Journal of Biomolecular NMR, **21:** 321–333, 2001.
KLUWER/ESCOM

Automated prediction of ^{15}N, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ chemical shifts in proteins using a density functional database

Xiao-Ping Xu & David A. Case*
Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037, U.S.A.

Received 10 July 2001; Accepted 13 September 2001

Abstract

A database of peptide chemical shifts, computed at the density functional level, has been used to develop an algorithm for prediction of ^{15}N and ^{13}C shifts in proteins from their structure; the method is incorporated into a program called SHIFTS (version 4.0). The database was built from the calculated chemical shift patterns of 1335 peptides whose backbone torsion angles are limited to areas of the Ramachandran map around helical and sheet configurations. For each tripeptide in these regions of regular secondary structure (which constitute about 40% of residues in globular proteins) SHIFTS also consults the database for information about sidechain torsion angle effects for the residue of interest and for the preceding residue, and estimates hydrogen bonding effects through an empirical formula that is also based on density functional calculations on peptides. The program optionally searches for alternate side-chain torsion angles that could significantly improve agreement between calculated and observed shifts. The application of the program on 20 proteins shows good consistency with experimental data, with correlation coefficients of 0.92, 0.98, 0.99 and 0.90 and r.m.s. deviations of 1.94, 0.97, 1.05, and 1.08 ppm for ^{15}N, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$, respectively. Reference shifts fit to protein data are in good agreement with 'random-coil' values derived from experimental measurements on peptides. This prediction algorithm should be helpful in NMR assignment, crystal and solution structure comparison, and structure refinement.

Introduction

Chemical shifts have been long recognized as an important part of the potential structural information contained in NMR spectra. Their high sensitivity to conformational variations and high accuracy of measurement make shifts attractive candidates for structural interpretation. Further, there is a large database of information to draw from (Seavey et al., 1991; Szilágyi, 1995), and interest in deciphering the structural information encoded in chemical shifts has rapidly increased in the past decade. Empirical methods (Spera and Bax, 1991; Wishart et al., 1991, 1997; Szilágyi, 1995; Grønwald et al., 1997; Iwadate et al., 1999; Cornilescu et al., 1999), semi-empirical models (Ösapay and Case, 1991; Williamson and Asakura, 1993) and *ab initio* quantum approaches (de Dios et al., 1993; de Dios, 1996; Sitkoff and Case, 1997; Ando et al., 1998) have all revealed promise for structural investigations of proteins.

Reliable and automated prediction of chemical shifts from structure would be an important step toward this goal. Chemical shift predictions of α-^{1}H and NH from crystal structures have been developed using semi-empirical methods (Ösapay and Case, 1991, 1994; Herranz et al., 1992; Williamson et al., 1992) with correlation coefficients of between 0.74 and 0.84 for $^{1}H^{\alpha}$ and between 0.57 and 0.71 for H^{N}. Predictions for side chain protons (especially in methyl groups) are more accurate. Empirical chemical shift surfaces have been used to predict $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ shifts for valine residues from X-ray structures with reasonable agreement between predicted and observed shifts, and for some valine fragments *ab initio* calculations with geometry optimization improved the prediction accuracy (Pearson et al., 1997). In addition to these structure-based prediction methods, Wishart et al. developed a sequence-based prediction method

*To whom correspondence should be addressed. E-mail: case@scripps.edu

(SHIFTY) to automatically predict the ^{1}H and ^{13}C (Wishart et al., 1977), and ^{15}N chemical shifts on the basis of sequence homology, with good accuracy when the sequence identity is high. Other empirical databases have been constructed to predict $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts (Iwadate et al., 1999), and to deduce backbone conformation from sequence and chemical shift homology (Cornilescu et al., 1999). Here we examine what sorts of results can be extracted from a database of quantum chemistry calculations on peptides.

Our recent density-functional investigation of structural effects on ^{15}N, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, and $^{13}C'$ shifts in peptides (Xu and Case, 2001) has shown that backbone conformation effect, side-chain orientation effect, neighbor residue effect and hydrogen bonding effect all contribute to the shifts in different ways. A combined consideration of these effects may provide more reliable structure information than any one of them in isolation. Here we propose a new structure-based prediction method of chemical shifts in proteins. The idea is based on an additive model of chemical shift contributions corresponding to conformational effects found in a database of density functional theory (DFT) calculations on peptides.

Materials and methods

Database

A calculated database was established using the model systems listed in Table 1. A hybrid density functional method (Becke's three-parameter hybrid method and employing the LYP correlation functions, B3LYP) (Becke, 1993; Lee et al., 1988; Miehlich et al., 1989; Pople et al., 1989) of the GIAO (Gauge Included Atomic Orbitals) (Wolinski et al., 1990)method in the Gaussian 98 program (Frisch et al., 1998) was used for the calculation. A standard Gaussian basis set 6-31G** was uniformly adopted for all atoms. As described in our previous work (Xu and Case, 2001), different conformational contributions to chemical shifts were computed separately. With different amino acids X and Y, various backbone torsion angles ϕ and ψ within energy favorable regions, and with several side-chain orientations, a total of 1335 peptide sequences were studied. These calculated results have shown that significant conformational contributions to chemical shifts arise not only from the probe residue itself (i, or s) but also from the preceding residue ($i-1$, or p) as well as the following residue ($i+1$, or f). Therefore, we used chemical shift patterns corresponding to tripeptides of consecutive residues, with their backbone torsion angles, the types and side-chain orientations of residues $i-1$ and i, and hydrogen bonding being used to constitute the database. The available backbone conformation range of the database is outlined by rectangles in Figure 1. We found that the influence from χ_2 and χ_3 on the shifts of the neighboring residues is small in most cases. Therefore, we can address the neighboring residue effects, including χ_1 and residue type effects, (for all residues except Pro) using fixed χ_2 and χ_3 values based on their most probable orientations in the protein data bank. For the self side-chain orientation effect, different combinations of χ_1, χ_2 and χ_3 within favorable energy ranges were studied and included in the database for most residues. All amino acids except for Cys and Pro can be predicted, but the restriction that three consecutive backbone ϕ,ψ pairs be in the regions outlined in Figure 1 limits us to regions of regular secondary structure, in effect about 40% of all residues. Potential methods to overcome this limitation are discussed below.

