Journal of Biomolecular NMR, **26:** 215–240, 2003.
KLUWER/ESCOM

Rapid and accurate calculation of protein ^{1}H, ^{13}C and ^{15}N chemical shifts

Stephen Neal, Alex M. Nip, Haiyan Zhang & David S. Wishart
Faculty of Pharmacy & Pharmaceutical Sciences, University of Alberta, Edmonton, AB, T6G 2N8, Canada

Received 22 November 2002; Accepted 6 March 2003

Key words: calculation, chemical shift, data mining, NMR, prediction, protein

Abstract

A computer program (SHIFTX) is described which rapidly and accurately calculates the diamagnetic ^{1}H, ^{13}C and ^{15}N chemical shifts of both backbone and sidechain atoms in proteins. The program uses a hybrid predictive approach that employs pre-calculated, empirically derived chemical shift hypersurfaces in combination with classical or semi-classical equations (for ring current, electric field, hydrogen bond and solvent effects) to calculate ^{1}H, ^{13}C and ^{15}N chemical shifts from atomic coordinates. The chemical shift hypersurfaces capture dihedral angle, sidechain orientation, secondary structure and nearest neighbor effects that cannot easily be translated to analytical formulae or predicted via classical means. The chemical shift hypersurfaces were generated using a database of IUPAC-referenced protein chemical shifts – RefDB (Zhang et al., 2003), and a corresponding set of high resolution ($<$2.1 Å) X-ray structures. Data mining techniques were used to extract the largest pairwise contributors (from a list of $\sim$20 derived geometric, sequential and structural parameters) to generate the necessary hypersurfaces. SHIFTX is rapid ($<$ 1 CPU second for a complete shift calculation of 100 residues) and accurate. Overall, the program was able to attain a correlation coefficient (r) between observed and calculated shifts of 0.911 (^{1}Hα), 0.980 (^{13}Cα), 0.996 (^{13}Cβ), 0.863 (^{13}CO), 0.909 (^{15}N), 0.741 (^{1}HN), and 0.907 (sidechain ^{1}H) with RMS errors of 0.23, 0.98, 1.10, 1.16, 2.43, 0.49, and 0.30 ppm, respectively on test data sets. We further show that the agreement between observed and SHIFTX calculated chemical shifts can be an extremely sensitive measure of the quality of protein structures. Our results suggest that if NMR-derived structures could be refined using heteronuclear chemical shifts calculated by SHIFTX, their precision could approach that of the highest resolution X-ray structures. SHIFTX is freely available as a web server at http://redpoll.pharmacy.ualberta.ca.

Introduction

Chemical shifts are the 'mileposts' of NMR spectroscopy. Not only are they important as spectral markers, but their dependency on multiple electronic and geometric factors means that chemical shifts can potentially provide a rich source of structural information. However, these multiple dependencies make both the interpretation and accurate prediction of chemical shifts exceedingly difficult – particularly for large molecules such as proteins. Fortunately, over the past decade, significant progress in chemical shift prediction has been made, both through computational advances (Williamson and Asakura, 1997; Case, 1998, 2000; Wishart and Case, 2001) and through the rapid expansion of biomolecular chemical shift databases (Seavey et al., 1991; Zhang et al., 2003).

Currently there are three main approaches for calculating protein chemical shifts from atomic coordinates: (1) Quantum mechanical, (2) classical, and (3) empirical. Quantum mechanical (QM) approaches employing density functional theory (DFT) have been used to very accurately calculate ^{1}H, ^{13}C and ^{15}N shifts for selected classes of residues in proteins (de-Dios et al., 1993; Le et al., 1995; Xu and Case, 2001, 2002). Classical approaches, which employ simplified or empirical equations derived from classical physics and experimental data, have been used to accurately calculate ^{1}H shifts for quite some time (Wagner et al., 1983; Dalgarno et al., 1983; Osapay and Case, 1991,

*To whom correspondence should be addressed. E-mail: david.wishart@ualberta.ca

1994; Wishart et al., 1991; Herranz et al., 1992; Williamson et al., 1992). Empirical approaches, which rely on chemical shift 'hypersurfaces' calculated from databases of observed chemical shifts, are capable of rapid, but only modestly accurate calculation of ^{1}H, ^{13}C, and ^{15}N shifts (Spera and Bax, 1991; Le and Oldfield, 1994; Beger and Bolton, 1997; Wishart and Nip, 1998; Iwadate et al., 1999). These hypersurfaces relate chemical shifts to various empirical parameters (backbone angles, nearest neighbors, sidechain angles, secondary structure, etc.). Pre-calculated chemical shift hyper-surfaces are also used in QM approaches to greatly accelerate the speed of their calculations (Xu and Case, 2001, 2002; Le et al., 1995).

As of yet, none of the three approaches has developed to a stage where it can offer a rapid, accurate method to calculate all (i.e., sidechain and backbone ^{1}H, ^{13}C and ^{15}N) chemical shifts for all residues under all conditions (diamagnetic and paramagnetic). Ideally, if one could combine the speed and predictive breadth of the empirical approaches with the accuracy of the classical or QM approaches, then it might be possible to achieve this goal. Here we describe a hybrid predictive method that attempts to combine the empirical hypersurface approach with the classical approach to accurately and rapidly calculate essentially all diamagnetic shifts for all 20 amino acid residues. The program, called SHIFTX, takes as input a protein structure in Protein Data Bank format, and predicts the diamagnetic ^{1}H, ^{13}C and ^{15}N chemical shifts of the protein's backbone and sidechain atoms. Tests indicate that SHIFTX is rapid (about 1 sec on a 2.2 GHz Pentium IV CPU for a complete shift calculation of 100 residues) and accurate. Overall, the program was able to attain a correlation coefficient (r) between observed and calculated shifts of 0.911 (^{1}Hα), 0.980 (^{13}Cα), 0.996 (^{13}Cβ), 0.863 (^{13}CO), 0.909 (^{15}N), 0.741 (^{1}HN), and 0.907 (sidechain ^{1}H) with an RMS error of 0.23, 0.98, 1.10, 1.16, 2.43, 0.49, and 0.30 ppm, respectively. We further show that the agreement between observed and SHIFTX calculated chemical shifts can be an extremely sensitive measure of the accuracy and precision of protein structures. We believe SHIFTX could serve as a valuable tool for refining and assessing protein structures, for validating and adjusting chemical shift assignments (Zhang et al., 2003) and potentially for generating 3D protein structures using only chemical shift information (Wishart and Case, 2001). A complete description of SHIFTX, its performance, applications and limitations follows.

Methods

SHIFTX employs a hybrid predictive protocol that uses a combination of classical equations and empirical 'hypersurfaces' to calculate chemical shifts from atomic coordinate data. Classical equations are used to calculate the effects of well-characterized physical phenomena such as ring currents, H-bonds and electric field effects. The chemical shift hypersurfaces (described in more detail below) are derived from observed data and are fundamentally statistical in nature. These hypersurfaces serve as a simple method for capturing complex, nonlinear, and multiparametric interactions that do not lend themselves to simple analytical expressions. A SHIFTX chemical shift calculation is therefore the sum of several components:

$$\delta_{calc} = \delta_{coil} + \delta_{RC} + \delta_{EF} + \delta_{HB} + \delta_{HS}, \quad (1)$$

where δ_{coil} is the random coil ^{1}H, ^{13}C or ^{15}N chemical shift (relative to DSS) of the amino acid as given by Wishart et al. (1995b), δ_{RC} is the ring current shift, δ_{EF} is the electric field contribution, δ_{HB} is the hydrogen bond contribution, and δ_{HS} is the contribution from the chemical shift hypersurfaces for the nucleus of interest (primarily the backbone dihedral angles). A SHIFTX prediction is composed of four phases: (1) Reading the PDB file; (2) checking and adding hydrogen atoms (if necessary); (3) calculating the classical contributions (ring currents, electric field effects, etc.); and 4) calculating the chemical shift hypersurface contributions and summing them with the results of phase 3. These calculations are performed for essentially all atoms that can yield measurable chemical shifts. A more detailed description of each of the four phases, including the specific formulae, criteria and protocols is given below.

File input

SHIFTX reads standard PDB files using file I/O methods originally developed for VADAR (Wishart et al., 1994). The program will read in a specified chain from a multi-chain file (defaulting to the first chain if another is not specified) and ignores non-standard amino acids, non-water heteroatoms and ligands (heme rings, metals etc.). Conditions that may impact the accuracy of predictions (missing atoms, numbering irregularities, and chain breaks) are noted in the program output. The I/O portion of the program also loads the amino

acid sequence, determines nearest neighbors, calculates dihedral angles, secondary structures, H-bond partners, salt bridges and charge pairs. These parameters are all used in evaluating the chemical shift hypersurfaces.

Hydrogen placement

SHIFTX initially determines if there are hydrogen atom coordinates provided in the PDB file. If not, the position of HN atoms is calculated using the plane formed by the N, CA, and CO_{n-1} atoms. The proton is placed in this plane 0.86 Å from the N atom such that the angle formed by the H-N bond and the N-CO_{n-1} bond is 118.9 degrees. The HA atom (for non-glycine residues) is placed 1.0 Å from CA, such that it fills the 120 degree tetrahedron formed by CA, CB, N, and CO. For glycine, the two alpha hydrogens are placed by rotating the vector formed by CA and N around the vector formed by CA and C by +120 and −120 degrees. All other hydrogens are added using the program REDUCE (Word et al., 1999; http://kinemage.biochem.duke.edu) which has been incorporated into the SHIFTX web server.

Ring current effects

The presence of aromatic rings and their associated ring currents can have a profound effect on the chemical shifts of nearby nuclei. As Osapay and Case (1991) have shown the effects of these currents are best calculated using the semi-classical methods of Haigh and Mallion (1980). Our findings indicate that all hydrogen nuclei can be affected by ring currents, as can the CA, CB, CO, and N atoms. In calculating the ring current contributions for a given protein, SHIFTX first generates a list of susceptible atoms and a list of rings, and then calculates the influence of each ring on each such atom. This influence is the product of a geometrical factor **G**, a target-specific constant **F**, and a ring-specific intensity **I** (see Table 1). With the values of the latter two constants being determined through parameter fitting.

For each ring, a normal is computed from the cross-product of two vectors originating from the first ring member and extending to the second and last ring members, respectively. The ring is deemed to lie in a plane perpendicular to this normal. Next, the projection of the target atom onto the ring plane is found; this is designated point **O** for purposes of the following calculations. Finally, the areas of a series of triangles are computed, with the vertices of each triangle being adjacent points on the ring and the point **O**. Because each pair of adjacent ring atoms is considered (including the first and last atoms), there will be one triangle for each ring member. These areas are *algebraic* and may be positive or negative. For instance, consider vectors $\mathbf{R}_i$ and $\mathbf{R}_j$ from **O** to the i^{th} and j^{th} ring members. The area of the triangle formed by these three points is negative if the cross product $\mathbf{R_i} \times \mathbf{R_j}$ is parallel to the ring normal, and positive if it is antiparallel. The area of each triangle is then multiplied by a distance factor, given by

$$d_{ij} = \frac{1}{r_i^3} + \frac{1}{r_j^3}, \tag{2}$$

where r_i = length of $\mathbf{R_i}$ and r_j = length of $\mathbf{R_j}$ Thus the geometrical factor **G** is given by:

$$G = \sum_{\substack{ring \\ members}} d_{ik}\, area_{ijO}. \tag{3}$$

G, the ring-specific intensity **I**, and the target nucleus factor **F** are multiplied together to yield the total effect on the target nucleus' chemical shift due to the given ring:

$$\delta_{RC} = GIF. \tag{4}$$

The ring-specific intensity factor (**I**) and target nucleus factors (**F**) were determined empirically using a simple grid search optimization protocol (step size of 0.01) on the training database. The residue-specific least square values (for **I** and **F**) suggested by Osapay and Case (1991) were used as starting values. Note that in this formulation, **F** corresponds to Osapay and Case's parameter B but with the implicit assumption that **F** will vary for different target nuclei (^{15}N, ^{13}C, ^{1}HN, etc.) as a result of their different shielding or differing electron cloud 'mobility'. Initially, the ring-current optimization was performed only on the HA shifts wherein all five ring-specific intensity parameters and the target nucleus factor (a total of six numbers) were allowed to vary. The resulting ring-specific intensity factors (**I**) are shown in Table 1 and were found to be quite similar to those reported by Case and Osapay. The resulting value for **F** (5.13×10^{-6}) also corresponds closely to the B value of 5.45×10^{-6} determined by Osapay and Case (1991). The ring current optimization process was then repeated for other target nuclei by holding the **I** values constant and allowing **F** to vary. Note that for heavy nuclei (^{15}N and ^{13}C) the grid step was changed to 0.10. The resulting target nucleus factors ($\times 10^6$) were: 7.06 for HN, 5.13 for HA and all other hydrogen atoms,

Table 1. Residue types containing aromatic rings, the number of ring members, and the associated 'intensity factor'. Note that Trp has two rings

Residue type	# of Ring members	Intensity factor	Ordered list of members
Phe	6	1.05	CG CD2 CE2 CZ CE1 CD1
Tyr	6	0.92	CG CD2 CE2 CZ CE1 CD1
Trp	6	1.04	CD2 CE3 CZ3 CH2 CZ2 CE2
Trp	5	0.90	CG CD2 CE2 NE1 CD1
His	5	0.43	CG ND1 CE1 NE2 CD2

1.50 for CA and 1.00 for CB, CO and N. These variations in **F** seem to reflect the susceptibility of different nuclei to ring current effects, with amide protons being most susceptible and CB, CO and N being least susceptible.

