BIOINFORMATICS

Vol. 19 no. 16 2003, pages 2054–2064
DOI: 10.1093/bioinformatics/btg280

Protein structural class identification directly from NMR spectra using averaged chemical shifts*

S.P. Mielke[1,2,§] and V.V. Krishnan[2,*,§]

[1]Biophysics Graduate Group, University of California, Davis, CA 95616, USA and
[2]Molecular Biophysics Group, L-448 Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94551, USA

Received on January 6, 2003; revised on March 19, 2003; accepted on March 28, 2003

ABSTRACT

Knowledge of the three-dimensional structure of proteins is integral to understanding their functions, and a necessity in the era of proteomics. A wide range of computational methods is employed to estimate the secondary, tertiary, and quaternary structures of proteins. Comprehensive experimental methods, on the other hand, are limited to nuclear magnetic resonance (NMR) and X-ray crystallography. The full characterization of individual structures, using either of these techniques, is extremely time intensive. The demands of high throughput proteomics necessitate the development of new, faster experimental methods for providing structural information. As a first step toward such a method, we explore the possibility of determining the structural classes of proteins directly from their NMR spectra, prior to resonance assignment, using averaged chemical shifts. This is achieved by correlating NMR-based information with empirical structure-based information available in widely used electronic databases. The results are analyzed statistically for their significance. The robustness of the method as a structure predictor is probed by applying it to a set of proteins of unknown structure. Our results show that this NMR-based method can be used as a low-resolution tool for protein structural class identification.

Contact: krish@llnl.gov

1 INTRODUCTION

The structural class of a protein lies at the top of any hierarchical characterization of its fold. The concept of protein structural class was first introduced by Levitt and Chothia, based on the visual inspection of polypeptide chain topologies in a data set of 31 proteins (Levitt and Chothia, 1976). In the last decade, the designation of class based on secondary structure content has been extremely useful from both experimental and theoretical points of view (Klein and Delisi, 1986; Klein, 1986; Zhang and Chou, 1992; Zhou *et al.*, 1992; Metfessel *et al.*, 1993; Boberg *et al.*, 1995; Chou and Zhang, 1995; Zhou, 1998; Wang and Yuan, 2000; Wang, 2001; Cai *et al.*, 2001; Li and Lu, 2001; Luo *et al.*, 2002). The structural class presents an intuitive description of overall protein fold. Knowledge of class can significantly increase the quality of secondary structure prediction from amino acid sequence (Chou, 1989; Deleage *et al.*, 1987; Deleage and Roux, 1987; Deleage and Dixon, 1989; Kneller *et al.*, 1990; Muggleton *et al.*, 1992; Cohen *et al.*, 1993), reduce the scope of conformational searches during energy optimization (Cohen and Kuntz, 1987; Carlacci and Englander, 1993), and improve the calculation of hydrophobicity coefficients (Cid *et al.*, 1992). In addition, such knowledge can provide information about other functional properties of proteins, such as cellular location (i.e. whether the molecule is an intracellular, extracellular, or membrane protein) and the presence of disulfide bonds (Nishikawa and Ooi, 1982, 1986a; Nishikawa *et al.*, 1986b).

Proteins are generally placed into one of three major classes: 'mainly-α' (α), 'mainly-β' (β) and $\alpha\beta$ (including 'α/β' and '$\alpha+\beta$'). α and β proteins are defined as those composed of predominantly α-helices or β-strands, respectively. The $\alpha+\beta$ class consists of proteins in which α and β regions are largely separated, and β-strands are often antiparallel, while the α/β class consists of proteins in which helices and strands are mixed, and β-strands are parallel. These definitions are generally accepted, and widely used in the literature (Chou and Zhang, 1995). In recent years, several methods have been proposed to identify and predict the structural classes of globular proteins (Dietmann and Holm, 2001; Taylor, 2002). These methods range from simple computational estimation from primary sequence information to full characterization from high-resolution three-dimensional structural information. Two well known databases providing the latter, available on the World Wide Web, are CATH (Class-Architecture-Topology-Homologous Superfamily, http://www.biochem.ucl.ac.uk/bsm/cath_new/)

[*] To whom correspondence should be addressed.

[§] The authors dedicate this manuscript with affection and admiration to Prof. Robert E. Feeney, Department of Food Sciences and Technology, University of California, Davis, on his 90th birthday.

(Orengo *et al.*, 1997) and SCOP (Structural Classification of Proteins, http://scop.berkeley.edu/) (Lo Conte *et al.*, 2002).

Since their first observation in nuclear magnetic resonance (NMR) spectra in 1957 (Gutowsky *et al.*, 1957), nuclear chemical shifts have proven to be powerful indicators of the types of structures that biopolymers can adopt. The development of modern NMR experiments is driven predominantly by the goal of increasing the resolution and sensitivity with which the chemical shift of a nucleus can be measured. In addition to structural information, chemical shifts provide detailed information about the nature of hydrogen exchange dynamics, ionization and oxidation states, the ring current influence of aromatic residues, and hydrogen bonding interactions (Szilagyi, 1995). Several excellent recent review articles describe a wide variety of experimental and computational methods for correlating chemical shifts with protein three-dimensional structure (Szilagyi, 1995; Ando, 2001; Wishart and Case, 2001). However, these methods rely on the chemical shift assignment of each resonance belonging to a particular atom in the molecule (Wüthrich, 1986), which remains a time consuming procedure, despite efforts to automate (or semi-automate) the process (Koradi *et al.*, 1998; Moseley and Montelione, 1999).

We present here an empirical approach for estimating protein structural class directly from NMR spectra, prior to the arduous task of resonance assignment. This approach is parallel in spirit to other spectroscopic methods, such as circular dichroism (CD) spectroscopy in the UV absorption range (Johnson, 1990; Perczel *et al.*, 1991; Sreerama and Woody, 1994) and IR Raman spectroscopy (Williams *et al.*, 1986; Bussian and Sander, 1989; Chi *et al.*, 1998; Sanders *et al.*, 1993) that do not require full knowledge of three-dimensional structures for the estimation of protein structural classes. We have extensively used chemical shift information available in the Biological Magnetic Resonance Bank [BioMagResBank (BMRB), http://www.bmrb.wisc.edu/] and Protein Data Bank (PDB, http://www.rcsb.org/pdb/), in combination with the aforementioned structure-based protein classification tools, CATH and SCOP. Averaged chemical shift (ACS) values are calculated for a set of proteins whose chemical shift data are available in the BMRB, and separated according to protein structural class designations into different data sets: α, β, and $\alpha\beta$ (including both α/β and $\alpha+\beta$) according to CATH, and α, β, and $\alpha\beta$ according to SCOP. Using rigorous statistical methods, several of these data sets are shown to represent independent distributions: $^1H_\alpha$ ACS values are observed to differentiate protein class with high sensitivity, ^1H_N values with somewhat less sensitivity, and values associated with the heteronuclei, $^{13}C_\alpha$ and ^{15}N, not at all. BMRB entries with chemical shift information, but with no experimental three-dimensional structural information (i.e. no corresponding PDB entries), are used to test the ability of the method to predict structural class. These predictions correlate well with those obtained using only amino acid sequence information. Our results demonstrate the feasibility of obtaining fast, low-resolution protein structural information directly from NMR spectra, in the absence of resonance assignments. Such information provides valuable insight prior to the time-intensive process of complete, high-resolution determination of three-dimensional structure.