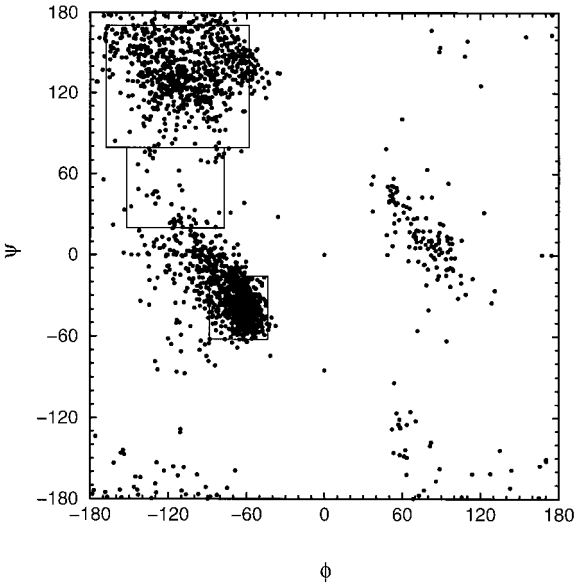

Figure 1. Dots show the distribution of backbone torsion angles for the proteins analyzed here (Table 2); the backbone conformation range of the DFT database is shown by the rectangular regions.

Proteins used for prediction

The proteins studied here are listed in Table 2. We selected twenty proteins based on the availability of as-

Table 1. Model systems used for establishing the calculated database

System conformation	Sequence		Purpose
1	$G_1GGGGGG_7$	β	Hydrogen bonding effects
	$G_1GGGGGG_7$-duplex	β	
	$G_1GGGGGG_7$-triplex	β	
	$G_1GGGGGGGG_9$	β	
2	G_1AYAG_5[a]	α, β	Backbone effects
3	G_1XYG_4[a]	β	Side-chain orientation and
	$G_1GGXYGGG_8$[a]	α	neighboring residue effects

*X, Y = all amino acids except for Pro and Cys.

Table 2. Proteins used for the predictions

Protein	PDB code	Resolution (Å)	No. of residues	Chemical shift ref. (BioMagResBank)
Alpha-lytic protease	2alp	1.70	198	Davis et al., 1997[a]
Calmodulin	1cll	1.70	148	Ikura et al. 1990 (bmr547)[a]
Calmodulin/W-7	1mux (NMR)		148	Osawa et al., 1998 (bmr4056)[c]
Calmodulin/M13	1cdl	2.20	147	Ikura et al., 1991 (bmr1634)[a]
Che Y	1chn	1.80	126	Moy etal., 1994 (bmr4083)
Cutinase	1cex	1.00	214	Pompers et al., 1997 (bmr4101)
Cutinase	1cug	1.75	197	Pompers et al., 1997 (bmr4101)[a]
Cyclophilin	2cpl	1.63	165	Ottiger et al., 1997[a]
Dehydrase	1mka	2.00	171	Copie et al., 1996[a]
Human carbonic anhydrase I	1hcb	1.60	260	Sethson et al., 1996 (bmr4022)[a]
Human HIV-1	1hvr	1.80	99	(bmr4356)[a]
Human thioredoxin in reduced form	1ert	1.70	105	Qin et al., 1996[a]
III-glc	1f3g	2.10	168	Pelton et al., 1991[a]
Profilin	1acf	2.00	125	Archer et al., 1994[a]
Profilin Ia	1prq	2.50	125	Archer et al., 1994[a]
Ribonuclease H	2rn2	1.48	155	Yamazaki et al., 1993 (bmr1657)[b]
Serine protease PB 92	1svn	1.40	269	Fogh et al., 1995
Ubiquitin	1ubi	1.80	76	Wang et al., 1995[a]
Ubiquitin	1ubq	1.80	76	Wang et al., 1995[a]
Ubiquitin	1d3z (NMR)		76	Wang et al., 1995[a]

[a]Corrected data from TALOS database (Cornilescu et al. 1999).
[b]Corrected data (Iwadate et al. 1999).
[c]Corrected in this work: −0.5 ppm for $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$.

signed resonances and with high resolution (< 2.5 Å) crystal structures. The protein coordinates were taken from the Brookhaven Protein Data Bank (Bernstein et al., 1977). Most of the experimental chemical shifts were taken from BioMagResBank or the database used for TALOS, with some chemical shift referencing corrections (Cornilescu et al., 1999). Other chemical shifts were obtained from the original literature using the ^{13}C shift referencing corrections made by Iwadate et al. (1999). Details are provided in Table 2.

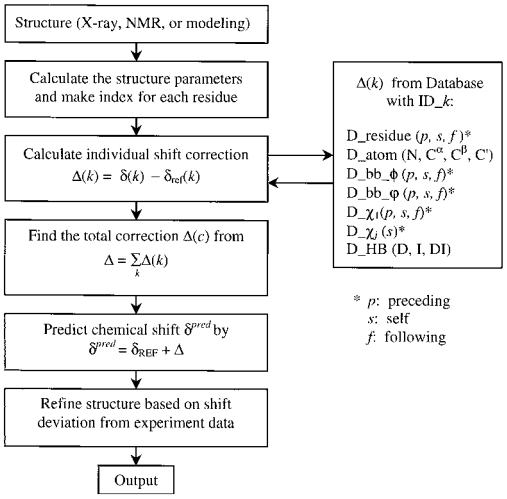

Figure 2. Flow chart of the part of ^{15}N and ^{13}C shift prediction in SHIFTS program.

Results and discussion

Algorithms and program

Figure 2 presents an outline of the prediction procedure used in our SHIFTS (version 4.0) program. The program takes an input of a structural file in pdb format, and an optional file of observed shifts. The latter is only used to prepare a table that compares calculated and observed shifts; its format is like a subset of the data in an 'NMR-star' file from BMRB (see http://www.bmrb.wisc.edu). The program first calculates the structural parameters for all amino acids in the protein, including backbone conformation, side-chain orientation, and hydrogen bonding geometries. From our density functional calculations we have identified eight potential contributions to the shifts, which are given in Table 3; we use the notation $\Delta(k,c)$, where k identifies one of the eight contributions in Table 3, and c is either helix or sheet. The first three contributions (BB-p, BB-s and BB-f) estimate the contribution from the backbone ϕ and ψ torsion angles of the preceding (p), self (s) and following (f) amino acids; the next two consider the influence of the type of sidechain and the χ_1 torsion angle for the preceding and self residue; and the final three consider hydrogen-bonding effects for ^{15}N shifts, either 'direct' (to the NH group), 'indirect' (to the carbonyl group), for both (for amide groups with hydrogen bonds at both NH and carbonyl). The contributions that depend upon torsion angles are interpolated directly from the DFT peptide results, which were calculated at 10° increments in the helical and sheet regions; hydrogen bond contributions are empirical functions of the hydrogen bond length. In the present model, these contributions are assumed to be additive (see Equation 1, below).

