Journal of Biomolecular NMR, 2 (1992) 83–98
ESCOM



J-Bio NMR 027

# A method for the calculation of protein α-CH chemical shifts

Michael P. Williamson[a,*], Tetsuo Asakura[b], Eiji Nakamura[b] and Makoto Demura[b]

[a]*Krebs Institute, Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield S10 2UH, U.K.*

[b]*Department of Biotechnology, Tokyo University of Agriculture and Technology, Koganei, Tokyo 184, Japan*

Received 22 April 1991
Accepted 5 July 1991

*Keywords:* Protein; NMR; Chemical shift

## SUMMARY

The chemical shifts of $C^{\alpha}H$ protons have been calculated for 9 proteins, based on coordinates taken from high-resolution crystal structures. Chemical shifts were calculated using ring-current shifts, shifts arising from magnetic anisotropies of bonds, and shifts arising from the polarizing effect of polar atoms on the $C^{\alpha}$-H bond. The parameters used were refined iteratively to give the best fit to (experimental − random coil) shifts over the set of 9 proteins. A further small correction was made to the averaged Gly $C^{\alpha}H$ shift. The calculated shifts match observed shifts with correlation coefficients varying between 0.45 and 0.86, with a standard deviation of about 0.3 ppm. The differences between calculated and observed shifts have been studied in detail, including an analysis of different crystal structures of the same protein, and indicate that most of the differences can be accounted for by small differences between the structure in solution and in the crystal. Calculations using NMR-derived structures give a poor fit. The calculations reproduce the experimentally observed differences between chemical shifts for $C^{\alpha}H$ in α-helix and β-sheet. Most of the differentiation in secondary-structure-dependent shifts arises from electric field effects, although magnetic anisotropy also makes a large contribution to the net shift. Applications of the calculations to assignment (including stereospecific assignment) and structure determination are discussed.

## INTRODUCTION

There is a long history of attempts to calculate chemical shifts in proteins. Considerable success has been achieved with the calculation of aromatic ring currents, based either on the classical Johnson-Bovey ring current loop model (Johnson and Bovey, 1958) or on semiclassical models (Hall et al., 1966; Haigh and Mallion, 1972; Memory, 1977); some improvements on the original

* To whom correspondence should be addressed.

0925-2738/$ 5.00 © 1992 ESCOM Science Publishers B.V.

method have utilised altered intensities for ring currents, particularly in histidine and tryptophan (Perkins and Dwek, 1980; Giessner-Prettre and Pullman, 1981). These models all give fairly similar results, and have been used widely and successfully, particularly for the calculation of methyl resonance frequencies in proteins.

Apart from ring current effects, there has been little obvious success in the calculation of protein chemical shifts, despite their ease of measurement and their ready application to structure calculation (Clayden and Williams, 1982). A significant result has been the demonstration of a correlation between the chemical shift of NH or $C^{\alpha}H$ protons and the inverse third power of the distance to its nearest carbonyl oxygen (Pardi et al., 1983; Wagner et al., 1983), but the results from this calculation are not sufficiently accurate to be of much use for structure calculations (but see Kline et al., 1988). Several authors (Sternlicht and Wilson, 1969; Asakura et al., 1977b; Asakura, 1981; Pastore and Saudek, 1990; Williamson, 1990) have noted a tendency for amide and $C^{\alpha}H$ protons to resonate to high field of their random coil values in helices and to low field in sheets, but until recently there has been no explanation for the effect.

We have shown (Asakura et al., 1991; Williamson and Asakura, 1991) that chemical shifts of $C^{\alpha}H$ protons in proteins can be calculated to good accuracy using the well-established methods for ring-current shifts, magnetic anisotropy effects and electric field effects, and that these calculations can account (among other things) for the secondary-structure-dependent effects noted above. The calculations have been made possible partly because of the number of assigned signals in proteins of well-defined three-dimensional structure, which has allowed a refinement of the parameters used. The purpose of this paper is to describe the calculational methods used, and to analyse their effectiveness and possible applications.

## METHODS

In general, chemical shifts of protons may be calculated as

$$\sigma = \sigma^{dia} + \sigma^{ani} + \sigma^{E} + \sigma^{ring} \tag{1}$$

where $\sigma^{dia}$ is the diamagnetic shift, $\sigma^{ani}$ is the shift due to anisotropy arising from locally induced currents on neighboring atoms, $\sigma^{E}$ is the polar effect arising from the electric fields created by polar groups in the molecule, and $\sigma^{ring}$ is the ring current effect from aromatic systems. $\sigma^{dia}$ depends on the atomic charge on the proton. All $C^{\alpha}H$ protons carry roughly the same charge (Asakura et al., 1977b); therefore the term $\sigma^{dia}$ has also been ignored. The calculation of $C^{\alpha}H$ shifts therefore reduces to

$$\sigma = \sigma^{ani} + \sigma^{E} + \sigma^{ring} \tag{2}$$

For $\sigma^{ring}$, we have used the equation of Johnson and Bovey (1958), with the ring-current intensity factors of Giessner-Prettre and Pullman (1981).

The term $\sigma^{ani}$ was calculated as the sum of the magnetic anisotropies arising from the C=O bond, the (O=)C-N bond, and the $C^{\alpha}$-N bond (Asakura et al., 1977a); alternatively, it is possible to treat the anisotropy of the amide group as a whole (Tigelaar and Flygare, 1972). We have so far ignored effects arising from the less common bonds, such as the C-O bond in serine and

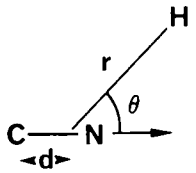


Fig. 1. Definition of coordinates and parameters for calculation of $\sigma^{ani}$ using the $C^{\alpha}$-N bond.

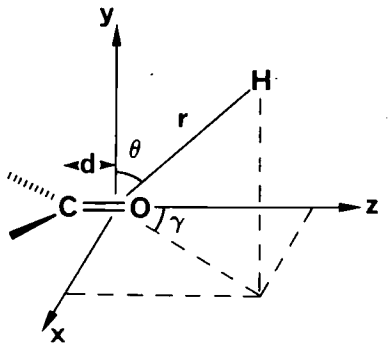


Fig. 2. Definition of coordinates and parameters for calculation of $\sigma^{ani}$ using $sp^2$-hybridised bonds. For C=O, the distance d is 1.1 Å. For (O=)C-N, d is 85% of the C-N distance (roughly 1.13 Å).