2 MATERIALS AND METHODS

2.1 Chemical shift information

Chemical shift values corresponding to the protein atoms ^1H_N, ^{15}N, $^1H_\alpha$ and $^{13}C_\alpha$ were obtained from BMRB star files (Seavey *et al.*, 1991). If information on the structure of a protein was also present in the corresponding star file, this information was extracted, as was information on the amino acid sequence. 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, since our goal was to establish a correlation between the chemical shift value averaged over the entire molecule, and the molecule's structural class.

The ACS of a nuclear species 'i' was calculated using:

$$\mathrm{ACS}_i \equiv (1/N) \sum_{k=1,M} \omega_k, \qquad (1)$$

where $i = {}^1H_N$, ^{15}N, $^1H_\alpha$ or $^{13}C_\alpha$, N is the total number of residues in the protein sequence, M is the total number of residues with a chemical shift value assigned for species i, and ω_k is the chemical shift of the kth resonance. We choose to divide by N, rather than M, in order to ensure that ACS values characterize entire molecules. Drastic underestimation of ACS values is circumvented by choosing only those molecules with 70% or greater assignment; for this choice, for the majority of proteins, $M \approx N$, and no significant differences are observed in the correlation (Sibley *et al.*, 2003). Typically, chemical shifts correspond to a single bond-correlated spectrum, such as a heteronuclear single quantum correlation, HSQC (Ernst *et al.*, 1990). BMRB chemical shifts are referenced using the widely accepted standard procedure recommended by Wishart *et al.* (1995), so no rereferencing of the values taken from the star files was necessary. Finally, a master list containing 378 proteins was generated. The complete list is available from the authors, upon request.

Three-dimensional structural information Structure files were obtained from Rutgers Center for Structural Biology (RCSB) (PDB format, http://www.rcsb.org/pdb/) (Berman *et al.*, 2000). Since most BMRB star files reference several corresponding PDB structures, it was necessary to examine each entry, and choose 'by hand' the most appropriate PDB ID number. When possible, the PDB ID corresponding to the 'best' NMR structure was chosen. Entries with no corresponding PDB structure were noted, and subsequently used in the test set for the classification scheme.

Protein classification Once one PDB ID number was designated, when possible, for each BMRB accession number in the master list, the CATH (Orengo *et al*., 1997) and SCOP (Lo Conte *et al*., 2002) secondary structure classifications for each entry were obtained from http://www.biochem.ucl.ac.uk/bsm/cath_new/ and http://scop.mrc-lmb.cam.ac.uk/scop/pdb.cgi/, respectively. The CATH and SCOP class-level designations were noted separately for each protein for which they were available, and summarized in both cases by the general categories, 'Alpha Proteins' (α), 'Alpha/Beta Proteins' (αβ) and 'Beta Proteins' (β).

Statistical analysis Upon completion of these class assignments, we sought to test the degree to which, for a given nuclear species, the distribution of ACS values in each class can be said to differ from that in the other two. Separately for CATH and SCOP, three data files containing ACS values were generated for each species, one for each protein class. (Proteins for which either no class had been assigned, or there was no corresponding PDB entry, were omitted.) Next, each pair of files was statistically compared, using the Komolgorov–Smirnov (KS) algorithm provided by (Press, 1988). The KS 'D' statistic provides a straightforward criterion for determining whether two data sets are drawn from different continuous distributions of a single independent variable. This standard test is based upon a comparison of the cumulative distribution functions of the two data sets. All analyses were performed using Perl scripts and C codes on a Silicon Graphics UNIX workstation.

3 RESULTS

Averaged chemical shifts are sensitive to protein structural class Figures 1A and B plot the $^{13}C_{\alpha}$ versus $^{1}H_{\alpha}$, and ^{15}N versus $^{1}H_{N}$, ACS values, respectively, for the proteins in our data set. Values corresponding to molecules deemed α-class, according to their CATH classification (see Methods), are represented by circles. Values corresponding to molecules deemed β-class are represented by squares. Inspection of the figures immediately suggests the possibility of a correlation between designated structural class and ACS. The degree to which each data set can be said actually to represent two distinct sets, each drawn from a different distribution, was first probed by comparison with results from NMR analysis of specific proteins of known structural class. Figures 1a and b plot the simulated HSQC spectra of $^{13}C_{\alpha}$ and $^{1}H_{\alpha}$ nuclei (^{13}C-HSQC or COSY) for histidine kinase (PDB code 1A0B, BMRB number 4857) (Ikegami *et al*., 2001) and liver fatty acid binding protein (PDB code 1LFO, BMRB number 4098) (Wang *et al*., 1998), respectively. Figure 1c and d plot the HSQC ^{15}N and $^{1}H_{N}$ spectra for the same proteins. Histidine kinase is known to be predominantly α-helical, and liver fatty acid binding protein predominantly β-sheet. The three-dimensional structures of these proteins are shown in Figure 1 (above and below Figure 1A and B). Each point in Figure 1a–d (circles for histidine kinase and squares for liver fatty acid binding protein) represents the crosspeak between nuclei due to the presence of one-bond J-coupling. Such NMR spectra are typically among the first sets of experimental results obtained for isotopically enriched proteins. Calculated ACS values for the $^{1}H_{\alpha}$–$^{13}C_{\alpha}$ plots of Figures 1a and b are 4.15–58.84 ppm and 4.73–57.55 ppm, respectively. Those for the $^{1}H_{N}$–^{15}N correlation of Figures 1c and d are 8.13–119.82 and 8.64–124.70 ppm, respectively. These values are indicated as large, black circles (for the histidine kinase data) and squares (for the binding protein data). These same data points are duplicated in Figure 1A and B, where it is apparent that they fall in the expected cluster within the appropriate larger data set.