All calculated contributions are relative to a reference shift given in Table 3. For example, the DFT shifts of the third residue in GAAAG with $\phi = -139°$ and $\psi = 135°$ were used as the reference value for backbone conformation effects for residues in a sheet. The value for Δ(*BB-s*,sheet) is then an estimate of the difference in the shift at the actual values of ϕ and ψ, compared to that computed for $(\phi, \psi) = (-139, 135)$. Similarly, $\Delta(\chi_1$ & *R-p*, $\alpha)$ for Ile in a helical conformation would be determined from the difference of its DFT shift in GGGXIGGG (where the chemical identity and torsion angles of residue 'X' match those in the protein of interest), compared to the DFT shift in the sequence GGGAIGGG with standard helical structure parameters. Details about these chemical shift patterns are described in our previous report (Xu and Case, 2001). The total contribution for either sheet or helix is then assumed to be the sum of the individual ones:

$$\Delta(c) = \sum_k \Delta(k, c) \tag{1}$$

and the final predicted chemical shift, $\delta^{\mathrm{pred}}(c)$, is given by

$$\delta^{\mathrm{pred}}(c) = \delta_{\mathrm{REF}}(c) + \Delta(c). \tag{2}$$

Here $\delta_{\mathrm{REF}}(c)$ is a chemical shift reference for each amino acid, with $c = \beta$ for sheet and $c = \alpha$ for helix.

Ideally $\delta_{\mathrm{REF}}(c)$ would be determined by DFT calculations using the standard structure parameters given in the footnote of Table 3. However, there are two practical problems with this prescription that led us instead to allow some empirical adjustment of the $\delta_{\mathrm{REF}}(c)$ values. First, the absolute chemical shifts of various types of nuclei are not uniformly well calculated at the level of theory we have used. For example, Cα and Cβ shifts (relative to TMS) are close to experiment, but computed shifts for C$'$ are about 16 ppm too low, and computed shifts for N are about 7 ppm to large in these quantum calculations. Secondly, we found that the basis set dependence of the chemical shifts varies for different residues by amounts up to 3 ppm). Much larger basis sets could dramatically improve this, but are not computationally feasible; further, trends in shifts (which changes in single bond

Table 3. Individual chemical shift references and corrections for residue Ala

k^a	$\Delta^k = \delta - \delta^k_{\text{ref}}$
BB-p	δ and δ^k_{ref} values from the A_4 backbone shift pattern[a] with actual φ, ψ for δ and standard[b] ϕ_s, ψ_s for δ^k_{ref}.
BB-s	δ and δ^k_{ref} values from the A_3 backbone shift pattern[a] with actual φ, ψ for δ and standard[b] ϕ_s, ψ_s for δ^k_{ref}.
BB-f	δ and δ^k_{ref} values from the A_2 backbone shift pattern[a] with actual φ, ψ for δ and standard[b] ϕ_s, ψ_s for δ^k_{ref}.
	χ_1&R-p δ and δ^k_{ref} values from the A_4 side-chain orientation shift pattern[c] at X = A
χ_1&R-p	δ and δ^{ref}_k values from the A_2 side-chain orientation shift pattern[c] at X = A
χ_1&R-s	δ and δ^k_{ref} values from the A_2 side-chain orientation shift pattern[c] at X = A
HB-D, HB-I, HB-DI	Formulas derived from the calculated dependence of shifts on hydrogen bond length[d]

[a]Figure 9 in Xu and Case (submitted).
[b]$\phi = -139°$ & $\psi = 135°$ for β-sheet and $\phi = -58°$ and $\psi = -47°$ for α-helix.
[c]Figure 12 in Xu and Case (submitted).
[d]Figure 5b in that one.

Table 4a. Chemical shift references $\delta_{\text{REF}}(\beta)$ used in the program SHIFTS

Residue	^{15}N			$^{13}C^{\alpha}$			$^{13}C^{\beta}$			$^{13}C'$		
	Calc.	Modif.	Diff.	Calc.	Modif.	Diff.	Calc.	Modif.	Diff.	Calc.	Modif.	Diff.
Ala	128.76	121.76	7.0	51.96	51.96	0.0	24.62	23.12	1.5	163.26	176.26	−13.0
Gly	115.32	107.32	8.0	44.72	44.72	0.0	–	–	–	159.10	173.10	−14.0
Met	125.30	119.3	6.0	55.69	55.69	0.0	36.83	34.83	2.0	160.02	175.02	−15.0
Gln	125.66	119.66	6.0	55.36	55.36	0.0	35.01	33.01	2.0	160.29	175.29	−15.0
Glu	124.70	118.70	6.0	54.97	54.97	0.0	36.13	34.13	2.0	160.26	175.26	−15.0
Lys	125.74	119.74	6.0	55.97	55.97	0.0	39.83	37.83	2.0	162.35	175.35	−13.0
Arg	124.95	118.95	6.0	55.73	55.73	0.0	36.95	34.95	2.0	162.30	175.80	−13.5
Phe	123.98	117.98	6.0	56.93	56.93	0.0	45.02	43.02	2.0	162.22	175.22	−13.0
Tyr	124.35	117.35	7.0	57.07	57.07	0.0	43.96	41.96	2.0	162.55	175.55	−13.0
Trp	127.94	118.94	9.0	56.10	56.10	0.0	31.47	33.47	−2.0	161.95	176.95	−15.0
His	124.29	115.29	9.0	55.06	55.56	0.0	33.58	32.58	1.0	161.74	174.74	−13.0
Ser	120.7	113.70	7.0	55.95	55.95	0.0	66.02	66.02	0.0	160.24	174.24	−14.0
Asn	120.94	114.94	6.0	52.86	52.86	0.0	40.25	41.25	−1.0	164.10	177.10	−13.0
Asp	120.67	116.67	4.0	52.55	52.55	0.0	39.92	42.92	−3.0	162.17	175.17	−13.0
Ile	127.20	121.20	6.0	60.65	60.65	0.0	44.96	42.46	2.5	161.94	174.94	−13.0
Val	126.40	120.40	6.0	60.98	61.98	−1.0	37.38	36.38	1.0	162.10	176.10	−14.0
Leu	125.68	118.68	9.0	54.13	54.13	0.0	44.69	46.19	−1.5	162.35	176.35	−14.0
Thr	122.14	115.64	6.5	56.71	59.71	−3.0	70.50	72.50	−2.0	160.62	174.62	−14.0

Table 4b. Chemical shift references $\delta_{REF}(\alpha)$ used in the program SHIFTS