Electric field effects

SHIFTX uses the method of Buckingham (1960) to calculate the effects of electrostatic fields on chemical shifts. The shifts of alpha carbons and all hydrogens ('target' atoms) are subject to electrostatic effects; these effects may be caused by CO, O, OD_n, OE_n, or N atoms ('source' atoms). Apart from the types and coordinates of the source and target atoms, this calculation also requires the coordinates of the target's 'partner' atom, to which it is bonded. A list of all the target atoms and their partners is available at the SHIFTX web site. All sources influence all targets within 3.0 Å, with the following exceptions: (1) No source atom influence targets on its own or adjacent residues; (2) O (carbonyl oxygen) atoms do not influence HN (amide hydrogen) atoms and (3) solvent (i.e. water) atoms do not act as sources. Each source atom has an associated partial charge, i.e., -0.9612×10^{-10} esu for O, OD_n and OE_n atoms, 1.3937×10^{-10} esu for C atoms and 0.7209×10^{-10} esu for N atoms. Given this information, the effect on the shift of each target by each source can be calculated as:

$$\delta_{EF} = \frac{1 \times 10^{22} q \varepsilon \cos\theta}{d^2}, \quad (5)$$

where $\varepsilon = 1 \times 10^{-12}$, q = source charge (in esu, see above), θ = angle formed by source-target-partner and d = distance from source to target (Å). The total effect on each target atom is the sum of the effects of each source atom.

Hydrogen bond effects

While it might be expected that electrostatic effects should account for most of the chemical shift perturbations brought on by nearby polar or charged atoms, we found that the explicit inclusion of hydrogen bond effects improved the overall performance in SHIFTX. The methods used to determine the presence of hydrogen bonds are modeled after those of VADAR (Wishart et al., 1994) with possible hydrogen bond donors being amide and alpha hydrogens (H and HA). The acceptors may be the carbonyl oxygens on the backbone (O), sidechain oxygens (OD_n, OE_n, OG_n, OH_n), or oxygen atoms from water in the solvent. SHIFTX compiles lists of possible donors and acceptors, and considers the possible existence of a hydrogen bond between each donor-acceptor pair.

For a bond to exist, the donor and acceptor must be on different residues, and if the acceptor is a solvent oxygen, the donor must not be an HA. Also, the oxygen-hydrogen separation must be less than an empirically determined distance; 3.50 Å for HNs and 2.77 Å for HAs. Bond geometry is also considered; specifically, the angle between the N-H bond vector and the C=O bond vector must be 90 degrees or more, computed with the vectors translated such that the C and N occupy the same point (Kabsch and Sander, 1983).

Having applied these rules to each donor-acceptor pair, SHIFTX then sorts the list of possible bonds by the O–H separation distance, shortest to longest. The list is then processed so that only the single 'strongest' hydrogen bond is identified for each donor-acceptor pair. More specifically, the process is as follows: The first bond on the list is deemed to exist, and to preclude the existence of any bonds involving the same donor or receptor; any such bonds are removed from the list. SHIFTX then moves on to the next bond on the now-culled list, and similarly removes any bond made

redundant, repeating this procedure until the end of the list is reached. Tests conducted on a large portion of the training data revealed that the inclusion of multiple acceptors (i.e., bifurcated H-bonds) in the calculation of hydrogen bond effects actually diminished the overall performance of SHIFTX. This finding is consistent with the idea that hydrogen bonds are pseudo-sigma (i.e., covalent) bonds involving partial transfer of the hydrogen donor's single electron to its acceptor atom (Baker and Hubbard, 1984).

The formulae developed by Wagner et al. (1983) and Wishart et al. (1991) for calculating the influence of hydrogen bonds on HA and HN shifts was adopted for this work. These workers found that a r^{-3} dependency (distance between donor and acceptor) was most consistent with their experimental data. We optimized their δ_{HB} parameters for amide hydrogens through a simple grid search of the two variable parameters (step size of 0.01) using our much larger training database. For amide protons, the best fit formula accounting for δ_{HB} shifts is given by:

$$\delta_{HB} = \frac{0.75}{r^3} - 0.99, \tag{6}$$

where r = the hyrodgen-oxygen separation in Å. The above formula is valid for hydrogen bond lengths between 1.5–3.5 Å. While the vast majority (>95%) of hydrogen bonds involve amide protons, the work of Wagner et al. (1983), Wishart et al. (1991) and Derewenda et al. (1995) clearly shows that alpha protons can also be involved in hydrogen bonds and that this bonding can influence their chemical shifts. Consequently we included a second equation in SHIFTX to account for HA atoms which acted as hydrogen bond donors. For these hydrogen bonds a pseudo-sigmoidal potential was found to work best where the H-bond shift (δ_{HB}) is assumed to be constant between 2.61–2.77 Å (long H-bonds) and 2.00-2.27 Å (short bonds). For H-bonds between 2.61–2.27 Å the hydrogen bond contribution to the HA chemical shift is given by:

$$\delta_{HB} = \frac{15.69}{r^3} - 0.67. \tag{7}$$

If a glycine residue happens to have H-bonds on both its HA's, the resulting contribution to the δ_{HB} chemical shift is the mean of the individual δ_{HB} shifts.

Empirical chemical shift hypersurfaces

Previous studies (for example, Spera and Bax, 1991; Le and Oldfield, 1994; Wishart and Nip, 1998) have shown the utility of torsion angle chemical shift hypersurfaces in predicting chemical shifts; the essential idea is to use to a residue's backbone angles as parameters to a lookup table which returns a chemical shift value based on those parameters. The empirical lookup tables described herein are an extension of the same idea to parameters other than local backbone torsion angles. To develop these tables or hypersurfaces, a special data-mining program, called MINER, was created to search through a large set of protein structural parameters thought to be important for chemical shift determination. MINER's task was to find which of those parameters or pairs of parameters was most effective in calculating atom-specific chemical shifts. 'Effective' in this context meant minimizing root-mean-squared errors or maximizing correlation between the predicted and observed shifts; in practice, the two metrics produced essentially the same result.

In operation, MINER is supplied with a database in which each row is a residue, and each column is a physical parameter of that residue; the parameters supplied are listed in Table 2. In addition, each record contains the observed chemical shift for the nucleus of interest, and a 'best guess' at that shift, the latter consisting of the sum of the classically calculable shift contributions (random coil, electrostatic, ring current, and hydrogen bond factors) for that nucleus. The input database is further divided into a 'training set' and one or more 'test sets'; any predictions generated from the training set were evaluated against the test set to prevent overfitting.

MINER's fundamental task is building chemical shift hypersurfaces (or 2D matrices), with vertical and horizontal axes each corresponding to one of the precalculated protein structural or sequence parameters. Each residue in the data set was assigned to one or more cells in the table in accordance with the table's axes and the residue's parameter values. For example, if the table's axes were 'phi' and 'psi' respectively, each residue with phi and psi angles in the range $-180°$ to $-170°$ would be assigned to the top-left cell in the table; if instead the residue had a psi angle of $-165°$, it would instead be assigned to the second cell in the top row. The mean shift difference (the difference between the observed and predicted shift, $\Delta\delta$) of the group of residues in each cell was calculated, forming a table of $\Delta\delta$ values, which was then used to compute a preliminary chemical shift hypersurface for each residue. In the case of continuous quantities (torsion angles and bond lengths), cubic splines were used to interpolate between table rows and columns; for dis-

Table 2. Physical or derived property parameters used by MINER

Factor	Previous residue	Next residue	Values/discretization
First residue	Y	N	True/False
Amino acid type	Y	Y	1 of 20
Phi angle	Y	Y	2 (of 18) 40°-wide bins centered every 20° from 0–340°
Psi angle	Y	Y	2 (of 18) 40°-wide bins centered every 20°from 0–340°
Chi angle	Y	Y	One of 0–120°, 120–240°, or 240–360°
Chi2 angle	N	N	One of 0–120°, 120–240°, or 240–360°
Secondary structure	Y	Y	One of coil, helix, or Beta sheet
Length of HA H-bond	Y	Y	0, or one of 20 equal-length bins
Length of HA2 H-bond	Y	Y	0, or one of 20 equal-length bins
Length of HN H-bond	Y	Y	0, or one of 20 equal-length bins
Length of O H-bond	Y	Y	0, or one of 20 equal-length bins
Disulfide bond	Y	Y	True/False
HA hydrogen bond	N	N	True/False
HA2 hydrogen bond	N	N	True/False
HN hydrogen bond	N	N	True/False
O hydrogen bond	N	N	True/False
Hydrogen bond status	N	N	Concatenation of previous four values

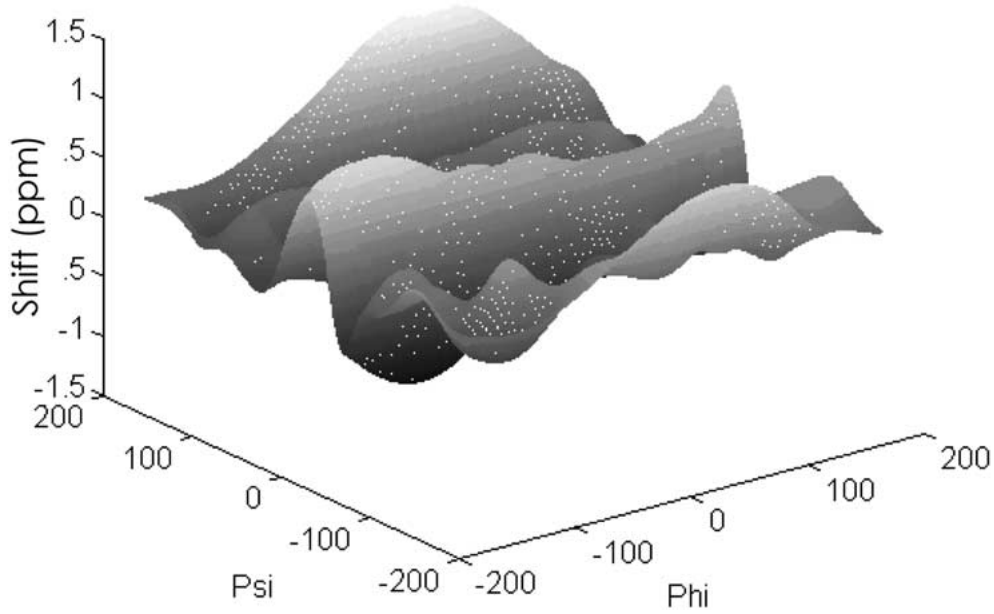


Figure 1. An example hypersurface, in this case the phi/psi surface for alpha hydrogens. Cubic splines have been used to interpolate across regions in which no data points exist (the 'forbidden regions' of a Ramachandran plot). The irregularity of the surface highlights the difficulty of capturing the effects of the backbone angles with an analytical formula.

crete quantities (such as amino acid type) the surface was actually a simple lookup table. An example of a SHIFTX hypersurface is shown in Figure 1.

Using the above technique, MINER generated and evaluated more than 400 possible hypersurfaces, noting those that were useful in predicting chemical shifts. Furthermore, because some of the structural parameters used in these hypersurface calculations were correlated to other parameters (both structural and chemical shift parameters), it was important to eliminate these co-dependencies. Consequently a hypersurface refinement procedure was implemented in MINER to select and optimize the best set of hypersurfaces. Specifically, the refinement process involved five steps: First, for each possible pair of parameters, MINER constructed a table from the 'training' set using the procedure outlined above, and computed $\Delta\delta$ values for all residues. Secondly, a trial 'new best guess' for each residue in the test set was computed from the sum of the 'best guess' and results of the previous step. The correlation of this 'new best guess' to the observed chemical shifts was computed, as was the RMSD between the two. Third, after each possible pairing of parameters was evaluated in this way, MINER selected the pair yielding the highest correlation and/or lowest RMSD, and used the hypersurface thus generated to compute a new 'best guess'. Fourth, an iterative optimization procedure was applied to correct previous hypersurfaces, which may have 'absorbed' some of the error that rightfully 'belongs' to the new hypersurface. Finally, this procedure was repeated until no further improvement in RMSD or correlation was observed.