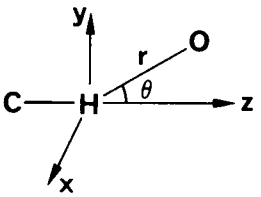


Fig. 3. Definition of coordinates and parameters for calculation of $\sigma^{E}$.

threonine side chains, since they will contribute little to the shifts of most $C^{\alpha}H$ protons. We have also ignored the C-C bond, as its anisotropy has been estimated to be small. The value of $\sigma^{ani}$ at proton A was calculated using McConnell's equation (McConnell, 1951), modified by ApSimon et al. (1967): for $C^{\alpha}$-N

$$\sigma^{ani} = (1/3r^3)\Delta\chi^{C\alpha N}(1 - 3\cos^2\theta) \tag{3}$$

where r is the distance between the centre of the anisotropy (B) and proton A, $\Delta\chi^{C\alpha N}$ is the magnetic anisotropy between axes parallel and perpendicular to the $C^{\alpha}$-N bond, and θ is the angle between the $C^{\alpha}$-N bond and the vector AB (Fig. 1). For the $sp^2$-hybridised C=O and (O=)C-N bonds,

$$\sigma^{ani} = (1/3r^3)[\Delta\chi_1(1 - 3\cos^2\theta) + \Delta\chi_2(1 - 3\sin^2\theta\cdot\sin^2\gamma)] \tag{4}$$

where the z axis is defined along the bond direction, the y axis lies perpendicular to the nodal plane of the π orbitals, and the x axis lies in the nodal plane; $\Delta\chi_1 = \chi_{yy} - \chi_{zz}$ is the magnetic anisotropy between the y and z axes, and $\Delta\chi_2 = \chi_{xx} - \chi_{zz}$ is the magnetic anisotropy between the x and z axes; θ is the angle between $r_{AB}$ and the y axis, and γ is the angle between the projection of r in the xz plane and the z axis (Fig. 2).

The term $\sigma^{E}$ calculated for proton $H_A$ arises from the polarisation of the C-$H_A$ bond caused by the intramolecular electric field arising from electric charges on atoms i in the molecule, and has been approximated as (Buckingham, 1960)

$$\sigma^{E} = \varepsilon_1 E_z + \varepsilon_2 E^2 \tag{5}$$

where the z direction is defined along the C-$H_A$ bond, the x and y directions are perpendicular to

it, $\theta_i$ is the angle between the i-$H_A$ vector and the C-$H_A$ bond, $Q_i$ is the charge on atom i, $r_{iA}$ is the distance between atom i and $H_A$ (Fig. 3), and

$$E_z = \Sigma(Q_i/r_{iA}^2)\cos\theta_i;\ E_y = \Sigma(Q_i/r_{iA}^2)\sin\theta_i;\ E^2 = E_z^2 + E_y^2 \quad (6)$$

In this work, the values of Q were those calculated by Momany et al. (1971): $Q_{C'} = +2.16$, $Q_N = -1.71$, and $Q_O = -1.84 \times 10^{-10}$ esu.

Calculation of the $C^\alpha H$ chemical shift thus involves calculation of the ring-current shift, followed by application of Eqs. 3, 4 and 5, using a total of 10 constants: $\Delta\chi_1^{C=O}$, $\Delta\chi_2^{C=O}$, $\Delta\chi_1^{C-N}$, $\Delta\chi_2^{C-N}$, and $\Delta\chi^{C\alpha N}$, the positions of the centres of magnetic anisotropy of the C=O, (O=)C-N and $C^\alpha$-N bonds, and $\varepsilon_1$ and $\varepsilon_2$. The values of these constants have been previously estimated as $\Delta\chi_1^{C=O} = -25.7$, $\Delta\chi_2^{C=O} = -13.5 \times 10^{-30}$ cm$^3$ (Zürcher, 1967), $\Delta\chi_1^{C-N} = -20.6$, $\Delta\chi_2^{C-N} = -13.2$, and $\Delta\chi^{C\alpha-N} = -10.4 \times 10^{-30}$ cm$^3$ (Asakura et al., 1977a), with the centres of anisotropy 1.0 Å from the C atom in the C=O bond, and in the centre of the bond for both C-N bonds; $\varepsilon_1 = -2.0 \times 10^{-12}$, $\varepsilon_2 = -10^{-18}$ (Asakura et al., 1977a) or $\varepsilon_1 = -2.9 \times 10^{-12}$, $\varepsilon_2 = -7.4 \times 10^{-19}$ (Pullman et al., 1978). However, many of these values are very crude estimates, and the values were therefore optimised, in an iterative manner (and altered extensively as a result), as described in the next section.

All calculations were carried out on a μVAX II computer. The programs were written in Fortran, and are available on receipt of a floppy disk. Two criteria were used to follow the optimisation process: the correlation coefficient R between calculated and experimental shifts, and the standard deviation sd of (experimental − mean experimental) − (calculated − mean calculated). In these calculations, the experimental shifts were corrected by subtraction of the random-coil shifts tabulated for that residue (Bundi and Wüthrich, 1979). Calculations were carried out on a set of 9 proteins, chosen because their three-dimensional structure is well characterised by high-resolution crystal structures, and their $^1H$ spectra are assigned and have indicated closely similar solution and crystal structures. The proteins used have been described previously (Williamson and Asakura, 1991) and are ubiquitin, ribonuclease $T_1$, eglin c, BPTI, human lysozyme, tendamistat, hen lysozyme, turkey ovomucoid third domain, and potato carboxypeptidase inhibitor; the Brookhaven protein databank structure names used for these were respectively 1UBQ, 1RNT, 1CSE, 6PTI, 1LZ1, 1HOE, 2LZT, 3SGB, and 4CPA. The references for the chemical shift values used for these proteins are given in our previous paper (Williamson and Asakura, 1991): for Gly $C^\alpha H$ shifts, the value taken was the average value, unless only one Gly $C^\alpha H$ shift was listed, in which case the value was omitted. Following our previous practice (Williamson, 1990), calculated $C^\alpha H$ shifts were not used in the calculation when the calculated ring-current shift exceeded 0.5 ppm, because of the likelihood of significant errors arising from small displacements of the ring position between crystal and solution structures.

