J Biomol NMR (2007) 38:289–302
DOI 10.1007/s10858-007-9166-6

ARTICLE

# Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology

**Yang Shen · Ad Bax**

Received: 19 April 2007 / Accepted: 16 May 2007 / Published online: 4 July 2007


**Abstract** Chemical shifts of nuclei in or attached to a protein backbone are exquisitely sensitive to their local environment. A computer program, SPARTA, is described that uses this correlation with local structure to predict protein backbone chemical shifts, given an input three-dimensional structure, by searching a newly generated database for triplets of adjacent residues that provide the best match in $\phi/\psi/\chi^1$ torsion angles and sequence similarity to the query triplet of interest. The database contains $^{15}N$, $^1H^N$, $^1H^\alpha$, $^{13}C^\alpha$, $^{13}C^\beta$ and $^{13}C'$ chemical shifts for 200 proteins for which a high resolution X-ray ( $\leq 2.4$ Å) structure is available. The relative importance of the weighting factors for the $\phi/\psi/\chi^1$ angles and sequence similarity was optimized empirically. The weighted, average secondary shifts of the central residues in the 20 best-matching triplets, after inclusion of nearest neighbor, ring current, and hydrogen bonding effects, are used to predict chemical shifts for the protein of known structure. Validation shows good agreement between the SPARTA-predicted and experimental shifts, with standard deviations of 2.52, 0.51, 0.27, 0.98, 1.07 and 1.08 ppm for $^{15}N$, $^1H^N$, $^1H^\alpha$, $^{13}C^\alpha$, $^{13}C^\beta$ and $^{13}C'$, respectively, including outliers.

**Keywords** Backbone · Chemical shift · Database · Empirical · Prediction · TALOS

**Electronic supplementary material** The online version of this article (doi:10.1007/s10858-007-9166-6) contains supplementary material, which is available to authorized users.

Y. Shen (✉) · A. Bax (✉)
Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892-0520, USA
e-mail: shenyang@niddk.nih.gov

A. Bax
e-mail: bax@nih.gov

## Introduction

Chemical shifts have long been recognized as an important source of structural information for proteins. However, their dependencies on multiple factors, including backbone and sidechain torsion angles, electric fields, ring currents, hydrogen bonding, and local strain have thwarted attempts to separately quantify the relation between each of those parameters and chemical shift. For protons, earlier studies show clear correlations between $^1H^\alpha$ chemical shift and secondary structure (Pastore and Saudek 1990; Williamson 1990; Wishart et al. 1991; Ösapay and Case 1994; Wishart and Sykes 1994; Szilágyi 1995), and between $^1H^N$ chemical shift and both hydrogen bonding and secondary structure (Pardi et al. 1983; Wagner et al. 1983; Williamson 1990; Wishart et al. 1991). For $^{13}C$, the secondary $^{13}C^\alpha$ and $^{13}C^\beta$ chemical shifts depend most strongly on the intraresidue backbone torsion angles $\phi$ and $\psi$ (Ando et al. 1984; Saitô 1986; Spera and Bax 1991; Wishart et al. 1991; de Dios et al. 1993; Iwadate et al. 1999; Wishart and Case 2002; Neal et al. 2003), while the secondary $^{15}N$ chemical shifts correlate with the $\psi_{i-1}/\phi_i$ torsion angles (Glushka et al. 1989; de Dios et al. 1993; Le and Oldfield 1994; Wishart and Case 2002; Neal et al. 2003; Wang and Jardetzky 2004). The effects of various other factors on protein backbone chemical shifts, such as side chain $\chi^1$ angles and neighboring residue type, have also been investigated (de Dios et al. 1993; Wang and Jardetzky 2002; Wishart and Case 2002; Neal et al. 2003; Wang and Jardetzky 2004; Villegas et al. 2007).

Currently there are multiple approaches for predicting chemical shifts for a given protein structure, including those based on (1) ab initio quantum mechanical (QM) calculations (de Dios et al. 1993; Xu and Case 2001; 2002), (2) empirical $\Delta(\phi,\psi)$ shielding surface analysis

(Spera and Bax 1991; Le and Oldfield 1994; Beger and Bolton 1997; Wishart and Nip 1998; Iwadate et al. 1999; Wang and Jardetzky 2004), (3) secondary structure and hydrogen bonding (Wagner et al. 1983; Ösapay and Case 1991; Wishart et al. 1991; Herranz et al. 1992; Ösapay and Case 1994; Williamson et al. 1995), (4) sequence homology (Gronwald et al. 1997; Wishart et al. 1997), and (5) artificial neural networks (Meiler 2003). All of these are capable of predicting protein chemical shifts with reasonable accuracy and have their individual strengths and weaknesses. For example, the empirical approaches are relative fast and can cover chemical shifts over a wide range, but with modest accuracy, whereas the QM approach potentially offers relatively high accuracy but can require extreme computation times and is very sensitive to assumptions about precise local geometry. A relatively recent, hybrid predictive method, SHIFTX, combines the empirical hypersurface approach with a classical analysis in terms of hydrogen bonding and secondary structure, and appears to yield the best compromise between prediction accuracy, speed, and completeness (Neal et al. 2003).

It has been well recognized that homologous proteins show quite similar patterns of secondary chemical shifts (Redfield and Dobson 1990). This similarity has been utilized during resonance assignment and for chemical shift prediction process of proteins with a minimum of ~30% sequence identity (Bartels et al. 1996; Gronwald et al. 1997; Wishart et al. 1997). Cornilescu et al. (1999) developed a database searching program, TALOS, which utilizes the inverse of this relation to extract structural information for proteins with known chemical shift assignments. TALOS searches a pre-defined database for triplets of adjacent residues that have the closest similarity in backbone secondary chemical shifts ($^{15}$N, $^{1}$H$^{\alpha}$, $^{13}$C$^{\alpha}$, $^{13}$C$^{\beta}$ and $^{13}$C′) and amino acid sequence to those of the query triplet. Backbone $\phi$ and $\psi$ angular restraints for the central residue of the query triplet are then derived from the central residues of the best-matched triplets, provided they exhibit consensus on the values of the $\phi$ and $\psi$ angles.

In the present study, we describe a TALOS-like database searching procedure, which utilizes both protein sequence and structural homology, to predict the backbone $^{15}$N, $^{1}$H$^{N}$, $^{1}$H$^{\alpha}$, $^{13}$C$^{\alpha}$, $^{13}$C$^{\beta}$ and $^{13}$C′ chemical shifts for a protein of known structure. This approach not only returns the predicted chemical shifts, but also the individual standard deviations observed for the best fitting fragments. These standard deviations are shown to correlate with the chemical shift prediction error, and therefore provide an important additional parameter when using the predicted chemical shifts for a wide range of potential purposes. The program, named SPARTA (*S*hift *P*rediction from *A*nalogy in *R*esidue type and *T*orsion *A*ngle), searches an expanded, TALOS-like database for triplets of adjacent residues that are most similar to the query triplet in terms of structure ($\phi$, $\psi$, $\chi^1$) and amino acid sequence. With the rapid database growth of proteins that have both accurate chemical shift assignments and high resolution 3D structures, the continuing increase in the number of reference proteins available to SPARTA is expected to yield further improvements in its performance with time. Even small improvements in the accuracy of predicted chemical shifts can be important, in particular when using molecular fragment replacement (MFR) searches (Kontaxis et al. 2005) of the protein structure database (RCSB) (Berman et al. 2000) where typically the chemical shifts and other parameters of 5–10 residue fragments are used for the search. For example, a 10% improvement in chemical shift prediction accuracy narrows the search over 35 chemical shifts of a 7-residue fragment by a factor $(1/0.9)^{35} \approx 40$.

## Methods

### Database

A database was created which contains nearly complete $^{15}$N, $^{1}$H$^{N}$, $^{1}$H$^{\alpha}$, $^{13}$C$^{\alpha}$, $^{13}$C$^{\beta}$ and $^{13}$C′ chemical shift assignments (22,952 $^{15}$N shifts, 20,369 $^{1}$H$^{N}$, 16,959 $^{1}$H$^{\alpha}$, 24,021 $^{13}$C$^{\alpha}$, 21,401 $^{13}$C$^{\beta}$ and 19,803 $^{13}$C′) of 200 proteins (Supplementary Material Table 1), together with the backbone $\phi$, $\psi$ and sidechain $\chi^1$ angles. The experimentally observed chemical shifts are adjusted by subtracting the calculated ring current shift contribution and the effect of nearest neighbor residue type (see below). The structural information of these 200 proteins (24,166 residues) is derived from their X-ray crystal coordinates, all available in the RCSB protein structure database (Berman et al. 2000) at a resolution $\leq 2.4$ Å, whereas nearly complete chemical shift assignments from the BioMagResBank (BMRB) (Jurgen et al. 2005) were selected and processed using the same criteria as previously used for the TALOS database (Cornilescu et al. 1999). For Gly residues, the averaged $^{1}$H$^{\alpha 2}$ and $^{1}$H$^{\alpha 3}$ shifts are used for $^{1}$H$^{\alpha}$ shifts in the database. The chemical shifts in the BMRB database were converted to secondary chemical shifts by subtracting their corresponding random coil chemical shifts values (Supplementary Material Tables 4 and 5) and the adjustments values arising from the effects of neighboring residues (Supplementary Material Tables 6 and 7), using mainly the values of the TALOS program. A re-optimization of the effect of nearest neighbors over the 200-protein database did not yield any significant improvement over the values previously derived for TALOS (Supplementary Material) and shows that the impact of these effects essentially is restricted to the $^{15}$N shifts. So, even though, for example, the presence of a Pro at position *i* decreases the

experimentally observed random coil $^{13}C^{\alpha}$ chemical shift of residue $i$–1 by nearly 2 ppm, no nearest residue effect is observed during our optimization. This result is attributed to a non-random coil distribution of the $\phi$, $\psi$ and $\chi^1$ angles of residue $i$ in the experimental study of random coil peptides, and therefore constitutes an effect already fully accounted for in terms of the $\phi$, $\psi$ and $\chi^1$ angle information used by SPARTA.