The above is a simplification of the actual procedure, which incorporated a number of features to preserve the statistical validity of the results, and selected against pairing factors that do not actually interact. Objections may be raised to the presence of the observed secondary structure as a parameter, when it is in fact derived from more fundamental parameters. Its inclusion here is a concession to the limitations of the two dimensional surfaces which MINER generates. The secondary structure in this case operates as a mechanism by which MINER can select a subset of the records in the database and generate a different curve or predictive rule for each such subset.

Training and testing databases

Two databases of protein structures and the corresponding chemical shifts were used as input to MINER to generate the empirical constants, torsion angle surfaces, and lookup tables used by SHIFTX. One database was for backbone ^{1}H and heteronuclear (^{13}C and ^{15}N) shifts, the other was for ^{1}H sidechain shifts. The database used to calibrate backbone ^{1}H, ^{13}C and ^{15}N shift predictions consisted of 37 diamagnetic proteins assembled from an extensive literature and BioMagResBank (Seavey et al., 1991) search. These proteins, their BMRB and PDB accession numbers as well as their resolution are shown in Table 3. In preparing this database every effort was made to find those proteins which had (a) been reliably referenced (as indicated from the literature or BMRB data), (b) spanned a variety of structural classes (all α, all β, mixed α/β), (c) were well-structured, (d) had no 'shift-significant' ligands or paramagnetic moieties, and (e) had high resolution X-ray structures ($<$2.1 Å). To prevent any instrumental or operator bias from creeping into the chemical shift calculations and the refinement process, assignment data was selected from a variety of labs. Additionally, several sets of protein shifts (esp. ^{13}C and ^{15}N) were re-referenced to conform to IUPAC recommendations (Wishart et al., 1995b; Markley et al., 1998). Note that for backbone ^{1}H calculations and comparisons, all glycine ^{1}HA shifts were averaged and treated as a single shift because of the limited information on stereospecific assignments.

The other database, used to analyze the sidechain hydrogens, was developed by searching the BMRB for chemical shift data for the nuclei of interest, and using that subset of the records for which there was a corresponding high-resolution X-ray crystal structure ($<$2.0 Å) in the Protein Data Bank (Berman et al., 2000). The selection criteria as described above was also employed for this database, however chemical shift referencing was not an issue for these shifts. The BMRB and PDB files used for the sidechain ^{1}H shift database are shown in Table 4.

The resulting databases were divided into equal-sized test and training sets, with every other residue being assigned to the test set. As a further check against overfitting, all optimization steps were evaluated by testing their results against randomly-chosen samples from the databases. Several (usually twenty) such samples would be generated and evaluated (in terms of correlation or RMSD) against the optimized surfaces. Given two hypersurfaces yielding similar correlations or RMSDs, preference was given to the surface yielding the smallest variance of correlations/RMSDs among the random samples.

Table 3. PDB and BMRB files used to calibrate backbone and heteronuclear shift predictions

PDB ID	Protein name	Resolution (Å)	BMRB accession / reference
2ALP	Alpha-lytic protease (L. enzymogenes)	1.70	(Cornilescu et al., 1999)
1GZI	Antifreeze protein (Ocean pout)	1.80	(Wishart et al., 1997)
1A6K	Myoglobin (Sperm whale)	1.10	4061
1A2P	Barnase (B. amyloliquefaciens)	1.50	975
4ICB	Calbindin D9K, minor A form (Pig)	1.60	390
1CLL	Calmodulin (Drosophila)	1.70	547
1ROP	ColE1 repressor protein (*E. coli*)	1.70	4072
1CEX	Cutinase (F. solani)	1.00	4101
3EZM	Cyanovirin-N (Nostoc ellipsosporum)	1.50	(Cornilescu et al., 1999)
2CPL	Cyclophilin-A (Human)	1.63	(Cornilescu et al., 1999)
1HCB	Carbonic anhydrase I (Human)	1.60	4022
1DMB	D-maltodextrin-binding protein (*E. coli*)	1.80	4354
1ICM	Fatty acid binding protein (Rat)	1.20	(Wishart et al., 1997)
1HFC	Fibroblast collagenase (Human)	1.56	4064
4FGF	Fibroblast growth factor (Human)	1.60	4091
1BKF	FK506 binding protein (Human)	1.60	4077
1HVR	HIV protease (HIV)	1.80	(Cornilescu et al., 1999)
4I1B	Interleukin 1β (Human)	2.00	1061
3LZT	Lysozyme (Chicken)	0.92	4562
1LZ1	Lysozyme (Human)	1.50	(Wishart et al., 1997)
1ONC	P-30 protein (Northern leopard frog)	1.70	4371
5PTI	Pancreatic trypsin inhibitor (Bovine)	1.00	46, 262, 485
1F3G	Phosphocarrier protein III glc (*E. coli*)	2.10	(Cornilescu et al., 1999)
1ACF	Profilin I (A. castellanii)	2.00	(Cornilescu et al., 1999)
1HKA	Pyrophosphokinase (*E. coli*)	1.50	4299
5P21	RAS P21 (Human)	1.35	(Wishart et al., 1997)
1RUV	Ribonuclease A (Bovine)	1.30	4031
2RN2	Ribonuclease H (*E. coli*)	1.48	(Wishart et al., 1997)
1RGE	Ribonuclease S (S. aureofaciens)	1.15	4259
2RNT	Ribonuclease T1 (Aspergillus oryzae)	1.80	(Wishart et al., 1997)
1SVN	Savinase (Bacillus lentus)	1.40	(Cornilescu et al., 1999)
1SNC	Staphylococcal nuclease (S. aureus)	1.65	(Cornilescu et al., 1999)
1MKA	Thiol ester dehydrase (*E. coli*)	2.00	(Cornilescu et al., 1999)
2TRX	Thioredoxin (*E. coli*)	1.68	(Wishart et al., 1997)
1ERT	Thioredoxin – reduced (Human)	1.70	(Cornilescu et al., 1999)
1TOP	Troponin C (Chicken)	1.78	4401
1UBQ	Ubiquitin (Human)	1.80	(Cornilescu et al., 1999)

Results and discussion

The correlation coefficients and RMS errors between the observed and calculated shifts for each atom type (^{1}H, ^{13}C, ^{15}N) as measured over all test proteins are listed in Tables 5 and 6. A more complete listing that breaks down the results in Tables 5 and 6 according to residue type and nucleus is provided at the SHIFTX web site. These tables list the correlation between the observed values with the 'best guess' values (incorporating the classical effects) and full SHIFTX predictions (which incorporate the hypersurfaces). To minimize the influence of probable mis-assignments and typographical errors in the input data, points for which the error (predicted minus observed) was greater than three standard deviations from the mean error were removed. Nuclei for which no observations were available (HH, HH11, HH12, HH21, and HH22) are not listed. Of the 24 sidechain protons for which

Table 4. PDB and BMRB files used to calibrate the sidechain hydrogen shift prediction

PDB ID	Resolution	BMRB accession	PDB ID	Resolution	BMRB accession
5PTI	1.00	48 1156 1179	1BM8	1.71	4254 4256
2SCP	2.00	4129	1PID	1.30	4266
1EPF	1.85	4162 4143	1AVS	1.75	4232
1R69	2.00	2539 195	1MJC	2.00	4296
1BRF	0.95	1991	1AIL	1.90	4317
1AAP	1.50	2024	1QST	1.70	4321
1HIP	2.00	2219	1EKG	1.80	4342
1FNF	2.00	2281	1ANF	1.67	4354
2PSP	1.95	2384	1EDH	2.00	4380
2MLT	2.00	245 606 36	1F2L	2.00	4397
4ICB	1.60	247 325	2WRP	1.65	916
1DOI	1.90	2472	1TGJ	2.00	4411
1F41	1.30	2476	1CTF	1.70	4429
1NOT	1.20	250 422 423	1AB1	0.89	4509
1IGD	1.10	2575	1AAZ	2.00	4459
1QHJ	1.90	2580	1RZL	1.60	4917
1IOB	2.00	2718 2719	1CY5	1.30	4661
3IL8	2.00	280	1BEN	1.40	554
3CHY	1.66	3440 4472	1HOE	2.00	60 1816
1CKU	1.20	2999 3000	1UBQ	1.80	68
1DUZ	1.80	3078 3079	1BYF	2.00	4782
1ET1	0.90	3427 3449 1666	1BWI	1.80	1093
1PVA	1.65	144 3471 3472	1ZNI	1.50	1444 884 883
2ZTA	1.80	371	1FTG	2.00	5011
1RSY	1.90	4039 4041	1E65	1.85	1210
1BQU	2.00	4150	2OVO	1.50	1374
1ROP	1.70	4072	1NOA	1.50	1766
1CLL	1.70	4174	1PID	1.30	4266
1CBS	1.80	4186	1TN4	1.90	1553
1ORC	1.54	4207	1RCF	1.40	1580
1MOL	1.70	4222			

Table 5. The correlation between the observed chemical shifts of the backbone atoms and SHIFTX predictions

Nucleus	Correlation (physical factors)	Correlation (all factors)	RMSD (ppm)	Number of data points
CA	0.897	0.980	0.98	4323
CB	0.990	0.996	1.10	3281
CO	0.511	0.863	1.16	3135
N	0.626	0.909	2.43	4204
H	0.557	0.741	0.49	2993
HA	0.621	0.911	0.23	4437

Table 6. The correlation between the observed chemical shifts of the sidechain protons and SHIFTX's predictions

Nucleus	Correlation (physical factors)	SHIFTX correlation (RMSD-ppm)	Number of data points	Notes
HB	0.9740	0.982 (0.21)	1172	For Ala, correlation is between the single observed value HB and the mean of the predicted values for HB1, HB2, and HB3
HB2	0.8788	0.924 (0.30)	2927	
HB3	0.8638	0.917 (0.31)	2854	
HD1	0.9906	0.992 (0.39)	987	For Leu and Ile, correlation is between the single observed value HD1 and the mean of the predicted values for HD11, HD12, HD13
HD2	0.9779	0.992 (0.33)	1216	For Leu, the correlation is between the single observed value HD2 and the mean of the predicted values for HD21, HD22, and HD23
HD21	0.2435	0.708 (0.26)	140	There appears to be some inconsistency in the nomenclature for HD21 and HD22. This was resolved by assigning the lower of the two to HD21 and the higher to HD22.
HD22	0.3278	0.801 (0.28)	140	See above
HD3	0.9277	0.944 (0.33)	471	
HE	0.9817	0.987 (0.46)	127	For Met, correlation is between the single observed valued HE and the mean of the predicted values for HE1, HE2, and HE3
HE1	0.6575	0.912 (0.51)	444	
HE2	0.9902	0.985 (0.35)	507	
HE21	0.3051	0.778 (0.24)	113	Inconsistency in the nomenclature for HE21 and HE22. This was resolved by assigning the lower to HE22 and the higher to HE21.
HE22	0.4374	0.743 (0.26)	113	See above
HE3	0.9926	0.994 (0.20)	260	
HG	0.5878	0.700 (0.26)	318	
HG1	0.6178	0.696 (0.20)	333	For Val, correlation is between the single observed valued HG1 and the mean of the predicted values for HG11, HG12, and HG13
HG12	0.3778	0.464 (0.37)	215	
HG13	0.4591	0.566 (0.35)	208	
HG2	0.9201	0.928 (0.25)	1925	For Ile, Val, and Thr, correlation is between the single observed valued HG2 and the mean of the predicted values for HG21, HG22, HG23
HG3	0.8027	0.846 (0.29)	1041	
HH2	0.5820	0.858 (0.19)	61	
HZ	0.4862	0.674 (0.40)	162	
HZ2	0.6329	0.851 (0.23)	60	
HZ3	0.4197	0.868 (0.18)	59	

data was available, 14 had correlations greater than 0.85, and all but seven had correlations greater than 0.75. Five of the seven poor performers were nuclei for which stereoscopic labeling was questionable; all seven were protons for which limited amounts of data (fewer than 350 shifts) were available.

Hypersurfaces

The utility of the hypersurfaces in improving predictions is obvious when the 'physical factors only' predictions are compared with those made using the hypersurfaces. The improvement is particularly notable in cases such as backbone ^{15}N shifts, where the correlation increased from 0.626 to 0.909 when the hypersurfaces were applied. As with all of the backbone nuclei, most of the improvement was due to the application of a hypersurface utilizing the backbone torsion angles. Other minor improvements in ^{15}N prediction were possible with the addition of hypersurfaces indexed by residue type/chi angle and predicted secondary structure/preceding residue; this latter captures the 'nearest neighbour' effect noted by earlier workers (Wishart et al., 1995a; Braun et al., 1994).

