Journal of Biomolecular NMR, 4 (1994) 807–826
ESCOM

J-Bio NMR 215

An evaluation of least-squares fits to COSY spectra as a means of estimating proton–proton coupling constants. I. Simulated test problems

Ju-Xing Yang and Timothy F. Havel*

Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 240 Longwood Avenue, Boston, MA 02115, U.S.A.

Received 24 November 1993
Accepted 6 July 1994

Keywords: 2D NMR; Simulation of spectra; COSY; Correlated spectroscopy; Proton–proton coupling constants; Scalar coupling; Parameter estimation

SUMMARY

A computational method is described that takes an initial estimate of the chemical shifts, line widths and scalar coupling constants for the protons in a molecule, and refines this estimate so as to improve the least-squares fit between an experimental COSY spectrum and the spectrum simulated from these parameters in the weak-coupling approximation. In order to evaluate the potential of such refinements for estimating these parameters from COSY experiments, the method has been applied to a large number of sample problems which were themselves simulated from standard conformations of the amino acids, along with 25 near-native conformations of the protein bovine pancreatic trypsin inhibitor. The results of this evaluation show that: (i) if the chemical shifts are known to within ca. 0.01 ppm and no noise or artifacts are present in the data, the method is capable of recovering the correct coupling constants, starting from essentially arbitrary values, to within 0.1 Hz in almost all cases. (ii) Although the precision of these estimates of the coupling constants is degraded by the limited resolution, noise and artifacts present in most experimental spectra, the large majority of coupling constants can still be recovered to within 1.0 Hz; the local minimum problem is not made significantly worse by such defects in the data. (iii) The method assigns an 'effective' line width to all the resonances, and in the process can resolve overlapping cross peaks. (iv) The method is not capable of determining the chemical shifts a priori, due to the presence of numerous local minima in the least-squares residual as a function of these parameters.

INTRODUCTION

Numerous procedures have been developed for estimating scalar coupling constants from homonuclear COSY-like experiments. These can be divided into roughly three classes. In the first class, special pulse sequences are used to make it relatively easy to measure the coupling constants

*To whom correspondence should be addressed.

from the fine structures of the cross peaks. For example, E.COSY (Griesinger et al., 1986) and z-COSY (Oschkinat et al., 1986) experiments yield spectra with simplified in-phase multiplet patterns for all the cross peaks. In the second class, data processing methods are used to extract the coupling constants from the spectra. Examples here include the DISCO method (Kessler et al., 1985), a related method involving NOESY spectra (Ludvigsen et al., 1991), and elimination of the peak width between absorptive and dispersive components (Kim and Prestegard, 1989). In the third class, simulations of various sorts are used to find values for the coupling constants that reproduce the observed fine structures. These methods have usually been developed to work in combination with other experiments and/or data processing techniques, for example TOCSY (Titman and Keeler, 1990), E.COSY (Smith et al., 1991), NOESY (Szyperski et al., 1992), and the 'J-doubling' method (Jones et al., 1993). Of course, if the molecule of interest can be isotopically labeled many heteronuclear methods are also applicable, but in this paper we limit ourselves to homonuclear experiments.

Once one has the correct coupling constants, density matrix or product operator calculations (Ernst et al., 1987) can be used to simulate the experimental spectra and thereby confirm them. Alternatively, one can carry out the simulation in the frequency domain, by diagonalizing the Hamiltonian directly (Widmer and Wüthrich, 1986). In principle, it should also be possible to refine the coupling constants so that the spectra obtained in this way agree with the experimental spectra, and perhaps even to automatically extract the coupling constants from the data. This has successfully been done with the E.COSY experiment (Mádi and Ernst, 1988), but to date the method has been too computationally demanding to be applied to entire spectra of large molecules such as proteins.

In this paper we describe and validate a simplified and computationally more efficient version of the frequency domain approach. This purely phenomenological procedure simulates 2D COSY spectra directly from the coupling constants, chemical shifts and line widths, without attempting to reproduce the full quantum-mechanical evolution of the spin systems involved. The efficiency of the procedure is greatly improved by using a novel matrix decomposition of 2D NMR spectra (Havel et al., 1994), which also makes it relatively easy to compute the derivative of the spectrum with respect to these parameters. The availability of these derivatives, in turn, makes it possible to minimize the sum over all points of the squared differences between the observed and simulated spectra with respect to the parameters, using a standard conjugate gradient algorithm. This algorithm has the advantage of requiring computer storage that grows only linearly with the number of spins, and is computationally quite efficient.

One obvious defect of the procedure for the purpose of estimating the coupling constants is that, as currently implemented, it uses the weak-coupling approximation for all pairs of spins. This is, of course, true for most existing methods of estimating homonuclear coupling constants. Our primary goal in this paper, however, is simply to evaluate the severity of three other, potentially more serious, obstacles to the procedure. The first of these stems from the fact that functions measuring the difference between observed and calculated spectra might possess numerous *local* minima, so that it could be very difficult to find reasonably good fits by minimizing them. The second lies in the fact that least-squares methods are only assured of being able to alleviate the effects of *random* errors in the data, whereas actual 2D NMR spectra, particularly of macromolecules in aqueous solution, are afflicted by numerous artifacts. Finally, the least-squares residual may change slowly with changes in the coupling constants and other parameters, so that merely random noise in the data can render even the globally optimum parameters unreliable. Such parameter estimation problems are called *ill-conditioned.*

The evaluation itself consists of an extensive set of test problems, designed to mimic the situations that occur in solution NMR studies of proteins. These test problems were generated by simulating the spectra from realistic values of the coupling constants and other parameters, in most cases with noise and typical data processing artifacts added. It is worth stressing that the use of simulated data is an *absolute* necessity when validating a new method, since only then does one know for sure what errors are present in the data as well as what the correct answer is. It is of course possible that unforeseen problems will arise when the method is subsequently applied to experimental data, and hence the results presented here should be viewed as a 'best-case' scenario. In our companion paper (Yang et al., 1994), we present further results obtained with actual experimental data which demonstrate that, with high-quality data and suitable processing techniques, the scenario presented here is not wildly optimistic.

If it were possible to converge to a set of chemical shifts that fit the target spectra starting from any reasonable initial guess at their correct values, of course, we would be able to automatically identify the spin systems by this procedure. Although this has not proved possible, at least with the optimization procedure we have used here, because the measured chemical shifts usually have small errors in them and may vary with pH and temperature, it is nevertheless necessary to treat the chemical shifts and line widths as variables in addition to the coupling constants in order to obtain reasonably good fits to actual experimental spectra.

COMPUTATIONAL METHODS

The computer program we have developed is based upon the fact that, if one assumes an absorptive line shape for all the peaks, a 2D COSY spectrum can be decomposed into a product of matrices (Havel et al., 1994)

$$\mathbf{S} = \mathbf{P}\mathbf{C}\mathbf{Q}^{\mathrm{T}} \tag{1}$$

where the matrix **S** of size M by N points contains the digitized spectrum, and the matrix **C** ($K \times L$) contains the signed volumes of the individual peaks in each cross peak of the spectrum. The columns of the $M \times K$ *peak matrix* **P** contain the digitized contours of the individual peaks in the 1D spectrum whose chemical shifts lie in the range of the ω_1 axis, while the $N \times L$ peak matrix **Q** contains the contours of the peaks in the range of the ω_2 axis. When the full symmetric spectrum is simulated, $\mathbf{Q} = \mathbf{P}$ and $\mathbf{C} = \mathbf{C}^{\mathrm{T}}$.

In order to compute the matrices **P** and **Q**, it is necessary to assume a line-shape function $L(\mu)$ for the spectrum. Although basic theory implies that this line shape should be Lorentzian, in practice (because of field inhomogeneity, truncation and apodization) the line shapes in the 2D spectra of large molecules are closer to Gaussian. Hence, in the current implementation of the program we have used a Gaussian of the form

$$G(\mu) = \frac{a}{\sqrt{2\pi}\sigma}\, e^{-(\mu-\nu)^2/2\sigma^2} \tag{2}$$

where a is an intensity constant, ν is the chemical shift of a peak in the (1D) multiplet, σ is a measure of the peak width and μ is a frequency variable. It should be clearly understood, however, that our method is not intrinsically limited to a Gaussian line shape.