## RESULTS

### *(a) Optimisation of parameters*

Of the initial set of parameters used, those for $\Delta\chi_1^{C=O}$ and $\Delta\chi_2^{C=O}$ were thought to be the most reliable; in justification of this, our previous calculations using these values alone (Asakura et al., 1991) had shown reasonable agreement between experimental and calculated shifts (Table

1, columns 2 and 3). Therefore the first round of optimisation kept these parameters fixed at the values given above (calculating main-chain carbonyl effects only), and varied those for $\Delta\chi_1^{C\text{-}N}$ and $\Delta\chi_2^{C\text{-}N}$, while setting to zero $\Delta\chi^{C\alpha N}$ and the electric field effect. For all optimisations, chemical shift calculations were performed using a range of values of the parameters to be optimised, and evaluated using three or four proteins. Both R and sd were assessed, although in the earlier stages of the optimisation, the behaviour of R was more regular, and R was used in preference to sd. In later stages sd was also well behaved, and refinement concentrated on improving sd without deterioration in R; by the end of the second stage of refinement, both sd and R had improved. Only if all proteins showed similar behaviour was the optimisation treated as meaningful, following which a best set of values was selected, and all proteins were recalculated using this best set. For most optimisations, all proteins showed remarkable consistency in the best values; this behaviour improved our confidence in the significance of the results, and the lack of any such consistency was a major factor in our rejection of $\Delta\chi^{C\alpha N}$ and $\varepsilon_2$ as useful parameters (see below).

This first optimisation of $\Delta\chi_1^{C\text{-}N}$ and $\Delta\chi_2^{C\text{-}N}$ resulted in very little change in the value of $\Delta\chi_1^{C\text{-}N}$, but a major change in $\Delta\chi_2^{C\text{-}N}$, from $-13.2$ to $+2.5 \times 10^{-30}$ cm$^3$. Following this, all four anisotropy values were optimised iteratively, together with the positions of the centres of anisotropy, leading to fairly small changes in parameters. At this point, introduction of $\Delta\chi^{C\alpha N}$ gave no improvement in the agreement between experimental and calculated shifts, for any value of $\Delta\chi^{C\alpha N}$.

The electric field effect was then introduced, and all values were reoptimised, using extensive iteration and comparison between the values obtained for different proteins. It was found that introduction of the $E^2$ term ($\varepsilon_2$) sometimes led to improvements in R, but almost invariably led to deter-

TABLE I
PROGRESS OF PARAMETER OPTIMISATION

| Protein | Zürcher values[a] | | Best, no side chains[b] | | Adding side chains[c] | | Final values[d] | | No. of protons used for fit[e] |
|---|---|---|---|---|---|---|---|---|---|
| | R | sd | R | sd | R | sd | R | sd | |
| Ubiquitin | 0.753 | 0.34 | 0.850 | 0.30 | 0.852 | 0.30 | 0.857 | 0.25 | 75 |
| RNAse $T_1$ | 0.774 | 0.35 | 0.844 | 0.30 | 0.840 | 0.31 | 0.859 | 0.27 | 89 |
| Eglin c | 0.713 | 0.32 | 0.806 | 0.34 | 0.803 | 0.34 | 0.807 | 0.28 | 62 |
| BPTI | 0.708 | 0.28 | 0.784 | 0.30 | 0.770 | 0.31 | 0.791 | 0.25 | 54 |
| Lysozyme (human) | 0.616 | 0.35 | 0.767 | 0.34 | 0.770 | 0.34 | 0.795 | 0.27 | 125 |
| Tendamistat | 0.579 | 0.34 | 0.736 | 0.38 | 0.732 | 0.38 | 0.733 | 0.30 | 70 |
| Lysozyme (hen) | 0.544 | 0.39 | 0.669 | 0.41 | 0.669 | 0.41 | 0.708 | 0.34 | 119 |
| Turkey ovomucoid | 0.560 | 0.29 | 0.606 | 0.33 | 0.636 | 0.33 | 0.696 | 0.26 | 48 |
| CPase inhibitor | 0.297 | 0.40 | 0.461 | 0.45 | 0.455 | 0.44 | 0.446 | 0.39 | 35 |

[a] Calculated using $\sigma^{ring} + \sigma^{ani}$ only: $\sigma^{ani}$ is calculated using only $\Delta\chi^{C=O}$, with Zürcher's values (1967).

[b] Parameters optimised without side chains ($\Delta\chi_1^{C=O} = -27.0$, $\Delta\chi_2^{C=O} = -9.0$, $\Delta\chi_1^{C\text{-}N} = -15.0$, $\Delta\chi_2^{C\text{-}N} = +1.3 \times 10^{-30}$ cm$^3$), side-chain anisotropies not calculated, no correction for glycine random coil shifts.

[c] Same parameters, side-chain calculations added.

[d] Final optimised parameters, including side chains and glycine random coil shifts.

[e] These numbers are less than the total number of $C^\alpha H$ in the protein, because values are not used where the assignments were not made, or where the calculated ring-current shift exceeded 0.5 ppm. In addition, the two C-terminal residues of BPTI are disordered in the crystal structure used, and therefore no calculated shifts are available.

ioration of sd. The term was therefore set to zero, and further optimisation justified the decision, by leading to better fits between experimental and calculated data. Similarly, attempts to include $\Delta\chi^{C\alpha N}$ at various points in the calculation always proved fruitless, and it was therefore omitted. The results using these parameters are given in Table 1, columns 4 and 5.

Up to this time, carbonyl groups in side chains had been ignored. They were then incorporated into the calculation, in the following way: for Asn and Gln, the C=O anisotropy was calculated, in the same way as a backbone carbonyl. For Asp and Glu, the effect from each of the two carbonyl oxygens was calculated in turn, and the result was averaged. Addition of the side-chain calculations resulted in very little change to the fit (Table 1, columns 6 and 7); however, subsequent refinement of parameters gave a significant improvement in fit (Table 1, columns 8 and 9).

*(b) Use of random coil values*

As noted above, the final calculated shift $\sigma$ was obtained by addition of $\sigma^{ring}$, $\sigma^{ani}$ and $\sigma^{E}$. The question then arises, to what value should $\sigma$ be compared – to the experimental $C^{\alpha}H$ shift, or to the experimental $C^{\alpha}H$ shift less the 'random coil' value of the $C^{\alpha}H$ shift for that residue type?

In principle, chemical shift calculations should be able to reproduce experimental shifts accurately, without the use of empirical corrections as represented by random coil shifts. However, many of the effects that contribute to the differences between the random coil shift values for different amino acid residues have not been included in these calculations. These would include the effects of neighbouring electronegative atoms, and the bond anisotropy arising from the different side chains, as well as the effect on $\sigma^{dia}$ coming from the differing charge on the $C^{\alpha}H$ in different residues. Therefore, we have taken the view that in the present state of the calculations it is better to compare our calculated values with (experimental – random coil) values, rather than directly with experimental values. The difference is substantial but not enormous; on comparison with experimental values directly, the fit for UBQ changes to R = 0.767, sd = 0.32, while that for BPTI changes to R = 0.635, sd = 0.37.