Up to 14 hypersurfaces were generated for each nucleus. Space limitations preclude a detailed listing in this paper, but the complete tables are available on the SHIFTX web site. To gain a better understanding of which factors (both physical and empirical hypersurfaces) most influenced backbone chemical shifts we have tabulated and enumerated these effects in Table 7. In this table we have identified the most dominant physical factors (electrostatic, ring current, hydrogen bond, and disulfide bond effects) along the most prominent hypersurfaces and enumerated their percent contribution to the calculated chemical shifts for each backbone nucleus. To compute these values, the 'total influence' for each shift prediction was defined to be the sum of the absolute values (in ppm) contributed by each factor, and 'relative influence' to be the ratio of the absolute value (in ppm) of each factor to the 'total influence'. These relative influences were averaged over all residues, and multiplied by one hundred to express them as a percentage in Table 7. A more complete listing of these tables (broken down according to residue type and including all hypersurfaces) is available at the SHIFTX web site.

In most cases, the effects of the backbone torsion angles are dominant. For instance, looking at the HA and CA predictions, three torsion-related factors contribute almost 60% to any given HA or CA shift. On the other hand, amide proton shifts appear to be more heavily influenced by physical factors such as hydrogen bonding and ring currents. Further inspection of this tables indicates that each nucleus is influenced somewhat differently by different components (physical factors or hypersurface factors). It is worth noting that these factors and their relative contribution for each of these backbone nuclei roughly correspond to the values originally proposed by Wishart and Case (2001).

As might be expected from statistically condensing a complex geometrical phenomena into a two dimensional table, the torsion angle hypersurface for HA (as seen in Figure 1) is quite complex. This aspect of SHIFTX – approximating an extremely complex function as a sum of simpler functions – naturally raises the question of how independent these simple components are of one another. To investigate this question further, principal component analysis (PCA) was applied to a subset of the SHIFTX training/testing data. PCA is a robust statistical technique for correlation analysis that allows one to independently and unambiguously identify the most prominent contributions to a given phenomenon or calculation. It also allows co-dependencies on certain parameters (i.e., hypersurface components or physical factors) to be identified. Shown in Table 8 is an example of a principal component analysis conducted on lysine HA chemical shift data taken from the SHIFTX training/testing database. This table lists the seven most prominent principal components (a total of 13 were tabulated) among SHIFTX parameters used in calculating lysine HA chemical shifts. Notably all but two of the 13 hypersurface and/or physical factors listed in the first column are make significant contributions (i.e., have individual weightings or loadings greater than 0.1) to these seven principal components. Looking at these results more closely, we see that the first principal component consists primarily of contributions from phi/psi or backbone dihedral angle variations. The second principal component is composed primarily of ring current contributions while the third component incorporates electric field and psi/hydrogen bond effects. These PCA loadings appear to be quite consistent with the results shown in Table 7 which demonstrate a similar weighting scheme for HA shift contributions. PCA analysis conducted on other nuclei (^{13}CA, ^{15}N, etc.) yields similar results but with different components being weighted more heavily. The key point from this analysis is that the factor space did not collapse into a small number of columns, nor did it zero-out the

Table 7. Relative influences of various factors and hypersurfaces on secondary shifts. The values shown are average percentage of influence on secondary shift for all residue types. The subscripts 'i+1' and 'i−1' indicate 'the following residue' and 'the preceding residue', respectively. AA corresponds to 'amino acid' and SS corresponds to 'secondary structure'

Factors	^{1}HA (%)	^{1}HN (%)	^{13}CA (%)	^{13}CB (%)	^{13}CO (%)	^{15}N (%)
Φ/Ψ	38.7	0	46.2	30.7	0	6.9
Ring current	13.8	14.5	4.8	0	0.7	0.2
AA/Φ	11.7	0	0	0	13.4	0
Ψ/O-Bond$_{i-1}$	10.8	0	0	0	0	0
Electric Field	4.8	3.9	1.6	0	0	0
Ψ_{i-1}/HN bond	0	25.3	0	0	0	0
H-bond	1.2	16.7	0	0	0	0
Φ_{i-1}/O-bond	0	10.4	0	0	0	5.2
Φ/Φ_{i+1}	0	7.9	0.9	0	0	0
Ψ/Ψ_{i-1}	0	0	0	0	0	29.2
χ_{i-1}/Ψ_{i-1}	0	0	4.3	0	0	15.0
AA/χ	0	0	12.8	14.3	0	13.6
AA_{i-1}/SS_{i-1}	0	0	0	0	0	11.9
Ψ/Φ_{i+1}	0	0	9.6	0	0	0
Φ/HA1-bond	0	0	5.0	0	0	0
Ψ/SS	0	0	0	13.3	0	0
Ψ/χ	0	0	0	12.0	32.1	0
χ/Ψ_{i-1}	0	0	0	0	10.7	0
Φ_{i+1}/Ψ_{i+1}	0	0	0	0	9.6	0
SS/Ψ_{i-1}	0	0	0	0	5.3	0

Table 8. Principal component analysis of the hypersurface predictive factors for lysine HA shifts. Each column is a 'synthetic' component created by PCA. The top row shows the percentage of the variance accounted for by the given component. The subscripts 'i+1' and 'i−1' indicate 'the following residue' and 'the preceding residue', respectively. AA corresponds to 'amino acid' and SS corresponds to 'secondary structure'

	Comp 1 (%)	Comp 2 (%)	Comp (%)	Comp 4 (%)	Comp 5 (%)	Comp 6 (%)	Comp 7 (%)
Percent variation	57.1393	27.4464	5.6123	4.7552	1.7781	0.8454	0.6477
Φ/Ψ	0.9608	−0.2448	0.0132	0.0485	−0.1033	0.0177	−0.0493
Ring current	0.2501	0.9673	0.0022	−0.0101	0.0307	0.0023	−0.0097
AA/Φ	0.0515	−0.0067	0.0082	0.0443	−0.0202	−0.1963	0.9161
Ψ/O-bond$_{i-1}$	−0.0668	0.0316	0.5376	0.8076	−0.2238	0.0397	−0.0329
Electric field	0.0753	−0.0487	0.3712	0.0072	0.821	−0.3981	−0.088
Φ_{i+1}/HA1-bond	0.0149	0.0024	−0.3048	0.2897	0.3687	0.2541	0.1404
SS/χ_{i+1}	0.0261	−0.0213	0.0381	0.0342	0.314	0.7626	0.0643
SS/Ψ_{i+1}	0.0103	−0.008	0.021	0.0148	0.1003	0.237	0.2784
χ/HA2-bond	0.0075	0.0128	0.0003	−0.0037	−0.0369	−0.1518	0.1847
χ_{i-1}/SS_{i-1}	−0.001	0.0138	−0.0057	−0.0002	−0.0012	−0.1134	0.0193
Ψ/Disulfide$_{i-1}$	−0.0203	0.0063	0.0041	−0.0053	−0.0357	−0.041	0.0293
Hydrogen bond	−0.0032	0.0058	0.6914	−0.5078	−0.122	0.2469	0.1127
1stRes/HA2-bond	−0.0058	0.0004	0.0031	0.0072	−0.0411	0.0118	−0.007

loadings on individual hypersurface/physical factors – both of which would indicate a strong co-dependence among the input factors. Nevertheless, these PCA results (which are quite consistent with the results shown in Table 7) indicate that most chemical shifts for most nuclei depend primarily on four or five major physical factors or hypersurface contributions. Typically the minor contributions just add a few percentage points to the overall quality of the fit.

Performance of ^{1}H predictions

Overall, ^{1}H shift predictions are seen to be highly correlated with experimentally observed shifts, especially for the side chain ^{1}H's. As expected, the weakest correlation is obtained for amide protons. These results are highly consistent with previously reported ^{1}H shift prediction methods (Williamson et al., 1992; Osapay and Case, 1991; Herranz et al., 1992). For instance the HA performance reported here of 0.911 compares favorably to the value of 0.849 reported for Osapay and Case (1991), 0.747 reported for Asakura et al. (1992) and 0.730 reported for Herranz et al. (1992). Likewise the HN results of 0.741 reported here compares well to the value of 0.575 reported for Osapay and Case (1991) and 0.711 reported for Herranz et al (1992). Similarly, the correlation coefficient of 0.907 reported here for side chain ^{1}H's compare favorably to the value of 0.899 reported by Osapay and Case (1991).

To conduct a more controlled comparison, we used the publicly available programs from Williamson (TOTAL) and Case (SHIFTS) to predict the backbone ^{1}H shifts for a selection of five proteins (PDB accessions: 193L, 4ICB, 1POH, 5PTI, and 5TNC) covering 520 residues. We obtain correlations for HA predictions of 0.796 (TOTAL) and 0.742 (SHIFTS) versus 0.917 for SHIFTX. For HN predictions we get 0.548 (TOTAL) and 0.336 (SHIFTS) versus 0.756 for SHIFTX. Clearly SHIFTX does comparatively well in proton shift calculation, however, there is still room for improvement – particularly for amide protons. This suggests that either we still have an imperfect understanding of all the contributions that lead to protein ^{1}H chemical shift variation (especially HN shifts) or that the prediction methods have reached their limit because of coordinate imprecision in the training or testing data. Our inclination is to believe it is more the latter than the former.

Performance of ^{13}C and ^{15}N predictions

Figure 2 shows scatter plots illustrating the correlation coefficient (r) for ^{13}C and ^{15}N shift predictions based on the analysis of 4323 ^{13}CA, and 4204 ^{15}N shifts collected from 37 proteins with X-ray structures having a resolution less than 2.1 Å. Overall, the predictions are seen to be highly correlated with the range of experimentally observed shifts, especially for the ^{13}CB's. As expected, the weakest correlation is obtained for ^{13}CO shifts. Earlier studies on ^{13}C and ^{15}N shift prediction are somewhat limited as they only reported results from a small or well-defined subset of residue types. For instance, de Dios et al. (1993) reported correlations of 0.97, 0.93 and 0.97 for ^{13}CA, ^{13}CB and ^{15}N shifts for ∼20 alanine and valine residues in calmodulin and staphylococcal nuclease. Wishart and Nip (1998) report correlations of 0.97 and 0.80 for the complete set of ^{13}CA and ^{15}N shifts for calmodulin (148 residues) using empirically derived chemical shift hypersurfaces relating backbone dihedral angles to ^{13}C and ^{15}N shifts. An unpublished neural network server called PROSHIFT (located at http://www.jens-meiler.de/proshift.html) has also been described, which claims to be able to predict ^{1}H, ^{13}C and ^{15}N shifts with RMSD errors of 0.22 pm, 0.98 and 2.08 ppm, respectively. However, it is not clear which ^{1}H and ^{13}C shifts were evaluated and what or how many proteins were included the training/testing data. Nevertheless, the use of neural networks appears to be a promising avenue of research for chemical shift prediction.

More recently Xu and Case (2001), reported correlations of 0.98, 0.99, 0.90 and 0.92 for ^{13}CA, ^{13}CB, ^{13}CO and ^{15}N shifts for residues (excluding His and Cys) from a set of 20 proteins found in well-defined helices and beta strands (which accounts for about 40% of all residues in proteins). While this manuscript was under review, Xu and Case (2002) published an extension to their earlier paper wherein they expanded their analysis to include residues from unstructured regions for the same 20 proteins (excluding cysteines as well as C and N terminal residues). They now report correlation coefficients of 0.97, 0.99, 0.83 and 0.90 for ^{13}CA, ^{13}CB, ^{13}CO and ^{15}N shifts with RMSD errors of 1.22, 1.31, 1.28 and 2.71 ppm, respectively. Currently their method is able to predict ^{13}C and ^{15}N chemical shifts for about 89% of all residues in diamagnetic proteins.