The computation of **Q** is the same as that of **P**, and hence in what follows we will only discuss the procedure used to compute **P**. Supposing there are S spins in the molecule, **P** can be written as a concatenation $\mathbf{P} = (\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_S)$ of $M \times K_i$ matrices, where the K_i columns of each $\mathbf{P}_i$ contain the shapes of the peaks in the multiplet for the ith spin. If we know the chemical shifts of the spins, the peaks in the multiplet of each spin can be computed as follows. Each coupling constant J splits each peak of a spin into two peaks, displaced from the original chemical shift by an amount $\Delta^{\pm} = \pm J/2$. Thus, if a spin is coupled to I other spins, the centers of the peaks in its multiplet are given by

$$\nu\left(\Delta_1^{\varepsilon_1}, \Delta_2^{\varepsilon_2}, \ldots, \Delta_I^{\varepsilon_I}\right) = \bar{\nu} + \Delta_1^{\varepsilon_1} + \Delta_2^{\varepsilon_2} + \ldots + \Delta_I^{\varepsilon_I} \tag{3}$$

where $\bar{\nu}$ is the chemical shift of the spin, ν is the chemical shift of a peak in the one-dimensional multiplet for that spin, $\Delta_i^{\varepsilon_i}$ is the amount of split due to J_i, and $\varepsilon_i = \pm 1$. The 2^I new peaks that result have intensities equal to 2^{-I} times the intensity of the peak before splitting.

Since each peak is limited to only a small region of the spectrum (particularly for a Gaussian line shape), it is possible to use a cutoff technique to reduce both the computation time and memory required for the simulation. In order to obtain continuous first derivatives (see below), we use a cubic spline cutoff, which leads to a line-shape function of the form

$$L(\mu) = \begin{cases} G(\mu) & \text{if } G(\mu) \geq C_1 \\ G(\mu)H(G(\mu)) & \text{if } C_1 > |\,G(\mu)\,| \geq C_2 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where

$$H(G) = A(G - C_2)^3 + B(G - C_2)^2 \tag{5}$$

with

$$A = -2/(C_1 - C_2)^3 \tag{6}$$

and

$$B = 3/(C_1 - C_2)^2 \tag{7}$$

The constants A and B ensure that H satisfies the conditions $H(C_1) = 1$, $H'(C_1) = 0$, $H(C_2) = 0$, and $H'(C_2) = 0$. With this cutoff function, most of the elements of **P** become zero and do not need to be stored, computed or, most importantly, used in computing the matrix products. For example, if a 500 MHz spectrum has a width of 10 ppm and a typical line width of 10 Hz, only about 1% of the points need to be computed, assuming $C_2 = G(5\sigma)$. This leads to a nearly 100-fold decrease in the computation time required.

The matrix **C** is computed according to the sign rules described in Neuhaus et al. (1985). If two spins are actively coupled, the split gives a plus sign for the $+J/2$ peak and a minus sign for the $-J/2$ peak, whereas if two spins are passively coupled, the split does not change the sign. The matrix **C** can be made to express this rule by decomposing it into a matrix of $K_i \times K_j$ submatrices, written as

$$\begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} & \cdots & \mathbf{C}_{1S} \\ \mathbf{C}_{21} & \mathbf{C}_{22} & \cdots & \mathbf{C}_{2S} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{C}_{S1} & \mathbf{C}_{S2} & \cdots & \mathbf{C}_{SS} \end{bmatrix} \tag{8}$$

The way in which the diagonal submatrices $\mathbf{C}_{ii}$ corresponding to the diagonal multiplets are computed will depend on the exact type of COSY experiment being simulated*. Since these multiplets are heavily overlapped, they contain relatively little usable information and hence should not usually be included when fitting simulated to experimental spectra. For these reasons, we have set $\mathbf{C}_{ii} = 0$ in this work.

Since a spin is only coupled with a few other spins in the same residue, the matrix **C** will also be very sparse. Because **C** is determined by how the spins are coupled, i.e., actively or passively, rather than by the coupling constants, it is independent of the coupling constants and does not vary in the course of our computations. Each submatrix $\mathbf{C}_{ij}$ can be written as a dyadic product of two vectors

$$\mathbf{C}_{ij} = \mathbf{v}_i(j)\mathbf{v}_j^{\mathrm{T}}(i) \tag{9}$$

where i and j are spin indices and $\mathbf{v}_i(j)$ is a vector of size K_i whose elements can only be +1, −1 or 0. To illustrate how it is computed, we consider the case where a spin i is coupled with a spin j by a coupling constant $J(i,j)$ and with a spin k by a coupling constant $J(i,k)$. Before the coupling is applied, $\mathbf{P}_i$ has one column, $K_i = 1$ and $\mathbf{v}_i(j) = +1$. When $J(i,j)$ is applied, $\mathbf{P}_i$ is split into two columns, $(\mathbf{P}_i^1(ij), \mathbf{P}_i^2(ij))$, where ij is used to emphasize that the split is due to the active coupling between spins i and j. Due to this split, $\mathbf{v}_i(j)$ becomes $(+1,-1)^{\mathrm{T}}$, whereas $\mathbf{v}_i(k)$ becomes $(+1,+1)^{\mathrm{T}}$, since k is passively coupled to j. On applying the second coupling constant $J(i,k)$, each column of $\mathbf{P}_i$ is further split into two columns, $\mathbf{P}_i = (\mathbf{P}_i^1(ij,ik), \mathbf{P}_i^2(ij,ik), \mathbf{P}_i^3(ij,ik), \mathbf{P}_i^4(ij,ik))$, $\mathbf{v}_i(j)$ is passively split into $(+1,+1,-1,-1)^{\mathrm{T}}$ and $v_i(k)$ is actively split into $(+1,-1,+1,-1)^{\mathrm{T}}$. It may be observed that the signs of the components of these vectors vary as the bits in an integer as it is incremented from zero to $K_i - 1$.

Thus, the matrices $\mathbf{C}_{ij}$ and **C** need not be explicitly stored, computed or used in the calculation of the matrix products, which saves further time and memory. In fact, Eq. 1 can be rewritten as

$$\mathbf{S} = \sum_{i,j}^{S} \mathbf{P}_i \mathbf{C}_{ij} \mathbf{Q}_j^{\mathrm{T}} = \sum_{i,j}^{S} (\mathbf{P}_i \mathbf{v}_i(j))\, (\mathbf{Q}_j \mathbf{v}_j(i))^{\mathrm{T}} \tag{10}$$

Therefore, the two matrix–matrix multiplications can be reduced to the sum of two matrix–vector and one dyadic vector multiplication. Since $\mathbf{v}_i(j) = 0$ if the ith and jth spins are not coupled, the number of nonzero terms in this sum is only linear in the number of spins S, which leads to a further reduction in the time required to compute it. The amount of memory required can also be further reduced by taking advantage of magnetic equivalence (Havel et al., 1994).

To estimate the coupling constants and other parameters, we minimize the following target function:

*If the diagonal contains dispersive peaks, it will actually be necessary to use different **P** and **Q** matrices for the diagonal (Havel et al., 1994).

$$F(\alpha, \{\sigma\}, \{\bar{\nu}\}, \{J\}) = \frac{1}{2}||\alpha \mathbf{S}_c - \mathbf{S}_o||^2 = \frac{1}{2}||\alpha \mathbf{PCQ}^T - \mathbf{S}_o||^2 = \frac{1}{2}Tr\left((\alpha \mathbf{S}_c - \mathbf{S}_o)(\alpha \mathbf{S}_c - \mathbf{S}_o)^T\right) \qquad (11)$$

where $\mathbf{S}_c$ is the simulated spectrum, $\mathbf{S}_o$ is the experimental spectrum and α is a scaling parameter which can be treated as a variable during the minimization. In practice, it may be advisable to restrict the indices in this sum of squares, so as to avoid uninteresting regions of the spectrum such as the main diagonal and the water signal. The variables used in the minimization can be any combination of α, σ_1, $\bar{\nu}_1$, σ_2, $\bar{\nu}_2$, ..., σ_S, $\bar{\nu}_S$, and J_1, J_2, ..., J_I, where σ_i and $\bar{\nu}_i$ are the line width and chemical shift for spin i, and J_j are the coupling constants. Note that whenever a set of spins have the same chemical shifts, they are treated as a single magnetically equivalent group, for which only one 'average' coupling constant is used as a variable for each other group to which they are coupled. This is the behavior expected in the extreme strong-coupling limit, and by taking it into account we decrease both the number of variables and our dependence on the weak-coupling approximation.

Because it is reasonably efficient and requires memory that is only linear in the number of variables, the minimization algorithm used in our computer program is the conjugate gradient method, which necessitates the computation of the derivatives of F with respect to the variables. These derivatives are computed analytically as follows.