Residue	^{15}N			$^{13}C^{\alpha}$			$^{13}C^{\beta}$			$^{13}C'$		
	Calc.	Modif.	Diff.	Calc.	Modif.	Diff.	Calc.	Modif.	Diff.	Calc.	Modif.	Diff.
Ala	123.26	119.26	4.0	54.91	54.91	0.0	20.29	17.79	2.5	164.25	179.25	−15.0
Gly	111.07	103.07	8.0	47.33	47.33	0.0	–	–	–	159.54	176.54	−17.0
Met	120.93	114.93	6.0	59.70	59.7	0.0	31.60	31.6	0.0	160.41	176.91	−16.5
Gln	121.11	115.11	6.0	59.65	58.65	1.0	29.48	27.48	2.0	160.83	178.83	−18.0
Glu	120.58	114.58	6.0	59.36	59.36	0.0	30.75	28.75	2.0	160.51	179.01	−19.0
Lys	122.57	116.57	6.0	60.65	60.65	0.0	34.23	32.23	2.0	162.45	179.45	−17.0
Arg	120.38	114.38	6.0	59.98	59.98	0.0	31.67	29.67	2.0	162.97	179.47	−16.5
Phe	123.94	116.94	7.0	56.97	57.97	−1.0	36.95	35.95	1.0	162.31	178.31	−16.0
Tyr	124.97	116.97	8.0	57.73	58.73	−1.0	35.73	36.73	−1.0	162.16	178.16	−16.0
Trp	122.95	115.95	7.0	56.89	58.89	−2.0	29.16	29.16	0.0	162.44	178.44	−16.0
His	121.15	115.15	6.0	60.87	60.37	0.0	29.41	27.41	2.0	161.93	177.93	−16.0
Ser	119.55	113.55	6.0	60.29	61.79	−1.5	62.90	62.4	0.5	160.33	175.83	−15.0
Asn	124.58	115.58	9.0	56.30	56.3	0.0	37.32	38.32	−1.0	162.93	177.43	−15.0
Asp	122.42	116.42	6.0	56.03	57.03	−1.0	37.33	39.33	−2.0	162.24	179.24	−17.0
Ile	122.51	117.51	5.0	64.86	64.86	0.0	39.82	36.82	3.0	162.15	178.15	−16.0
Val	123.25	116.25	7.0	65.70	66.7	−1.0	29.29	31.29	−2.0	162.57	177.57	−15.0
Leu	120.31	115.31	5.0	56.88	57.88	−1.0	39.16	40.66	−1.0	162.44	178.44	−16.0
Thr	118.11	109.11	9.0	63.72	64.72	−1.0	65.15	68.15	−3.0	160.52	175.52	−15.0

[a]Without side-chain torsion angle modification.
[b]With side-chain torsion angle modification.
[c] Calculated using SYBYL without protons.

torsions) are adequately calculated at the level of theory we use (Xu and Case, 2001). Even after making a global correction for the absolute shifts, we found smaller problems related to differences in amino acid side chains. For example, $^{13}C'$ shifts for alanine in helices are observed to be about 3 ppm higher than for glycine in a similar environment, but the DFT difference is about 5 ppm. Hence we allowed additional small (<4 ppm) adjustments in the $\delta_{REF}(c)$ values to account for these problems. Our final values are listed in Table 4, where they are compared to the 'raw' predictions from the DFT calculations; the 'modified' values were optimized manually by comparison of fits to experiment for subsets of the 20 proteins given in Table 2; some improvement could probably be made by a further automatic adjustment of these parameters, but this has not been done.

Although the introduction these adjustments adds some empiricism to the prediction scheme, the final values are quite reasonable. For example, if we define δ_{av-REF} as a simple average of $\delta_{REF}(\alpha)$ and $\delta_{REF}(\beta)$, the values of δ_{av-REF}, are close to the corresponding experimental random coil shifts from peptides. [This simple average of helical and sheet results proved useful earlier in comparing calculated proton shifts to those of random coil peptides (Ösapay and Case, 1994).] A comparison between δ_{av-REF} and the experimental shifts obtained from the $AcGGXGGNH_2$ sequence (Schwarzinger et al., 2000) is shown in Figure 3. (Here we have added an additional correction of −4.3 ppm for nitrogens and −0.30 ppm for carbonyl carbons [derived from our earlier DFT calculations] to account for the fact that the residue preceding 'X' is glycine in the experimental values and alanine in our δ_{av-REF} calculated values. For $^{13}C^{\alpha}$ and $^{13}C^{\beta}$, the preceding residue effect predicted by DFT is negligible, so that no correction was made for them.) In general, the reference values we obtained by optimizing our SHIFTS predictions on proteins are close to the 'random-coil' values extracted from measurements on peptides, and the latter values could be used in place of the former with only a small degradation in prediction quality (data not shown). This consistency supports the notion that our empirical adjustments to the reference shifts indeed compensate for systematic errors in relatively small-basis-set DFT calculations, and are not simply making the final data 'look good' by optimization of adjustable parameters. We consider this point more fully in our discussion of Figure 8 (below).

Table 5. Examples of the improvement on the prediction accuracy via the side chain orientation modification

Residue (PDB ID)		χ_1	χ_2	N^H	C^α	C^β	C'	Energy (kcals/mol)
L116	Without M[a]	−138.46	−141.49	120.84	59.98	41.49	180.62	312.796
(1acf)	With M[b]	180	60	122.23	58.10	40.63	178.55	311.779
	Expt.			122.11	57.54	40.26	178.50	
T107	Without M[a]	66.02	–	109.62	63.89	66.78	176.23	166.859
(1cex)	With M[b]	−60	–	113.20	65.58	68.51	176.33	167.673
	Expt.			111.50	64.77	69.07	176.74	
E11	Without M[a]	−57.96	176.87	112.09	59.00	28.72	179.31	161.882
(1cll)	With M[b]	180	60	116.25	56.82	27.86	179.97	162.001
	Expt.			117.10	55.40	29.10	180.20	

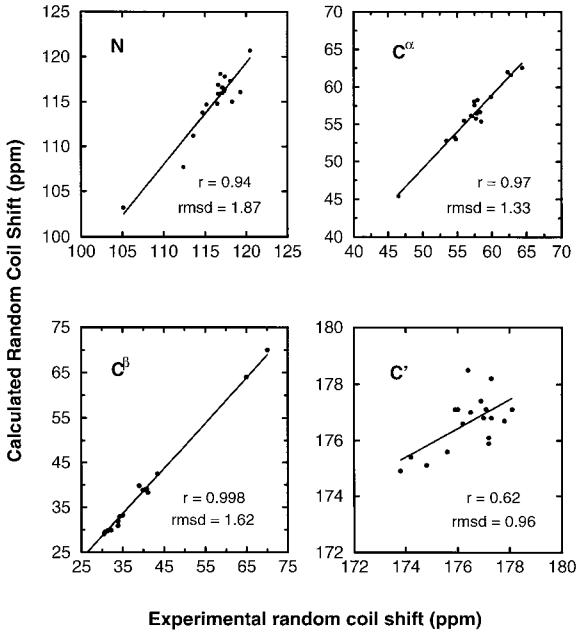

Figure 3. Comparsion of reference shifts from Table 4 with observed shifts from $AcGGXGGNH_2$ peptides.

The shift prediction program SHIFTS is written in the NAB (Nucleic Acid Builder) language (Macke and Case, 1998). The whole package includes the functionality of our earlier SHIFTS code (Ösapay and Case, 1991), for calculations of proton chemical shifts, plus the new material described here for ^{15}N and ^{13}C shifts in proteins. There are two output files being formed after running SHIFTS. One of them is a structural parameter file with backbone conformational parameters (ϕ and ψ), side-chain orientation parameters (χ_1 and χ_2), and hydrogen bonding information. The another output is a predicted chemical shifts file giving a breakdown of the detailed contributions. Comparison with experimental data is also given in the output file if the data is available. The running time for prediction is in seconds. Both NAB and SHIFTS are distributed under the terms of the GNU General Public License (GPL), and may be obtained from http://www.scripps.edu/case.