As a comparision, the correlation coefficients we obtained for SHIFTX predictions over a set of 37 pro-

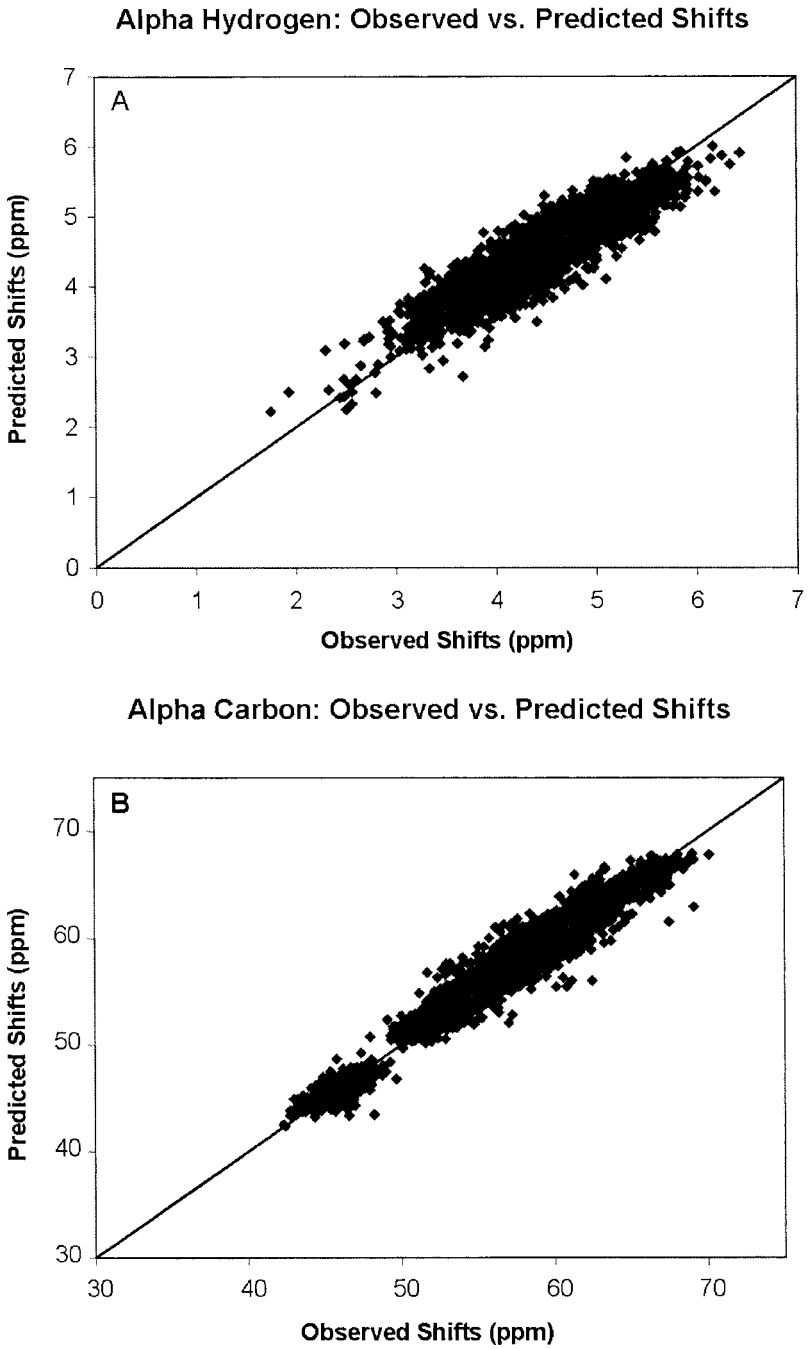
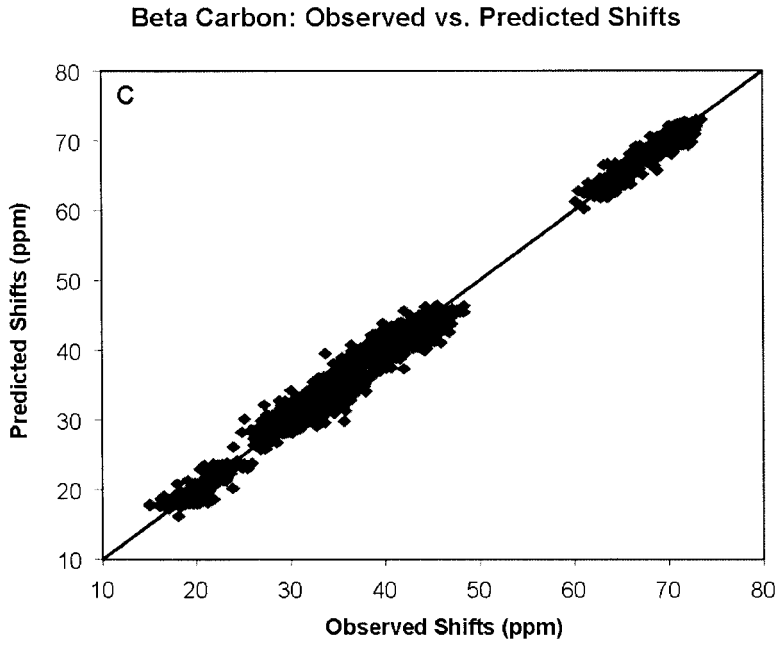


Figure 2. Scatterplots comparing observed vs. SHIFTX predicted shifts for backbone nuclei including (A) ^{1}Hα; (B) ^{13}Cα; (C) ^{13}Cβ; (D) ^{13}CO; (E) ^{15}N; and (F) ^{1}HN. A line with a slope of one is drawn in each graph for comparison purposes.



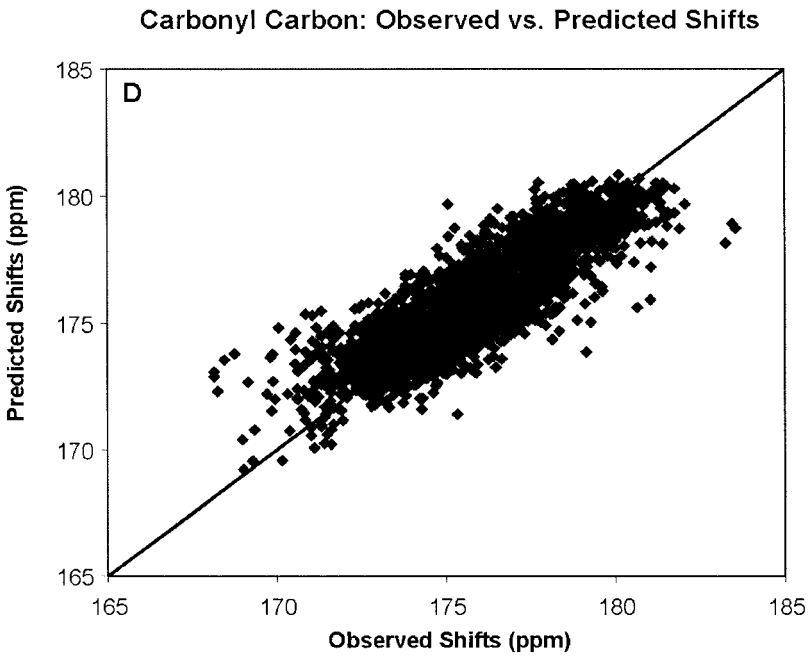
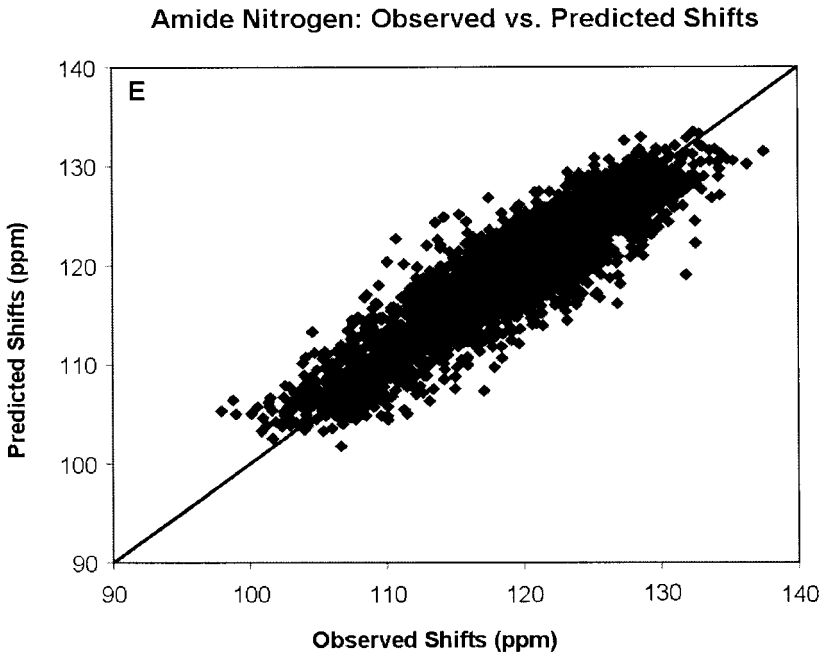


Figure 2. Continued.



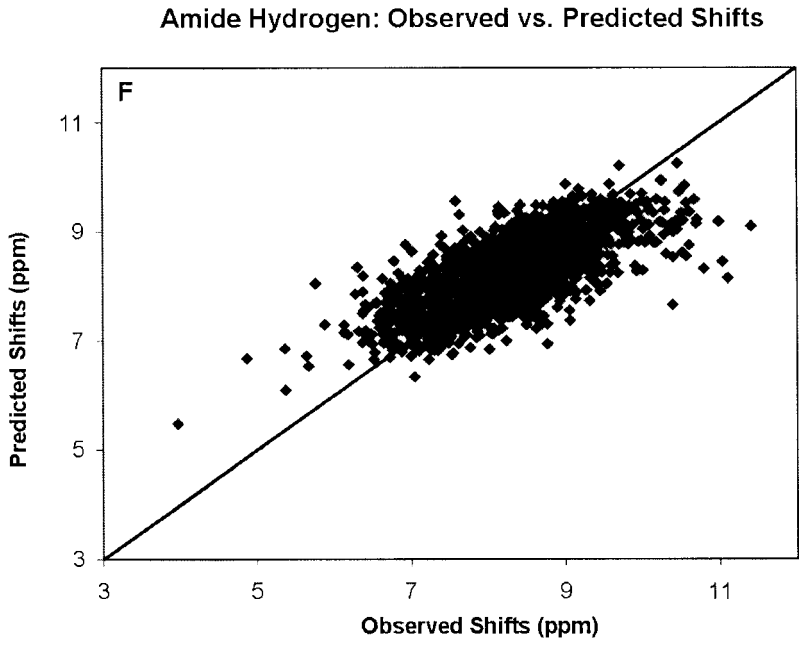


Figure 2. Continued.

teins were 0.980, 0.996, 0.863 and 0.909 for ^{13}CA, ^{13}CB, ^{13}CO and ^{15}N shifts with RMSD errors of 0.98, 1.10, 1.16 and 2.43 ppm respectively for all residues (100% coverage). Despite their differences in derivation and testing, the overall, the performance for both SHIFTX and SHIFTS (Xu and Case, 2002) for heteronuclear chemical shift calculation appears to be similar. It is possible that by combining SHIFTX and SHIFTS predictions together, one may be able to modestly improve the overall quality of heteronuclear chemical shift calculations. Efforts are underway in our laboratory to investigate this possibility.

The key point we would like to make here is that the ^{1}H, ^{13}C and ^{15}N shifts calculated by SHIFTX were for all residues, in all conformations and for all proteins in the test set. No prior geometrical optimization or energy minimization were performed on the input PDB file and the calculations for all ^{1}H, ^{13}C and ^{15}N shifts for each protein were done in less than 1 CPU second. We believe these performance characteristics are essential for using the program in practical situations pertaining to structure refinement, protein assignment, assignment validation and structure validation.

Secondary validation

A frequent complaint about many predictive methods is that they work well on the training and test data, but fail miserably when tried on a 'novel' set of previously unseen data. This is a manifestation of the all-too-familiar problem of over-training. As a further check on the validity and generality of SHIFTX we applied the program to predicting the ^{1}H, ^{13}C and/or ^{15}N chemical shifts of a set of comparable high resolution X-ray structures which were not included in the original SHIFTX test/training set. Ten such proteins (Table 9) were identified. The correlation and RMSDs between the experimental and SHIFTX predicted shifts for all backbone nuclei are shown in Table 10, along with the corresponding values for the training data. The results for the novel proteins are clearly comparable to those for the training data, and in the case of the amide hydrogens, the correlation is actually better for the previously-unseen proteins. We believe that these results adequately demonstrate the generality of SHIFTX, which is to say that the methods and formulas used by SHIFTX to make predictions are valid outside of the proteins used to arrive at those methods and formulas.

Applications

Chemical shift calculations, if sufficiently accurate, can have a wide range of practical applications in protein NMR. These include (1) aiding chemical shift assignment; (2) chemical shift assignment validation; (3) chemical shift reference checking; (4) structure refinement; (5) structure evaluation; and (6) structure generation. For instance, Williamson et al. (1995) have shown how ^{1}H shift predictions based on the structure of the G domain of protein B1 were clearly able to identify two experimental mis-assignments. This paper also demonstrated that there is a good correlation between the accuracy of predicted ^{1}H shifts and the resolution of the corresponding structure.. Subsequent studies by Kuszewski et al. (1995a, b) and Osapay et al. (1994) showed that chemical shift refinement using ^{1}H, and then later, ^{13}C shift 'calculators' could be used to improve the quality of NMR-derived protein structures.