The derivative with respect to α is easily seen to be

$$\frac{dF}{d\alpha} = Tr[(\mathbf{PCQ}^T)\,\mathbf{D}^T] \qquad (12)$$

where $\mathbf{D} = \alpha \mathbf{PCQ}^T - \mathbf{S}_o$. To obtain the remaining derivatives, let x represent any one of the variables other than α. Then

$$\frac{dF}{dx} = \alpha \cdot Tr[(\mathbf{P}_x\mathbf{CQ}^T + \mathbf{PCQ}_x^T)\,\mathbf{D}^T] = \alpha \cdot Tr[(\mathbf{Q}^T\mathbf{D}^T\mathbf{P}_x + \mathbf{Q}_x^T\mathbf{D}^T\mathbf{P})\,\mathbf{C}] \qquad (13)$$

where $\mathbf{P}_x = d\mathbf{P}/dx$ and $\mathbf{Q}_x = d\mathbf{Q}/dx$. If the spectrum is symmetric, i.e., $\mathbf{Q} = \mathbf{P}$ and $\mathbf{S}_o^T = \mathbf{S}_o$, this simplifies to

$$\frac{dF}{dx} = 2\alpha \cdot Tr[\mathbf{P}_x\mathbf{CP}^T\mathbf{D}] = 2\alpha \cdot Tr[\mathbf{P}_x^T\mathbf{DPC}] \qquad (14)$$

To compute $\mathbf{P}_x$, we need to compute the derivative of the elements of each column with respect to x. From Eq. 4, we have

$$\frac{dL}{dx} = \begin{cases} \dfrac{dG}{dx} & \text{if } G(\mu) \geq C_1 \\ \left(H(G) + G\dfrac{dH}{dG}\right)\dfrac{dG}{dx} & \text{if } C_1 > |\,G(\mu)\,| \geq C_2 \\ 0 & \text{otherwise} \end{cases} \qquad (15)$$

where

$$\frac{dH}{dG} = 3A(G - C_2)^2 + 2B(G - C_2) \qquad (16)$$

and dG/dx is computed as follows. If $x = \sigma$, we have from Eq. 2

$$\frac{\mathrm{d}G}{\mathrm{d}\sigma} = \left(\frac{(\mu - \nu)^2}{\sigma^3} - \frac{1}{\sigma}\right)G \tag{17}$$

while for $x = \nu$

$$\frac{\mathrm{d}G}{\mathrm{d}\nu} = \frac{(\mu - \nu)}{\sigma^2}G \tag{18}$$

and for $x = J$

$$\frac{\mathrm{d}G}{\mathrm{d}J} = \frac{(\mu - \nu)}{\sigma^2}\frac{\mathrm{d}\nu}{\mathrm{d}J}G \tag{19}$$

where dν/dJ = ± 1/2, depending on the sign of Δ in Eq. 3.

Once again, the products **DQC**, $\mathbf{D}^{\mathrm{T}}\mathbf{PC}$ and **DPC** in Eqs. 13 and 14 can be computed very efficiently by taking into account the sparseness of the matrices, and using the simplifications described in Eqs. 9 and 10. The total amount of time required to evaluate the function F and its gradient, in fact, is $O(MNS)$. On the BPTI sample problems described below, wherein $M = N = 2048$ and 928 variables were used, the absolute amount of time required was about 18 s on a 50 MHz Silicon Graphics Indigo R4000.

RESULTS AND DISCUSSION

To systematically evaluate the proposed procedure, we designed two types of test problems which focus on the particularly important case of polypeptides. The first is a realistic set of conformations for the 20 amino acids, which was obtained by combining the standard side-chain rotomers described in Ponder and Richards (1987) with the standard α-helix and β-sheet ϕ angles. For each of these 222 amino acid conformations, the three-bond coupling constants were calculated from the Karplus relation (Karplus, 1959), using the published parameters (De Marco et al., 1978; Pardi et al., 1984); the two-bond couplings were set to −14 Hz (which is accurate enough for our purposes). The chemical shifts were obtained by taking the random-coil values (Wüthrich, 1986) and arbitrarily shifting them upfield so that they collectively spanned at most 2 ppm, but in such a way that their order was preserved and no cross-peak overlaps were either created or eliminated relative to the random-coil values. Thus, the 1024 points used for the ω_1 and ω_2 dimensions correspond to a very high digital resolution. Two sets of such test problems were created, one with a (full) line width of 6 Hz and another with a line width of 20 Hz.

The second type of problem was obtained by computing 25 BPTI conformations from the simulated NMR distance constraints denoted as **B-I** in Havel (1991), using the DG-II distance geometry program described there. The rms coordinate deviation among these structures averaged 1.76 Å (0.94 Å among the α-carbons alone). In this case the published chemical shifts (Wagner et al., 1987; Berndt et al., 1992) were used where available; the peaks corresponding to the few missing assignments were omitted from the spectra. Once again, the coupling constants were obtained from the Karplus relation, and a realistic line width of 7 Hz was employed. These spectra spanned a range of −0.5 to 11 ppm, with a digital resolution of 2048 in both the ω_1 and ω_2 dimensions. The simulated field strength used for all test problems was 500 MHz.

In both types of test problem, the simulated spectra were treated as 'experimental spectra', and an attempt was made to recover the parameters used to simulate them by the minimization procedure described above. Since we are only attempting to correct for small errors in their values, the initial chemical shifts used for these minimizations were obtained by randomly perturbing the target values by a small quantity, i.e.

$$\bar{\nu}_{init} = \bar{\nu} \pm \delta_\nu r \tag{20}$$

where $\bar{\nu}$ is the target chemical shift, δ_ν is a constant, and r is a uniform random number between [−1,1]. The initial line widths used, on the other hand, were

$$\sigma_{init}\sqrt{2log2} = w_{init} = w \pm \delta_w r \tag{21}$$

where w is the target line width at half height and δ_w is a constant. Finally, the starting values of the three-bond coupling constants were uniformly set to 7 ± 0.1 Hz, since this is their expected average value; two-bond couplings were started from their target value of −14 Hz. The two-bond couplings were also allowed to vary in the course of the computation because it may be necessary to do this with actual experimental data, first in order to obtain the best possible fits to the data, and second to serve as internal controls on the reliability of the results.

Because it could have a drastic effect upon the accuracy and precision of the resulting optimized parameter values, these evaluations were performed both with and without added noise and data processing artifacts. In order to simulate the noise and artifacts as realistically as possible, each target spectrum was transformed to a set of hypercomplex FIDs. These FIDs were truncated and their initial points filtered by doubling the value of the first point in t_1 and altering the first P points along t_2 by multiplication with

$$p + (1 - p)B \tag{22}$$

where B is a constant, $p = i/P$ and i is the point index (starting from 0; the effects of such artifacts on experimental spectra have been discussed extensively by Otting et al. (1986)). Gaussian noise was then added to the real and imaginary parts of the FIDs with a standard deviation in intensity equal to a given percentage of their maximum values, a cos^2 window function applied, the resulting time-domain spectrum zero-filled to the original size, and transformed back to the frequency domain. Finally, the resulting spectrum was symmetrized by adding it to its matrix transpose (which could also be done with experimental spectra, provided that the water signal is not included in the sum of squares being minimized). As may be seen in Figs. 1 and 2, this procedure resulted in simulated spectra that are similar in appearance to actual experimental protein spectra*.

Single amino acid test problems

For this class of test problems we set w to either 6 or 20 Hz and δ_w to 1.0 or 4.0 Hz. The initial errors in the chemical shifts δ_ν were 0.01 ppm, and the coupling constants were set to 7 ± 0.1 Hz, as previously described. For each set of 'target' coupling constants computed from the amino acid conformations used (as described above), eight sets of target spectra were generated. The first two

*The '+' shape of the cross peaks is due to the unequal number of FIDs retained in t_1 and t_2, together with the fact that the spectra have been symmetrized.

were the 'ideal' spectra directly from the simulation program described above, using 6 and 20 Hz line widths. In the second pair of target spectra, the corresponding FIDs were truncated to 32 points in t_1 and 128 points in t_2, and 2% Gaussian random noise was added, as described above ('high resolution/low noise'). In the third, the FIDs were truncated to 32 points in t_1 and 128 points in t_2, and 20% noise added ('high resolution/high noise'). Finally, in the fourth, the FIDs were truncated to 16 points in t_1 and 64 points in t_2, and 2.0% noise was added ('low resolution/low noise'). In the latter three pairs of sets, the linear filter was used with $B = 0.5$ and $P = 3$. Except for the fact that our simulated spectra contained about 500 points/ppm, the quality of these spectra covers the range encountered in biological applications. Figures 1a, c, e and g show typical examples of these target spectra for the parameters obtained from the conformer of glutamine denoted by B6, while Figs. 1b, d, f and h show the corresponding spectra simulated from the parameter values obtained by minimizing the difference between the calculated and target spectra.