We note that the random coil shifts represent averages over all conformations allowed to each amino acid type in small peptides, and that each amino acid has a different ($\varphi$,$\psi$) conformational distribution. Therefore ideally, a complete calculation would calculate random coil shifts for the ($\varphi$,$\psi$) distribution of each amino acid, as

$$\sigma_{\text{random coil}} = \frac{\sum_{\varphi}\sum_{\psi}\sigma(\varphi,\psi)\exp[-E(\varphi,\psi)/RT]}{\sum_{\varphi}\sum_{\psi}\exp[-E(\varphi,\psi)/RT]} \qquad (7)$$

and subtract this from the tabulated value (Bundi and Wüthrich, 1979). The lack of any generally agreed ($\varphi$,$\psi$) distribution for each amino acid makes this approach difficult; however, we plan to attempt this calculation shortly. We anticipate that the effects should not be large, because (1) Comparison of the existing ($\varphi$,$\psi$) plots for different amino acids (e.g. Pullman and Pullman, 1974) shows little variation with residue type (apart from glycine); moreover, except for glycine, there are very few residue-specific conformational differences seen in crystal structures of proteins (J.M. Thornton, personal communication). (2) We have looked at the difference between calculated and (experimental – random coil) shifts for each amino acid, to look for systematic errors. Such a

search should reveal effects such as those discussed here, as well as effects due to residue charge and surface exposure. No such systematic effects were observed (except for glycine).

Glycine is something of a special case, since the $C^{\alpha}H$ are methylene, rather than methine, protons. This means that there is no $C^{\alpha}$-$C^{\beta}$ bond to affect the $C^{\alpha}H$ shift by its magnetic anisotropy. One might also expect an effect on $\sigma^{dia}$ due to the different charge distribution in the residue. We therefore reasoned that it might be necessary to apply a further correction to the random coil values for glycine in order to obtain a good fit between experimental and calculated values. After trying a range of values, it was found that an addition of 0.28 ppm to the calculated chemical shift for the averaged Gly $C^{\alpha}H$ shift gave a markedly better fit, particularly for sd.

To summarise, our method for $C^{\alpha}H$ chemical shift calculation runs as follows:

(a) For each proton, calculate $\sigma^{ring}$ using the Johnson-Bovey method;

(b) For each proton, go through all backbone and side-chain carbonyls, calculating $\sigma^{ani}$ using Eq. 4, with the parameters $\Delta\chi_1^{C=O} = -18.0$, $\Delta\chi_2^{C=O} = -8.0$, $\Delta\chi_1^{C-N} = -12.0$, $\Delta\chi_2^{C-N} = +1.3 \times 10^{-30}$ cm$^3$, with the centres of magnetic anisotropy 1.1 Å from the C atom in the C=O bond, and 85% of the way from C to N for the C-N bond;

(c) For each proton, calculate $\sigma^{E}$ using Eq. 5, with $\varepsilon_1 = -1.1 \times 10^{-12}$ and $\varepsilon_2 = 0$;

(d) Average each Gly $C^{\alpha}H$ pair, and add 0.28 ppm.

(e) This calculation gives values that can be compared to (experimental − random coil), with the results detailed in Table 1. In addition, the mean of the calculated values is 0.76 ppm higher than the mean of the (experimental − random coil) values. For a calculation of chemical shift values to correspond with the experimental values, it is therefore necessary to subtract 0.76 ppm and add the random coil values. The 0.76 ppm correction presumably arises in large part from the $\sigma^{dia}$ term, which we have ignored here (c.f. Eqs. 1 and 2)

Comparisons of calculated to experimental values are shown in Fig. 4.

*(c) Calculation of Gly $C^{\alpha}H$ shifts*

The calculations of Gly $C^{\alpha}H$ shifts described above gave results on both $C^{\alpha}H$ protons in a stereospecific manner, but were then averaged and compared to averaged experimental shifts, because of the difficulty in stereospecifically assigning Gly $C^{\alpha}H$ protons. Indeed, of the proteins in the table, only tendamistat has any Gly $C^{\alpha}H$ stereospecific assignments. We therefore carried out a study to investigate the possibility of using calculated shifts to stereospecifically assign Gly $C^{\alpha}H$ protons.

Calculations of Gly $C^{\alpha}H$ shifts were of comparable accuracy to those for other $C^{\alpha}H$, i.e. they

TABLE 2
CHEMICAL SHIFT VALUES FOR STEREOSPECIFICALLY ASSIGNED $C^{\alpha}H$ PROTONS IN TENDAMISTAT[a]

| Residue | Experimental Δδ(HA1 − HA2) | Calculated Δδ(HA1 − HA2) |
|---|---|---|
| 51 | − 0.39 | − 0.02 |
| 59 | + 0.67 | − 0.63 |
| 65 | − 0.30 | + 0.11 |

[a] Assignments from Kline et al. (1988).