Here we wish to demonstrate how SHIFTX can be used in three specific applications: (a) Chemical shift reference checking; (b) chemical shift assignment validation and (c) structure evaluation. The first application concerns a key issue in heteronuclear NMR – that is the inconsistency of chemical shift referencing for ^{13}C and ^{15}N nuclei. This subject has been discussed extensively in a number of recent reviews (Wishart and Case, 2001; Wishart and Sykes, 1994). While clearly defined protocols do exist (Wishart et al., 1995b; Markley et al., 1998), it is estimated that nearly 25% of newly reported protein ^{13}C and ^{15}N shifts are improperly referenced (Zhang et al., 2003). A key issue is how to identify and correct these chemical shift referencing errors. Here we show that by using the 3D structure (either X-ray or NMR) of the protein of interest and calculating its ^{13}C or ^{15}N shifts with SHIFTX it is relatively easy to detect and correct these referencing errors. Illustrated in Figure 3 is a plot of the observed versus SHIFTX-calculated ^{13}CA shifts for ribonuclease H (BMRB-1657; PDB 2RN2). A best fit line (solid line) drawn through the scatter plot shows that the ^{13}CA shifts have been systematically shifted upfield by 1.6 ppm, relative to their properly referenced values (dashed line). The same kind of plot could be generated for ^{1}H and ^{15}N chemical shifts to calculate their offset values too. Simple scatter plots such as these could certainly help NMR spectroscopists identify and correct referencing errors – as long as an appropriate 3D structure is available. Alternately, by simply calculating the difference between

Table 9. Proteins used to validate SHIFTX performance on previously unseen data

Name	PDB file	Chain #	Resolution (Å)	BMRB accession
Bucandin (B. candidus)	1F94	A	0.97	5097
Ribonucleoprotein A1 (HNRP-Human)	1L3K	A	1.10	4084
T cell transduction protein (Human)	1D4T	A	1.10	5211
Plastocyanin (Poplar)	1PLC		1.33	4019
Beta-defensin-2 (Human)	1FD3	D	1.35	4642
Anitfreeze protein (T. molitor)	1EZG	B	1.40	5323
HPt domain of ArcB (*E. coli*)	2A0B		1.57	4857
Retinoic acid binding protein (Human)	1CBS		1.80	4186
Glur2 ligand binding core (Rat)	1M5E	A	1.46	5182
IIB Cellobiose (*E. coli*)	1IIB	B	1.80	4955

Table 10. The correlation, RMSD, and number of observed shifts for ten proteins not used in the training data, and the corresponding values for the training data

Nucleus	Correlation	Validation set		Correlation	Training set	
		RMSD error (ppm)	# Observed shifts		RMSD error (ppm)	# Observed shifts
HA	0.895	0.26	748	0.911	0.23	4437
H	0.746	0.52	1001	0.741	0.49	2993
N	0.901	2.53	904	0.909	2.43	4204
CA	0.979	1.02	866	0.980	0.98	4323
CB	0.996	1.10	778	0.996	1.10	3281
CO	0.856	1.17	758	0.863	1.16	3135

the observed shifts and SHIFTX-calculated shifts for a given nucleus and then averaging these differences, it is also possible to determine the required chemical shift offset. This is precisely what is done in the reference correction program SHIFTCOR (Zhang et al., 2003), available at http://redpoll.pharmacy.ualberta.ca

Another obvious application of SHIFTX is in chemical shift assignment checking or assignment validation. As anyone who has assigned spectra from larger proteins knows, there is always some uncertainty in the correctness of the chemical shift assignments. Many of these are corrected during the course of subsequent structure determination steps, but some (particularly ^{13}C and ^{15}N shift errors) may not be so easily detected. Shown in Table 11 is a list of chemical shift assignments for a small thioredoxin-like protein from Methanobaterium thermoautotrophicum, (Mt0807) which is being studied in our laboratory. Listed in this table are three sets of assignments for ^{15}N, ^{13}CA, ^{13}CB and ^{13}CO shifts for two regions (the N terminus, near the active site, and the C terminus). The first set corresponds to the initial set of assignments which was used in the subsequent structure generation. The second set corresponds to SHIFTX-calculated shifts generated from the initial Mt0807 structure. The last set of shifts corresponds to the corrected, final set of shifts made after a series of manual and automated checks and corrections. Highlighted in bold are the changes that were made, partly as a consequence of the shift predictions generated by SHIFTX. Inspection of this table will indicate that SHIFTX was able to flag a number of residues with large discrepancies between observed and predicted shift values. Furthermore, in many cases the final 'corrected' shifts ended up being much closer to the SHIFTX-calculated shifts. In this way SHIFTX was able to guide the assignment process and clarify a number of assignment ambiguities. Certainly, if a previously existing homologous structure or even an exact X-ray structure is available, it should be possible to use SHIFTX predictions in a similar manner to help guide even the initial assignment process.

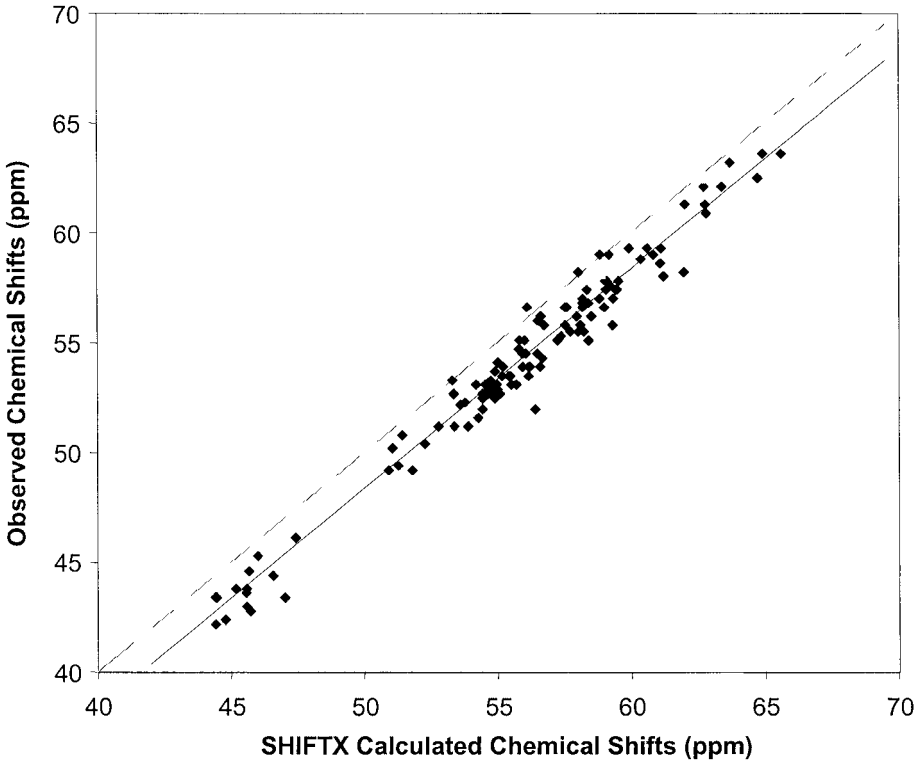


Figure 3. Plot of the observed vs. SHIFTX predicted $^{13}C\alpha$ shift of ribonuclease H (BMRB1657). A dashed line is drawn with slope 1 and intercept of 0 to indicate how the best-fit line is offset by 1.63 ppm.

As a final example of an application of SHIFTX, we wish to show how chemical shift calculations can be used to evaluate the relative quality of X-ray and NMR protein structures. As has been pointed out by a number of authors (Williamson et al., 1995; Kuszewski et al., 1995a; Laskowski et al., 1996), even the best NMR structures do not appear to have the same quality or effective resolution as most X-ray structures. While the addition of more global restraints (i.e., residual dipolar couplings) or more precisely measured constraints (J-couplings) is certainly helping the situation, there is still a long way to go. The discrepancy between the quality of NMR structures versus X-ray structures can be made particularly evident if we plot out the correlation coefficient between SHIFTX-predicted and observed chemical shifts for ^{1}HA, ^{1}HN, ^{15}N and ^{13}CA nuclei for X-ray structures of varying resolution. These plots, which cover 123 (^{13}CA) to 157 (^{1}HN) proteins each and which exclude most proteins from the SHIFTX training set, are illustrated in Figure 4. As can be seen in all four plots, there is a modest, but obvious trend ($r \sim 0.6$) showing that the agreement between SHIFTX-predicted and observed shifts falls with decreasing resolution. A best-fit trend line is drawn through each of the four distributions. Now if we take all the NMR structures (ranging from 149 structures with ^{13}CA shifts to 229 structures with ^{1}HA shifts) and calculate their average correlation coefficient for each of the four nuclei we find that, on average, NMR structures exhibit relatively poor agreement between SHIFTX-calculated and observed chemical shifts. In fact, if we place a large dot (corresponding to the respective average NMR correlation coefficients) on the trend lines shown in Figure 4, we can extrapolate that the average protein NMR structure is equivalent to an X-ray structure of 3.0 to 3.5 Å resolution! This resolution estimate is slightly worse than what has been suggested by other authors using different approaches (Williamson et al., 1995; Doreleijers et al., 1998; Laskowski et al., 1996), but given that we are effectively using four independent shift measures, our conclusion appears to be sound.

Interestingly, of all the shifts studied (including, side chain ^{1}H, ^{13}CB and ^{13}CO) the best indicators of structure quality appear to be ^{1}HN and ^{15}N shifts.

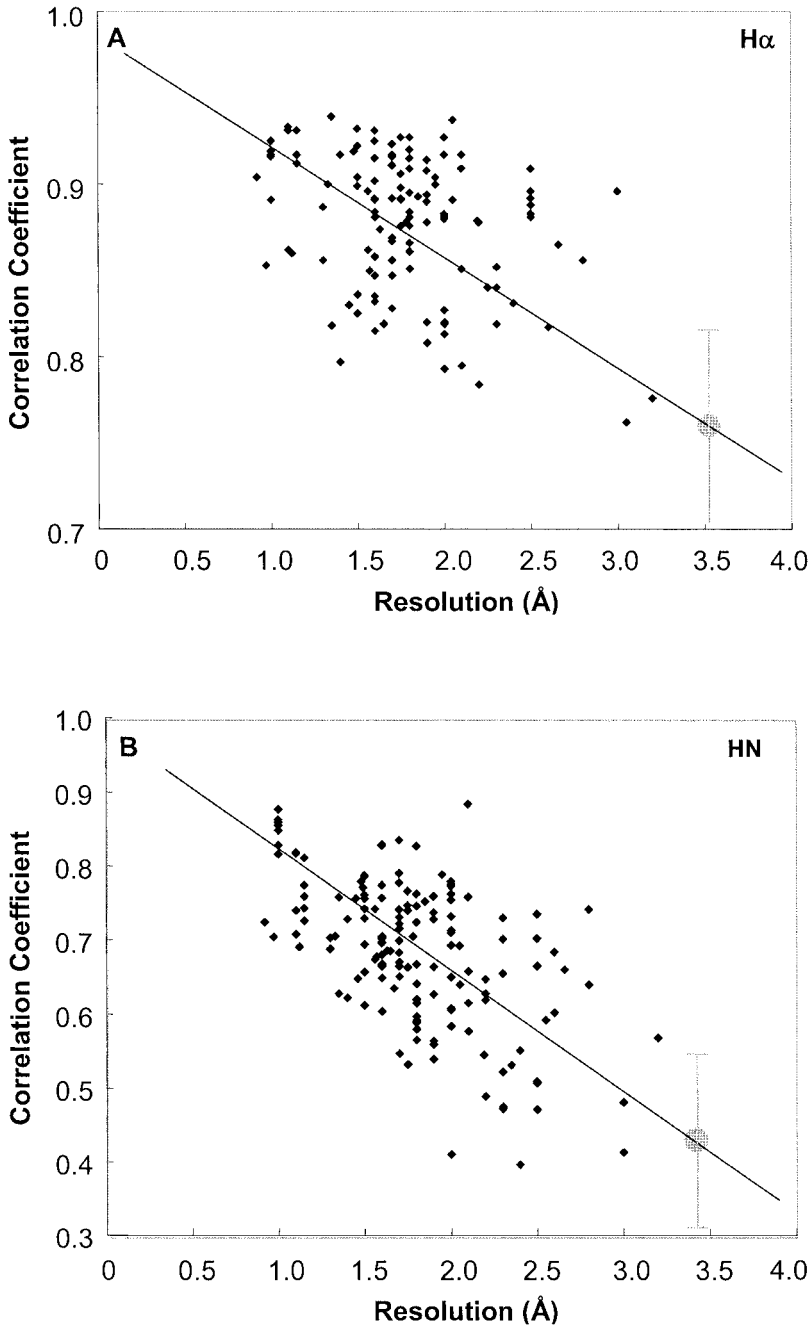
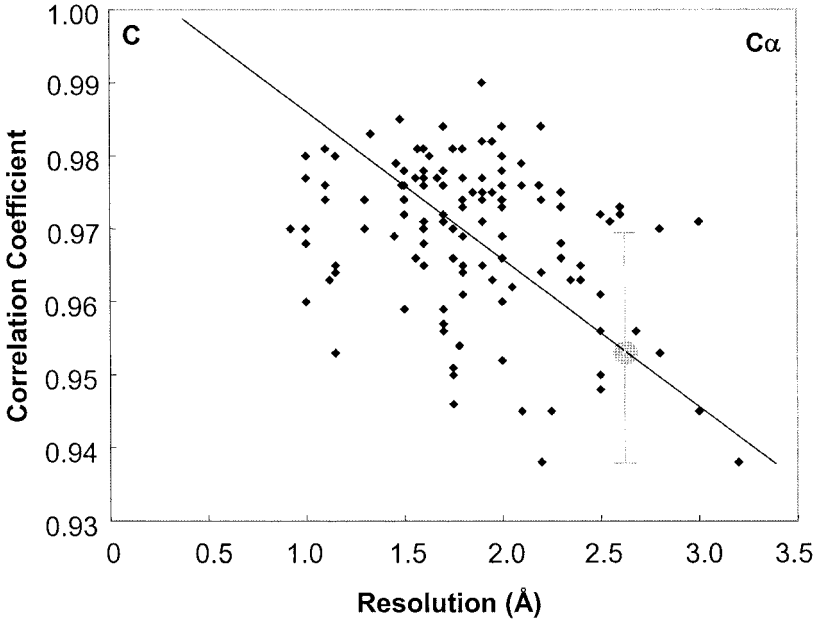