Several distinct features can be observed in Figure 1. Even though the overall features of the HSQC (^{15}N and ^{13}C) spectra of the helical and sheet proteins look very similar, the average chemical shifts of the spectra are distinctly different, as illustrated by the vertical lines in Figure 1a–d. The $^{1}H_{\alpha}$ ACS values of the proteins classified as mainly-α (circles) in Figure 1A are shifted upfield (more shielded), while the $^{13}C_{\alpha}$ values are shifted downfield (less shielded) with respect to the proteins that are classified as mainly-β (squares). For the $^{1}H_{N}$–^{15}N correlation of Figure 1B, though the $^{1}H_{N}$ ACS values show the same trend as the $^{1}H_{\alpha}$ values of Figure 1A (i.e. they are shifted upfield for the mainly α-helical proteins), the ^{15}N ACS values also shift upfield, in contrast to the $^{13}C_{\alpha}$ ACS values of Figure 1A, relative to the mainly-β proteins. The difference between the $^{1}H_{\alpha}$ ACS values of the highly helical and sheet proteins is $+0.59$ ppm, and that for the $^{1}H_{N}$ ACS values is $+0.51$ ppm. It is this relative shifting of the ACS values that leads to the apparent separation of the data by class identified by CATH.

Distribution of protein structural classes with respect to ACS values The distinctness of the data sets according to class was further probed by simple statistical analysis. Figures 2 and 3 show histograms of the protein distributions, separated according to SCOP and CATH classification protocols, respectively. Figure 2 shows the ACS values of the $^{1}H_{\alpha}$ nucleus (left panels) and $^{1}H_{N}$ nucleus (right panels), classified using SCOP as α, αβ, and β in panels a and d, b and e, and c and f, respectively. Figure 3 shows the distributions resulting from the CATH classification. For the histograms in Figures 2 and 3, a bin size of width 0.05 ppm was chosen, and points on the X-axes were chosen at the centers of the bins. Table 1 lists the total number of proteins in each protein class for both SCOP and CATH classification, and the standard deviation and standard deviation of the mean of the individual distributions. We make no a priori assumptions about the actual form of the distribution functions. The mean $^{1}H_{\alpha}$ ACS values of

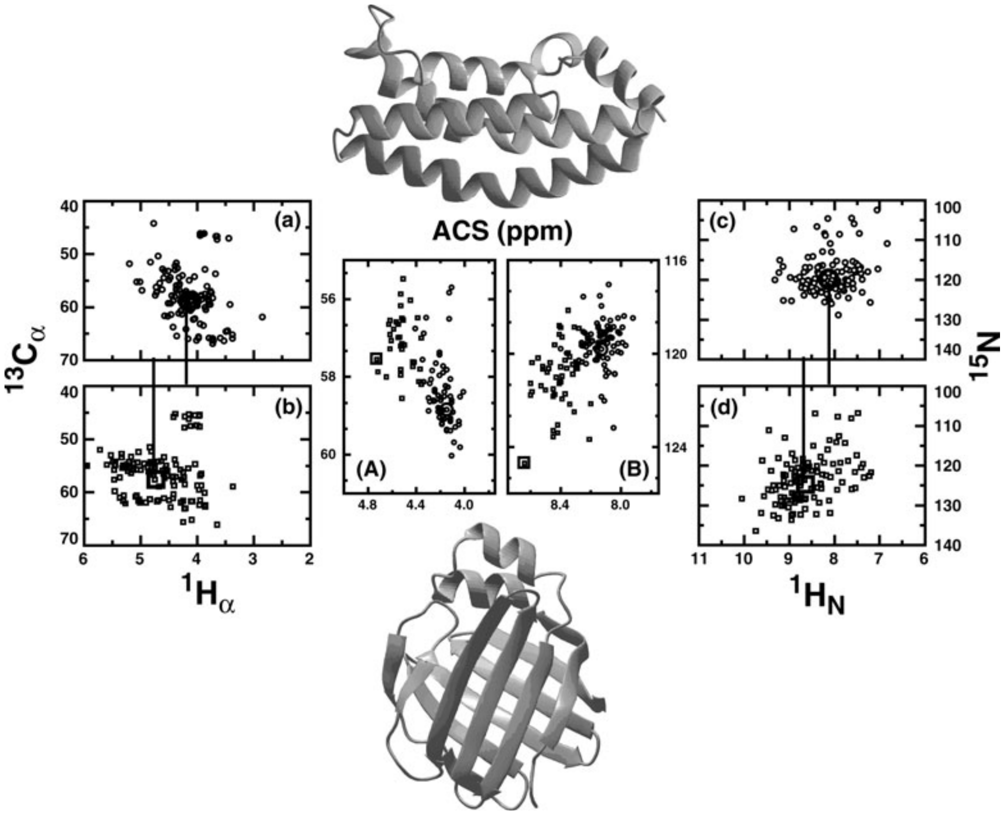

Fig. 1. Representative examples to show that ACS is a structural parameter directly obtainable from NMR spectra. (**a**) and (**c**): simulated ^{13}C and ^{15}N-HSQC spectra of an α-helical protein (Histidine kinase, PDB code 1A0B, BMRB number 4857), respectively. (**b**) and (**d**): simulated ^{13}C and ^{15}N-HSQC spectra of a β-sheet protein (Liver fatty acid binding protein, PDB code 1LFO, BMRB number 4098). The ACS calculated from each spectrum is noted by a black circle (helical protein) and square (sheet protein). (**A**) and (**B**): representative examples of the ACS values calculated from $^{13}C_\alpha$–$^1H_\alpha$ and ^{15}N–^1H_N correlations, respectively, for a set of proteins for which chemical shift information is obtained from BioMagResBank. The circles and squares correspond to proteins that are classified as mainly-α and mainly-β, respectively, under the CATH classification scheme. ACS values from (a) and (b), and (c) and (d), are reproduced in (A) and (B), respectively.

the α, αβ and β classes of the SCOP-classified proteins are 3.83, 3.94 and 4.05 ppm, respectively, while the corresponding CATH classification values are 3.79, 3.93 and 4.05 ppm. The mean values for the three different classes of proteins under both SCOP and CATH differ by at least 0.1 ppm, providing the justification for separating the ACS values into three categories: mainly-α, αβ and mainly-β. However, such a difference is not evident for the means of the ^1H_N ACS values [Figs 2 and 3 (right panels), and Table 1]. In this case, one expects the ACS values can be reasonably separated into only two major divisions, α and 'αβ/β', as the mean values of the αβ and β classes are too close to be distinguished. Similar analyses revealed no such distinction for the backbone heteroatoms, $^{13}C_\alpha$ and ^{15}N.