Side-chain orientation refinement

Using the SHIFTS program, we calculated the ^{15}N, $^{13}C^\alpha$, $^{13}C^\beta$, and $^{13}C'$ chemical shifts for twenty proteins with known shifts, as described below. Checking the data with large deviations for both ^{15}N and ^{13}C (see examples in Table 5), it was found that some modification on side-chain orientation is often helpful to improve the prediction accuracy We also checked the flexibility of side-chain orientation in some proteins. The example given here is the structures (PDB code 1ubq, 1ubi and 1d3z) of protein ubiquitin. 1ubq (Vijay-Kumar et al., 1987) and 1ubi (Ramage et al., 1994) are crystal structures determined by two different groups and 1d3z (Cornilescu et al., 1998) is the structure obtained by NMR. Taking 1ubq as reference, Figure 4 gives the relative variations in the backbone torsion angles ϕ and ψ, and side-chain torsion angles χ_1 and χ_2. For these structures, the variation in backbone conformation is negligible except for the region close to the C terminal. The significant differences in side-chain orientations (top two panels of Figure 4) suggest that the most representative side-chain conformation may not be present in any particular PDB entry. With this in mind, we added automatic side-chain ori-

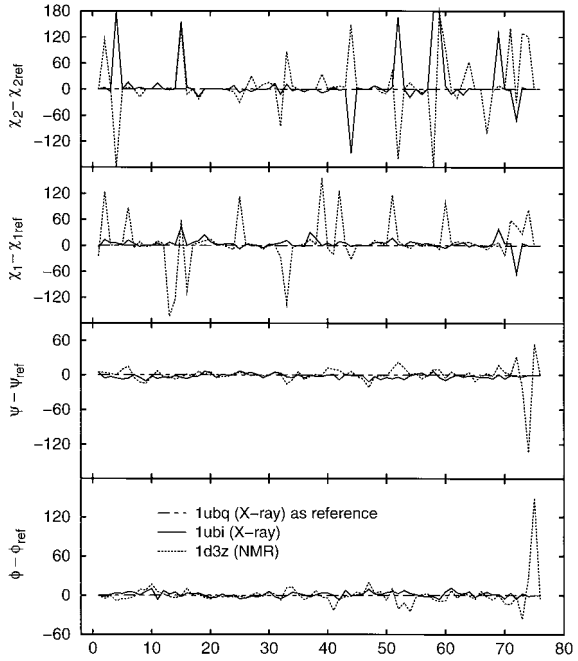

Figure 4. Relative variations in the backbone torsion angles ϕ and ψ, and side-chain torsion angles χ_1 and χ_2 for three ubiquitin structures 1ubq, 1ubi, and 1d3z with 1ubq as the reference.

entation refinement process with weighted criterion

$$\Omega = 0.9\Delta N + 1.0\Delta C^{\alpha} + 0.8\Delta C^{\beta} + 0.3\Delta C' \quad (3)$$

into SHIFTS. Here ΔN, ΔC^{α}, ΔC^{β} and $\Delta C'$ are the absolute differences between predicted and experimental shifts for these nuclei, and the coefficients were emprically determined, based roughly on the influence of side-chain torsion angles on each type of shift. When a new side-chain orientation is a low-energy region, and the Ω value is less than the former one more than 1.0 ppm, the former Ω and side-chain orientation are replaced by the new ones and all related effects are recalculated. The process is repeated until a minimum Ω is obtained. The examples in Table 5 show significant improvement in predicted results. For example, for residue E11 in calmodulin, use of the crystal structure geometry leads to prediction errors of 5.0, 3.6, 0.4 and 0.9 ppm for N, Cα, Cβ, and C$'$, respectively. Changing the side-chain torsion angles as shown reduces these deviations to 0.9, 1.4, 1.2 and 0.2 ppm, with the improvement especially noticeable at the N and Cα positions. The percentage of residues involved in such refinement is around 10%. It should be noted that since this procedure relies upon comparison to experimental shifts, it cannot be used for 'blind' predictions from structure alone, but it is useful in pointing to side chain orientations that may benefit from more careful attention in refinements where the shifts are known.

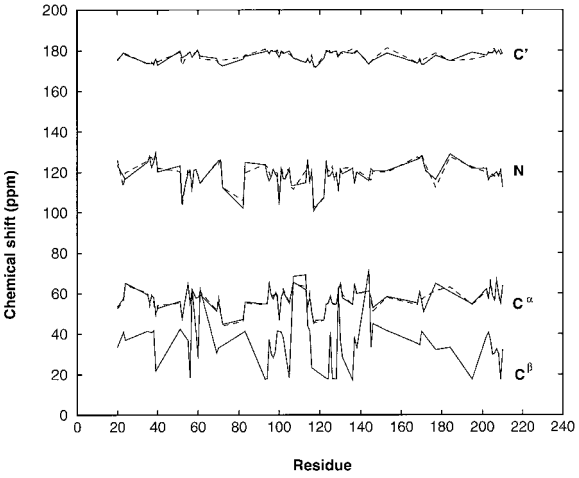

Figure 5. Results for the protein cutinase.

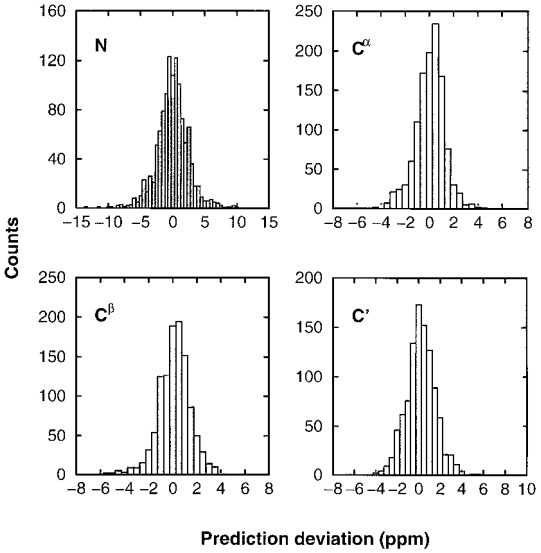

Figure 6. Distribution of deviations between predicted and observed shifts, with (clear bars) and without (shaded bars) sidechain torsion angle refinement.