These sets of test problems are summarized in Table 1 along with the abbreviations used, while Table 2 lists the results obtained with each set. As can readily be seen, when no noise or artifacts were present in the spectra (SAA(6, fr, nn) and SAA(20, fr, nn) in the tables), the correct chemical shifts, line widths and coupling constants could be accurately and reliably recovered from the simulated spectra. In fact, the coupling constants were correct to within 0.1 Hz in all but one case, namely the arginine conformation R2 at a line width of 6 Hz. This one problem was due to the minimizer being trapped in a local minimum, since it was always run until the gradient norm became small (ca. 1×10^{-6} of its initial value). Interestingly, in this case it turned out that the α to β_2 coupling constant was 9.5 Hz greater than the target value of 2.1 Hz, while the β_2 line width was 6.6 Hz greater than its 6.0 Hz target value. Thus, the program compensated for the lack of cancellation of the peaks in the multiplet due to the large coupling constant by making the line width too large. Although this problem did not occur very often, we would nevertheless recommend starting these minimizations from an underestimate of the line widths. In all other cases a very small value of the target function was attained, and in a rather small number of minimization iterations.

Although the final values of the target function increased substantially, the quality of the results was not significantly degraded by the presence of moderate spectral noise (2%), limited signal resolution (32 × 128) and spectral artifacts (SAA(6, hr, ln) and SAA(20, hr, ln)). A more complete description of the results obtained in these test problems, grouped by amino acid type, is given in Table 3*. These spectra are certainly of high quality, but are within the range of what can be obtained in favorable cases with biological macromolecules. The line widths obtained were systematically too high by amounts ranging from 1 to 10 Hz or more. In a few cases (e.g. TRP-R2 at a 20 Hz line width) this was due to local minima like that encountered for Arg-R2 above, but in most cases it was due to the fact that truncation, combined with the cos^2 window function used, systematically broadened all the peaks in the target spectra. This shows, incidently, that these minimizations have a very large convergence radius with respect to the line widths, so that the large variations in line widths found in experimental spectra should seldom be a problem in practice**. As expected, some difficulties (e.g. with phenylalanine and tryptophan) were encoun-

*Complete tables for each amino acid (or BPTI conformation) in each type of test problem are available as supplementary material to this paper.

**Because of the systematic broadening due to field inhomogeneity and truncation of the FIDs, these 'effective' line widths should not ordinarily be used to estimate T_2.

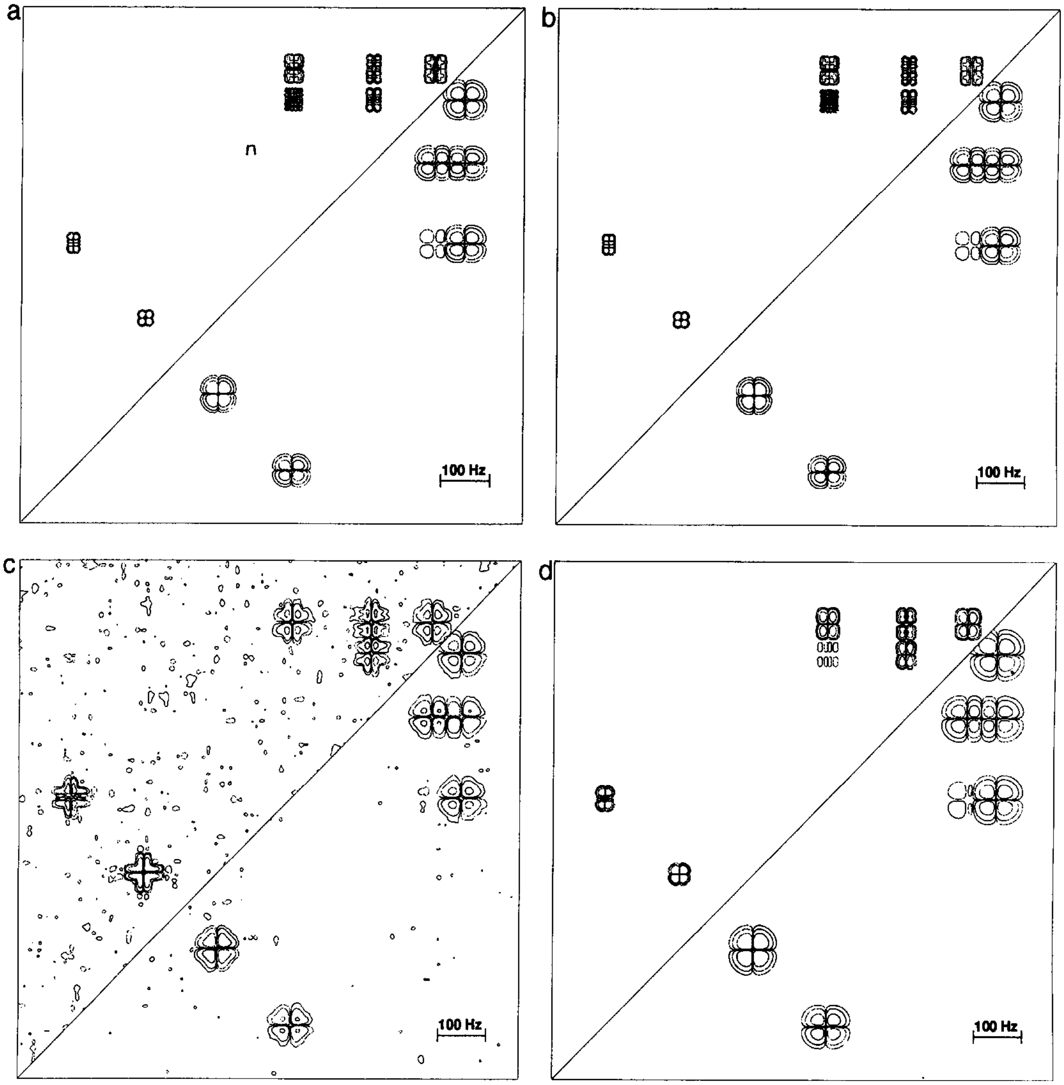
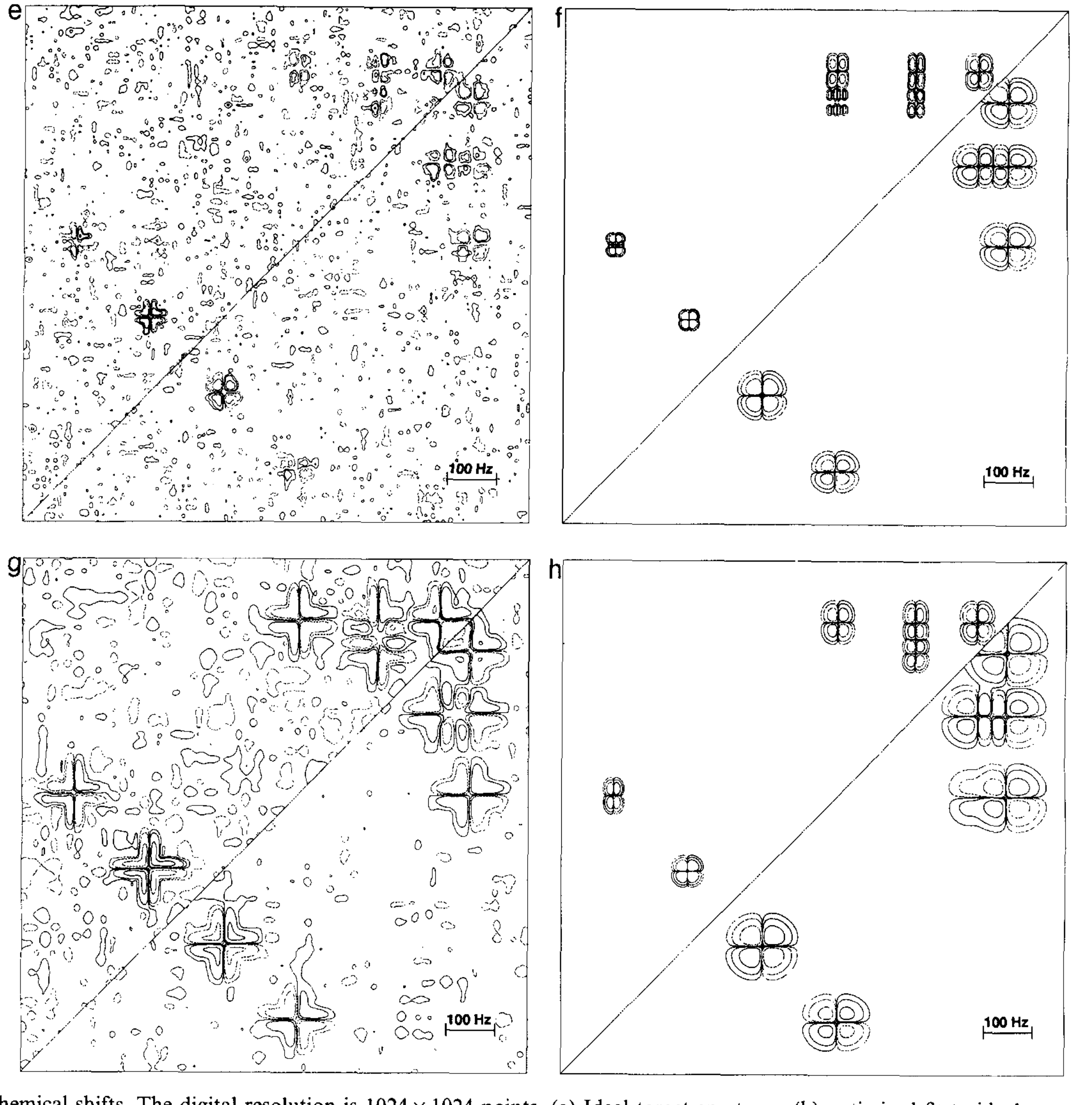