Kolmogorov–Smirnov (KS) tests In order to determine whether any reliable information could be obtained from the data presented in Figures 2 and 3, and Table 1, it was necessary to address the important question of whether the distributions of ACS values classified as α, β or αβ, are in fact independent of one another. In order to explore this issue, the KS test was performed. Table 2 lists the results of the test for all nuclei, and for comparisons of all potentially distinct classes, as designated by both CATH and SCOP. We consider two distributions to be independent for values of the 'significance' less than or equal to 0.05; i.e. if there is 5% or less probability of obtaining a value of the statistic D 'by chance alone', the two distributions being compared are deemed significantly different. Table 2 lists both values of the statistic, and the significance of those values for all comparisons. Only the separations by class, according to SCOP, of $^1H_\alpha$ ACS values are seen to be significant for all three comparisons. For CATH-based separations of $^1H_\alpha$ ACS values, the α-class is seen to be distinct from both αβ and β, but αβ and β just miss being distinct from each other, at the 5% significance cutoff. Both the $^{13}C_\alpha$ and ^1H_N separations also appear to be useful for distinguishing α from both αβ and β, but not αβ and β from each other, for both SCOP and CATH schemes. Table 2 reveals

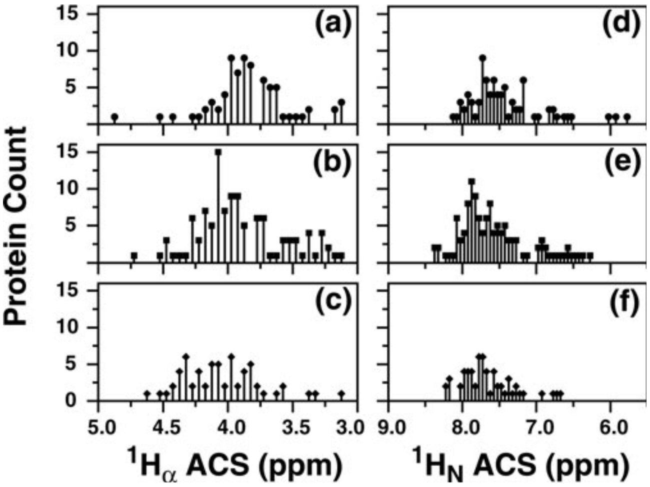

Fig. 2. ACS values versus number of proteins in the three major structural classes defined according to the SCOP method. (**a**), (**b**) and (**c**) display the $^1H_\alpha$ ACS values for proteins that are mainly-α, mainly-β, and a mixture of α and β(αβ) (both α/β and α + β), respectively. (**d**), (**e**) and (**f**) display the corresponding ^1H_N values for mainly-α, mainly-β and αβ (both α/β and α+β), respectively.

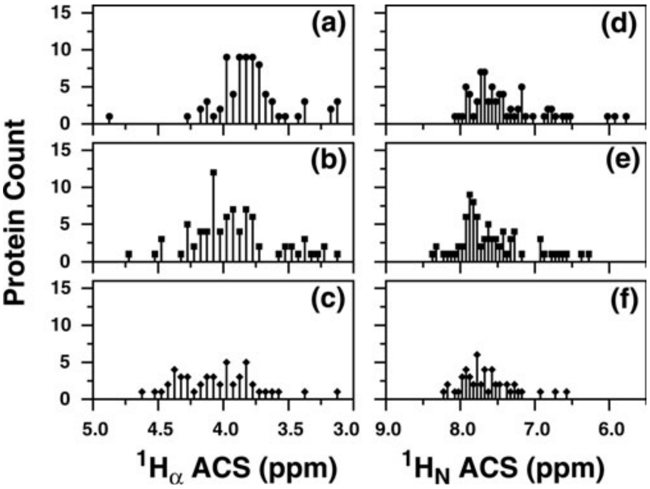

Fig. 3. ACS values versus number of proteins in the three major structural classes defined according to the CATH method. (**a**), (**b**), and (**c**) display the $^1H_\alpha$ ACS values for proteins that are α, β, and αβ (both α/β and α + β), respectively. (**d**), (**e**) and (**f**) display the corresponding ^1H_N values for α, β and αβ (both α/β and α+β), respectively.

the results for ^{15}N to be unreliable for all six comparisons. The KS test results, then, confirm quantitatively the trends suggested by the mean values of the data sets (Table 1).

Identification of the protein class from NMR data Rigorous statistical analysis of the data clearly suggests that only the

Table 1. Characterization of the statistical distribution of structural classes

Class	Total[b]	Mean	SD[c]	SDM[d]	2 ∗ SDM
SCOP[a] nucleus $^1H_\alpha$					
α	88	3.83	0.34	0.040	0.072
αβ	122	3.94	0.52	0.047	0.093
β	61	4.05	0.30	0.038	0.076
Nucleus ^1H_N					
α	87	7.54	0.64	0.069	0.14
αβ	122	7.68	0.99	0.089	0.18
β	60	7.70	0.41	0.053	0.11
CATH[e] nucleus $^1H_\alpha$					
α	77	3.79	0.29	0.033	0.066
αβ	83	3.93	0.32	0.035	0.070
β	49	4.05	0.30	0.043	0.086
Nucleus ^1H_N					
α	75	7.45	0.56	0.064	0.13
αβ	83	7.62	0.48	0.053	0.11
β	49	7.69	0.43	0.061	0.12

[a] SCOP (Structural Classification of Proteins).
[b] Total number of proteins.
[c] SD: standard deviation.
[d] SDM: standard deviation about the mean.
[e] CATH (Class-Architecture-Topology-Homologous Superfamily).