Figure 4. Scatter plots of the resolution vs. correlation coefficient for the backbone nuclei corresponding to (A) ^{1}Hα; (B) ^{1}HN; (C) ^{13}Cα; and (D) ^{15}N. A best-fit line has been drawn through each graph. The large dot (light gray) with the error bars shown in each graph indicates the position that an average NMR structure would sit on this best-fit line.



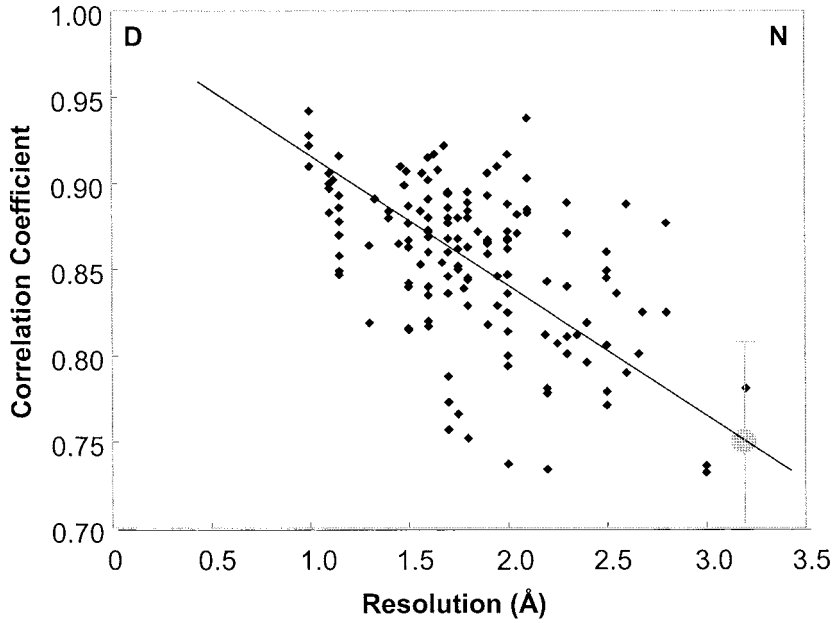


Figure 4. Continued.

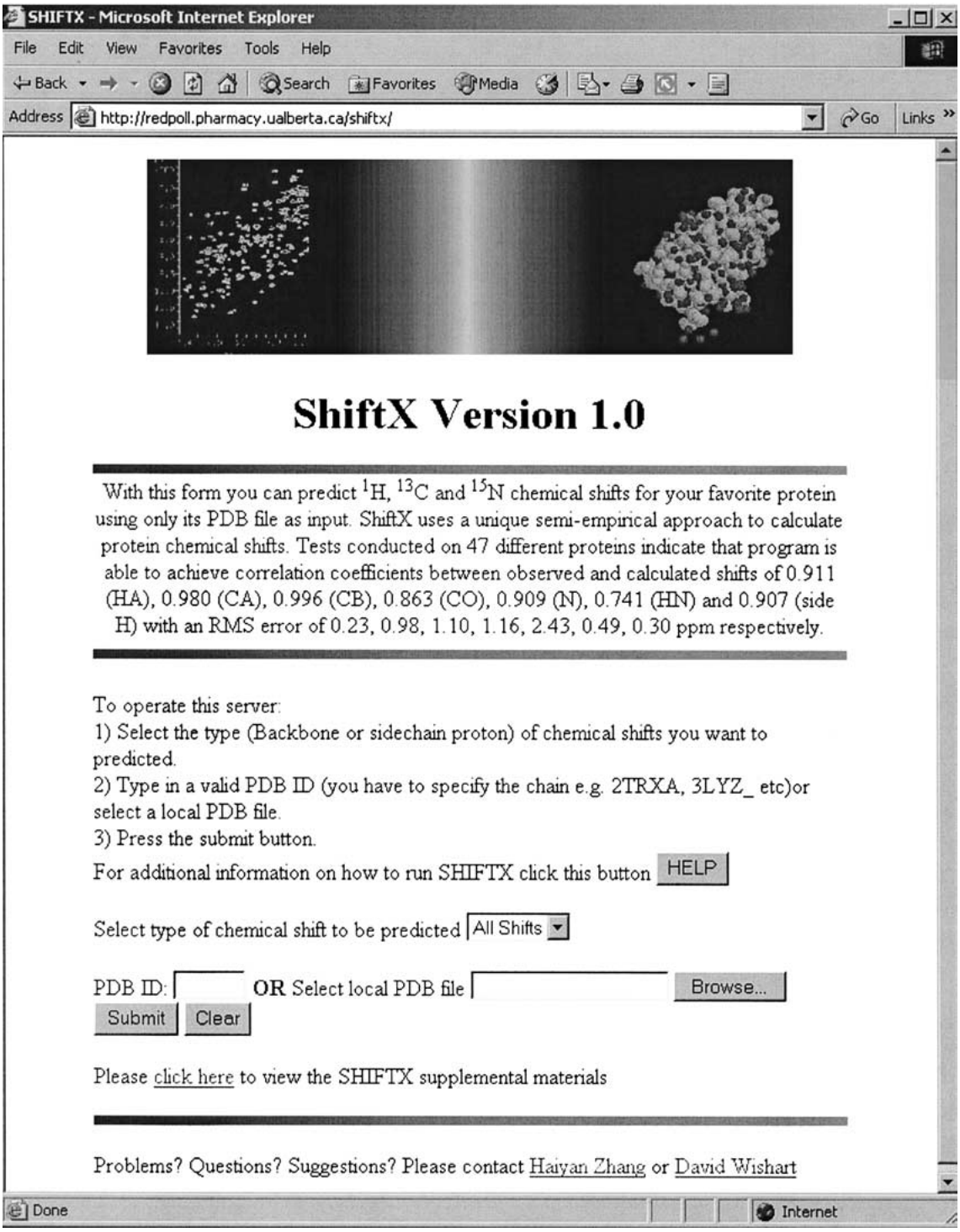


Figure 5. A screen shot of the SHIFTX web server.

Table 11. Comparison between initial assignments (Old), SHIFTX calculated assignments (Shiftx) and final assignments (New) for Mt0807. Sequence numbering begins from the leader peptide tag. Assignment changes precipitated by SHIFTX are highlighted in **BOLD**

Seq	^{15}N Old	^{15}N Shiftx	^{15}N New	^{13}CA Old	^{13}CA Shiftx	^{13}CA New	^{13}CB Old	^{13}CB Shiftx	^{13}CB New	^{13}CO Old	^{13}CO Shiftx	^{13}CO New
M 11	115.9	122.0	**121.3**	55.9	56.2	55.9	33.0	32.6	33.0	177.1	175.3	**176.1**
V 12	120.9	127.3	120.9	61.3	62.0	61.3	34.3	28.0	34.3	175.5	173.7	175.5
V 13	126.3	119.3	126.3	62.4	61.4	62.4	32.8	34.4	32.8	175.3	175.3	175.3
N 14	127.0	126.9	127.0	53.9	50.9	53.9	40.5	41.8	40.5	174.7	174.5	174.7
I 15	127.1	129.1	127.1	59.6	59.4	59.6	39.3	41.3	39.3	174.4	174.5	175.4
E 16	127.4	126.4	127.4	54.3	54.4	54.3	34.0	31.4	34.0	176.1	174.9	176.1
V 17	122.4	120.8	122.4	59.6	59.5	59.6	33.9	33.4	33.9	174.4	174.5	174.4
F 18	126.7	128.3	126.7	57.3	56.4	57.3	45.5	42.3	**40.9**	174.7	175.2	**175.5**
T 19	113.6	120.6	**110.9**	57.1	63.4	**59.6**	65.1	72.4	**68..5**	175.6	172.5	**176.7**
S 20	122.4	114.6	**113.6**	59.9	56.9	**56.8**	65.1	64.0	**65.3**	175.0	171.0	**172.0**
P 21	–	–	–	–	63.3	–	–	30.8	–	–	176.5	–
T 22	118.2	115.6	**115.1**	58.4	63.4	58.4	63.9	69.3	63.9	177.6	173.9	**174.7**
C 23	116.3	117.7	**124.8**	57.0	60.7	57.0	30.0	30.7	30.0	174.8	172.8	**177.1**
P 24	–	–	–	–	60.6	57.3	–	32.4	–	–	175.9	–
Y 25	112.0	122.3	112.0	59.6	61.9	59.6	40.9	37.0	40.9	174.2	174.4	174.2
C 26	126.7	114.8	**116.7**	57.3	60.1	**56.1**	45.5	29.6	**30.0**	174.7	171.7	**177.9**
S 82	115.9	115.1	**114.0**	58.4	58.7	**57.1**	63.9	62.0	**65.1**	176.3	175.2	**175.6**
R 83	119.7	119.1	**122.4**	55.4	59.2	**59.8**	29.2	28.8	**29.8**	179.0	178.6	**177.9**
E 84	121.3	118.8	**115.9**	55.7	59.2	**60.6**	33.0	29.2	**28.8**	178.9	178.5	**177.1**
E 85	118.6	117.1	118.6	58.9	59.9	58.9	30.0	29.2	30.0	179.6	179.6	179.6
L 86	120.5	117.3	120.5	57.9	57.8	57.9	40.9	41.3	40.9	179.8	178.9	179.8
F 87	119.8	120.1	119.8	60.5	60.9	60.5	36.7	38.4	36.7	180.4	176.8	180.4
E 88	119.0	120.8	119.0	59.8	58.4	59.8	29.5	29.7	29.5	177.1	179.0	178.7
A 89	120.1	121.8	120.1	54.9	54.9	54.9	18.3	17.9	18.3	177.8	179.6	177.8
I 90	117.8	119.0	117.8	66.2	64.4	66.2	37.9	37.0	37.9	180.4	178.0	180.4
N 91	112.0	117.0	**117.8**	57.8	55.6	**56.9**	37.4	38.6	**38.8**	178.2	176.3	**178.2**
D 92	120.9	119.4	120.9	51.8	56.4	51.8	42.7	40.9	42.7	177.2	178.1	177.2
E 93	124.0	120.3	124.0	59.9	58.6	59.9	32.6	29.3	32.5	176.6	178.6	**178.5**
M 94	116.6	121.3	**125.1**	56.1	57.9	**57.1**	30.4	33.0	**29.9**	177.9	176.4	**177.0**
E 95	118.2	122.4	118.2	56.0	56.7	56.0	33.9	30.4	33.9	176.9	176.2	176.9

These chemical shifts are exquisitely sensitive to H-bonds, aromatic rings, nearby charges and side chain torsion angles (Wishart and Case, 2001). Typically, such subtle structural parameters are not easily captured or measured by traditional NOE measurements, whereas they are more identifiable or at least refineable through higher resolution (<2.0 Å) X-ray crystallography.

What these data help illustrate is the enormous potential that chemical shifts could have in structure refinement. If NMR structures could be fully refined using at least some or even all of their ^{1}H, ^{13}C and ^{15}N chemical shifts, then as these graphs suggest, it does not seem unreasonable to expect that NMR structures could one day match the highest quality X-ray structures – both in terms of their effective resolution and in terms of their structural 'correctness'.