Hydrogen atoms were added to the X-ray coordinates using the program DYNAMO (Kontaxis et al. 2005) and secondary structure in the proteins was determined by the program STRIDE (Frishman and Argos 1995). Identification of hydrogen bonding interactions for backbone carbonyl oxygen, amide $^1H^N$ and $^1H^{\alpha}$ atoms was made using the Kabsch and Sander criteria (Kabsch and Sander 1983), using an electrostatic interaction energy of <–0.5 kcal/mol between two hydrogen bonding groups as a cut-off. The oxidation state of Cys residues is obtained by inspection of $C^{\beta}$ chemical shifts and the existence of an S–S bond in the X-ray structure (i.e., distance <2.5 Å between $S^{\gamma}$ atoms of two Cys residues that are at least 4 residues apart in the protein sequence).

Chemical shifts can be affected significantly by ring currents of nearby aromatic groups. In particular for $^1H^{\alpha}$ shifts, which in the absence of ring current shifts cluster between 5.5 ppm and 3.5 ppm, ring current contributions can extend this range considerably. To account for the effects of ring currents on the experimental chemical shifts, the ring current shifts were calculated for backbone $^{15}N$, $^1H^N$, $^1H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ nuclei using the Haigh–Mallion model (Haigh and Mallion 1979; Case 1995) on the basis of the protein X-ray coordinates. The calculated ring current contributions then were subtracted from the experimental chemical shifts in the database, using appropriate scaling that has been optimized for the best chemical shift prediction performance (see Results and Discussion), in order to obtain a ring-current-free database. Below, the ring-current-corrected secondary chemical shifts are referred as the (experimental) secondary chemical shifts, unless specified otherwise.

## Database search procedure

SPARTA is written in the standard C++ language. For a medium sized protein of 100 residues, SPARTA chemical shift prediction takes 25 s on a Linux computer with a 2.8 GHz CPU. A schematic view of the SPARTA prediction method is presented in Fig. 1.

SPARTA requires as input a standard RCSB coordinate file, and extracts from these the backbone $\phi$, $\psi$ and side-chain $\chi^1$ torsion angles as well as the hydrogen bond lengths for $^1H^N$ and $^1H^{\alpha}$ atoms if H-bonded, and it also calculates the ring current shifts for all backbone $^{15}N$, $^1H^N$, $^1H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ nuclei. For this query protein, SPARTA then evaluates the similarity in both amino acid type and $\phi/\psi/\chi^1$ torsion angles for each string of three sequential residues relative to all triplets of sequential residues contained in the database, and it retains the best 20 matches. For a nucleus of type, $r$, the similarity score, $S(i,j,r)$, between a triplet centered at $i$ in the query protein and a triplet centered at residue $j$ in the database is given by:

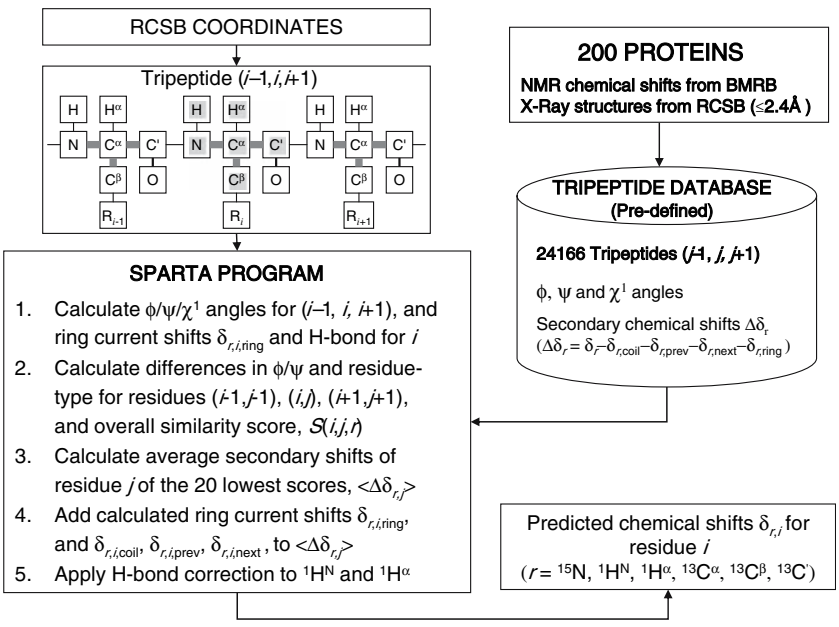


**Fig. 1** Schematic representation of the SPARTA program

$$S(i,j,r) = \sum_{n=-1}^{+1} \left[ k^{H}_{n,r} \times \Delta^{2}_{\mathrm{ResType}} + k^{\phi}_{n,r} \times (\phi_{i+n} - \phi_{j+n})^{2} + k^{\psi}_{n,r} \times (\psi_{i+n} - \psi_{j+n})^{2} + k^{\chi^1}_{n,r} \times \Delta\chi^{1}_{i+n,j+n} \right] \quad (1)$$

where $k^{H}_{n,r}$ denotes the weighting factors for residue type similarity, $k^{\phi}_{n,r}$, $k^{\psi}_{n,r}$ and $k^{\chi^1}_{n,r}$ denote the weighting factors for $\phi$, $\psi$ and $\chi^1$ angle similarities, and $\Delta_{\mathrm{ResType}}$ represents the residue-type similarity (Supplementary Table S3). Because each factor in Eq. (1) (e.g., neighboring residue effect) is of markedly different importance when predicting chemical shifts of different types of nuclei (Wishart and Case 2002; Neal et al. 2003), the weighting factors in Eq. (1) have been optimized separately for each of the six types of nuclei. Values for the weight factors, $k^{H}_{n,r}$, $k^{\phi}_{n,r}$, $k^{\psi}_{n,r}$ and $k^{\chi^1}_{n,r}$, optimized by grid searching, are presented in Table 1.

The above weight factors are optimized such that the chemical shifts of a protein that has been omitted from the database yields best agreement with the chemical shifts of the best 20 matches found in the remaining 199 proteins, a procedure repeated 200 times, such that each protein once serves as query protein. Note that the database triplet information includes the torsion angles, residue types, and secondary chemical shift information from which ring current effects and neighboring-residue-type effects have been removed. So, after finding the best 20 matches, ring current contributions need to be added to these, as do neighboring residue effects. In addition, as described in the Results and Discussion section, hydrogen bonding contributes considerably to $^{1}$H chemical shifts, and its effect can be included in chemical shift prediction since the structure of the query protein is known. Analogous to what is used by the TALOS program, the residue-type similarity, $\Delta_{\mathrm{ResType}}$, is derived from a 20 × 20 residue-type similarity matrix (Supplementary Material Table 3), which also has been iteratively adjusted during the optimization.

**Table 1** Empirical weighting factors used when deriving the similarity score of Eq. (1)

| Nucleus | $k^{H}_{-1}$ | $k^{H}_{0}$ | $k^{H}_{1}$ | $k^{\phi}_{-1}$ | $k^{\phi}_{0}$ | $k^{\phi}_{1}$ | $k^{\psi}_{-1}$ | $k^{\psi}_{0}$ | $k^{\psi}_{1}$ | $k^{\chi^1}_{-1}$ | $k^{\chi^1}_{1}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $^{15}$N | 9 | 16 | 1 | 4 | 16 | 1 | 32 | 4 | 1 | 1 | 0 |
| $^{1}$H$^{N}$ | 4 | 16 | 1 | 9 | 32 | 4 | 16 | 4 | 1 | 0 | 0 |
| $^{1}$H$^{\alpha}$ | 3 | 16 | 3 | 1 | 64 | 3 | 4 | 32 | 1 | 0 | 0 |
| $^{13}$C$^{\alpha}$ | 9 | 96 | 12 | 1 | 96 | 6 | 1 | 96 | 1 | 0 | 0 |
| $^{13}$C$^{\beta}$ | 1 | 32 | 3 | 1 | 96 | 6 | 1 | 96 | 1 | 0 | 0 |
| $^{13}$C′ | 1 | 16 | 4 | 1 | 32 | 32 | 1 | 32 | 16 | 0 | 1 |

$k^{H}_{n}$ ($n$ = –1, 0, 1): weighting factors for residue-type homology of preceding, current and next residue

$k^{\phi/\psi^1}_{n}$ ($n$ = –1, 0, 1): weighting factors for $\phi/\psi$ angles of first, center, and last residue of each triplet

$k^{\chi^1}_{n}$ ($n$ = –1, 1): options for $\chi^1$ angle filtering procedure of first and last residue of each triplet (1: on; 0: off)

It is well recognized that sidechain $\chi^1$ angles can impact the intraresidue backbone chemical shifts. In addition, $^{15}$N and $^{1}$H$^{N}$ chemical shifts are affected by the $\chi^1$ angle of the preceding residue (Le and Oldfield 1996; Wishart and Case 2002; Neal et al. 2003). However, considering the uncertainties in the $\chi^1$ angles obtained from X-ray crystal structures relative to the far more accurate backbone $\phi$ and $\psi$ angles, it proved not useful to incorporate the effect of small differences in $\chi^1$ angles in the same way as is done for $\phi$ and $\psi$. Instead, all residues in the database are grouped according to their staggered rotamers: *gauche*$^{+}$ ($\chi^1$ = –60° ± 60°), *trans* (180° ± 60°) and *gauche*$^{-}$ (60° ± 60°), and the above database search procedure is performed in a manner where only the triplets with the same $\chi^1$ conformation (*gauche*$^{+}$, *trans*, or *gauche*$^{-}$) for the center residue $j$ as that of center residue $i$ of the query triplet are selected from the database. To this extent, $\Delta\chi^{1}_{i+n,j+n}$ in Eq. (1) equals zero if the rotameric states are the same, and infinity otherwise. This ''$\chi^1$-rotamer filtering'' for the preceding and following residue is also found to be important for chemical shift prediction of $^{15}$N and $^{13}$C′ spins, respectively, which are referred as ''$\chi^{1}_{-1}$-rotamer filtering'' ($k^{\chi^1}_{-1}$) and ''$\chi^{1}_{1}$-rotamer filtering'' ($k^{\chi^1}_{1}$) options, and included in Table 1 along with other weighting factors.