Fig. 1. Contour plots of the target spectra, simulated using the coupling constants computed via the Karplus relation from the glutamine conformation denoted by B6, together with the optimized fits to them. The spectra obtained using 6 Hz line widths are shown above the diagonal, while those obtained using 20 Hz are shown below. The chemical shifts have been chosen arbitrarily for easy viewing, but without creating or eliminating any cross-peak overlap relative to the random-coil

tered with a 20 Hz line width. Nevertheless, it appears that moderate noise, imperfect resolution and typical spectral artifacts do not make the local-minimum problem much worse. More importantly, these results show that the least-squares method is capable of correcting for such artifacts, and that the ill-conditioning problem is not serious in these sorts of fits.

On further increasing the noise to the much higher level of 20%, the quality of the results obtained declined substantially (SAA(6, hr, hn) and SAA(20, hr, hn)). A similar effect was also obtained on merely halving the signal resolution to 16 x 64 (SAA(6, lr, ln) and SAA(20, lr, ln)). In order to determine the nature of the difficulties, these four runs were repeated starting from the



chemical shifts. The digital resolution is 1024 × 1024 points. (a) Ideal target spectrum; (b) optimized fit to ideal target spectrum; (c) high resolution/low noise target spectrum; (d) optimized fit to high resolution/low noise target spectrum; (e) high resolution/high noise target spectrum; (f) optimized fit to high resolution/high noise target spectrum; (g) low resolution/low noise target spectrum; (h) optimized fit to low resolution/low noise target spectrum.

target values of the coupling constants, chemical shifts and line widths (SAA(6, hr, hn, t), SAA(6, lr, ln, t), SAA(20, hr, hn, t) and SAA(20, lr, ln, t)), with results very similar to those obtained starting from random values. Thus, the difficulties appear to be primarily due to ill-conditioning, rather than local minima. Despite the difficulties, it should be stressed that these are, as a whole, *extremely* positive results. The number of FIDs/ppm used in the high-resolution spectra can easily be obtained in practice, while spectra containing 20% noise are of far lower quality than those routinely used for assignment purposes (as can be seen in Fig. 1e, it is difficult to even see many

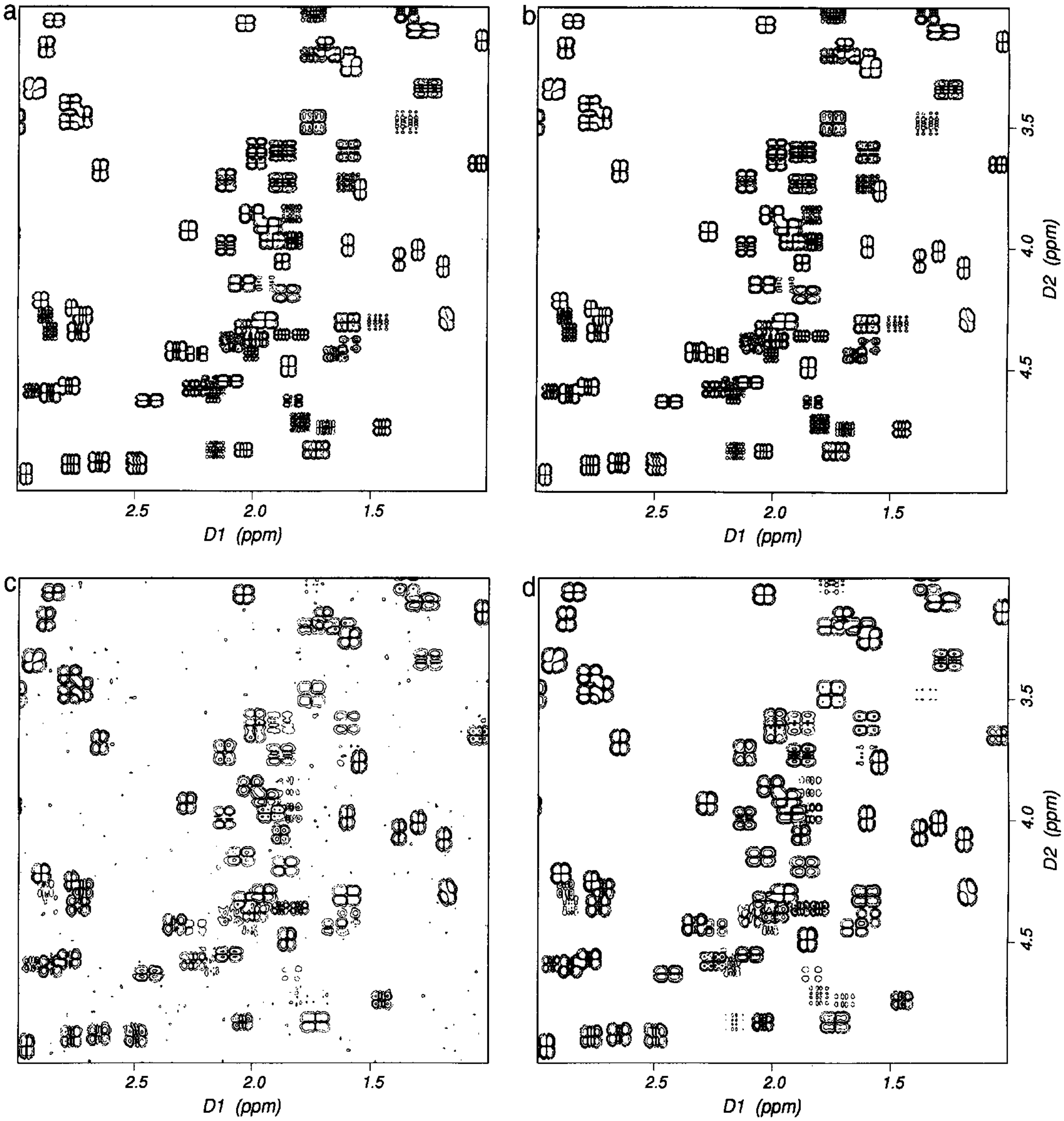
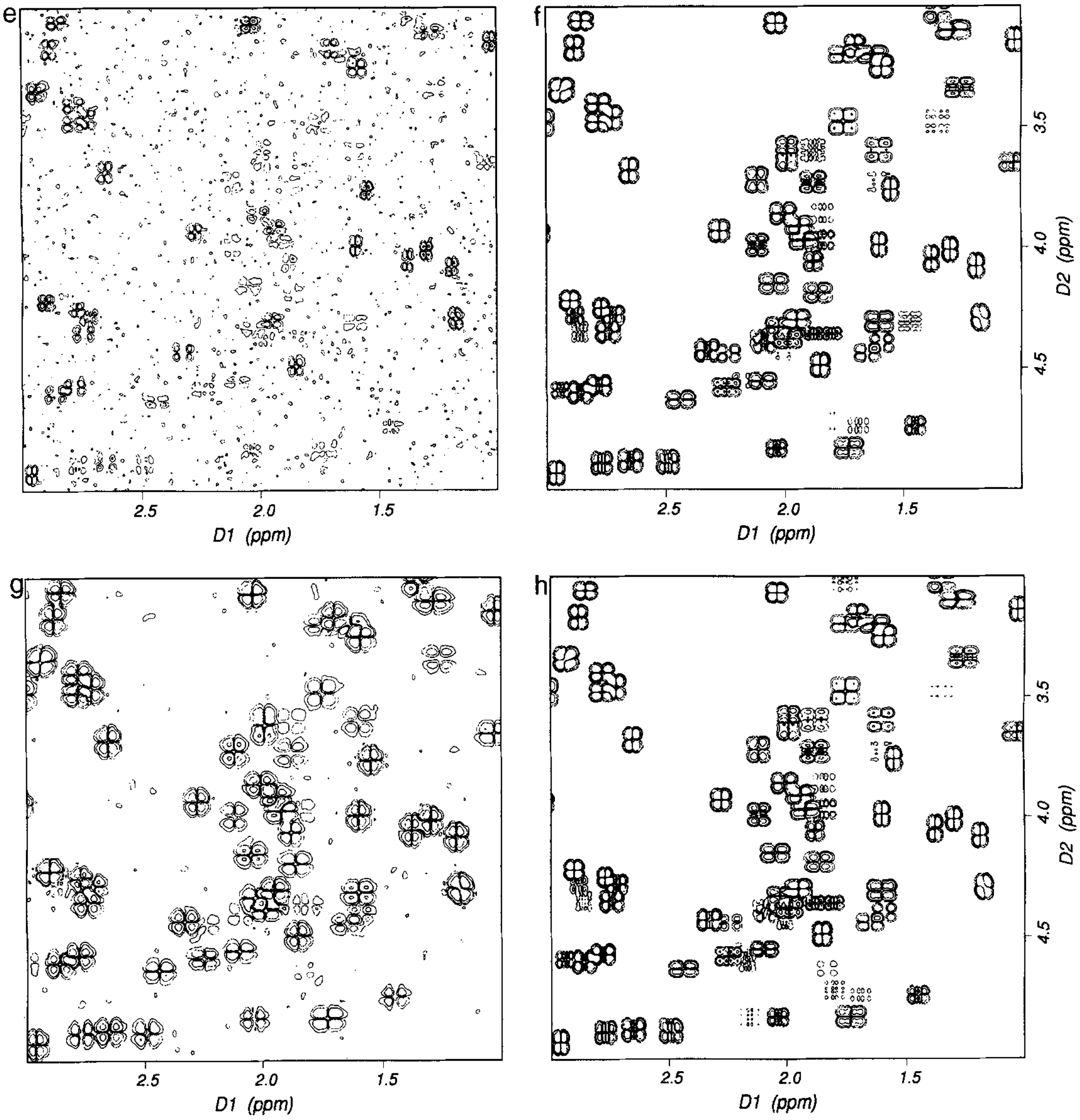