Limitations

SHIFTX is not yet capable of predicting all shifts for all nuclei in all proteins. For instance, some rarely measured ^{1}H shifts are not particularly well predicted (HG12 and HE21). This result is most likely due to poor statistical data (needed to model the hypersurfaces) or to atom mislabelling. Likewise SHIFTX does not predict side chain (i.e., beyond CB) ^{13}C and ^{15}N

shifts. However, these shifts are infrequently measured or reported. Furthermore, they tend not to differ substantively from the random coil values reported previously (Wishart et al., 1995a). A more serious limitation for SHIFTX, however, is the fact that the program does not calculate paramagnetic effects, nor does it account for the presence of organic ligands (heme rings, aromatic substrates, etc.). Parameters and formulae do exist to account for many of these effects (particularly for ^{1}H shifts) for common ligands or metals (Osapay and Case, 1991; Banci et al., 1997; Wishart and Case, 2001) and efforts are underway to incorporate these into the next release of SHIFTX. The inclusion of rare or unique organic ligands (i.e., drug leads, specially developed inhibitors, etc.) will present some challenges and so the parameterization of their effects will likely only be crudely approximated.

Unlike QM approaches, SHIFTX is not particularly sensitive to bond lengths, non-torsional bond angles, bond hybridization state or partial charge distribution. This is both good and bad. At one level, by ignoring these difficult-to-observe effects, it is possible to take unmodified PDB coordinate files and use SHIFTX to predict chemical shifts quite accurately. On the other hand, chemical shifts are exquisitely sensitive to very small coordinate errors and as other workers have shown (Pearson et al., 1997; Le et al., 1995), when protein structures are 'regularized' or minimized to yield optimal covalent geometry, the agreement between QM calculated shifts and observed chemical shifts is substantially improved. While we have shown that SHIFTX is also quite sensitive to the quality of protein structures, this is most likely reflects the quality or precision of non-covalent effects (ring placement, H-bond lengths, torsion angles) rather than in covalent effects such as bond lengths and bond angles. This suggests that the use of QM methods such as those of Xu and Case (2001) or Le et al. (1995) could be of greater assistance in the very detailed refinement of protein structures.

In its present form SHIFTX is not capable of including chemical shift corrections arising from temperature effects, solvent pH effects, local variations in side chain pKa values or isotope effects. Temperature effects are most significant for amide protons and are frequently used to assess H-bond status (Baxter et al., 1997). Temperature does not appear to have a significant effect on other ^{1}H, ^{13}C and ^{15}N shifts. Efforts are underway to include a simple temperature correction term for amide protons in the next release of SHIFTX. In addition to temperature, variations in solvent pH can play a significant role in the quality of chemical shift predictions. This is particularly true for histidine and somewhat less so for other charged amino acids (aspartic acid, glutamic acid, lysine and arginine). It is likely that solvent pH affects the shifts of serine, threonine, cysteine and tyrosine as well. One barrier in modeling these effects has been the difficulty in obtaining reliable solvent pH values for both NMR and X-ray samples. Frequently NMR protein samples are assigned over a range of different pH's, solvents (H_2O, D_2O, DMSO-water) or temperatures and the reported shifts represent either an average shift or a set of heterogeneous shifts collected under different conditions. This makes it difficult to accurately discern clear pH trends in chemical shifts. On the other hand, for X-ray samples, the true pH of a protein crystal is often difficult to ascertain and is rarely reported in the PDB file. Additionally, the differences between X-ray structures (solved at one pH) and NMR assignments (collected at another pH) further complicate the situation. Modelling solvent pH effects is made even more difficult by the fact that amino acids in proteins will frequently have substantially different pKa values (and titration curves) than free amino acids or unstructured peptides. Given that the prediction of side chain pKa's in proteins is still a difficult computational problem (Gibas and Subramanian, 1996) and given the previously mentioned difficulties in ascertaining accurate pH values, we have chosen to ignore pH effects in this version of SHIFTX. Overall, the inclusion of pH and pKa effects in SHIFTX will require substantially more software development and careful re-measurement of many previously reported shifts under more defined conditions.

Isotope effects can and do affect ^{1}H, ^{13}C and ^{15}N chemical shifts in proteins. Currently SHIFTX does not include deuterium isotope effects in predicting ^{13}C and ^{15}N shifts (Bjorndahl et al., 2001; Gardner et al., 1997). These corrections (about 0.43 ppm for ^{13}CA, 0.82 ppm for ^{13}CB and 0.23 ppm for ^{15}N) will soon be added as an option to the SHIFTX server.

Another limitation of SHIFTX arises from its use of pre-defined sequence and structure parameters. These sequence/structure parameters were originally chosen in 1998 based on their previously reported correlations to chemical shifts (Wishart and Nip, 1998; Osapay and Case, 1991; de Dios et al., 1993). However, since that time, additional properties – including CO and CN bond lengths, backbone omega values, intrapeptide NH to CO bond distances, etc. have been found to have some impact on measured chemical

shifts (Xu and Case, 2001). Hopefully the inclusion of these structural parameters and their corresponding hypersurfaces in future releases of SHIFTX should improve its overall performance.

The fact that SHIFTX uses a hybrid approach drawing on closed form analytical expressions in conjunction with empirical hypersurfaces and look-up tables makes the method somewhat less elegant than QM methods or pure 'classical' approaches. It also makes SHIFTX less amenable to incorporation into more standard structure refinement packages, such as AMBER or XPLOR, which rely on having smoothly differentiable functions for conjugate gradient or Newton Raphson minimization. However, we have found that chemical shift optimization using SHIFTX can be done relatively easily and effectively using a Simplex minimizer (which doesn't require derivatives) or through Genetic algorithms or Monte Carlo searches in torsion angle space. A report on this work will be forthcoming shortly.

Availability

SHIFTX is available as a web server at http://redpoll.pharmacy.ualberta.ca. A screen shot of the web server is shown in Figure 5. Unsupported copies of the SHIFTX source and/or binary code (written in C, compiled on Linux, Solaris, Irix, and Win32) may be obtained on request from the authors.

Acknowledgements

We with to thank Godwin Amegbey for providing his assignment data on Mt0807 in advance of publication. Financial support by the Natural Sciences and Engineering Research Council (NSERC), Bristol–Myers Squibb and by the Protein Engineering Network of Centres of Excellence (PENCE Inc.) is gratefully acknowledged.

References

Baker, E.N. and Hubbard, R.E. (1984) *Prog. Biophys. Mol. Biol.*, **44**, 97–179.

Banci, L., Bertini, I., Savellini, G.G., Romagnoli, A., Turano, P., Cremonini, M.A., Luchinat, C. and Gray, H.B. (1997) *Prot. Struct. Funct. Gen.*, **29**, 68–76.

Baxter, N.J. and Williamson, M.P. (1997) *J. Biomol. NMR*, **9**, 359–369.

Beger, R.D. and Bolton, P.H. (1997) *J. Biomol. NMR*, **10**, 129–142.

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

Bjorndahl, T.C., Watson, M.S., Slupsky, C.M., Spyracopolous, L., Sykes, B.D. and Wishart, D.S. (2001) *J. Biomol. NMR*, **19**, 187–188.

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

Buckingham, A.D. (1960) *Can. J. Chem.*, **38**, 300–307.

Case, D.A. (1998) *Curr. Opin. Struct. Biol.*, **8**, 624–630.

Case, D.A. (2000) *Curr. Opin. Struct. Biol.*, **10**, 197–203.

Cornilescu, G., Delaglio, F. and Bax, A. (1999) *J. Biomol. NMR*, **13,** 289–302

Dalgarno, D.C., Levine, B.A. and Williams, R.J.P. (1983) *Biosci. Rep.,* **3,** 443–452.

de Dios, A.C., Pearson, J.G. and Oldfield, E. (1993) *Science*, **260,** 1491–1496.

Derewenda, Z.S., Lee, L. and Derewenda, U. (1995) *J. Mol. Biol.* **252**, 248–262.

Doreleijers, J.F., Rullmann, J.A. and Kaptein, R. (1998) *J. Mol. Biol.*, **281**, 149–164.

Gardner, K.H., Rosen, M.K. and Kay, L.E. (1997) *Biochemistry*, **36**, 1389–1401.

Gibas, C.J. and Subramanian, S. (1996) *Biophys. J.* **71**, 130–147.

Haigh, C.W. and Mallion, R.B. (1980) *Progr. NMR Spectrosc.*, **13**, 303–344.

Herranz, J., Gonzalez, C., Rico, M., Nieto, J.L., Santoro, J., Jimenez, M.A., Bruix, M., Neira, J.L. and Blanco, F.J. (1992) *Magn. Reson. Chem.*, **30**, 1012–1018.

Iwadate, M., Asakura, T. and Williamson, M.P. (1999) *J. Biomol. NMR*, **13**, 199–211.

Kabsch, W. and Sander, C. (1983) *Biopolymers*, **22**, 2577–2637.

Kuszewski, J., Qin, J., Gronenborn, A.M. and Clore, G.M. (1995a) *J. Magn. Reson. B*., **106,** 92–96.

Kuszewski, J., Gronenborn, A.M. and Clore, G.M. (1995b) *J. Magn. Reson.*, **B107**, 293–297.

Laskowski, R.A., Rullmannn, J.A., MacArthur, M.W., Kaptein, R. and Thornton, J.M. (1996) *J. Biomol. NMR*, **8**, 477–486.

Le, H. and Oldfield, E. (1994) *J. Biomol. NMR*, **4**, 341–348.

Le, H., Pearson, J.G., de Dios, A.C. and Oldfield, E. (1995). *J. Am. Chem. Soc*. **117**, 3800–3807.

Markley, J.L., Bax, A., Arata, Y., Hilbers, C.W., Kaptein, R., Sykes, B.D., Wright, P.E. and Wüthrich, K. (1998) *J. Biomol. NMR*, **12**, 1–23.

Osapay, K. and Case, D.A. (1991) *J. Am. Chem. Soc.*, **113**, 9436–9444.

Osapay, K. and Case, D.A. (1994) *J. Biomol. NMR*, **4**, 215–230.

Osapay, K., Theriault, Y., Wright, P.E. and Case, D.A. (1994) *J. Mol. Biol.*, **244**, 183–197.

Pearson, J.T., Le, H., Sanders, L.K., Godbout, N., Havlin, R.H. and Oldfield, E. (1997) *J. Am. Chem. Soc.* **119**, 11941–11950.

Seavey, B.R., Farr, E.A., Westler, W.M. and Markley, J.L. (1991) *J. Biomol. NMR*, **1**, 217–236.

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

Wagner, G., Pardi, A. and Wüthrich, K. (1983) *J. Am. Chem. Soc.*, **105**, 5948.

Williamson, M.P. and Asakura, T. (1997) *Meth. Mol. Biol.*, **60**, 53–69.

Williamson, M.P., Asakura, T., Nakamura, E. and Demura, M. (1992) *J. Biomol. NMR*, **2**, 93–98.

Williamson, M.P., Kikuchi, J. and Asakura, T. (1995) *J. Mol. Biol.*, **247**, 541–546.

Wishart, D.S. and Case, D.A. (2001) *Meth. Enzymol.*, **338**, 3–34.

Wishart, D.S. and Nip, A.M. (1998) *Biochem. Cell Biol.*, **76**, 153–163.
Wishart, D.S. and Sykes, B.D. (1994) *Meth. Enzymol.*, **239**, 363–392.
Wishart, D.S., Bigam, C.G., Holm, A., Hodges, R.S. and Sykes, B.D. (1995a) *J. Biomol. NMR*, **5**, 67–81.
Wishart, D.S., Bigam, C.G., Yao, J., Abildgaard, F., Dyson, H.J., Oldfield, E., Markley, J.L. and Sykes, B.D. (1995b) *J. Biomol. NMR*, **6**, 135–140.
Wishart, D.S., Sykes, B.D. and Richards, F.M. (1991) *J. Mol. Biol.*, **222**, 311–333.
Wishart, D.S., Willard, L., Richards, F.M. and Sykes, B.D. (1994) VADAR: A comprehensive program for protein structure evaluation. Version 1.2. Edmonton, Alberta, Canada.
Word, J.M., Lovell, S.C., Richardson, J.S. and Richardson, D.C. (1999) *J. Mol. Biol.*, **285**, 1733–1747.
Xu, X-P. and Case, D.A. (2001) *J. Biomol. NMR*, **21**, 321–333.
Xu, X-P. and Case, D.A. (2002) *Biopolymers*, **65**, 408–423.
Zhang, H., Neal, S. and Wishart, D.S. (2003) *J. Biomol. NMR*, **25**, 173–195.