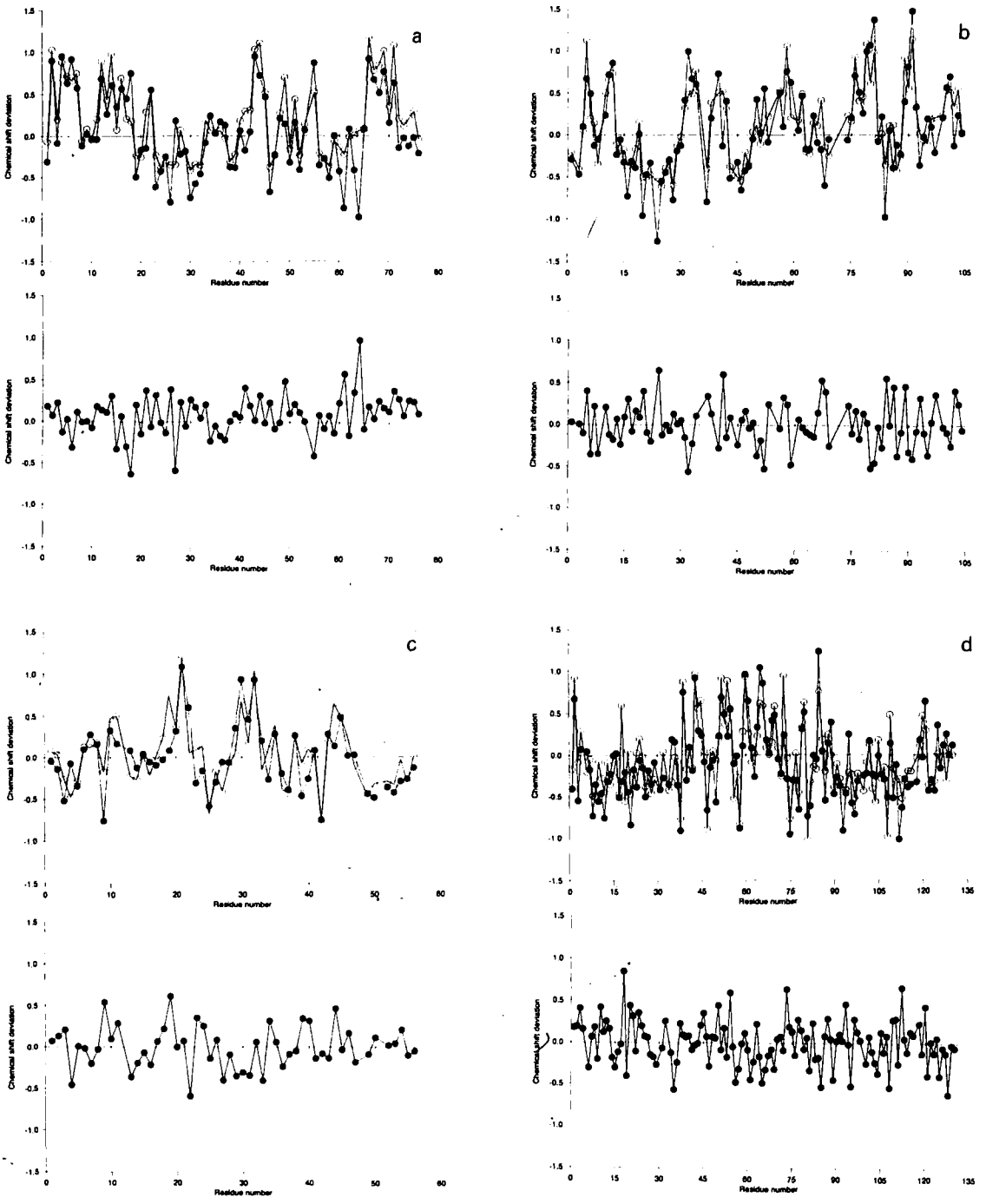


Fig. 4. Comparison of calculated and experimental $C^{\alpha}H$ shifts, by residue number. *Top panel:* Experimental (● - - - ●) and calculated (○ —— ○) shifts. The shifts are presented as the deviation from the random coil value. A positive value indicates a low field shift. *Lower panel:* Difference (calculated – experimental). The proteins shown are (a) ubiquitin, (b) ribonuclease $T_1$, (c) BPTI, (d) human lysozyme.

TABLE 3
PREDICTED GLY $C^{\alpha}H$ ASSIGNMENTS

| Protein | Residue | Experimental $\lvert\Delta\delta(HA1 - HA2)\rvert$ | Calculated $\Delta\delta(HA1 - HA2)$ |
|---|---|---|---|
| Ribonuclease | 7 | 0.39 | +0.50 |
| Eglin c | 40 | 0.68 | +0.34 |
| Eglin c | 70 | 0.92 | +0.71 |
| BPTI | 12 | 0.64 | +0.86 |
| Human lysozyme | 19 | 0.94 | +0.41 |
| Hen lysozyme | 4 | 0.33 | −0.33 |
| Ovomucoid | 54 | 1.04 | +0.30 |
| CPase inhibitor | 20 | 0.30 | −0.35 |
| CPase inhibitor | 35 | 1.45 | −0.81 |

had a value for the sd of between 0.25 and 0.30 ppm. We therefore reasoned that if the difference between the shifts for the two Gly $C^{\alpha}H$ protons was greater than 0.30 ppm in both the experimental and the calculated data, then the probability of an incorrect assignment based on chemical shift arguments should be 5% or less (corresponding to an error of 2 × sd). As shown in Table 2, the only stereospecifically assigned $C^{\alpha}H$ pair fitting this criterion is predicted incorrectly. Until more stereospecific assignments are available, it is not possible to test the hypothesis further. However, from the other proteins in this study, 9 pairs of Gly $C^{\alpha}H$ fit the criterion, and are listed in Table 3. We predict that for the pairs of protons listed with a positive number in the last column of Table 3, proton HA1 will prove to be the lower field of the two.

*(d) Comparisons of different structures*

In order to understand the reasons for the differences between calculated and experimental shifts, we decided to compare calculations for alternative structures of the same protein. This should give an indication of how much the results are dependent on the exact coordinates of the structure. As a first trial of the dependence of calculations on structure, we took the coordinates of various proteins and altered each coordinate by a fixed amount in a random direction. Some results are shown in Table 4, and show that, as expected, an increased magnitude in the random coordinate change decreases the correlations, both with experimental data and with the calculation done using the unchanged crystal coordinates.

A more meaningful result comes from a study of different crystal structures of the same protein. The Brookhaven protein database contains several structures for hen lysozyme, and several for BPTI. We therefore calculated $C^{\alpha}H$ shifts for each crystal structure, and compared the calculated shifts both with the experimental data and with each other. The results are shown in Table 5. Finally, we calculated shifts for each of the 9 NMR structures of tendamistat in the Brookhaven protein data bank. Individual structures had R and sd varying between 0.44–0.62 and 0.37–0.49 ppm, respectively. After averaging the calculated shifts for the 9 structures, the fit was R = 0.58, sd = 0.39 ppm.

TABLE 4
DEPENDENCE OF RESULTS ON RANDOM VARIATION OF COORDINATES

| Protein | Magnitude of random change (Å) | Correlation with experimental data | | Correlation with unchanged coordinate calculation | |
|---|---|---|---|---|---|
| | | R | sd | R | sd |
| UBQ | 0 | 0.857 | 0.25 | 1.000 | 0.00 |
| UBQ | 0.1 | 0.870 | 0.24 | 0.931 | 0.16 |
| UBQ | 0.15 | 0.828 | 0.28 | 0.845 | 0.24 |
| UBQ | 0.2 | 0.758 | 0.34 | 0.735 | 0.33 |
| PTI | 0 | 0.791 | 0.25 | 1.000 | 0.00 |
| PTI | 0.1 | 0.758 | 0.27 | 0.942 | 0.13 |
| PTI | 0.15 | 0.713 | 0.30 | 0.873 | 0.19 |
| PTI | 0.2 | 0.648 | 0.34 | 0.783 | 0.27 |