Fig. 2. Contour plots of the 2048 × 2048 points target spectra, simulated using the published chemical shifts and a 7.0 Hz line width, together with the coupling constants computed via the Karplus relation from the first of the 25 BPTI conformations. Only the portion of the aliphatic/aliphatic region from 1 to 3 ppm along ω_1 and 3 to 5 ppm along ω_2 is shown. (a) Ideal target spectrum; (b) optimized fit to ideal target spectrum; (c) high resolution/low noise target spectrum;

of the cross peaks when 20% noise is present!). Nevertheless, even in these highly unfavorable cases, the majority of the coupling constants could still be recovered to within 2.0 Hz.

It may be observed that, although the quality of the results obtained in both the low resolution/low noise and high resolution/high noise runs was similar, the values of the least-squares residual were far higher in the latter case. This is to be expected, since it is impossible to accurately fit random noise with a limited number of Gaussians. It should also be mentioned that the quality



(d) optimized fit to high resolution/low noise target spectrum; (e) high resolution/high noise target spectrum; (f) optimized fit to high resolution/high noise target spectrum; (g) low resolution/low noise target spectrum; (h) optimized fit to low resolution/low noise target spectrum.

of the results obtained in the low-resolution case was significantly worse if the target spectra were not symmetrized. This raises the question concerning how seriously the accuracy of the method will be affected by our use of a simple Gaussian line shape in practice. To find out, we made an additional pair of runs, at 6 and 20 Hz line widths, respectively, but using target spectra generated with a 50:50 mix of Gaussian and Lorentzian line shapes. These were 'ideal' target spectra, containing no noise or data processing artifacts, so that the errors incurred by using an imperfect line shape model alone could be assessed (SAA(6, fr, nn, 50) and SAA(20, fr, nn, 50)). As can be

TABLE 1
DESCRIPTION OF SIMULATED TARGET SPECTRA[a]

Abbreviation	Molecular structure	Full line width (Hz)	Percent Gaussian	Signal resolution	Thermal noise	Digital resolution (Hz/point)	Starting parameters
SAA(6, fr, nn)	222 Single amino acid conformations, containing a total of 1344 coupling constants	6.0	100	Full	None	0.12 through 0.57	Random
SAA(6, hr, ln)		6.0	100	High	Low		Random
SAA(6, hr, hn)		6.0	100	High	High		Random
SAA(6, lr, ln)		6.0	100	Low	Low		Random
SAA(6, hr, hn, t)		6.0	100	High	High		Target
SAA(6, lr, ln, t)		6.0	100	Low	Low		Target
SAA(6, fr, nn, 50)		6.0	50	Full	None		Random
SAA(20, fr, nn)	222 Single amino acid conformations, containing a total of 1344 coupling constants	20.0	100	Full	None	0.36 through 1.30	Random
SAA(20, hr, ln)		20.0	100	High	Low		Random
SAA(20, hr, hn)		20.0	100	High	High		Random
SAA(20, lr, ln)		20.0	100	Low	Low		Random
SAA(20, hr, hn, t)		20.0	100	High	High		Target
SAA(20, lr, ln, t)		20.0	100	Low	Low		Target
SAA(20, fr, nn, 50)		20.0	50	Full	None		Random
BPTI(7, fr, nn)	25 Near-native BPTI conformations, each containing 328 coupling constants	7.0	100	Full	None	2.81	Random
BPTI(7, hr, ln)		7.0	100	High	Low	2.81	Random
BPTI(7, hr, hn)		7.0	100	High	High	2.81	Random
BPTI(7, lr, ln)		7.0	100	Low	Low	2.81	Random

[a] The first column shows the abbreviation used in the text and in other tables for each type of spin system evaluated. The next column shows the kind of molecular structure from which the coupling constants were calculated via the Karplus relation. This is followed by the (full) line width that was used and the percentage Gaussian (versus Lorentzian) character. 'Full' signal resolution means that the simulated ideal spectra were used as targets, 'high' resolution means that only 32 x 128 points (in the SSA problems) or 512 x 1024 points (in BPTI) were retained, while 'low' resolution means that 16 x 64 points (in SSA) or 256 x 512 points (in BPTI) were retained (see text for further details). A 'low' noise level means that Gaussian noise with a standard deviation equal to 2% of the maximum of any FID was added to all the FIDs, while 'high' noise means that 20% of the maximum was used. The digital resolution varies in the SAA problems, because the chemical shifts were adjusted as described in the text. 'Random' starting parameters means that the minimizations were started from randomly perturbed values (see text), while 'target' means that the actual target parameters were used.

seen in Table 2, the quality of the results obtained in the 6 Hz case was still reasonably good, with 95% of the coupling constants being recovered to within 1.0 Hz. At a 20 Hz line width, however, the results were even worse than with the low-resolution or high-noise spectra.

In a final set of runs (SAA(*,*,*, vsf) in Table 4), we addressed the question of whether the scale factor α could be determined automatically and reliably by treating it as a variable during the minimization. Surprisingly, it was found that the addition of this one further degree of freedom made these fits substantially more ill-conditioned, especially at a 20 Hz line width. It therefore appears that the scale factor will have to be determined independently in actual applications to experimental data (Yang et al., 1994).

Multiple overlapping spin systems in BPTI

The purpose of these test problems was to see how peak overlap affects the fits, as well as how many more variables affect the computation time and local-minimum problem. BPTI contains a total of 328 independent coupling constants (63 of which are two-bond couplings) among 300 (magnetically equivalent groups of) spins. To obtain the starting parameters, we set w = 7 Hz, δ_w = 2 Hz, and δ_v = 0.01 ppm. Four sets of test problems were generated for each set of 'target' coupling constants computed from the 25 BPTI conformations. The first used the 2048 × 2048 points spectra simulated from the target parameters as their target spectra (ideal), the second used spectra with 2% added noise, 512 FIDs in t_1 and 1024 in t_2 as their targets (high resolution/low noise), the third used spectra with 20% added noise and 512 × 1024 points FIDs (high resolution/ high noise) while the fourth used 2% noise and 256 × 512 points FIDs (low resolution/low noise). As above, the latter three sets employed a linear filter with B = 0.5 and P = 10.