Table 2. Results of Kolmogorov–Smirnov test

Classes compared	SCOP[a]		CATH[b]	
	KS *D* Statistic[c]	Significance[d]	KS *D* Statistic[c]	Significance[d]
$^1H_\alpha$				
α ↔ αβ	0.24	**0.0039**	0.32	**0.00042**
α ↔ β	0.41	**0.0000060**	0.41	**0.000042**
αβ ↔ β	0.24	**0.018**	0.23	0.058
$^{13}C_\alpha$				
α ↔ αβ	0.29	**0.00030**	0.29	**0.0021**
α ↔ β	0.41	**0.0000090**	0.34	**0.0015**
αβ ↔ β	0.18	0.15	0.21	0.11
^1H_N				
α ↔ αβ	0.22	**0.012**	0.28	**0.0029**
α ↔ β	0.26	**0.015**	0.27	**0.021**
αβ ↔ β	0.11	0.65	0.092	0.94
^{15}N				
α ↔ αβ	0.082	0.88	0.10	0.79
α ↔ β	0.11	0.77	0.14	0.59
αβ ↔ β	0.13	0.47	0.13	0.65

[a] Proteins classified using SCOP.
[b] Proteins classified using CATH.
[c] Maximum value of absolute difference between cumulative distribution functions.
[d] Significance: values less than/equal to 0.05 are considered significant (numbers in bold print).

$^1H_\alpha$ ACS values are capable of distinguishing the three different structural classes of the proteins, as designated using either SCOP or CATH protocols, in a statistically significant way. Following the success of the KS test for $^1H_\alpha$, we sought to define the range of $^1H_\alpha$ ACS values corresponding to each class. Considering the overlap between the distributions (Figs 2 and 3, left panels, and Table 1), a conservative approach has been considered. As the relative mean values of the three classes of the proteins differ from each other by no more than about 0.14 ppm for $^1H_\alpha$, we have chosen to consider twice the standard deviation of the mean as the width for a particular class (and still there is overlap between the bins; see Table 1). Though this definition is not rigorously justified, it is reasonable as a first-order approximation of the expected distribution of ACS values about a mean assumed to represent a best estimate of the value characterizing a particular protein class. Our results, then, are as follows. For protein structural classes, α, αβ and β, defined by SCOP, the centers of the ACS values are 3.83 ± 0.072, 3.94 ± 0.093 and 4.05 ± 0.076 ppm, respectively. The corresponding values for the CATH-classified proteins are 3.79 ± 0.066, 3.93 ± 0.070 and 4.05 ± 0.086 ppm, respectively.

Using these criteria, we have predicted the structural classes of a set of proteins, lacking experimental three-dimensional structural information, from the corresponding chemical shift information. The results for a total of 37 proteins, predicted using both CATH- and SCOP-derived empirical relations, are summarized in Table 3. Of the 14 proteins that are predicted to be in the mainly-α class based on either the CATH- or SCOP-derived correlations, one falls in the overlapping region between the α and αβ classes, according to the SCOP-derived correlation. This protein (BMRB 4698) is designated 'α/αβ' in Table 3. A similar scenario is observed in the αβ and β classes for CATH- and SCOP-derived predictions, where proteins falling in overlapping regions are again designated α/αβ, or αβ/β. It must be noted that predictions were not possible for a few proteins that do not have ACS values within the conservative widths defined (data not shown). Two proteins could be classified using the CATH-, but not the SCOP-based, relation. These are entered as 'NP' (No Prediction) in the SCOP column of Table 3. Table 3 also shows that there is no cross-prediction between α and β classes.

4 DISCUSSION

NMR spectroscopy plays a vital role in determining the structures of proteins in the solution state. In spite of advancement in the field during the past decade, determining the complete three-dimensional structure of any given protein is still a time-consuming proposition. Though the information content in the complete structure at atomic resolution is indisputable, several groups have recently begun exploring alternative high-resolution methods that are faster than conventional experiments (Atkinson and Saudek, 2002; Grishaev and Llinas, 2002).

Prior to collecting several days' worth of NMR spectra for structure determination, other biophysical methods are generally adopted to infer secondary structural information for the protein of interest. In particular, CD spectroscopy is extensively used to estimate the secondary structure content of medium-sized proteins. In CD spectroscopy, deconvolution of the experimental molar ellipticity at 222 nm is used to estimate secondary structure content. In the case of NMR, chemical shifts have been used as regular indicators of a particular secondary structure. For example, an $^1H_\alpha$ resonance that is shifted upfield with respect to the corresponding random coil value is considered to be α-helical, while one shifted downfield to be β-strand. This is a widely accepted procedure, and a large number of NMR studies have shown that such correlation is valid (Cornilescu *et al.*, 1999; Case, 2000). However, NMR spectral information has seldom been used to obtain relatively low-resolution structural information, such as secondary structure content. In some cases, the results of CD are used to determine whether it is feasible to obtain complete, three-dimensional structural information for a particular protein, using NMR. This suggests the critical importance of evaluating whether data obtained from NMR itself can be used to estimate secondary structure content. Lee and Cao have addressed this question extensively in their comprehensive study (Lee and Cao, 1996), and have shown that the correlation between NMR- and CD-based secondary structure estimation is poor. Further, while CD spectroscopy is more suitable for studying relatively small proteins and polypeptides, the characterization of larger molecules requires NMR.

Computational methods often play a primary role in initial predictions of protein structure; for example, in predictions of protein structural class. These methods are typically invoked even before a protein is expressed or extracted for any biophysical characterization. Secondary structure estimations from CD are often inconsistent with such computational predictions from NMR. On the other hand, to date, estimations from NMR have required the time-consuming process of resonance assignment. A method such as that proposed here could essentially fulfill the need for an empirical, NMR-based estimator of protein structural class that is both accurate and efficient.

Our results show that $^1H_\alpha$ ACS values clearly distinguish the three different protein classes, α, mixed αβ and β, when the proteins are classified either by CATH or SCOP. The SCOP-based distribution shows a better statistical quantification, as the total number of proteins in that case is higher. Though an intuitive difference between the various structural classes with respect to ACS values is evident (Fig. 1), we quantify the validity of the estimation using KS tests. The KS statistic, D, is defined simply as the maximum value of the absolute difference between two cumulative distribution functions, and is insensitive to the actual form of those functions. The KS

Table 3. Prediction of structural class from NMR data for proteins of undetermined three-dimensional structure