As described above, for each query triplet, the similarity score $S(i,j,r)$ is calculated for all triplets in the NMR database and the 20 best matched triplets, based on their $S(i,j,r)$ scores, are retained. Note that, in general, the selection of database triplets will be different for the various types of nuclei, $r$, whose chemical shifts are being predicted. The averaged value of the secondary chemical shifts of the central residue, weighted by the inverse of $S(i,j,r)$, is calculated over this set of 20 triplets, and used as the raw predicted secondary chemical shift, $\Delta\delta(r)$, for residue $i$. The adjusted predicted chemical shift $\delta(r)$ for residue $i$ is then obtained by adding the random coil chemical shift $\delta^{\mathrm{rc}}(r)$ of residue $i$ (Supplementary Table S4 and S5), using adjustment values to account for neighboring residues $i$ – 1 and $i$ + 1 (Supplementary Tables S6 and S7) and scaled ring currents shifts calculated from the 3D coordinates, as well as the above mentioned hydrogen bonding contribution (see Results and Discussion). In the SPARTA output, for each query nucleus the program reports, besides the predicted chemical shifts and estimated errors (see Results and discussion), also the protein name and sequence position from which each of the database triplets originates, $\Delta\delta(r)$ of the center residue, and the similarity value, $S(i,j,r)$.

## Removal of abnormal chemical shifts

Outlier chemical shifts in the database, which may be real but also include erroneous chemical shift assignments and

typographical errors, can have a disproportionate effect when calculating averages of chemical shifts for the selected triplets, adversely affecting the overall SPARTA chemical shifts predictions. Most $^{1}H^{\alpha}$ chemical shift outliers result from large ring current contributions from nearby aromatic rings (Wishart and Case 2002), which typically cannot be calculated at sufficient accuracy from the limited precision of the X-ray atomic coordinates, and therefore cannot be completely removed from the database values. On the other hand, for $^{15}N$ and $^{13}C'$ chemical shifts, most outliers appear to result from aliasing errors. To minimize the influence of these outliers, chemical shifts that deviate by more than five standard deviations from their predicted values were removed from the database. About 300 such chemical shifts were identified. Considering that for our 2.4-Å cut-off database the atomic precision of the X-ray coordinates of aromatic rings and the absence of quantitative information on their internal dynamics limits the accuracy at which ring current shift corrections can be calculated, errors in these calculated ring currents increase with the size of the ring current effect. Therefore, $^{1}H^{\alpha}$ shift outliers that have larger than 1.5 ppm calculated ring current contributions and deviate from the SPARTA-predicted $^{1}H^{\alpha}$ shifts by more than three standard deviations, are also removed from the original database.

### Correction for chemical shift referencing and deuterium isotope shifts

Even today, despite explicit IUPAC guidelines for chemical shift referencing (Markley et al. 1998), many of the deposited BMRB chemical shifts have systematic uniform offsets relative to the recommended chemical shift standards, all based on the methyl signal of internal 2,2-dimethylsilapentane-5-sulfonic acid or DSS. Various procedures have been proposed to identify such referencing problems (Cornilescu et al. 1999; Zhang et al. 2003; Wang et al. 2005; Wang and Wishart 2005). It is important to take such offset problems into consideration because, after the 20 best-matching triplet fragments have been selected, the effect of referencing errors would be the same as that of random chemical shift errors. In our study, for a given protein in the database, a chemical shift reference correction was applied to the $^{15}N$, $^{13}C'$, $^{13}C^{\alpha}(^{13}C^{\beta})$ and $^{1}H^{\alpha}(^{1}H^{N})$ shifts, respectively, if the average error between the experimental and SPARTA predicted chemical shifts is larger than a tolerance of 0.6, 0.3, 0.2, and 0.08 ppm. The same reference corrections are used for all $^{1}H^{\alpha}$ and $^{1}H^{N}$ shifts in a given protein, and the same applies for $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ shifts in fully protonated proteins. For deuterated proteins, the deuterium isotope shifts are significant for $^{15}N$ and $^{13}C$ spins. In particular, perdeuteration affects backbone $^{15}N$, $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ shifts on the order of –0.3, –0.5 to –0.7 and –0.6 to –1.0 ppm, respectively (Venters et al. 1996; Gardner et al. 1997; Gardner and Kay 1998; Neal et al. 2003; Moseley et al. 2004). In SPARTA, the assumption of a uniform isotope shift for $^{15}N$ as well as for $^{13}C^{\alpha}$ and for $^{13}C^{\beta}$ is used. The isotope shift correction then is applied automatically during the above described shift referencing correction, but for deuterated proteins SPARTA no longer requires the correction for $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ nuclei to be identical, thereby allowing for the, on average, somewhat larger isotope effect on $^{13}C^{\beta}$.

## Results and discussion

It is well recognized that chemical shifts in proteins are highly sensitive to local conformation, and can be interpreted in terms of a summation of various contributions, including backbone and side chain torsion angles, hydrogen bonding, ring currents, electric fields, bond angle distortion, steric clashing, etc. Quantum chemical calculations suggest that local strain, as manifested in bond angle distortions, also can play an important role. In practice, however, the resolution at which the structure is known is generally insufficient to quantify such bond angle distortions. It is also important to note that although the above mentioned factors impact all $^{1}H$, $^{13}C$ and $^{15}N$ chemical shifts, their relative contributions vary widely with the type of nucleus (Wishart and Case 2002; Neal et al. 2003). In this work, we only consider the contributions from backbone $\phi$ and $\psi$ angles, side chain $\chi^{1}$ angles, hydrogen bonding, and ring current shifts. The contribution of each of these four factors on $^{15}N$, $^{1}H^{N}$, $^{1}H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ chemical shifts is taken into account during SPARTA shift prediction. The performance of SPARTA has been optimized by iterative adjustment of weight factors that yield the best prediction in terms or root-mean-square (RMS) deviation, averaged over all the database proteins. The final RMS deviations between the experimental and SPARTA-predicted secondary chemical shifts are shown in Fig. 2. With the exception of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ shifts in disulfide-linked Cys residues, values are very similar for the different residue types and do not vary significantly with secondary structure (Supplementary Figures S2 and S3). SPARTA has also been evaluated for nine proteins for which no chemical shift lists were originally found in the BMRB and which were not used during the optimization process, and showed very similar chemical shift prediction accuracy.

### Accuracy of SPARTA chemical shift predictions

*$^{15}N$ and $^{1}H^{N}$ shifts.* Backbone $^{15}N$ shifts of proteins are known to be quite sensitive to the preceding residue type

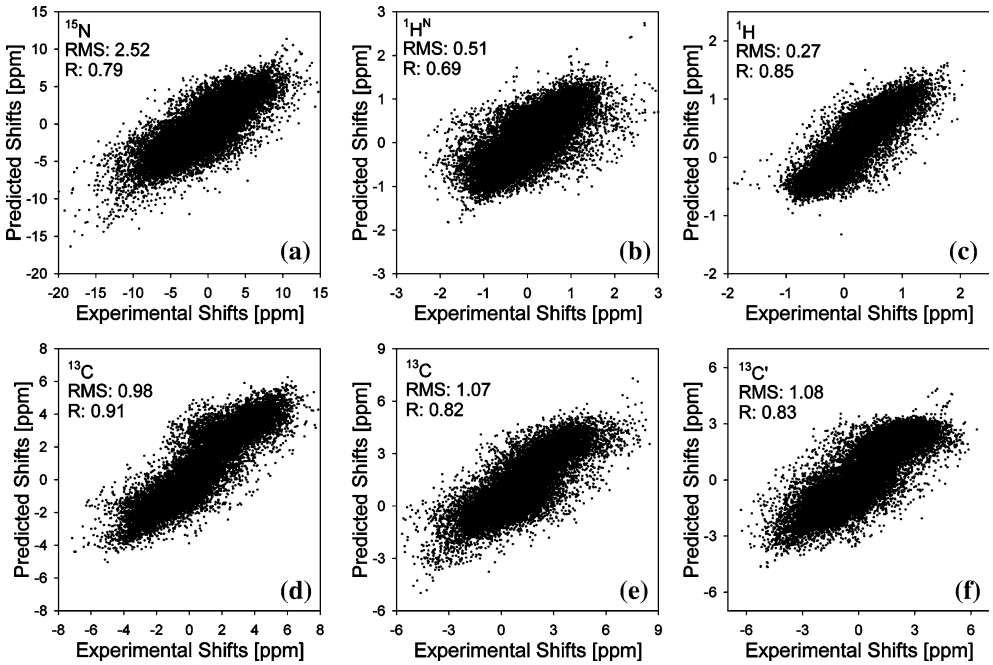


**Fig. 2** Scatter plots comparing experimental and SPARTA-predicted secondary chemical shifts for backbone $^{15}N$, $^{1}H^{N}$, $^{1}H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ nuclei. The RMS deviations (in ppm) and Pearson correlation coefficients (R) between experimental and SPARTA-predicted shifts are indicated. For $^{1}H^{N}$ and $^{1}H^{\alpha}$, the SPARTA-predicted shifts include hydrogen bond corrections for atoms that are engaged in intramolecular hydrogen bonds

(Glushka et al. 1989; de Dios et al. 1993; Le and Oldfield 1994; Wishart and Case 2002; Neal et al. 2003; Wang and Jardetzky 2004). Indeed, our empirical optimization confirms that $^{15}N$ shifts are significantly affected by the amino acid type as well as by the hydrogen bonding interaction of its $^{1}H^{N}$ and the $\psi_{i-1}/\phi$ backbone torsion angles that bracket the peptide bond. In addition, the residue type and $\chi^1$ rotameric state of both residues $i$ and $i-1$ are important (Table 1). In our database, $^{15}N$ shifts are found to have an average upfield shift of –2.4 ± 3.1 ppm in $\alpha$-helices (8519 residues) and a downfield shift of 1.0 ± 4.3 ppm in $\beta$-sheets (5,686 residues) (Supplementary Figure S1). Averaged over the entire database, the RMS deviation between experimental and SPARTA predicted $^{15}N$ chemical shifts equals 2.52 ppm. If one were to exclude predicted shifts that are in error by more than three standard deviations, as is often done in the evaluation of analogous programs, the RMS error decreases to 2.36 ppm.