In Figs. 2a, c, e and g, we show a portion of the aliphatic/aliphatic region of the target spectra for one of the BPTI conformations, while in Figs. 2b, d, f and h we show the same region of the corresponding final minimized spectra. In the fits to the ideal spectra, the total overlap of 28% of the cross peaks was more than 10%, while 15% of the cross peaks had a total overlap with other

TABLE 2
SUMMARY OF RESULTS OBTAINED BY FITTING SIMULATED TARGET SPECTRA[a]

Abbreviation	Total/maximum frequency of coupling constant violations (%)			Average rms deviation		Average of final		Mean no. of itera-tions
	>0.5 Hz	>1.0 Hz	>2.0 Hz	C.S. (ppb)	L.W. (Hz)	Function	Gradient	
SAA(6, fr, nn)	0.2/2.7 (R)	0.1/0.9 (R)	0.1/0.9 (R)	1.22	0.21	1.0	0.7	36
SAA(6, hr, ln)	0.7/2.7 (W)	0.0/0.0	0.0/0.0	1.31	3.71	89.4	0.3	44
SAA(6, hr, hn)	22.4/37.5 (W)	9.2/16.1 (R)	3.0/6.2 (R)	1.40	3.85	3650.6	0.0	55
SAA(6, lr, ln)	45.1/96.4 (W)	18.2/75.0 (W)	2.4/17.0 (W)	1.20	10.52	463.2	0.3	56
SAA(6, hr, hn, t)	23.6/35.7 (W)	8.8/15.6 (D)	3.9/12.5 (D)	0.48	3.79	3690.8	0.0	49
SAA(6, lr, ln, t)	46.1/96.4 (W)	17.6/84.8 (W)	2.3/17.9 (W)	0.28	10.51	463.7	2.0	41
SAA(6, fr, nn, 50)	26.0/100.0 (G)	5.3/22.0 (Y)	0.9/4.9 (E)	1.26	2.14	43.7	0.0	78
SAA(20, fr, nn)	0.0/0.0	0.0/0.0	0.0/0.0	1.28	0.26	0.0	0.2	43
SAA(20, hr, ln)	3.3/13.4 (W)	0.3/2.7 (W)	0.0/0.0	1.34	7.53	70.4	0.2	51
SAA(20, hr, hn)	45.1/62.5 (V)	22.6/40.6 (V)	9.2/25.0 (V)	2.39	8.27	3648.2	0.1	56
SAA(20, lr, ln)	57.0/82.0 (Y)	29.0/67.0 (W)	5.4/33.3 (F)	3.59	22.46	844.1	3.3	75
SAA(20, hr, hn, t)	46.7/70.0 (F)	22.8/43.8 (P)	8.2/18.8 (P)	2.09	8.04	3604.6	0.1	53
SAA(20, lr, ln, t)	57.4/77.7 (W)	27.9/60.7 (W)	4.4/33.3 (F)	3.42	22.56	847.2	2.9	67
SAA(20, fr, nn, 50)	94.8/100.0 (A)	81.4/100.0 (A)	35.6/66.0 (Y)	1.18	2.66	6.1	0.0	71
BPTI(7, fr, nn)	0.1/0.6 (R)	0.1/0.3 (R)	0.0/0.3 (R)	0.51	0.19	0.1	0.0	483
BPTI(7, hr, ln)	0.8/2.8 (I)	0.2/2.2 (R)	0.0/0.5 (I)	0.46	3.59	8.9	0.0	534
BPTI(7, hr, hn)	19.2/32.2 (P)	8.3/15.2 (R)	2.9/5.6 (P)	0.72	3.65	873.6	0.0	753
BPTI(7, lr, ln)	5.7/25.0 (V)	1.7/6.8 (I)	0.3/5.0 (I)	0.70	10.08	30.5	0.0	374

[a] The abbreviations are described in Table 1. The total frequency of the coupling constant violations greater than 0.5, 1.0 and 2.0 Hz are given in percent to the left of each '/' in the next three columns, while the maximum frequency of violations in any one of the 20 types of amino acids is given to the right of each '/', together with the one-letter code of the amino acid for which that maximum was attained in parentheses. The average rms violations of the target chemical shifts and line widths are given in the next two columns, followed by the average over all minimizations of the final function value (residual), its rms gradient norm, and the number of iterations required.

TABLE 3
RESULTS FOR AMINO ACID SPIN SYSTEMS (SAA(6, hr, ln)/SAA(20, hr, ln))[a]

Amino acid	Count		Frequency of coupling constant violations (%)			Average rms deviation		Average of final		Mean no. of iterations
	S.S.	C.C.	>0.5 Hz	>1.0 Hz	>2.0 Hz	C.S. (ppb)	L.W. (Hz)	Function	Gradient	
Ala	2	2	0.0/0.0	0.0/0.0	0.0/0.0	0.1/0.1	1.0/1.4	253.1/178.5	0.2/0.1	27/25
Arg	16	7	1.8/2.7	0.0/0.0	0.0/0.0	2.3/1.9	3.8/6.7	64.9/42.6	0.1/0.1	58/53
Asn	14	5	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.1	2.8/6.8	42.4/47.0	0.1/0.1	38/45
Asp	8	4	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.1	1.5/3.2	127.1/56.9	0.2/0.1	26/35
Cys	8	4	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.2	1.7/4.2	122.5/66.1	0.2/0.1	23/35
Gln	28	7	0.0/1.5	0.0/0.0	0.0/0.0	0.1/0.1	3.7/8.5	37.1/65.3	0.1/0.2	56/43
Glu	16	9	0.0/6.2	0.0/0.7	0.0/0.0	0.0/0.1	2.6/4.8	160.0/72.9	0.2/0.1	52/59
Gly	2	1	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.1	0.5/1.1	152.5/95.1	0.4/0.5	13/13
His	14	4	0.0/0.0	0.0/0.0	0.0/0.0	3.7/3.1	2.1/5.3	50.4/10.7	0.2/0.0	26/35
Ile	12	8	2.1/3.1	0.0/0.0	0.0/0.0	0.0/0.1	4.1/8.4	87.0/210.6	0.5/0.3	87/51
Leu	10	5	0.0/2.0	0.0/0.0	0.0/0.0	0.1/0.6	2.3/4.8	100.6/44.7	0.5/0.1	30/73
Lys	16	9	1.4/0.0	0.0/0.0	0.0/0.0	0.1/0.1	3.8/8.7	105.6/59.4	0.2/0.2	58/67
Met	16	6	0.0/1.0	0.0/0.0	0.0/0.0	1.4/1.9	2.8/4.7	104.8/82.0	0.3/0.1	37/40
Phe	10	6	0.0/13.3	0.0/0.0	0.0/0.0	0.2/1.0	3.7/6.5	66.9/14.2	0.2/0.0	40/96
Pro	2	8	0.0/6.2	0.0/0.0	0.0/0.0	0.1/0.2	2.4/4.6	45.6/54.2	0.0/0.1	37/68
Ser	8	2	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.1	1.0/2.4	107.3/72.1	0.6/0.6	21/22
Thr	8	3	0.0/0.0	0.0/0.0	0.0/0.0	1.3/2.7	1.6/3.1	86.4/43.0	0.4/0.1	35/29
Trp	14	8	2.7/13.4	0.0/2.7	0.0/0.0	0.0/0.2	6.7/12.9	171.7/167.8	0.7/0.3	42/61
Tyr	10	5	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.1	4.6/8.3	38.9/32.1	0.1/0.1	44/56
Val	8	4	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.3	2.7/3.5	73.7/66.4	0.4/0.1	35/72

[a] In this table, numbers to the left of each slash were obtained by fitting the 6 Hz line width target spectra, while numbers to the right of each slash were obtained by fitting the 20 Hz line width target spectra. The rows are the cumulative results for all the spin systems evaluated for each of the 20 amino acids. The column labeled 'count of S.S.' is the total number of spin systems (or equivalently, conformations) used for that amino acid, while 'count of C.C.' contains the total number of independent two- and three-bond coupling constants in each amino acid. This is followed by three columns counting the percentage of coupling constants in all of the spin systems for each amino acid which differred from their target values by more than 0.5, more than 1.0 and more than 2.0 Hz. The next two columns report the average over all spin systems of the rms violations of the chemical shifts (C.S.) in ppb and line widths (L.W.) in Hz. The two subsequent columns report the average final values of the least-squares residual (Function) and its gradient (Gradient), in arbitrary units. The last column gives the average number of conjugate gradient iterations that were required to achieve these values.

cross peaks exceeding 50%*. These numbers were about the same in the fits to the high resolution/low noise and high resolution/high noise spectra, but in the fits to the low resolution/low noise spectra 39% of the cross peaks were overlapped by 10% or more, and 23% had a total overlap exceeding 50%. Considerably more overlap, of course, would be expected if the diagonal peaks are included in the simulation, but for this same reason the diagonal region probably should not be included in the sum of squares being minimized, even in actual applications to experimental data.