BMRB[a]	ACS $(^1\mathrm{H}_\alpha)$[b]	Protein name	Structural class (using CATH-based correlation)[c]	Structural class (using SCOP-based correlation)[d]
4664	3.818	Lipocalin Q83	α	α
4688	3.899	L18	αβ	α/αβ
4698	3.846	Transforming Growth Factor β type II receptor	α	α/αβ
4722	3.823	Shikimate Kinase	α	α
4752	3.819	Gpnu1-E68	α	α
4771	3.726	Tola3	α	NP
4791	3.808	HCV NS3 RNA helicase	α	α
4792	3.778	ParD dimer	α	α
4829	3.841	Interleukin enhancer binding factor	α	α
4834	3.766	S. aureus peptide deformylase	α	α
4908	3.769	α′-domain of ERp57	α	α
5014	3.724	MyBP-C cC5	α	NP
5040	3.778	I1(I29T) monomer	α	α
5093	3.881	RbfADelta25	αβ	α/αβ
5107	3.826	Sensor & Substrate Binding Domain from Lon (La) Protease	α	α
5316	3.781	Gag	α	α
4113	3.931	Vaccinia Glutaredoxin-1	αβ	αβ
4132	4.015	Human ubiquitin-conjugating enzyme	β	αβ/β
4719	3.922	Ras binding domain of rat AF6	αβ	αβ
4802	3.968	N-terminal domain of H-NS	αβ/β	αβ
4881	3.983	Azotobacter vinelandii C69A holoflavodoxin II	αβ	αβ
4901	3.991	p62 N-terminal domain	αβ	αβ
4940	3.933	Antennal Specific Protein 1	αβ	αβ
4965	3.925	L11	αβ	αβ
5030	3.937	Honeybee antennal specific Protein 2	αβ	αβ
5093	3.881	RbfADelta25	αβ	αβ
4302	4.010	Protein disulfide isomerase a′ domain	β	αβ/β
4720	4.066	Inhibitor-2 monomer	β	β
4870	4.094	region 4.2 of sigma70 of *Escherichia coli* RNA polymerase holoenzyme	β	β
4881	3.983	Azotobacter vinelandii C69A holoflavodoxin II	β	β
4901	3.991	p62 N-terminal domain	β	β
4913	4.046	cAMP-regulated phosphoprotein-19 monomer	β	β
4929	4.090	Tctex1 dimer	β	β
4956	4.013	YajQ from *E. coli*	β	αβ/β
4973	4.100	Saratin	β	β
4999	3.979	Nucleocapsid binding domain of the sendai virus phosphoprotein	β	β
5049	4.053	Extracellular domain of subunit 2 of the human receptor	β	β

[a] BioMagResBank (BMRB) accession number (http://www.bmrb.wisc.edu/).
[b] Averaged chemical shift (ACS) calculated for the $^1\mathrm{H}_\alpha$ nuclei.
[c] Structural class estimation based on the empirical distribution obtained by CATH classification.
[d] Structural class estimation based on the empirical distribution obtained by SCOP classification.

test is a standard tool, and provides a straightforward, reliable measure of the degree to which two unbinned distributions that are functions of one independent variable differ. For a more detailed description, see, e.g. Press, 1988.

The empirical correlation presented here provides a way to determine directly the structural classes of proteins in the absence of resonance assignments. It can be easily incorporated into any commercial or academic software package that employs manual or automated peak picking routines to reduce an HSQC spectrum into a single ACS value. We have also investigated other possible parameters to represent collectively a statistical distribution of data, such as skewness, variance, and kurtosis (Press, 1988). Though these lead to similar results, for the sake of simplicity only the average [Methods, Equation (1)] is considered. Further, ACS is expressed in the same unit as chemical shift (ppm). Instead of using the absolute chemical shift values to determine the averages, we have also explored definitions such as chemical shift index (CSI) (Wishart and Sykes, 1994), which determines the relative change in the chemical shift with respect to the corresponding random coil value. CSI may better distinguish proteins that are comprised primarily of either helices or sheets; the αβ proteins cannot be defined, because the values of α and β segments are opposite in sign, and tend to cancel each other.

Determination of the structural classes of proteins with no available experimental three-dimensional structure information (from NMR or X-ray), using $^1H_\alpha$ ACS values, provides an internal test of the reliability factor (Table 3). The structural classes of these proteins were also estimated using prediction algorithms that utilize only amino acid sequences. For many of these proteins, the sequence-based class prediction approach provided similar results for the mainly-α class, while larger differences were observed for mainly-β class proteins. However, considering the variability and confidence limits associated with such predictions (http://cubic.bioc.columbia.edu/eva/ and references therein), it is difficult to define a suitable control set for comparison. In some cases, using the sequence-based prediction method (http://www.bork.embl-heidelberg.de/SSCP/) (Eisenhaber *et al*., 1996), we have observed large variations in the estimation of sheet and helical classes for the same amino acid sequence (data not shown).

In general, the quality of structural predictions based on specific algorithms is examined either by redistribution test or jack-knife test (Chou, 1989). However, in this manuscript we have considered neither of these methods, for the following reasons. First, our method is not algorithm-based; our results are strictly the outcome of an empirical correlation between known protein structural classes and averaged chemical shifts. Second, in self-consistency tests (Chou, 1989), it is necessary to define a training set of proteins that obey a particular criterion; for example, the resolution of three-dimensional structure. Though it is possible to define such criteria for protein classes, use of chemical shift information as the test criterion must be considered premature, as there is currently no consensus definition of the 'accuracy' of such information (Wishart and Case, 2001).

Although we have shown that ACS values can be used to identify directly the structural classes of proteins, thereby providing a first, low-resolution structural estimate from experiments, critical questions still remain. For example, what is the reliability of the estimates? As the number of proteins that we add into our correlations of ACS with protein class increases, we expect the reliability of the method to improve. In the empirical correlation derived between secondary structure content and ACS values, we have determined a reliability factor of 80% when $^1H_\alpha$ nuclei are used (Sibley *et al*., 2003). Notwithstanding the limited number of proteins in the current study, and that we have defined the relative regions of ACS values demarcating the structural classes in a conservative manner, we suggest the reliability of this method is also about 80%. Though crude, this result is indirectly supported by the results of Table 3. Another factor that may affect the results is the definition of the protein classes themselves. These definitions can be highly variable (Zhang and Zhang, 1998), and differences are observed even between CATH and SCOP protocols. Though the number of such cases is small, this discrepancy cannot be ruled out as a source of diminished reliability.

Another remaining question is whether it is possible that certain amino acids bias the current estimates, since the method is based on an average of the chemical shifts. The distribution of chemical shifts for each of the amino acids found in the BMRB database suggests that no particular amino acid dominates the ACS values. In a recent paper (Sharman *et al*., 2001) rigorous statistical analyses of $^1H_\alpha$ chemical shifts are used to show that there is no correlation between amino acid type and propensity to fall within helical or sheet regions. The exact nature of the chemical shift dependence on secondary structure for a specific amino acid residue remains to be determined (Sharman *et al*., 2001; Havlin *et al*., 2001). In addition, long range and context-dependent effects on protein structural class definition are still not clearly understood (Sharman *et al*., 2001), and may also play important roles in influencing chemical shifts.