As applies for $^{15}N$, $^{1}H^{N}$ shifts are also affected significantly by the $\phi/\psi_{i-1}$ backbone torsion angles and by the preceding residue type. But unlike $^{15}N$ shifts, $^{1}H^{N}$ shifts are not found to be sensitive to the side chain conformation of the preceding residue (Table 1) and, as has long been known, the effect of hydrogen bonding on $^{1}H^{N}$ shifts is quite significant. An $r^{-3}$ dependence on hydrogen bond length was found to be most consistent with experimental chemical shifts (Pardi et al. 1983; Wagner et al. 1983; Wishart et al. 1991). In the present work, this dependence is evaluated by using the much larger $^{1}H^{N}$ shifts dataset from the SPARTA database, which contains chemical shifts for 20,369 $^{1}H^{N}$ atoms, of which 14,789 are engaged in intramolecular hydrogen bonds with lengths in the 1.6–2.8 Å range. Due to the absence of solvent molecules in most of the RCSB entries, hydrogen bonds with the solvent are not considered here. When including all intramolecular H-bonded amide protons to oxygen in the SPARTA database, an optimal fit relative to $r_{H..O}^{-3}$ yields (Fig. 3a and b):

$$\Delta\delta(^{1}H^{N}) = 13.26 \times r_{H..O}^{-3} - 1.39 \quad \text{ppm} \qquad (2)$$

with a Pearson's correlation coefficient of $R = 0.58$. Interestingly, in part this $r_{H..O}^{-3}$ dependency on hydrogen bond length is already reflected in the raw SPARTA-predicted $^{1}H^{N}$ secondary chemical shifts, which do not use H-bond input information (Fig. 3b). Since the SPARTA-predicted shifts are obtained from a database search based on the local $\phi/\psi/\chi^1$ angles and residue type, the correlation between the predicted $^{1}H^{N}$ secondary chemical shifts and the hydrogen bond length suggests that the hydrogen bond length is correlated with the local torsion angles. Indeed, upon inspection of all 1,197 proteins from the RCSB solved at an X-ray resolution $\leq 1.6$ Å, clear correlations with the adjacent torsion angles are observed, even when considering simple one-dimensional relations (Fig. 4).

For residues at the protein surface, chemical shifts are also affected by the solvent (Avbelj et al. 2004). Therefore, the hydrogen-bonded $^{1}H^{N}$ atoms in the SPARTA database are further grouped according to their residual solvent exposure. We calculate the solvent exposure of residue X as the ratio between the solvent-accessible surface area (Lee and Richards 1971) of residue X in the protein and its solvent-accessible surface in an extended Gly-X-Gly

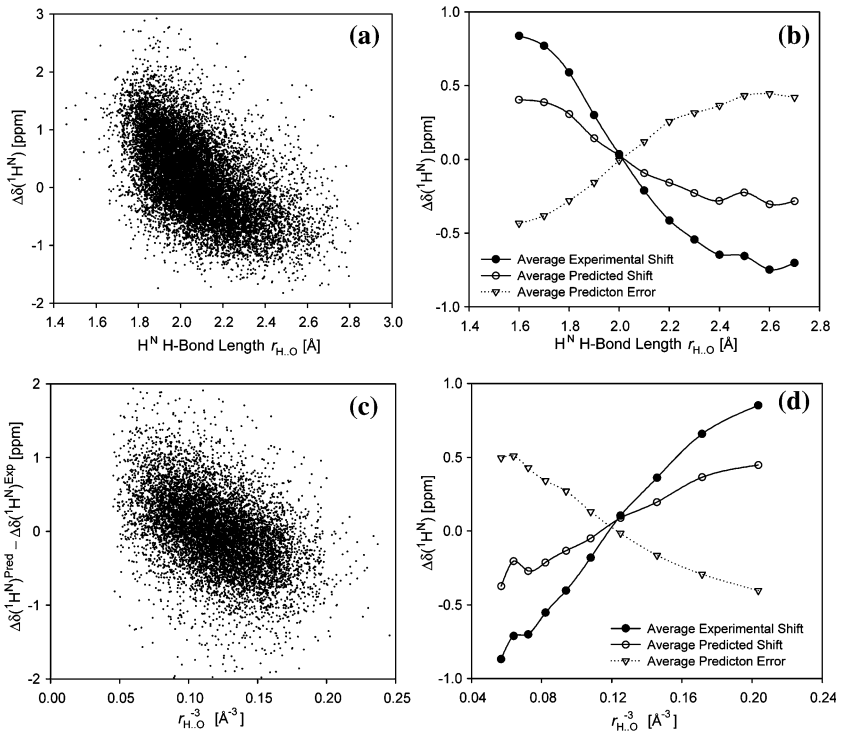


**Fig. 3** Plots of $^1H^N$ secondary chemical shift versus hydrogen bond length. (**a**) Scatter plot of experimental secondary chemical shifts, $\Delta\delta(^1H^N)^{Exp}$, versus H-bond length, $r_{H..O}$, for all intramolecularly H-bonded $^1H^N$ atoms in the SPARTA database. (**b**): Plot of the average $^1H^N$ secondary chemical shifts, $\Delta\delta(^1H^N)$, binned according to hydrogen bond lengths, $r_{H..O}$. The hydrogen bond length bin size is 0.1 Å, and only bins with >50 data are plotted. The predicted shifts (○) do not yet include a hydrogen bond correction term. (**c**) and (**d**) are analogous to (**a**) and (**b**), but include only the secondary chemical shifts of the $^1H^N$ atoms with residual solvent exposure <0.3, and are plotted with respect to $r_{H..O}^{-3}$

tripeptide (Shrake and Rupley 1973). Using a solvent exposure value of 0.3 as a cutoff, 9473 H-bonded amide protons (64%) are identified as not solvent exposed. The correlation between the experimental secondary chemical shifts of these $^1H^N$ atoms and their hydrogen bond lengths (Fig. 3d) is

$$\Delta\delta(^1H^N) = 14.43 \times r_{H..O}^{-3} - 1.63 \quad \text{ppm} \tag{3}$$

with $R = 0.63$. Moreover, there is a significant correlation,

$$\Delta\delta(^1H^N)^{Pred} - \Delta\delta(^1H^N)^{Exp} = -7.76 \times r_{H..O}^{-3} + 0.92 \quad \text{ppm} \tag{4}$$

between the average SPARTA prediction error and the applicable hydrogen bond length (Fig. 3c and d). Therefore, using the Eq. (4), the raw SPARTA-predicted secondary shifts of the buried and hydrogen-bonded amide protons in query proteins can easily be corrected. This H-bond correction of the $^1H^N$ shift prediction decreases the RMS deviation between SPARTA-predicted and the experimental $^1H^N$ shifts in the database decreased by ~10% (from 0.53 ppm to 0.47 ppm) for the buried and hydrogen-bonded amide protons. Averaged over the entire database, the RMS difference between experimental $^1H^N$ shifts and SPARTA predicted shifts equals 0.51 ppm (0.46 ppm when removing predictions that deviate by more than three standard deviations).

A similar but much weaker correlation between the experimental chemical shifts and hydrogen bond lengths of the attached amide protons is also observed for $^{15}N$ (Supplementary Figure S4) with a best fitting of

$$\Delta\delta(^{15}N) = 37.32 \times r_{H..O}^{-3} - 4.75 \quad \text{ppm} \tag{5}$$

and $R = 0.28$. However, for $^{15}N$ the H-bond dependence of secondary chemical shift is almost completely accounted for by the raw SPARTA results, which show only very small residual influence of H-bond length on the averaged prediction error (Supplementary Figure S4b). Application of an H-bond correction was found to have no effect for $^{15}N$ predicted and therefore is not included in the program.

*$^1H^\alpha$ shifts.* $^1H^\alpha$ shifts have long been used as reliable indicators of secondary structure. Indeed, our database shows a pronounced upfield shift (–0.28 ± 0.29 ppm, Supplementary Figure S1) in α-helix, and a downfield shift (0.59 ± 0.44 ppm) in β-sheet. Unlike $^1H^N$ shifts, which are strongly affected by the H-bond interaction of $^1H^N$, $^1H^\alpha$ shifts tend to be more affected by ring currents (Wishart and Case 2002; Neal et al. 2003). Indeed, a large