Chemical shift predictions

Figure 5 presents an example result for the chemical shift prediction of protein cutinase (PDB code: 1cex) using SHIFTS. The predicted chemical shift pattern

Table 6. Percentage of predictable residues and good prediction

Protein	PDB code	No. of residues	Predictable residue # (%)[a]	Good percentage[b]			
				^{15}N	$^{13}C^{\alpha}$	$^{13}C^{\beta}$	$^{13}C'$
Alpha-lytic protease	2alp	198	57 (28.8)	67.9	84.6	78.8	78.8
Calmodulin	1cll	148	83 (56.1)	83.6	98.5	93.9	94.0
Calmodulin/W-7	1mux (NMR)	148	71 (48.0)	80.0	93.8	90.6	93.8
Calmodulin/M13	1cdl	147	76 (51.7)	89.1	95.3	90.5	84.4
Che Y	1chn	126	66 (52.40	92.3	94.3	92.5	na[c]
Cutinase	1cex	214	79 (36.9)	85.9	92.3	88.6	89.7
Cutinase	1cug	197	78 (39.6)	89.5	92.1	88.2	89.5
Cyclophilin	2cpl	165	46 (27.9)	79.5	93.2	95.5	na[c]
Dehydrase	1mka	171	66 (38.6)	86.4	91.7	80.0	90.0
Human carbonic anhydrase I	1hcb	260	81 (31.2)	70.4	80.2	80.0	78.2
Human HIV-1	1hvr	99	39 (39.4)	71.1	92.1	92.1	81.6
Human thioredoxin in reduced form	1ert	105	53 (50.5)	92.3	90.4	94.2	na[c]
III-glc	1f3g	168	34 (20.2)	87.9	90.9	90.6	93.9
Profilin	1acf	125	58 (46.4)	84.3	90.2	89.1	85.4
Profilin Ia	1prq	125	53 (42.4)	81.2	83.3	88.6	90.9
Ribonuclease H	2rn2	155	46 (29.7)	84.0	92.6	86.1	92.6
Serine protease PB 92	1svn	269	97 (36.1)	75.3	84.5	79.1	87.6
Ubiquitin	1ubi	76	33 (43.4)	87.5	93.8	93.8	93.8
Ubiquitin	1ubq	76	33 (43.4)	90.6	93.8	100	87.5
Ubiquitin	1d3z (NMR)	76	32 (42.1)	81.3	92.1	100	90.6
Average		(40.7)	83.0	91.0	89.6	88.4	

[a]Determined by the available size of the database.
[b]Criterions: $|\delta_{pred} - \delta_{expt}| < 3.0$ ppm for ^{15}N and $|\delta_{pred} - \delta_{expt}| < 2.0$ ppm for $^{13}C^{\alpha}$, $^{13}C^{\beta}$, and $^{13}C'$.
[c]No experimental data available.

for both ^{15}N and ^{13}C are very close to that from experiment. Figure 6 shows the distribution of prediction errors with and without side-chain orientation modification. These distributions are nearly symmetric about zero error, and are made slightly better by side-chain optimization, but the differences are not great. Table 6 lists the percentage of predictable residue and 'good prediction' for all of the proteins in Table 2. Here the number of predictable residues is determined by the available size of our DFT database and 'good prediction' is defined as a deviation from the corresponding experimental value by less than 3.0 ppm for $^{15}N^{H}$, 2.0 ppm for $^{13}C^{\alpha}$, $^{13}C^{\beta}$, and $^{13}C'$. The average value for predictable residue percentage is 41%. It is expected to be improved by the extension in the database size. The percentages of good prediction in average are 83, 91, 89 and 88% for ^{15}N, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ respectively after side-chain orientation refinement. Without such refinement, the corresponding percentages 76, 85, 89, and 86%. Prediction deviations greater than 5.0 ppm for ^{15}N and 2.5 ppm for ^{13}C raise the real possibility that the residue in question either has a different conformation in solution from that in the crystal, or has an error in the published shifts (Iwadate et al., 1999). The percentage of residues with deviations of this size was 6.2% for ^{15}N, 5.4% for $^{13}C^{\alpha}$, 6.2% for $^{13}C^{\beta}$, and 6.4% for $^{13}C'$ from 1135 N^{H}, 1061 C^{α}, 1006 C^{β} and 917 C' shifts in 1181 amino acids with available experimental data. For comparison, Iwadate et al. (1999) excluded about 9% of observed C^{α} and C^{β} shifts in constructing an empirical database.

Figures 7 and 8 give overall results for these four nuclei after the exclusion on this criterion. Figure 7 compares the final calculated and observed shifts, whereas Figure 8 subtracts the reference shift from both, so that secondary shifts are being compared. Linear correlation coefficients are shown in the plots, showing that the general behavior of chemical shift dispersion is being captured in this model. Note that the values plotted in Figure 8 are independent of the adjustments to the reference shifts discussed above;

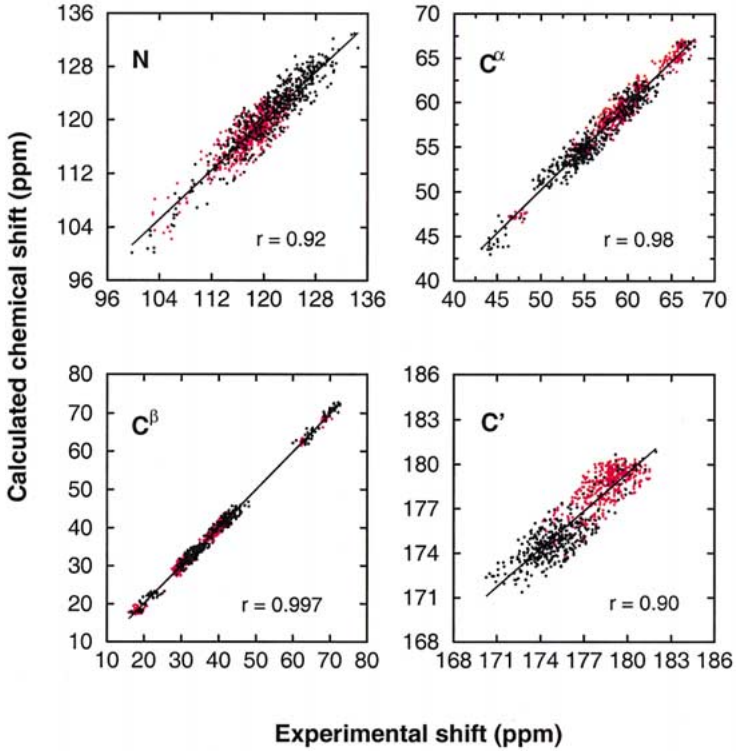

Figure 7. Comparison between predicted and experimental $^{15}N^{H}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, and $^{13}C'$ chemical shifts for the 20 proteins in Table 2. Red symbols for helix, black for sheet.