TABLE 5
RESULTS OF CALCULATIONS ON ALTERNATIVE CRYSTAL STRUCTURES

(a) Hen lysozyme

| Structure | Comparison with exptl. values | | Pairwise comparison with | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 2LZT | | 2LYM | | 1LYM(1) | | 1LYM(2) | |
| | R | sd | R | sd | R | sd | R | sd | R | sd |
| 1LZT | 0.80 | 0.27 | 0.77 | 0.28 | 0.79 | 0.27 | 0.61 | 0.39 | 0.60 | 0.39 |
| 2LZT | 0.79 | 0.28 | | | 0.91 | 0.15 | 0.70 | 0.32 | 0.69 | 0.33 |
| 2LYM | 0.78 | 0.29 | | | | | 0.74 | 0.30 | 0.65 | 0.34 |
| 1LYM(1)[a] | 0.64 | 0.39 | | | | | | | 0.51 | 0.44 |
| 1LYM(2) | 0.67 | 0.36 | | | | | | | | |

[a]Structure 1LYM has two nonequivalent molecules in the unit cell.

(b) BPTI

| Structure | Comparison with exptl. values | | Pairwise comparison with | | | |
|---|---|---|---|---|---|---|
| | | | 5PTI | | 4PTI | |
| | R | sd | R | sd | R | sd |
| 6PTI | 0.79 | 0.25 | 0.85 | 0.21 | 0.95 | 0.12 |
| 5PTI | 0.73 | 0.28 | | | 0.81 | 0.23 |
| 4PTI | 0.77 | 0.26 | | | | |

## DISCUSSION

*(a) Can the calculations be improved?*

The fits between calculated and experimental shifts reported above are good but not excellent. There are several possible explanations for this: the parametrisation may be inadequate, both in terms of the optimised values used for the parameters, and the equations adopted; or it may be that the solution structures of the proteins are different from the crystal structures, and thus no matter how accurate the calculational method, it will never be able to reproduce experimental NMR shifts in solution from crystal structures. We shall consider these points in turn.

There is probably scope for improvement in the parametrisation. With the calculational scheme described here, increases in the magnitude of $\Delta\chi_1^{C=O}$ and $\Delta\chi_1^{C-N}$ lead to a general (but small) improvement in R, but a deterioration in sd, indicating that the parameters may not be at their optimum values. Substantial improvements will probably have to await more NMR assignments of proteins with high-resolution crystal structures. As described above, there are several possible contributions to the chemical shift that we have not calculated, including the bond anisotropies of C-C, C-O, C-H and N-H bonds, as well as the $C^{\alpha}$-N bonds that we calculated but rejected, because they did not give any general improvement in the fit. We have also only calculated electric field effects from the charges of backbone C′, O and N atoms; other atoms could be considered. Semiempirical force-field calculations have shown variations in atomic charges in peptides depending on the local backbone conformation. This would affect electric field calculations, and would alter $\sigma^{dia}$. Inclusion of additional effects such as these would add considerably to the complexity of the calculation, but may result in real improvements. They could also remove the somewhat clumsy expedient of using random coil shifts as part of the calculation. However, for reasons discussed below, we do not consider that it is worth implementing most of these sophistications at present, although work on $C^{\beta}H$ and NH protons, to begin shortly, may throw further light on their importance.

We consider that the main reason for the lack of accuracy in our calculations is the fact that the solution structure differs from the crystal structure. It differs in two respects: firstly, in that the crystal structure may have small residual errors, which, although small in terms of atomic coordinates, can produce significant differences in the calculated shift; and secondly, in that the protein in solution may have different mean atomic positions from the protein in the crystal, and furthermore will be much more mobile than in the crystal.

The results presented in the previous section indicate that errors in the atomic coordinates of around 0.15–0.2 Å could account for almost all the discrepancies found between calculated and observed shifts. Errors of this magnitude are unlikely in any high-resolution structure, except in localised regions. However, as the resolution of a crystal structure gets worse, atomic coordinates become less precise. In particular, at resolutions of 2.5 Å or worse, it becomes difficult to identify the orientation of the peptide bond, with the result that the plane of the peptide bonds may become altered by 180°. This would lead to major errors in the calculated shift. Examination of the results for the proteins studied here shows that there is no discernible correlation between crystal structure resolution and goodness of fit, except that the three crystal structures with resolutions worse than 2.0 Å (potato carboxypeptidase inhibitor, and the two hen lysozyme structures 1LYM(1) and 1LYM(2), all with resolutions of 2.5 Å) gave markedly worse fits*. Thus, there are

clear indications that imprecisions in the crystal structure coordinates contribute to the discrepancies found between calculated and observed shift.

However, the more important factor would appear to be the difference between crystal and solution structure. Evidence for this comes from the comparisons between alternative crystal structures of the same protein, listed in Table 5. The solution structure of hen lysozyme has been shown to be similar to the crystal structure, but is presumably not identical to any individual structure. One might reasonably suppose that the difference between the solution structure and any one crystal structure would be of the same order of magnitude as that between one crystal structure and another, and therefore that the difference between calculated and experimental shifts would be of a similar magnitude to the difference between shifts calculated for different crystal structures, assuming a perfect calculational procedure. This is certainly the case for the hen lysozyme structures, and almost true for the BPTI structures, implying that much of the discrepancy, for these two proteins at least, arises from the difference between crystal and solution structures. Presumably, the reason for the differences in the R and sd values for the different proteins, shown in Table 1, is related to their different degrees of similarity between solution and crystal structure.

A further test of the theory is provided by comparisons of calculated chemical shifts only for those parts of the protein that are most likely to be similar in solution and in the crystal. For example, the N-terminal residues are frequently somewhat disordered in solution; omission of the three N-terminal residues from the comparison of experimental to calculated shift results in a significant improvement in most cases. For a more precise test, we deleted the results for $C^{\alpha}H$ when the crystallographic B-factor of the associated $C^{\alpha}$ was large. For all proteins tested, this resulted in a small improvement in the fit, providing a further indication that much of the discrepancy arises from the differences between crystal and solution structures. Inspection of Fig. 4 shows that many of the regions of large discrepancy between calculated and observed shift are found in loops and turns. These residues include, in ubiquitin, residues 14–21, 61–64 and the C-terminus; in ribonuclease $T_1$, residues 5–9, 67–69 and 92–97; in BPTI, residues 17–19, 39–40 and 44; and in lysozyme, residues 18–23, 54–57, 85–89, and the exposed loop/helix/turn section 104–118. Figure 5 shows the location of protons in ubiquitin whose calculated chemical shift differs by more than 0.3 ppm from the observed shift. These protons are clearly localised in certain regions of the protein in most cases, notably in the loops on the left of the figure, in the C-terminal strand, and in the helix section 23–30.