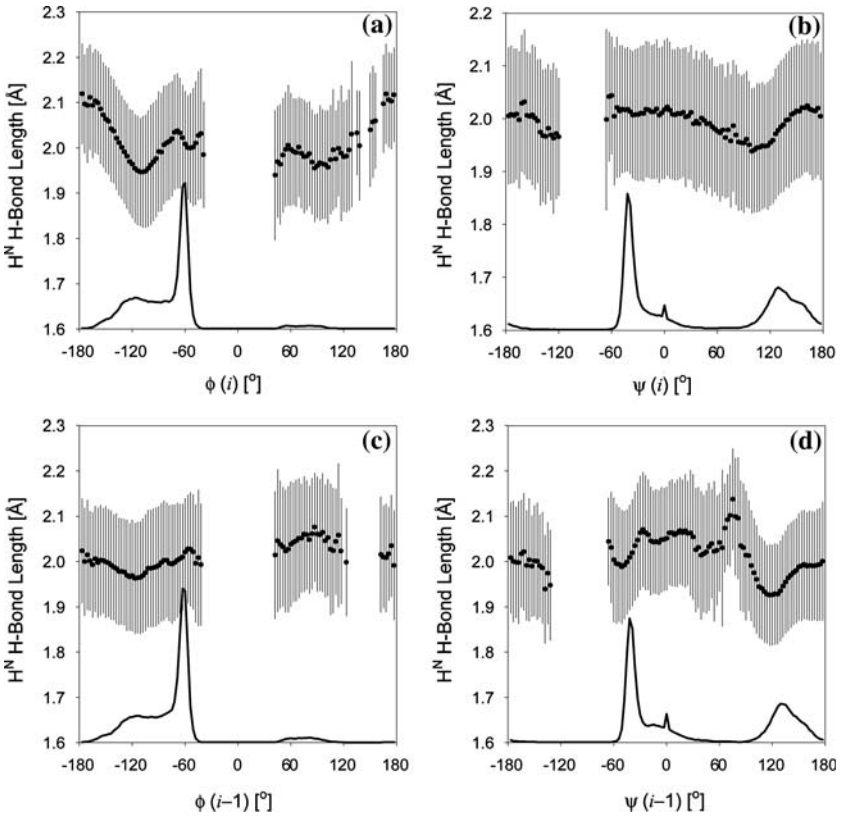


**Fig. 4** Plots of H-bond length involving $^1H^N$ atoms versus $\phi$ and $\psi$ torsion angles for 1197 proteins, solved at high resolution ( ≤ 1.6 Å) and taken from the RCSB. The horizontal axes represent the $\phi$ or $\psi$ torsion angles of the current ($i$) (**a** and **b**) or preceding ($i$ – 1) (**c** and **d**) residue. Vertical axes represent the average and the standard deviation (vertical bar) of the distribution of hydrogen bond lengths for $^1H^N$ atoms with residual solvent exposure <0.3 within each given bin of $\phi$ or $\psi$. The bin size is 3° for $\phi$ and $\psi$ angles, and only the bins with >30 data are plotted. The normalized density for each bin is shown by solid lines at the bottom of each panel

improvement in $^1H^\alpha$ shift prediction was obtained upon inclusion of ring current effects in the SPARTA program. This is accomplished by first subtracting the calculated ring current shifts from experimental secondary shifts for all entries in the database, and then adding the ring current shifts calculated from the coordinates of the query protein to the predicted shifts (see Material and methods). For best prediction performance, we find that a scaling factor of 0.6, obtained by a simple grid search, for all computed ring current shifts improves performance, and this scaling factor is used for all results presented here. Upon considering ring current effects, the RMS deviation between the experimental shifts and SPARTA-predicted shifts decreased by about ~13% for $^1H^\alpha$, but only ~0.1%, ~1%, and ~4% for $^{15}N$, $^{13}C$ and $^1H^N$ shifts, respectively.

$^1H^\alpha$ chemical shifts are also known to correlate with the $^1H^\alpha$ hydrogen bond length $r_{H..O}$ (Pardi et al. 1983; Wagner et al. 1983; Wishart et al. 1991), again following a $r^{-3}$ dependence. In our database, 3306 out of 16959 $^1H^\alpha$ atoms are engaged in intramolecular hydrogen bonds to oxygen, as defined by a Kabsch and Sander H-bond energy cutoff of –0.5 kcal/mole, with H-bond lengths range from 2.0 Å to 2.9 Å. An $r^{-3}$ dependency of $^1H^\alpha$ hydrogen bond length $r_{H..O}$ is found for secondary shifts of H-bonded $^1H^\alpha$ atoms present in the database (Fig. 5), even when disregarding all other factors, such as torsion angles. A best fit yields

$$\Delta\delta(^1H^\alpha) = 14.88 \times r_{H..O}^{-3} - 0.26 \quad \text{ppm} \tag{6}$$

with a correlation coefficient of 0.38. Remarkably, this correlation between the $^1H^\alpha$ hydrogen bond length $r_{H..O}$ and the predicted $^1H^\alpha$ shifts is already present prior to considering the effect of H-bonding (Fig. 5b), and must result from the correlation between the $^1H^\alpha$ hydrogen bond length and the $\phi$ and $\psi$ torsion angles of the center ($i$) and neighboring ($i$ – 1, $i$ + 1) residues (Figure S5). However, there remains a significant correlation of

$$\Delta\delta(^1H^\alpha)^{Pred} - \Delta\delta(^1H^\alpha)^{Exp} = -9.70 \times r_{H..O}^{-3} + 0.49 \quad \text{ppm} \tag{7}$$

between the SPARTA prediction errors of these $^1H^\alpha$ atoms and their H-bond lengths (Fig. 5b). Therefore, the accuracy of the SPARTA-predicted secondary shifts of the

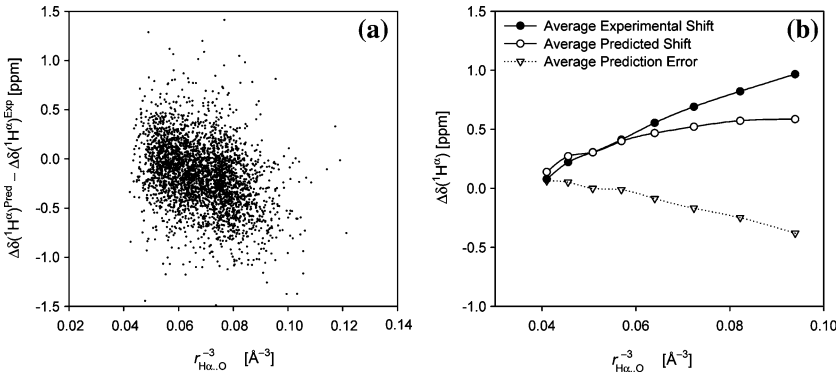


**Fig. 5** Plots of secondary $^1H^\alpha$chemical shift, $\Delta\delta(^1H^\alpha)$, versus H-bond lengths for $^1H^\alpha$ atoms. (**a**) Scatter plot of the difference between secondary chemical shift, predicted in the absence of an explicit H-bonding term and experimentally observed secondary chemical shift, versus the inverse cube of the H-bond length, $r^{-3}_{H..O}$, for all hydrogen-bonded $^1H^\alpha$ atoms in the SPARTA database. (**b**) Plot of the average $^1H^\alpha$ secondary chemical shifts $\Delta\delta(^1H^\alpha)$, binned according to H-bond length, using a bin size of 0.1 Å. Only bins with >50 data are plotted

H-bonded $^1H^\alpha$ could be improved by subtracting a correction of $-9.70 \times r^{-3}_{H..O} + 0.49$ from the SPARTA-predicted secondary shifts obtained from a direct database searching. The RMS deviation between experimental and SPARTA-predicted shifts is 0.37 ppm and 0.25 ppm, for hydrogen-bonded and non-hydrogen-bonded $^1H^\alpha$ atoms, respectively, without this correction. Application of the correction of Eq. (7) to the predicted shifts of H-bonded $^1H^\alpha$ atoms, reduces the RMS deviation between the predicted and the experimental $^1H^\alpha$ shifts by ~20% to 0.30 ppm. Overall, the RMS deviation between experimental $^1H^\alpha$ shifts and SPARTA-predicted values then becomes 0.27 ppm (Fig. 2). When removing predictions that deviate by more than three standard deviations, this number drops to 0.25 ppm.

*$^{13}C^\alpha$ and $^{13}C^\beta$ shifts.* $^{13}C^\alpha$ and $^{13}C^\beta$ chemical shifts in proteins are particularly sensitive to backbone $\phi$ and $\psi$ angles. Indeed, $^{13}C^\alpha$ and $^{13}C^\beta$ secondary chemical shifts have been used extensively for identification of secondary structure and prediction of protein backbone torsion angles. For our database, $^{13}C^\alpha$ and $^{13}C^\beta$ secondary chemical shifts in the database have distributions of 2.92 ± 1.47 and –0.27 ± 1.09 ppm, respectively, in α-helix, and –1.10 ± 1.38 and 2.34 ± 1.80 ppm, respectively, in $\beta$-sheet (Supplementary Figure S1). The results of the SPARTA prediction for $^{13}C^\alpha$ and $^{13}C^\beta$ chemical shifts show a good correlation between predicted and observed chemical shifts (Fig. 2). The RMS deviation between all experimental chemical shifts in the database and SPARTA-predicted chemical shifts are 0.98 and 1.07 ppm for $^{13}C^\alpha$ and $^{13}C^\beta$, respectively, which decreases to 0.88 ppm ($^{13}C^\alpha$) and 0.97 ppm ($^{13}C^\beta$) when removing outliers beyond three standard deviations. As can be seen from the optimized weighting factors in Table 1, the $^{13}C^\alpha$ and $^{13}C^\beta$ chemical shift predictions are dominated by the similarities of the intraresidue $\phi$ and $\psi$ angles, but similarity in residue type and $\chi^1$ angle for the center residue of the triplet are also important. The effect from neighboring residues on $^{13}C^\alpha$ and $^{13}C^\beta$ chemical shifts is relative small, which presumably contributes to the success of empirical methods (Wishart and Case 2002). The hydrogen bonds of the attached $^1H^\alpha$ atoms are likely to affect the $^{13}C^\alpha$ chemical shifts. In our database, 4,391 out of 24,021 $^{13}C^\alpha$ chemical shifts are from the $^{13}C^\alpha$ atoms where the attached $C^\alpha$ atoms are engaged in intramolecular H-bonds. A correlation between those $^{13}C^\alpha$ secondary chemical shifts and the strength of the $H^\alpha$ hydrogen bond is observed (Supplementary Figure S6). However, considering that the H-bond dependence of $^{13}C^\alpha$ secondary chemical shift is almost completely accounted for by the raw SPARTA results (Supplementary Figure S6b), the H-bond correction for $^{13}C^\alpha$ chemical shifts was not included in the program.