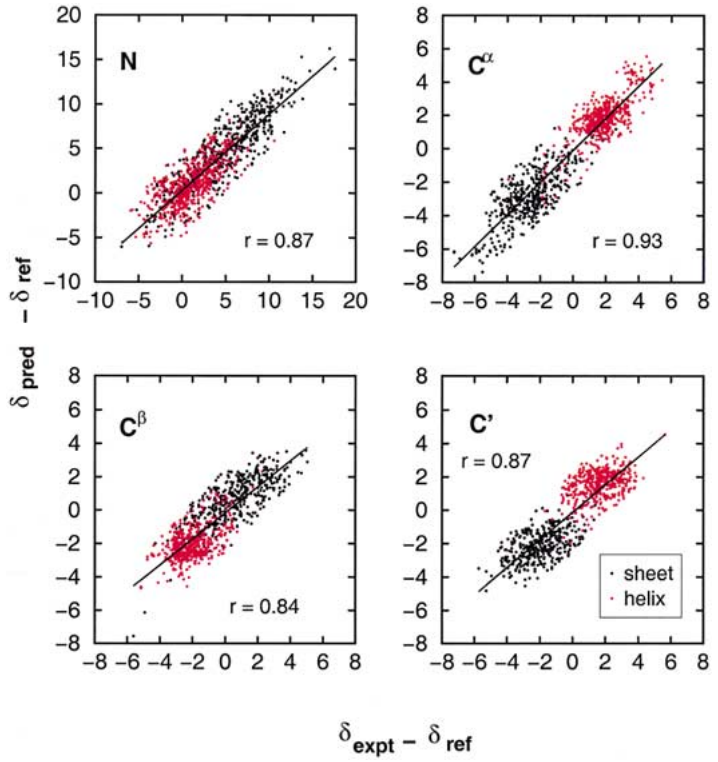

Figure 8. Comparison between predicted and experimental $^{15}N^{H}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, and $^{13}C'$ secondary chemical shifts for the 20 proteins in Table 2. Red symbols for helix, black for sheet.

Table 7. Comparison of predicted $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts using our computational database (SHIFTS) and an empirical database (TANSO, Iwadate et al., 1999)

PDB code	C^{α}				C^{α}			
	rmsd (ppm)		Good (%)		rmsd (ppm)		Good (%)	
	SHIFTS	TANSO	SHIFTS	TANSO	SHIFTS	TANSO	SHIFTS	TANSO
1acf	1.21	1.12	90.2	90	1.17	1.06	89.1	93.3
2alp	1.46	1.36	84.6	82.4	1.85	1.81	78.8	74.5
1cll	0.87	1.21	98.5	95.5	1.21	0.74	93.9	100
1mux	1.08	1.26	93.8	92.2	1.11	0.71	90.6	98.4
1cdl	1.02	1.05	95.3	93.7	1.13	0.86	90.5	95.2
1chn	0.86	1.02	94.3	96.2	1.00	1.23	92.5	86.8
1cex	1.06	1.53	92.3	82.1	1.29	1.16	88.6	90
1cug	1.08	1.49	92.1	82.7	1.43	1.25	88.2	88.1
2cpl	1.16	1.53	93.2	81.8	1.02	1.24	95.5	90.9
1mka	1.26	1.53	91.7	79.5	1.59	1.63	80.0	76.2
1hcb	1.46	1.32	80.2	85.3	1.70	1.80	80.0	68.9
1hvr	1.27	1.21	92.1	94.6	1.26	1.15	92.1	89.2
1ert	1.25	1.13	90.4	92.2	1.03	1.07	94.2	88.2
1f3g	1.09	1.07	90.9	93.9	1.49	1.31	90.6	93.8
1acf	1.21	1.12	90.2	90	1.17	1.06	89.1	93.3
1prq	1.34	1.18	83.3	91.5	1.46	1.37	88.6	86
2rn2	1.12	1.26	92.6	83.5	1.29	1.33	86.1	85.7
1svn	1.43	1.3	84.5	89.2	1.62	1.39	79.1	83.7
1ubi	0.98	0.98	93.8	100	0.87	1.06	93.8	93.5
1ubq	0.98	0.97	93.8	100	0.84	1.14	100	90.3
1d3z	0.83	0.92	96.9	100	0.87	1.13	100	96.8
Aver.	1.18	1.26	90.4	89.2	1.32	1.27	88.2	87.9

they are hence based completely on DFT calculations without adjustable parameters. (The use of the seven terms shown in Table 3 is of course somewhat arbitrary. Once this choice is made, though, there are no explicitly adjustable parameters involved in getting the data shown in Figure 8.) By contrast, the absolute shifts shown in Figure 7 would look much worse had we not made empirical adjustments to the DFT-predicted reference shifts, as discussed above and documented in Table 4. This behavior is consistent with our belief that our DFT calculations are reasonably reliable in giving trends in shifts as a function of conformation, but are not accurate enough to be useful in predicting changes in shifts when going from one side chain type to another.

The root-mean-square deviations of the predictions are of course the same for Figures 7 and 8. These are 1.94, 0.97, 1.05 and 1.08 ppm for N, Cα, Cβ, and C′, respectively. Without the exclusion of those 'bad' residues (deviations > 5 ppm, about 6%), the corresponding rms deviations are 2.55, 1.18, 1.32, and 1.31 ppm. As Figure 8 illustrates, these errors are roughly 1/10 of the observed range of secondary structural shifts for all four nuclei. The remaining errors are the result of uncertainties in the structures we have used, and of deficiencies in the computational model. As with similar predictions of proton shifts (Ösapay and Case, 1991), it is not possible at present to separate these two effects. The carbon shifts show clear differences between helix and sheet conformations, which have been recognized for a long time (Spera and Bax, 1991; Szilágyi, 1995). The nitrogen results are more complex, a fact that has also been recognized for some time.

It is of interest to compare the present method to other approaches for estimating shifts from structure. The most straightforward comparison is to Iwadate et al.'s (1999) empirical method TANSO, which can predict C^{α} and C^{β} shifts from structures. Table 7 shows results from both TANSO and SHIFTS on a set of 20 proteins. The general accuracy of prediction for these two approches is quite close, suggesting that the

DFT calculations on peptides have captured about the same information content as this empirical survey. A second comparison for $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ predictions can be made to quantum chemistry results from Pearson et al. (1997). We chose valine residues in the proteins calmodulin (1cll), SNase (1snc) and ubiquitin (1ubq) for comparison. Without any side-chain structure refinement, the correlation coefficient *r* for our prediction results is 0.93 for $^{13}C^{\alpha}$ and 0.89 for $^{13}C^{\beta}$, compared with 0.67 and 0.39 for Pearson et al. (1997). With geometry optimization combined with the selection of lowest energy χ_1 conformation, Pearson et al. (1997) significantly improved their *r* values to 0.96 and 0.95, which is somewhat better than the SHIFTS results. However, this model is only directly applicable to valines, and requires considerable computational effort for each prediction.

The SHIFTY program (Wishart et al., 1997) can be used to predict both ^{15}N and ^{13}C shifts. Its performance is strongly dependent on the sequence similarity to a protein with known shifts. For example, alpha-lytic protease (2alp) has 26% identity to proteins with known shifts; the correlation coefficient from SHIFTY for ^{15}N shifts is 0.39 and for $^{13}C^{\alpha}$ is 0.73, compared with SHIFTS values 0.83 and 0.96. These are impressive results for such low sequence identity, and the usefulness of a comparison program like SHIFTY will grow as more shifts are measured. Still, the current protocol should be a useful adjunct to homoogy-based models.