As the estimation of structural class from NMR is directly influenced by the quality of the data used, the method is most useful in cases in which the resolution of the corresponding HSQC spectrum is excellent. Experiments based on transverse relaxation optimized spectroscopy (TROSY) (Pervushin *et al*., 1997) provide an additional advantage in applicability to large proteins. From a practical point of view, the method would be most appropriate if a sufficient number of individual cross-peaks is observed in an HSQC spectrum. Further, since calculated ACS values are based on the total number of residues in a protein, and not on the total number of cross-peaks observed, we recommend a minimum of 70% of the total number of peaks expected be present in a given spectrum for determination of a reliable ACS value. As a final point, all amino acid residues have $^1H_\alpha$ resonances (glycine has two), so these will be fully represented in any calculation of the $^1H_\alpha$ ACS. In contrast, proline residues lack an amide proton resonance, and consequently are not observed in ^{15}N-HSQC spectra; an abundance of proline-rich proteins in a data set could conceivably lead to an underestimate of amide ACS values.

5 CONCLUSIONS

Progress in the structural biology of proteins comes from both experimental and theoretical efforts. Computational methods are capable of delivering fast structural information, ranging from low-resolution protein structural class definition to high-quality information based on homology modeling. Experimental methods that concentrate on obtaining high-resolution information are hampered by inherent time cost, and lack the capacity to provide low-resolution structural information expediently. NMR spectroscopy is a powerful tool for obtaining high-resolution structural and dynamical details of molecules in the solution state. In order to explore new experimental methods for the fast identification of protein structures using NMR, we have investigated the degree to which the ACS of a particular nuclear species in the protein backbone can be used as a low-resolution structural parameter that correlates with protein structural class. We have found that the differences between the ACS values characterizing various protein structural classes, though small, are in several cases statistically significant.

ACS-based methods do not provide an alternative to conventional NMR-based experiments. The former must only be considered initial predictors of protein class or secondary structure content. Nevertheless, the quality of estimates using these methods is comparable to or slightly better than that obtained using sequence-based structure prediction algorithms. ACS methods might provide a novel technique for monitoring protein structural changes in real time, such as in protein folding experiments. Such methods might also be used to detect major structural changes that occur upon protein–protein, protein–DNA/RNA, and other complex formations, to provide some direct experimental structural information in situations in which other techniques are incapable of doing so (e.g. in studies of large and/or highly disordered proteins), and to facilitate initial protein fold identification in high throughput proteomics applications.

SUPPORTING INFORMATION

A list of all the proteins, BMRB numbers, PDB codes, and calculated ACS values are available from the authors upon request.

ACKNOWLEDGEMENTS

Thanks to Drs C.J. Benham, and W.H. Fink for critical comments on the manuscript. 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 (V.V.K.).

SUPPLEMENTARY DATA

For Supplementary data, please refer to *Bioinformatics* online.

REFERENCES

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

Atkinson,R.A. and Saudek,V. (2002) The direct determination of protein structure by NMR without assignment. *FEBS Lett.*, **510**, 1–4.

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. *Nucleic Acids Res.*, **28**, 235–242.

Boberg,J., Salakoski,T. and Vihinen,M. (1995) Accurate prediction of protein secondary structural class with fuzzy structural vectors. *Protein Eng.*, **8**, 505–512.

Bussian,B.M. and Sander,C. (1989) How to determine protein secondary structure in solution by Raman spectroscopy: practical guide and test case DNase I. *Biochemistry*, **28**, 4271–4277.

Cai,Y.D., Liu,X.J., Xu Xb,X. and Zhou,G.P. (2001) Support Vector Machines for predicting protein structural class. *BMC Bioinformatics*, **2**, 3.

Carlacci,L. and Englander,S.W. (1993) The loop problem in proteins: a Monte Carlo simulated annealing approach. *Biopolymers*, **33**, 1271–1286.

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

Chi,Z., Chen,X.G., Holtz,J.S. and Asher,S.A. (1998) UV resonance Raman-selective amide vibrational enhancement: quantitative methodology for determining protein secondary structure. *Biochemistry*, **37**, 2854–2864.

Chou,K.C. and Zhang,C.T. (1995) Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.*, **30**, 275–349.

Chou,P.Y. (1989) Prediction of protein structural class from amino acid composition. In Fasman,G.D. (ed.), *Prediction of Protein Structure and Principles of Protein Conformation*, Plenum Press, New York, pp. 549–586.

Cid,H., Bunster,M., Canales,M. and Gazitua,F. (1992) Hydrophobicity and structural classes in proteins. *Protein Eng.*, **5**, 373–375.

Cohen,F.E. and Kuntz,I.D. (1987) Origins of structural diversity within sequentially identical hexapeptides. *Proteins*, **2**, 162–166.

Cohen,B.I., Presnell,S.R. and Cohen,F.E. (1993) Prediction of the three-dimensional structure of human growth hormone. *Protein Sci.*, **2**, 2134–2145.

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.

Deleage,G. and Dixon,J.S. (1989) Use of class prediction to improve protein secondary structure prediction. In Fasman,G.D. (ed.), *Prediction of Protein Structure and Principles of Protein Conformation*, Plenum Press, New York, pp. 587–597.

Deleage,G. and Roux,B. (1987) An algorithm for protein secondary structure prediction based on class prediction. *Protein Eng.*, **1**, 289–294.

Deleage,G., Tinland,B. and Roux,B. (1987) A computerized version of the Chou and Fasman method for predicting the secondary structure of proteins. *Anal. Biochem.*, **163**, 292–297.

Dietmann,S. and Holm,L. (2001) Identification of homology in protein structure classification. *Nat. Struct. Biol.*, **8**, 953–957.

Eisenhaber,F., Imperiale,F., Argos,P. and Frommel,C. (1996) Prediction of Secondary Structural Content of Proteins From Their Amino Acid Composition Alone .1. New Analytic Vector Decomposition Methods. *Proteins—Structure Function and Genetics*, **25**, 157–168.

Ernst,R.R., Bodenhausen,G. and Wokaun,A. (1990) *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*, The International series of monographs on chemistry; 14, Clarendon Press, Oxford.

Grishaev,A. and Llinas,M. (2002) Protein structure elucidation from NMR proton densities. *Proc. Natl Acad. Sci. USA*, **99**, 6713–6718.