*$^{13}C'$ shifts.* It has long been recognized that $^{13}C'$ shifts in α-helices experience a downfield shift (1.77 ± 1.38 ppm in our database, Supplementary Figure S1), and an upfield shift (–1.35 ± 1.39 ppm) in $\beta$-sheet, which makes the $^{13}C'$ shift another useful indicator for protein secondary structure. However, unlike $^{13}C^\alpha$ and $^{13}C^\beta$ shifts, $^{13}C'$ shifts are also rather sensitive to the nature of the following residue and to the hydrogen bond interaction of the carbonyl oxygen atom. Our iterative optimization procedure finds that for the $^{13}C'_i$ of residue $i$ not only the adjacent torsion angles, $\psi_i$ and $\phi_{i+1}$, are important but also $\phi_i$ and $\psi_{i+1}$. These latter angles presumably reflect more the impact of regular secondary structure, and thereby indirectly the effect of H-bonding, rather than direct effects on the $^{13}C'_i$ chemical shift. Parameters in Table 1 indicate that $^{13}C'_i$ chemical shifts are also sensitive to both the intraresidue $\chi^1_i$ and the sequential $\chi^1_{i+1}$ angle, which shows that the $^{13}C'$ shift dependence on structure is more complex than for $^{13}C^\alpha$ and $^{13}C^\beta$. The dependence of the $^{13}C'$ shift on the strength of the hydrogen bond, manifested primarily through its center $\sigma_{22}$ tensor element (Asakawa et al. 1992;

Zheng et al. 1997; Wei et al. 2001), is well known. A correlation between the strength of the hydrogen bond to the carbonyl oxygen and local conformation ($\phi$ and $\psi$ angles) is also present, however (Supplementary Figure S7). In fact, our analysis indicates that the H-bond dependence of the $^{13}C'$ shift is already accounted for by the raw SPARTA prediction (Supplementary Figure S8b), and application of an H-bond correction term to the predicted $^{13}C'$ shifts does not further improve the prediction. Averaged over all proteins in the database, the RMS deviation between experimental and SPARTA-predicted $^{13}C'$ chemical shifts (Fig. 2) is 1.08 ppm, which decreases to 1.01 ppm when removing outliers that deviate by more than three standard deviations.

## Precision of SPARTA chemical shift predictions

In SPARTA, the chemical shifts of the center residue in the 20 best-matched triplets are averaged, weighted by the inverse of the similarity score, to yield the raw predicted shifts, which is subsequently ''refined'' by adding the ring current (for all nuclei) and hydrogen bonding corrections (for H-bonded $^{1}H^{N}$ and $^{1}H^{\alpha}$ atoms only). Importantly, the standard deviation found for the secondary shifts of the center residue in these 20 triplets, which represents the prediction ''precision,'' correlates with the accuracy of the predicted shift (Fig. 6). This is an important result as it provides individual error bars for each prediction, which are included as output parameters of the SPARTA program and are derived using linear equations (Supplementary Material Eqs S1 to S6), obtained by best fitting the graphs of Fig. 6.

## SPARTA evaluation outside the database

As a secondary validation and further check on the general applicability of SPARTA, we used the program to predict the $^{15}N$, $^{1}H^{N}$, $^{1}H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ chemical shifts of nine proteins for which X-ray coordinates and BMRB chemical shifts are available but which were not included in our database. These proteins were either missed during the initial search when preparing our database, or concern newly released BMRB chemical shifts. The RMS deviation between the experimental and SPARTA predicted shifts for $^{15}N$, $^{1}H^{N}$, $^{1}H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ are listed in Table 2, along with the information for these nine proteins. The prediction results for these nine proteins are comparable to those of the proteins in the SPARTA database, confirming that the parameter optimization procedure used does not bias in favor of proteins included in the database.

## SPARTA prediction using NMR solution structures

It is often assumed that most NMR structures do not achieve the coordinate accuracy of high quality X-ray

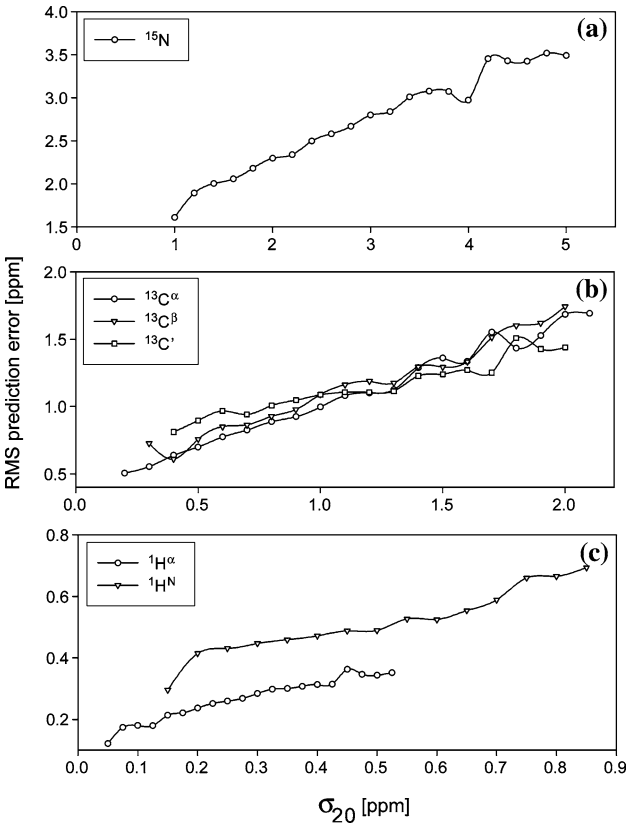


**Fig. 6** Correlation plot between the precision and accuracy of SPARTA-predicted (**a**) $^{15}N$, (**b**) $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, and (**c**) $^{1}H^{N}$, $^{1}H^{\alpha}$, chemical shifts. Horizontal axes represent the standard deviation of the chemical shifts for the center residue of the 20 selected database triplets, $\sigma_{20}$, or the precision of the prediction. Vertical axes represent the RMS difference between the SPARTA prediction and the experimental chemical shifts, which are binned according to the standard deviation, i.e., the precision of the prediction. The bin sizes for the precision are 0.25, 0.1, 0.05, 0.1, 0.1, and 0.1 ppm, for $^{15}N$, $^{1}H^{N}$, $^{1}H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$, respectively, and only bins with >50 data are plotted. The best-fit linear fitting parameters of the correlations are available as Supplementary Material Equations S1 to S6, and are used by SPARTA to calculate an ''estimated prediction error'' for each of the predicted shifts

structures (Williamson et al. 1995; Laskowski et al. 1996). For this reason, only X-ray structures were used in our present study as well as analogous analyses by others that require an empirical protein database. Here, we evaluate how chemical shift prediction with SPARTA is impacted by the type of input structure used by applying it to a set of 16 randomly chosen proteins for which NMR coordinates are available, as well as a complete set of $^{15}N$, $^{1}H^{N}$, $^{1}H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ chemical shifts (Table 3), several of which also had X-ray coordinates in the RCSB database. For proteins with a set of NMR conformers, the SPARTA chemical shift predictions were performed for each conformer, and the averaged predicted shifts were used as the final predicted chemical shifts. The RMS deviation between the experimental and SPARTA predicted shifts for

**Table 2** Summary of SPARTA results for nine test proteins, not included in the database

| BMRB code/protein name | RCSB code | Resolution (Å) | No. of residues | RMS deviation[a] [ppm] | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | $^{15}N$ | $^{1}H^{N}$ | $^{1}H^{\alpha}$ | $^{13}C^{\alpha}$ | $^{13}C^{\beta}$ | $^{13}C'$ |
| 4094 | 1IAR | 2.30 | 129 | 2.29 | 0.43 | 0.23 | 0.88 | 1.00 | 0.84 |
| 4186 | 3CBS | 2.00 | 137 | 1.97 | 0.35 | 0.21 | 0.95 | 1.19 | 0.85 |
| 4425 | 1BDO | 1.80 | 80 | 2.46 | 0.41 | 0.26 | 0.99 | 1.08 | – |
| 4472 | 1KMI | 2.90 | 129 | 2.57 | 0.62 | 0.28 | 0.97 | 1.02 | – |
| 5275 | 1KQR | 1.40 | 160 | 2.68 | 0.56 | 0.37 | 1.00 | 1.16 | 1.15 |
| 5513 | 1MMS | 2.57 | 140 | 2.82 | 0.53 | 0.31 | 1.36 | 1.42 | 1.05 |
| 7264 | 1FH9 | 1.72 | 312 | 2.70 | 0.61 | – | 1.05 | 1.26 | 1.26 |
| 7272 | 2IHB | 2.71 | 124 | 2.27 | 0.55 | 0.24 | 1.18 | 0.77 | 1.05 |
| GB3 | 1IGD | 1.10 | 56 | 2.60 | 0.43 | 0.28 | 1.13 | 1.16 | 1.08 |
| Average RMSD for test proteins | | | | 2.48 | 0.50 | 0.27 | 1.06 | 1.12 | 1.04 |

[a] RMS deviation between the predicted shifts and experimental shifts, which are reference-corrected by using the average prediction error if this error exceeds the referencing tolerance (see Methods)

**Table 3** Summary of SPARTA prediction accuracy when applied to predicting backbone chemical shifts of test proteins, using NMR coordinates