Conclusions

The chemical shift prediction method and program SHIFTS was developed and first applied to practical application for proteins in this work. Good results were obtained based on an additive model for multiple effects on chemical shift in proteins and the establishment of a computational database in which backbone conformation, side-chain orientation, sequence and hydrogen bonding are systematically varied. In some cases, the prediction accuracy may be improved by alterations in side-chain torsion angles; this analysis then provides specific suggestions for residues (mostly on the protein surface) where the predominant side-chain orientation in solution may differ from that found in crystal structures.

It is clear that there is still a lot of work to be done to understand the origins of chemical shift dispersion in proteins. The most obvious limitation of the current model is that it is limited to regions of regular secondary structure, i.e., to about 40% of residues among the proteins we studied. We are working to extend the number of residues for which predictions can be made by carrying out additional DFT calculations on peptides with a wider range of backbone torsion angles (cf. Figure 1). It is also likely that a combination of information derived from empirical and computational databases will in the end provide the most useful and reliable information. We hope that proton, nitrogen and carbon chemical shift data will be increasingly useful, in conjunction with other NMR data, in structural analyses of proteins in both high and low-resolution applications.

Acknowledgements

This work was supported by NIH grant GM 45811. We thank Frank Delaglio for advice on using the TALOS database, and David Wishart and Ad Bax for helpful comments and encouragement.

References

Ando, I., Kameda, T., Asakawa, N., Kuroki, S. and Kurosu, H. (1998) *J. Mol. Struct.*, **441**, 213–230.

Becke, A.D. (1993) *J. Chem. Phys.*, **98**, 5648–5652.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, F., Bryce, M.D., Rogers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) *J. Mol. Biol.*, **112**, 535–542.

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

Cornilescu, G., Marquardt, J.L., Ottiger, M. and Bax, A. (1998) *J. Am. Chem. Soc.*, **120**, 6836.

de Dios, A.C. (1996) *Prog. NMR Spectrosc.*, **97**, 229–278.

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

Frisch, M.J., Trucks, G.W., Schlegel, H.B., Gill, P.M.W., Johnson, B.G., Robb, M.A., Cheeseman, J.R., Keith, T., Petersson, G.A., Montgomery, J.A., Raghavachari, K., Al-Laham, M.A., Zakrzewski, V.G., Ortiz, J.V., Foresman, J.B., Cioslowski, J., Stefanov, B.B., Nanayakkara, A., Challacombe, M., Peng, C.Y., Ayala, P.Y., Chen, W., Wong, M.W., Andres, J.L., Replogle, E.S., Gomperts, R., Martin, R.L., Fox, D.J., Binkley, J.S., Defrees, D.J., Baker, J., Stewart, J.P., Head-Gordon, M., Gonzalez, C. and Pople, J.A. (1998) *Gaussian 98, Reversion A.6*, Gaussian, Inc., Pittsburgh, PA.

Gronwald, W., Boyko, R.F., Sönnichsen, R.D., Wishart, D.S. and Sykes, B.D. (1997) *J. Biomol. NMR*, **10**, 165–179.

Herranz, J., Gonzalez, C., Rico, M., Nieto, J.L., Santoro, J., Jimenez, M.A., Bruix, M., Neita, 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.

Lee, C., Yang, W. and Parr, R. (1988) *Phys. Rev.*, **B37**, 785–789.

Macke, T and Case, D.A. (1998) In *Molecular Modeling of Nucleic Acids,* N.B. Leontis and J. SantaLucia (Eds.), American Chemical Society, Washington, pp. 379–393.
Miehlich, B., Savin, A., Stoll, H. and Preuss, H. (1989) *Chem. Phys. Lett.*, **157**, 200.
Ösapay, K. and Case, D.A. (1991) *J. Am. Chem. Soc.*, **113**, 9436–9444.
Ösapay, K. and Case, D.A. (1994) *J. Biomol. NMR*, **4**, 215–230.
Osawa, M., Swindlls, M.B., Tanikawa, J., Tanaka, T., Mase, T., Furuya, T. and Ikura, M. (1998) *J. Mol. Biol.*, **276**, 165–176.
Pearson, J.G., Le, H., Sanders, L.K., Godbout, N., Havlin, R.H. and Oldfield, E. (1997) *J. Am. Chem. Soc.*, **119**, 11941–11950.
Pople, J.A., Head-Gordon, M., Fox, D. J., Raghavachari, K. and Curtiss, L.A. (1989) *J. Chem. Phys.*, **93**, 2537.
Ramage, R., Green, J., Muir, T.W., Ogunjobi, O.M., Love, S. and Shaw, K. (1994) *Biochem. J.*, **299**, 151.
Schwarzinger, S., Kroon, G.J.A, Foss, T.R., Chung, J., Wright, P.E. and Dyson, H.J. (2000) *J. Biomol. NMR*, **18**, 43–48.
Seavey, B., Farr, E.A., Westler, W.M. and Markley, J.A. (1991) *J.Biomol. NMR*, **1**, 217–236.
Sitkoff, D. and Case, D.A. (1997) *J. Am. Chem. Soc.*, **119**, 12262–12273.
Spera, S. and Bax, A. (1991) *J. Am. Chem. Soc.*, **113**, 5490–5492.
Szilágyi, L. (1995) *Prog. NMR Spectrosc.*, **27**, 325–443.
Szilágyi, L. and Jardetzky, O. (1989) *J. Magn. Reson.*, **83**, 441.
Vijay-Kumar, S., Bugg, C.E. and Cook, W.J. (1987) *J. Mol. Biol.*, **194**, 531.
Williamson, M.P. and Asakura, T. (1993) *J. Magn. Reson.*, **B101**, 63–71.
Williamson, M.P., Asakura, T., Nakamura, E. and Demura, M. (1992) *J. Biomol. NMR*, **2**, 83–98.
Wishart, D.S., Sykes, B.D. and Richards, F.M. (1991) *J. Mol. Biol.*, **222**, 311.
Wishart, D.S., Watson, M.S., Boyko, R.F. and Sykes, B.D. (1997) *J. Biomol. NMR*, **10**, 329–336.
Wolinski, K., Hilton, J.F. and Pulay, P. (1990) *J. Am. Chem. Soc.*, **112**, 8251.
Xu, X.-P and Case, D.A. (2001) submitted.
Yamazaki, T., Hinck, A.P., Wang, Y.-X., Nicholson, L.K., Torchia, D.A., Wingfield, P.T., Stahl, S.J., Kaufman, J.D., Chang, C.-H., Domaille, P.J. and Lam, P.Y.S. (1996) *Prot. Sci.*, **5**, 495–506.