (1995) Chemical shifts in proteins come of age. *Prog. Nucl. Magn. Reson. Spectrosc.*, **27**(P4): 325–443.

Szilagyi, L. and Jardetzky, O. (1989) α-Proton chemical shifts and secondary structure in proteins. *J. Magn. Reson.*, **83**: 441–449.

Thanabal, V., Omecinsky, D.O., Reily, M.D. and Cody, W.L. (1994) The 13C chemical shifts of amino acids in aqueous solution containing organic solvents: Application to the secondary structure characterization of peptides in aqueous trifluoroethanol solution. *J. Biomol. NMR*, **4**: 47–59.

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

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

Williamson, M.P. (1990) Secondary-structure dependent chemical shifts in proteins. *Biomolymers*, **29**(10–1): 1428–1431.

Williamson, M.P., Kikuchi, J. and Asakura, T. (1995) Application of H-1 NMR chemical shifts to measure the quality of protein structures. *J. Mol. Biol.*, **247**: 541–546.

Wishart, D.S. and Case, D.A. (2001) Use of chemical shifts in macromolecular structure determination. *Meth. Enzymol.*, **338**: 3–34.

Wishart, D.S. and Nip, A.M. (1998) Protein chemical shift analysis: A practical guide. *Biochem. Cell Biol.-Biochim. Biol. Cell.*, **76**: 153–163.

Wishart, D.S. and Sykes, B.D. (1994a) The C-13 chemical-shift index – A simple method for the identification of protein secondary structure using C-13 chemical-shift data. *J. Biomol. NMR*, **4**: 171–180.

Wishart, D.S. and Sykes, B.D. (1994b) Chemical shifts a tool for structure determination. *Meth Enzymol.*, **239**: 363–392.

Wishart, D.S., Bigam, C.G., Holm, A., Hodges, R.S. and Sykes, B.D. (1995a) H-1, C-13 and N-15 random coil NMR chemical shifts of the common amino acids. 1. Investigations of nearest-neighbor effects. *J. Biomol. NMR*, **5**: 67–81.

Wishart, D.S., Bigam, C.G., Yao, J., Abildgaard, F., Dyson, H.J., Oldfield, E., Markley, J.L. and Sykes, B.D. (1995b) H-1, C-13 and N-15 chemical shift referencing in biomolecular NMR. *J. Biomol. NMR*, **6**: 135–140.

Wishart, D.S., Sykes, B.D. and Richards, F.M. (1991a) Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *J. Mol. Biol.*, **222**: 311–333.

Wishart, D.S., Sykes, B.D. and Richards, F.M. (1991b) Simple techniques for the quantification of protein secondary structure by H-1 NMR spectroscopy. *FEBS Lett.*, **293**: 72–80.

Wishart, D.S., Sykes, B.D. and Richards, F.M. (1992) The chemical shift index – a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry*, **31**: 1647–1651.

Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*. Wiley, New York, xv, 292 pp.

Zhang, H.Y., Neal, S. and Wishart, D.S. (2003) RefDB: A database of uniformly referenced protein chemical shifts. *J. Biomol. NMR*, **25**: 173–195.