| BMRB code/protein name | No. of residues | RCSB code | Experimental method | RMS deviation[a] [ppm] | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | $^{15}N$ | $^{1}H^{N}$ | $^{1}H^{\alpha}$ | $^{13}C^{\alpha}$ | $^{13}C^{\beta}$ | $^{13}C'$ |
| 4094[b] | 129 | 2CYK | NMR | 2.89 | 0.49 | 0.30 | 1.37 | 1.37 | 1.07 |
| 4186[b] | 137 | 1BM5 | NMR | 3.28 | 0.52 | 0.36 | 1.40 | 1.71 | 1.16 |
| 4296[b] | 70 | 3MEF | NMR | 3.30 | 0.61 | 0.45 | 1.27 | 1.27 | 1.06 |
| 4425[b] | 80 | 2BDO | NMR | 3.92 | 0.58 | 0.40 | 1.50 | 1.73 | – |
| 4472[b] | 129 | 1CEY | NMR | 3.00 | 0.58 | 0.31 | 1.30 | 1.18 | – |
| 4876 | 130 | 1I56 | NMR | 3.53 | 0.61 | 0.77 | 1.60 | 1.55 | – |
| 5275[b] | 160 | 1KRI | NMR | 3.49 | 0.61 | 0.44 | 1.25 | 1.52 | 1.18 |
| 5513[b] | 140 | 1OLN | NMR | 2.82 | 0.52 | 0.30 | 1.36 | 1.42 | 1.05 |
| 6120[c] | 148 | 1T17 | NMR | 3.09 | 0.49 | 0.39 | 1.20 | 1.35 | 1.21 |
| 6364[c] | 113 | 1XNE | NMR | 2.58 | 0.60 | 0.33 | 1.24 | 1.18 | – |
| 6367[c] | 72 | 1XN7 | NMR | 2.21 | 0.63 | 0.32 | 1.04 | 1.29 | 1.21 |
| 6368[c] | 101 | 1XN9 | NMR | 2.42 | 0.48 | 0.29 | 1.07 | 1.10 | – |
| 6799[c] | 102 | 1YWX | NMR | 2.84 | 0.52 | 0.35 | 1.03 | 1.21 | 0.99 |
| BPTI[d] | 58 | 5PTI | X-ray | 2.45 | 0.46 | 0.27 | 1.21 | 1.51 | 1.10 |
| | | 1PIT | NMR | 2.58 | 0.47 | 0.36 | 1.28 | 1.43 | 1.19 |
| Ubiquitin[d] | 76 | 1UBQ | X-ray | 2.28 | 0.49 | 0.28 | 0.86 | 1.17 | 0.86 |
| | | 1D3Z | NMR | 2.19 | 0.47 | 0.29 | 0.82 | 1.11 | 0.86 |
| GB3 | 56 | 1IGD | X-ray | 2.60 | 0.43 | 0.28 | 1.13 | 1.16 | 1.08 |
| | | 2OED | NMR | 2.52 | 0.45 | 0.27 | 1.07 | 1.08 | 1.00 |
| Average RMSD using NMR coordinates | | | | 2.92 | 0.54 | 0.37 | 1.23 | 1.34 | 1.10 |

[a] RMS deviation between the predicted shifts and experimental shifts, which are reference-corrected by using the average prediction error if this error exceeds the referencing tolerance (see Methods)

[b] SPARTA shift prediction using X-ray coordinates are given in Table 2

[c] Structural genomics proteins

[d] Proteins contained in SPARTA database

$^{15}N$, $^{1}H^{N}$, $^{1}H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ are listed in Table 3. The results for these proteins show that, on average, NMR structures exhibit somewhat poorer agreement between SPARTA-predicted and experimental shifts. In particular, the accuracy of $^{1}H^{\alpha}$ shift prediction using NMR solution structure is markedly lower. This may result from the fact

that on average, of the six types of nuclei considered, $^{1}H^{\alpha}$ shifts are most sensitive to ring current shifts, and calculation of the ring current shift requires very accurate coordinates. For example for BPTI, which has 8 aromatic residues out of a total of 58, all predicted $^{1}H^{\alpha}$ shifts outliers calculated from the NMR solution structures are arising from incorrect ring current shifts. The same applies for canine milk lysozyme (RCSB code 1I56), which has 18 (out of 130) aromatic residues (Table 3). On the other hand, for the highly refined NMR structures of ubiquitin and GB3, SPARTA-predicted $^{1}H^{\alpha}$ chemical shifts are comparable in accuracy to those obtained from the X-ray structures, suggesting that the lower prediction accuracy is not an inherent property of NMR structures. Another example of an NMR structure whose chemical shifts are accurately predicted results from the structural genomics program (RCSB code 1XN9; Table 3).

## Comparison with SHIFTX and DC results

Several other approaches have been developed relatively recently to predict the chemical shifts for proteins with increased accuracy (Xu and Case 2002; Meiler 2003; Neal et al. 2003). Among those, the program SHIFTX (Neal et al. 2003), which is based on a hybrid predictive approach, yields the lowest reported RMS deviations between predicted and experimental chemical shifts. Therefore, we compare the SPARTA predictions with SHIFTX-predicted $^{15}N$, $^{1}H^{N}$, $^{1}H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ chemical shifts, obtained from the SHIFTX website (http://redpoll.pharmacy.ualberta.ca/shiftx/) for each protein in the SPARTA database, using the same X-ray coordinates. When using the SHIFTX program, the experimental chemical shifts of each protein are subjected to the same shift referencing correction method used in our evaluation of SPARTA, i.e., applying a correction of the average prediction error if this error is larger than a given tolerance (see Methods). The RMS deviations between the ''re-calibrated'' experimental chemical shifts and SHIFTX predicted chemical shifts are then calculated. Indeed, we find that SHIFTX-predicted shifts agree very well with experimental shifts, with RMS deviations of 2.87 ($^{15}N$), 0.54 ($^{1}H^{N}$), 0.29 ($^{1}H^{\alpha}$), 1.12 ($^{13}C^{\alpha}$), 1.25 ($^{13}C^{\beta}$) and 1.28 ppm ($^{13}C'$). When following the evaluation procedure used by Neal et al., and removing predicted shifts that deviate by more than three standard deviations, the respective RMS deviations are 2.70, 0.50, 0.25, 1.04, 1.12 and 1.21 ppm, and in close agreement with those reported by Neal et al. (2003).

The chemical shifts predicted by the empirical $\Delta(\phi,\psi)$-surfaces (Spera and Bax 1991), which are re-calculated based on the data in the SPARTA database, were also obtained using the DC program of the NMRPipe software package (Delaglio et al. 1995) for all $^{15}N$, $^{1}H^{N}$, $^{1}H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ nuclei of all proteins in the SPARTA database. The secondary chemical shifts taken from the $\Delta(\phi,\psi)$-surfaces also correlate well with the experimental shifts, albeit slightly less well than the SPARTA and SHIFTX results: the RMS deviations between the $\Delta(\phi,\psi)$-surface secondary chemical shifts and the experimental chemical shifts are 3.10, 0.67, 0.36, 1.12, 1.20 and 1.29 ppm, respectively.

Figure 7 compares the chemical shift prediction accuracy of SPARTA, SHIFTX, and $\Delta(\phi,\psi)$-surfaces for all 200 proteins in our database. Although SPARTA offers only modest improvements in chemical shift prediction relative to SHIFTX and the $\Delta(\phi,\psi)$-surface method, even these modest gains can have considerable impact on various novel approaches that aim to utilize chemical shifts in structure determination. For example, SPARTA can be used to ''assign'' chemical shifts to the entire set of RCSB proteins. Using a so-called molecular fragment replacement (MFR) approach (Kontaxis et al. 2005), this ''chemical-shift-assigned'' RCSB can then be searched for fragments that most closely match the chemical shift pattern of, for example, any 7-residue fragment of an NMR-assigned protein of unknown structure. When conducting such a search, a 10% improvement in chemical shift prediction accuracy narrows the RCSB search by *ca* $(0.9)^{-42} \approx 80$.

As an example of such an MFR application, we briefly compare the results of a standard MFR search for the 56-residue protein GB3, obtained using a ''SPARTA-chemical-shift-assigned'' library of 858 proteins, taken from the RCSB, with the output of the standard program, which relies on the $\Delta(\phi, \psi)$-surfaces to predict chemical shifts for proteins in the crystallographic database. For comparing the results, we report the average backbone rmsd between the 10 top fragments selected by MFR (Kontaxis et al. 2005), when using only the chemical shifts and amino acid

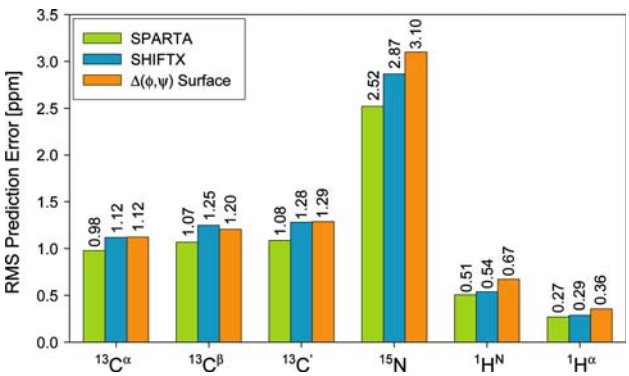


**Fig. 7** Comparison of accuracies of SPARTA-predicted, SHIFTX-predicted, and $\Delta(\phi,\psi)$-surface-predicted $^{15}N$, $^{1}H^{N}$, $^{1}H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ chemical shifts. The accuracies of predicted shifts are calculated as the RMS deviations between the predicted and experimental shifts

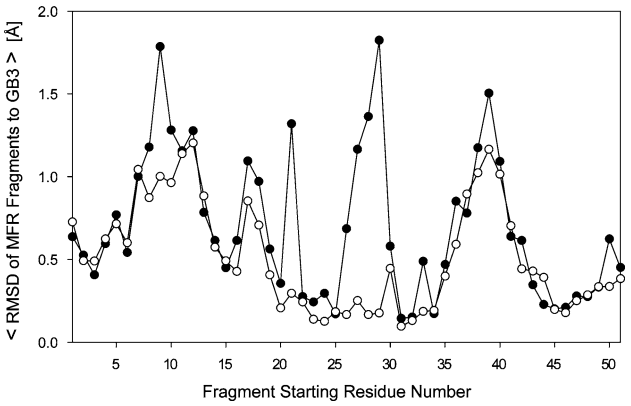


**Fig. 8** Averaged backbone (N, $C^{\alpha}$ and C′) coordinate RMS deviation between MFR-selected 6-residue fragments and the corresponding X-ray coordinates for protein GB3. Filled circles correspond to results of standard MFR, using an 858-protein library, and secondary chemical shifts for these derived from $\Delta(\phi,\psi)$-surfaces (Kontaxis et al. 2005). Open circles are the analogous result when using the SPARTA-predicted shifts for the same 858-protein library. The MFR program selects the 10 best-matched 6-residue fragments from the library on the basis of similarity between experimental chemical shifts ($^{15}N$, $^{1}H^{N}$, $^{1}H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$) and database chemical shifts, as well as residue type similarity between the GB3 fragment and fragments in the library

sequence as input values, and default relative weights of the various chemical shift types. As can be seen in Fig. 8, use of the ''SPARTA-assigned'' protein library results in considerable improvement in MFR prediction accuracy. Importantly, the improvements are largest in difficult regions, where outliers exist in the experimental chemical shift data. For example in the standard MFR search, all selected fragments containing residue F30, located near the middle of a long α-helix but with an extreme upfield $^{13}C^{\alpha}$ secondary chemical shift of –1.35 ppm, exhibit relatively large backbone coordinate RMS deviations relative to the true structure (Fig. 8). However, when using the ''SPARTA-chemical-shift-assigned'' library, these fragments fall much closer to the X-ray and NMR structures.