Rather than using crystal structures, one could use NMR structures, calculated for the protein in solution. The results for tendamistat, using the 9 structures calculated using distance geometry methods (Kline et al., 1988), are given in the previous section and are rather disappointing. Our interpretation of the results is that the NMR structures present a less precise picture of the solution structure than does the X-ray structure (although the NMR results may conceivably be more accurate); in other words, the effective 'X-ray resolution' of the NMR structures is probably only about 2.5 Å, but they may more fully represent the structure of the protein in solution than does the X-ray structure.

*This would imply that any calculation of chemical shifts on a protein with a resolution of 2.5 Å or worse is likely to be fruitless.

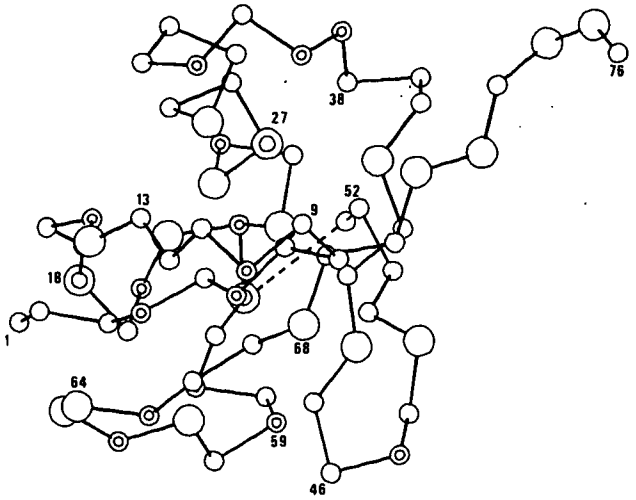


Fig. 5. Representation of the difference between calculated and observed $C^{\alpha}H$ shift for ubiquitin. The balls are drawn at the positions of the $C^{\alpha}$, with radii 0.5 Å for deviations less than 0.3 ppm, and 0.8 Å for deviations greater than 0.3 ppm. Balls are drawn with two circles for negative deviations. Residue 54 is omitted, because the $C^{\alpha}H$ was not assigned in the assignment list used (Di Stefano and Wand, 1987).

Finally, we note that solution structures are mobile. The mobility of the structures will lead to nonlinear chemical shift changes, which in general will be unpredictable. Nevertheless, one might hope that an approximation to the time-averaged shift might be available from a weighted ensemble of structures, produced for example by molecular dynamics or Monte Carlo methods. It remains to be seen how large an effect on chemical shift is played by mobility, but, judging by the rather small temperature dependence observed for $C^{\alpha}H$ protons in proteins over rather large temperature ranges, the effect may well be small. However, for amide protons, such effects could well be significant.

*(b) The origin of secondary-structure-dependent shifts*

We have previously shown (Williamson, 1990) that, after allowing for ring-current shifts, $C^{\alpha}H$ in helices tend to be shifted 0.30 ppm to high field of their random coil position, whereas $C^{\alpha}H$ in β-sheets tend to come 0.36 ppm to low field, with a high statistical significance. Our calculations reproduce this difference very closely (Williamson and Asakura, 1991), and we feel that it is of interest to look in detail at the origins of the difference in average shift between α-helix and β-sheet, because of the widespread interest in the problem (Asakura et al., 1991). Accordingly, we have studied the influences on the chemical shift of several protons in β-sheet and α-helix. Figure 6 shows the details for representative protons in β-sheet and α-helix, namely $Lys^{6}$ and $Asp^{32}$ from ubiquitin, respectively.

Figure 6 indicates that there are many contributors to the chemical shift of any one proton, particularly to the $\sigma^{E}$ terms, which, being $r^{-2}$-dependent, carry over long distances. Many more groups have a significant effect than there was space to include on Fig. 6. Only those groups are shown in Fig. 6 that contribute by over 0.1 ppm: there are 26 groups that affect $Lys^{6}$ by at least 0.02 ppm, and 21 that affect $Asp^{32}$ by the same amount. However, it is striking that most of the

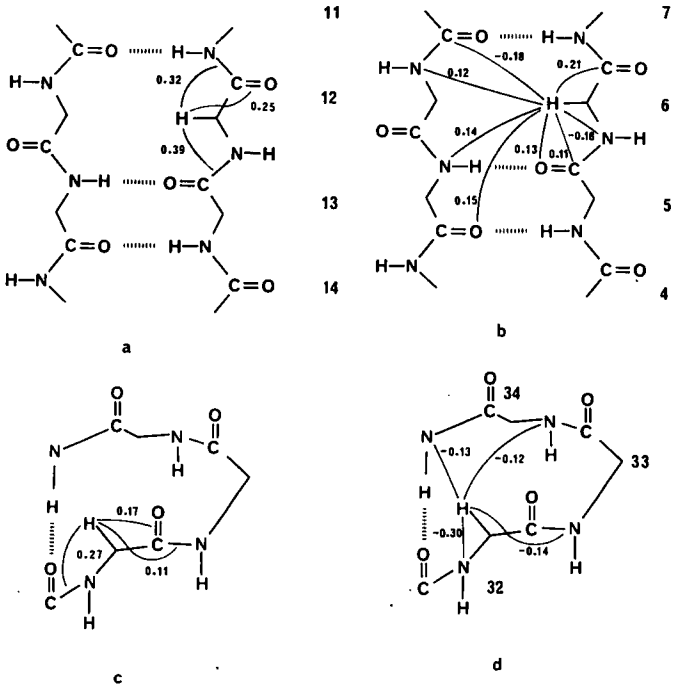


Fig. 6. Contributions to the chemical shift of protons in an α-helix and a β-sheet. a and b: Lys[6] of UBQ (β-sheet). c and d: Asp[32] of UBQ (α-helix). a and c: $\sigma^{ani}$. b and d: $\sigma^{E}$. Only contributions greater than 0.1 ppm are shown.

major contributors to the chemical shift, even in a β-sheet, are within 4 bonds of the $C^{\alpha}H$ affected; in other words, the chemical shift is determined largely by the local backbone conformation. The difference between chemical shifts in α-helices and β-sheets is seen to lie largely in $\sigma^{E}$, and arises from the different orientation of the $C^{\alpha}$-H vector relative to the carbonyl group in the different secondary structure types, which causes $\sigma^{E}$ to be positive for sheets and negative for helices.