Gutowsky,H.S., Saika,A., Takeda,M. and Woessner,D.E. (1957) Proton magnetic resonance studies on natural rubber. II. Line shape and T1 measurements. *J. Chem. Phys.*, **27**, 534–542.

Havlin,R.H., Laws,D.D., Bitter,H.M. L., Sanders,L.K., Sun,H. H., Grimley,J.S., Wemmer,D.E., Pines,A. and Oldfield,E. (2001) An experimental and theoretical investigation of the chemical shielding tensors of C-13(alpha) of alanine, valine, and leucine residues in solid peptides and in proteins in solution. *J. Am. Chem. Soc.*, **123**, 10362–10369.

Ikegami,T., Okada,T., Ohki,I., Hirayama,J., Mizuno,T. and Shirakawa,M. (2001) Solution structure and dynamic character of the histidine-containing phosphotransfer domain of anaerobic sensor kinase ArcB from Escherichia coli. *Biochemistry*, **40**, 375–386.

Johnson,W.C., Jr. (1990) Protein secondary structure and circular dichroism: a practical guide. *Proteins*, **7**, 205–214.

Klein,P. and Delisi,C. (1986) Prediction of protein structural class by discriminant analysis. *Biopolymers*, **25**, 1659–1672.

Klein,P. (1986) Prediction of protein structural class from the amino acid sequence. *Biochim. Biophys. Acta.*, **874**, 205–215.

Kneller,D.G., Cohen,F.E. and Langridge,R. (1990) Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.*, **214**, 171–182.

Koradi,R., Billeter,M., Engeli,M., Guntert,P. and Wuthrich,K. (1998) Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *J. Mag. Reson.*, **135**, 288–297.

Lee,M.S. and Cao,B. (1996) Nuclear magnetic resonance chemical shift: comparison of estimated secondary structures in peptides by nuclear magnetic resonance and circular dichroism. *Protein Eng.*, **9**, 15–25.

Levitt,M. and Chothia,C. (1976) Structural patterns in globular proteins. *Nature*, **261**, 552–558.

Li,Q.Z. and Lu,Z.Q. (2001) The prediction of the structural class of protein: application of the measure of diversity. *J. Theor. Biol.*, **213**, 493–502.

Lo Conte,L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.

Luo,R.Y., Feng,Z.P. and Liu,J.K. (2002) Prediction of protein structural class by amino acid and polypeptide composition. *Eur. J. Biochem.*, **269**, 4219–4225.

Metfessel,B.A., Saurugger,P.N., Connelly,D.P. and Rich,S.S. (1993) Cross-validation of protein structural class prediction using statistical clustering and neural networks. *Protein Sci.*, **2**, 1171–1182.

Moseley,H.N.B. and Montelione,G.T. (1999) Automated analysis of NMR assignments and structures for proteins. *Curr. Opin. Struct. Biol.*, **9**, 635–642.

Muggleton,S., King,R.D. and Sternberg,M.J. (1992) Protein secondary structure prediction using logic-based machine learning. *Protein Eng.*, **5**, 647–657.

Nakashima,H., Nishikawa,K. and Ooi,T. (1986) The folding type of a protein is relevant to the amino acid composition. *J. Biochem. (Tokyo)*, **99**, 153–162.

Nishikawa,K. and Ooi,T. (1982) Correlation of the amino acid composition of a protein to its structural and biological characters. *J. Biochem. (Tokyo)*, **91**, 1821–1824.

Nishikawa,K. and Ooi,T. (1986a) Amino acid sequence homology applied to the prediction of protein secondary structures, and joint prediction with existing methods. *Biochim. Biophys. Acta.*, **871**, 45–54.

Nishikawa,K. and Ooi,T. (1986b) Radial locations of amino acid residues in a globular protein: correlation with the sequence. *J. Biochem. (Tokyo)*, **100**, 1043–1047.

Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J. (1997) CATH – a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Perczel,A., Hollosi,M., Tusnady,G. and Fasman,G.D. (1991) Convex constraint analysis: a natural deconvolution of circular dichroism curves of proteins. *Protein Eng.*, **4**, 669–679.

Pervushin,K., Riek,R., Wider,G. and Wuthrich,K. (1997) Attenuated T-2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc. Natl Acad. Sci. USA*, **94**, 12366–12371.

Press,W.H. (1988) *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, Cambridge, New York.

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.

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. Am. 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**, 1223–1227.

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**, 325–443.

Taylor,W.R. (2002) A ‘periodic table’ for protein structures. *Nature*, **416**, 657–660.

Wang,H., He,Y., Hsu,K.T., Magliocca,J.F., Storch,J. and Stark,R.E. (1998) 1H, 15N and 13C resonance assignments and secondary structure of apo liver fatty acid-binding protein. *J. Biomol. NMR*, **12**, 197–199.

Wang,Z.X. (2001) The prediction accuracy for protein structural class by the component- coupled method is around 60%. *Proteins*, **43**, 339–340.

Wang,Z.X. and Yuan,Z. (2000) How good is prediction of protein structural class by the component- coupled method? *Proteins*, **38**, 165–175.

Williams,R.W., McIntyre,J.O., Gaber,B.P. and Fleischer,S. (1986) The secondary structure of calcium pump protein in light sarcoplasmic reticulum and reconstituted in a single lipid component as determined by Raman spectroscopy. *J. Biol. Chem.*, **261**, 14520–14524.

Wishart,D.S., Bigam,C.G., Yao,J., Abildgaard,F., Dyson,H.J., Oldfield,E., Markley,J.L. and Sykes,B.D. (1995) H-1, C-13 and N-15 Chemical Shift Referencing in Biomolecular NMR. *J. Biomol. NMR*, **6**, 135–140.

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

Wishart,D.S. and Sykes,B.D. (1994) 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.

Wüthrich,K. (1986) *NMR of Proteins and Nucleic Acids*. The George Fisher Baker non-resident lectureship in chemistry at Cornell University, Wiley, New York.

Zhang,C.T. and Chou,K.C. (1992) An optimization approach to predicting protein structural class from amino acid composition. *Protein Sci.*, **1**, 401–408.

Zhang,C.T. and Zhang,R. (1998) A new criterion to classify globular proteins based on their secondary structure contents. *Bioinformatics*, **14**, 857–865.

Zhou,G., Xu,X. and Zhang,C.T. (1992) A weighting method for predicting protein structural class from amino acid composition. *Eur. J. Biochem.*, **210**, 747–749.

Zhou,G.P. (1998) An intriguing controversy over protein structural class prediction. *J. Protein Chem.*, **17**, 729–738.