## Software availability

The SPARTA software package, which includes source code (written in C++), binary code (compiled for Linux, Win32, Solaris, Irix and Mac), our NMR protein database, installation instructions and examples, can be downloaded from http://spin.niddk.nih.gov/bax/.

**Acknowledgement** We thank Dr. Frank Delaglio for helpful discussions and comments on the coding, and testing of SPARTA, and Dr. Jinfa Ying for sharing the results of DFT calculations of the relation between chemical shift and geometry distortion. This work was supported by the Intramural Research Program of the NIDDK, NIH, and by the Intramural AIDS-Targeted Antiviral Program of the Office of the Director, NIH.

## References

Ando I, Saito H, Tabeta R, Shoji A, Ozaki T (1984) Conformation-dependent carbon-13 NMR chemical shifts of poly(L-Alanine) in the solid state: FPT INDO calculation of *N*-acetyl-*N*′-methyl-L-alanine amide as a model compound of poly(L-alanine). Macromolecules 17:457–461

Asakawa N, Kuroki S, Kurosu H, Ando I, Shoji A, Ozaki T (1992) Hydrogen-bonding effect on $^{13}C$ NMR chemical shifts of L-Alanine residue carbonyl carbons of peptides in the solid state. J Am Chem Soc 114:3261–3265

Avbelj F, Kocjan D, Baldwin RL (2004) Protein chemical shifts arising from α-helices and β-sheets depend on solvent exposure. Proc Natl Acad Sci USA 101:17394–17397

Bartels C, Billeter M, Güntert P, Wüthrich K (1996) Automated sequence-specific NMR assignment of homologous proteins using the program GARANT. J Biomol NMR 7:207–213

Beger RD, Bolton PH (1997) Protein $\phi$ and $\psi$ dihedral restraints determined from multidimensional hypersurface correlations of backbone chemical shifts and their use in the determination of protein tertiary structures. J Biomol NMR 10:129–142

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucl Acids Res 28:235–242

Case DA (1995) Calibration of ring-current effects in proteins and nucleic acids. J Biomol NMR 6:341

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

de Dios AC, Pearson JG, Oldfield E (1993) Secondary and tertiary structural effects on protein NMR chemical shifts: an *ab initio* approach. Science 260:1491–1496

Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR 6:277–293

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

Gardner KH, Kay LE (1998) The use of $^{2}H$, $^{13}C$, $^{15}N$ multidimensional NMR to study the structure and dynamics of proteins. Annu Rev Biophys Biomol Struct 27:357–406

Gardner KH, Rosen MK, Kay LE (1997) Global folds of highly deuterated, methyl-protonated proteins by multidimensional NMR. Biochemistry 36:1389–1401

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

Gronwald W, Boyko RF, Sönnichsen FD, Wishart DS, Sykes BD (1997) ORB, a homology-based program for the prediction of protein NMR chemical shifts. J Biomol NMR 10:165–179

Haigh CW, Mallion RB (1979) Ring current theories in Nuclear Magnetic Resonance. Progr NMR Spectrosc 13:303–344

Herranz J, González C, Rico M, Nieto JL, Santoro J, Jiménez MA, Bruix M, Neira JL, Blanco FJ (1992) Peptide group chemical shift computation. Magn Reson Chem 30:1012–1018

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

Jurgen FD, Aart JN, Wim V, Jundong L, Alexandre MJJB, Robert K, John LM, Eldon LU (2005) BioMagResBank databases DOCR and FRED containing converted and filtered sets of experimental NMR restraints and coordinates from over 500 protein PDB structures. J Biomol NMR 32:1–12

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

Kontaxis G, Delaglio F, Bax A (2005) Molecular fragment replacement approach to protein structure determination by chemical shift and dipolar homology database mining. Methods Enzymol 394:42–78

Laskowski RA, Rullmann JAC, MacArthur MW, Kaptein R, Thornton JM (1996) AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. J Biomol NMR 8:477–486

Le HB, Oldfield E (1994) Correlation between $^{15}$N NMR chemical shifts in proteins and secondary structure. J Biomol NMR 4: 341–348

Le HB, Oldfield E (1996) *Ab initio* studies of amide-N-15 chemical shifts in dipeptides: applications to protein NMR spectroscopy. J Phys Chem 100:16423–16428

Lee B, Richards FM (1971) The interpretation of protein structures: Estimation of static accessibility. J Mol Biol 55:379–380

Markley JL, Bax A, Arata Y, Hilbers CW, Kaptein R, Sykes BD, Wright PE, Wüthrich K (1998) Recommendations for the presentation of NMR structures of proteins and nucleic acids—IUPAC-IUBMB-IUPAB Inter-Union Task Group on the standardization of data bases of protein and nucleic acid structures determined by NMR spectroscopy. J Biomol NMR 12:1–23

Meiler J (2003) PROSHIFT: protein chemical shift prediction using artificial neural networks. J Biomol NMR 26:25–37

Moseley HNB, Sahota G, Montelione GT (2004) Assignment validation software suite for the evaluation and presentation of protein resonance assignment data. J Biomol NMR 28:341–355

Neal S, Nip AM, Zhang HY, Wishart DS (2003) Rapid and accurate calculation of protein $^{1}$H, $^{13}$C and $^{15}$N chemical shifts. J Biomol NMR 26:215–240

Ösapay K, Case DA (1991) A new analysis of proton chemical shifts in proteins. J Am Chem Soc 113:9436–9444

Ösapay K, Case DA (1994) Analysis of proton chemical shifts in regular secondary structure of proteins. J Biomol NMR 4: 215–230

Pardi A, Wagner G, Wüthrich K (1983) Protein conformation and proton NMR chemical shifts. Eur J Biochem 137:445–454

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

Redfield C, Dobson CM (1990) Proton NMR studies of human lysozyme: spectral assignment and comparison with hen lysozyme. Biochemistry 29:7201–7214

Saitô H (1986) Conformation-dependent $^{13}$C chemical shifts: a new means of conformational characterization as obtained by high-resolution solid-state $^{13}$C NMR. Magn Reson Chem 24:835–852

Shrake A, Rupley JA (1973) Environment and exposure to solvent of protein atoms. Lysozyme and Insulin. J Mol Biol 79:351–364

Spera S, Bax A (1991) Empirical correlation between protein backbone conformation and C$\alpha$ and C$\beta$ $^{13}$C Nuclear Magnetic Resonance chemical shifts. J Am Chem Soc 113:5490–5492

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

Venters RA, Farmer II BT, Fierke CA, Spicer LD (1996) Characterizing the use of perdeuteration in NMR studies of large proteins: $^{13}$C, $^{15}$N and $^{1}$H assignments of human carbonic anhydrase II. J Mol Biol 264:1101–1116

Villegas ME, Vila JA, Scheraga HA (2007) Effects of side-chain orientation on the $^{13}$C chemical shifts of antiparallel $\beta$-sheet model peptides. J Biomol NMR 37:137–146

Wagner G, Pardi A, Wüthrich K (1983) Hydrogen bond length and proton NMR chemical shifts in proteins. J Am Chem Soc 105:5948–5949

Wang LY, Eghbalnia HR, Bahrami A, Markley JL (2005) Linear analysis of carbon-13 chemical shift differences and its application to the detection and correction of errors in referencing and spin system identifications. J Biomol NMR 32:13

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

Wang YJ, Jardetzky O (2004) Predicting $^{15}$N chemical shifts in proteins using the preceding residue-specific individual shielding surfaces from $\phi$, $\psi^{i-1}$, and $\chi^{1}$ torsion angles. J Biomol NMR 28:327–340

Wang YJ, Wishart DS (2005) A simple method to adjust inconsistently referenced $^{13}$C and $^{15}$N chemical shift assignments of proteins. J Biomol NMR 31:143–148

Wei Y, Lee DK, Ramamoorthy A (2001) Solid-State $^{13}$C NMR chemical shift anisotropy tensors of polypeptides. J Am Chem Soc 123:6118–6126

Williamson MP (1990) Secondary-structure dependent chemical shifts in proteins. Biopolymers 29:1423–1431

Williamson MP, Kikuchi J, Asakura T (1995) Application of $^{1}$H NMR chemical shifts to measure the quality of protein structures. J Mol Biol 247:541–546

Wishart DS, Case DA (2002) Use of chemical shifts in macromolecular structure determination. Methods Enzymol 338:3–34

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

Wishart DS, Sykes BD (1994) The $^{13}$C Chemical-Shift index—a simple method for the identification of protein secondary structure using $^{13}$C chemical-shift data. J Biomol NMR 4:171–180

Wishart DS, Sykes BD, Richards FM (1991) Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. J Mol Biol 222:311–333

Wishart DS, Watson MS, Boyko RF, Sykes BD (1997) Automated $^{1}$H and $^{13}$C chemical shift prediction using the BioMagResBank. J Biomol NMR 10:329–336

Xu XP, Case DA (2001) Automated prediction of $^{15}$N, $^{13}$C$^{\alpha}$, $^{13}$C$^{\beta}$ and $^{13}$C$'$ chemical shifts in proteins using a density functional database. J Biomol NMR 21:321–333

Xu XP, Case DA (2002) Probing multiple effects on $^{15}$N, $^{13}$C$^{\alpha}$, $^{13}$C$^{\beta}$, and $^{13}$C$'$ chemical shifts in peptides using density functional theory. Biopolymers 65:408–423

Zhang HY, Neal S, Wishart DS (2003) RefDB: a database of uniformly referenced protein chemical shifts. J Biomol NMR 25:173–195

Zheng G, Wang L, Hu J, Zhang X, Shen L, Ye C, Webb GA (1997) Hydrogen bonding effects on the $^{13}$C NMR chemical shift tensors of some amino acids in the solid state. Magn Reson Chem 35:606–608