One further point to emerge from Fig. 6 is the relative importance of $\sigma^{ani}$ and $\sigma^{E}$ to the total chemical shift. It appears from our calculations that the two are roughly equally important (in contrast to Gresh and Giessner-Prettre, 1990, who found a predominant importance for $\sigma^{E}$). We note that $\sigma^{ani}$ is proportional to $r^{-3}$, whereas $\sigma^{E}$ is proportional to $r^{-2}$; an important contribution from $\sigma^{ani}$ at least partially explains the $r^{-3}$-dependence for chemical shift observed previously (Pardi et al., 1983; Wagner et al., 1983).

*(c) Future development*

We hope that this work will act as a stimulus to further research on the origins of chemical shift, both in proteins and elsewhere. However, there are several more directly useful applications. Perhaps the most obvious, and the most eagerly awaited, is that it may at last start to become possible to use $C^{\alpha}H$ shifts for protein assignment, particularly when the shifts are unusually different from the random coil value. This application will be particularly useful when some amino acids have been assigned by residue type, but not yet by location in the sequence. Thus, one can envisage that

for a 150-residue protein containing 8 alanine residues, perhaps three might have sufficiently unusual shifts for them to be assigned immediately*. Not quite so obvious, but also very useful, is that it may be possible to use chemical shifts to make *stereospecific* assignments, of $C^{\beta}H$ and the methyl groups of valine and leucine. The results presented above suggest that stereospecific assignment of Gly $C^{\alpha}H$ is not possible in many cases using chemical shifts; we shall shortly undertake calculations to test the feasibility of using chemical shifts for the stereospecific assignments of $C^{\beta}H$ and methyl groups.

With the use of protein $^{15}N$ labelling, it is becoming increasingly common to be able to assign backbone nuclei for ever larger proteins, but not to solve their structures in any detail. Given a good crystal structure, chemical shift calculations have several applications in such a situation. It should be possible, as shown above, to identify regions of the protein where solution and crystal structures are similar, and regions where they are not. As discussed above, the regions of low similarity are often those where mobility is evident in the crystal, by a high B-factor, but there may well be other regions (such as crystal contacts) where low similarity is not paralleled by a high B-factor. This brings us to the most difficult, but perhaps the most exciting, application, namely, that chemical shifts could be used as a structure refinement tool. This would at last make some use of the chemical shift parameter, one of the most easily measured and (so far) least useful NMR parameters. Such an application, along the lines of recent research being carried out on ring-current shifts, would involve extensive calculation, probably using Monte Carlo methods, but could be of enormous benefit in locating and correcting local deviations from the correct solution structure.

## ACKNOWLEDGEMENTS

We thank Dr. Peter J. Artymiuk (Krebs Institute) for advice, and Professor Timothy A. Cross (Florida State University, U.S.A.) for helpful discussions.

*Note added in proof*

We have recently become aware of an independent study of protein chemical shifts that reaches very similar conclusions to ours (Ösapay, K. and Case, D.A. (1991) *J. Am. Chem. Soc.*, in press).

## REFERENCES

ApSimon, J.W., Craig, W.G., Demarco, P.V., Mathieson, D.W. and Saunders, L. (1967) *Tetrahedron*, **23**, 2357–2373.
Asakura, T. (1981) *Makromol. Chem.*, **182**, 1097–1109.
Asakura, T., Ando, I. and Nishioka, A. (1977a) *Makromol. Chem.*, **178**, 1111–1132.
Asakura, T., Ando, I. and Nishioka, A. (1977b) *Makromol. Chem.*, **178**, 1521–1533.
Asakura, T., Nakamura, E., Asakawa, H. and Demura, M. (1991) *J. Magn. Reson.*, **93**, 355–360.
Buckingham, A.D. (1960) *Can. J. Chem.*, **38**, 300–307.
Bundi, A. and Wüthrich, K. (1979) *Biopolymers*, **18**, 299–311.
Clayden, N.J. and Williams, R.J.P. (1982) *J. Magn. Reson.*, **49**, 383–396.
Di Stefano, D.L. and Wand, A.J. (1987) *Biochemistry*, **26**, 7272–7281.
Giessner-Prettre, C. and Pullman, B. (1981) *Biochem. Biophys. Res. Commun.*, **101**, 921–926.

*It should be obvious that a high-resolution (2 Å or better) crystal structure must be available first.

Gresh, N. and Giessner-Prettre, C. (1990) *Biochem. Biophys. Res. Commun.*, **171**, 1211–1216.
Haigh, C.W. and Mallion, R.B. (1972) *Org. Magn. Reson.*, **4**, 203–228.
Hall, G.G., Hardisson, A. and Jackman, L.M. (1966) *Tetrahedron*, **19**, (Suppl. 2), 101–109.
Johnson, C.E. and Bovey, F.A. (1958) *J. Chem. Phys.*, **29**, 1012–1014.
Kline, A.D., Braun, W. and Wüthrich, K. (1988) *J. Mol. Biol.*, **204**, 675–724.
McConnell, H.M. (1951) *J. Chem. Phys.*, **27**, 226–229.
Memory, J.D. (1977) *J. Magn. Reson.*, **27**, 241–244.
Momany, F.A., McGuire, R.F., Yan, J.F. and Scheraga, H.A. (1971) *J. Phys. Chem.*, **75**, 2286–2297.
Pardi, A., Wagner, G. and Wüthrich, K. (1983) *Eur. J. Biochem.*, **137**, 445–454.
Pastore, A. and Saudek, V. (1990) *J. Magn. Reson.*, **90**, 165–176.
Perkins, S.J. and Dwek, R.A. (1980) *Biochemistry*, **19**, 245–248.
Pullman, A., Berthod, H., Giessner-Prettre, C., Hinton, J.F. and Harpool, D. (1978) *J. Am. Chem. Soc.*, **100**, 3991–3994
Pullman, B. and Pullman, A. (1974) *Adv. Protein Chem.*, **28**, 347–524.
Sternlicht, H. and Wilson, D. (1969) *Biochemistry*, **6**, 2881–2892.
Tigelaar, H.L. and Flygare, W.H. (1972) *J. Am. Chem. Soc.*, **94**, 343–346.
Wagner, G., Pardi, A. and Wüthrich, K. (1983) *J. Am. Chem. Soc.*, **105**, 5948–5949.
Williamson, M.P. (1990) *Biopolymers*, **29**, 1423–1431.
Williamson, M.P. and Asakura, T. (1991) *J. Magn. Reson.*, **94**, 557–562.
Zürcher, R.F. (1967) *Progr. Nucl. Magn. Reson. Spectrosc.*, **2**, 205–257.