When the ideal spectra were used as the targets, the results were very similar to those obtained

*The overlap of the *ij*th cross peak is defined as the integral (sum over the discretized spectrum) of the product of the positive part of that cross peak with the positive part of every other cross peak $\alpha^{++}_{ij,kl}$, and similarly for $\alpha^{+-}_{ij,kl}$, $\alpha^{-+}_{ij,kl}$ and $\alpha^{--}_{ij,kl}$. The total overlap is then $\Sigma_{kl}(\alpha^{++}_{ij,kl} + \alpha^{--}_{ij,kl} - \alpha^{+-}_{ij,kl} - \alpha^{-+}_{ij,kl})$, which is normalized by dividing it by $\alpha^{++}_{ij,ij} + \alpha^{--}_{ij,ij} - \alpha^{+-}_{ij,ij} - \alpha^{-+}_{ij,ij}$.

TABLE 4
SUMMARY OF RESULTS OBTAINED BY FITTING SIMULATED TARGET SPECTRA WITH A VARIABLE SCALING FACTOR[a]

Abbreviation	Total/maximum frequency of coupling constant violations (%)			Average rms deviation		Average of final		Mean no. of itera-tions
	>0.5 Hz	>1.0 Hz	>2.0 Hz	C.S. (ppb)	L.W. (Hz)	Func-tion	Gradient	
SAA(6, fr, nn, vsf)	0.7/4.2 (I)	0.4/4.2 (I)	0.4/4.2 (I)	1.58	0.75	7.1	0.0	89
SAA(6, hr, ln, vsf)	3.2/17.9 (W)	0.4/2.0 (Q)	0.1/0.5 (Q)	1.28	3.43	110.7	0.1	86
SAA(6, hr, hn, vsf)	24.3/35.7 (W)	9.2/25.0 (C)	4.5/25.0 (C)	1.27	3.53	3713.2	0.1	96
SAA(6, lr, ln, vsf)	51.0/98.2 (W)	26.6/85.7 (W)	10.0/60.0 (Y)	1.32	10.42	446.6	4.9	91
SAA(20, fr, nn, vsf)	0.0/0.0	0.0/0.0	0.0/0.0	1.18	0.23	0.0	0.0	76
SAA(20, hr, ln, vsf)	51.6/100.0 (G)	34.9/100.0 (G)	16.9/83.0 (W)	1.37	8.25	69.8	1.6	123
SAA(20, hr, hn, vsf)	77.8/100.0 (G)	56.7/100.0 (G)	32.9/100.0 (G)	2.69	9.01	3657.3	6.3	133
SAA(20, lr, ln, vsf)	87.5/100.0 (D)	74.0/100.0 (G)	51.6/100.0 (G)	22.19	62.48	2604.9	87.5	149

[a] See footnote to Table 2.

with the single spin systems (BPTI(7, fr, nn) in the tables). In only two out of the 25 cases were any local minima encountered, both of which were of the same 'coupling constant versus line width' variety. The results with the high resolution/low noise (BPTI(7, hr, ln)) spectra were nearly as good as what was obtained in the ideal case, whereas in the high resolution/high noise (BPTI(7, hr, hn)) and low resolution/low noise (BPTI(7, lr, ln)) spectra, serious problems were again encountered. In the low resolution/low noise case, an additional run starting from the values of the parameters used to simulate the target spectra again gave very similar results (data not shown in tables), implying once again that the problems were due to ill-conditioning rather than local minima. In Fig. 3, we plot the predicted versus the correct values of all three-bond couplings in all 25 runs for BPTI(7, hr, ln), BPTI(7, hr, hn) and BPTI(7, lr, ln). These plots show, in particular, that the precision of the coupling constants estimated by our procedure is lower for small couplings.

We conclude that moderate amounts of peak overlap do not greatly affect the quality of the fits obtained, and in particular do not make the ill-conditioning problem significantly worse (at least as long as the correct line shape is used). It also appears that increasing the number of variables by attempting to fit all the spin systems in a small protein together does not, by itself, worsen the local minimum problem (although it does, of course, substantially increase the computation time required). Thus, the results obtained above with single spin systems should be readily applicable at least to small proteins. Table 5 shows a breakdown by type of the errors in the coupling constants that were obtained on fitting the high resolution/high noise and low resolution/low noise spectra (BPTI(7, hr, hn) and BPTI(7, lr, ln)). The relatively high frequency of problems encountered with the nondegenerate protons was caused by near-degeneracies in the chemical shifts of many methylene protons (in fitting actual spectra, further problems would be created in such cases by overlap with the diagonal peaks and strong coupling effects). Otherwise, the errors do not appear to vary systematically with the type of coupling constant.

In order to determine how accurate the chemical shifts used as input to the program have to be, we also performed a number of BPTI runs using the ideal spectra as the targets and starting from

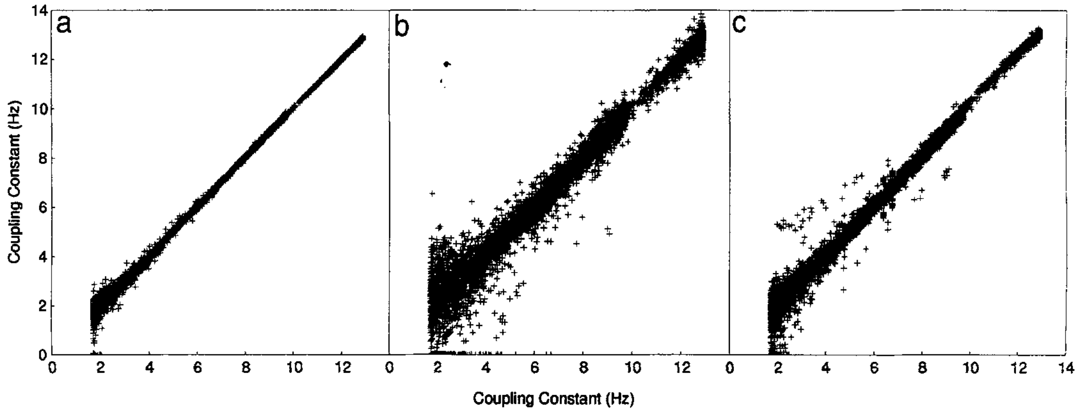


Fig. 3. Plots of the predicted (vertical) versus the target (horizontal) values of the three-bond couplings in the test problems (a) BPTI(7, hr, ln); (b) BPTI(7, hr, hn); and (c) BPTI(7, lr, ln) (see Table 1).

chemical shifts that were obtained by randomly perturbing the target values by $\delta_v = 0.05$ ppm (data not shown in tables). This gives initial chemical shift errors of up to 25 Hz, or roughly seven times the half line width of 3.5 Hz. Because these optimizations converged relatively slowly, requiring 2000 iterations to attain rms gradient norms of order 0.1, these runs were performed only on the test problems derived from the first five BPTI conformations. In these five problems, between 38 and 49% of the chemical shifts could be recovered to within 0.01 ppm, while only

TABLE 5

TOTAL COUPLING CONSTANT VIOLATIONS BY TYPE FOR BPTI TEST PROBLEMS (BPTI(7, hr, hn)/BPTI(7, lr, ln))[a]

Coupling type	Total C.C.	Percentage of couplings violated by		
		>0.5 Hz	>1.0 Hz	>2.0 Hz
NH-C$^\alpha$H	1175	8.4/0.4	0.1/0.2	0.0/0.0
NH-C$^\alpha$H*	225	4.4/0.0	0.9/0.0	0.0/0.0
C$^\alpha$H-C$^\beta$H	1900	29.2/5.4	13.8/0.4	4.1/0.0
C$^\alpha$H-C$^\beta$H*	275	0.7/1.5	0.0/1.5	0.0/0.0
C$^\beta$H-C$^\gamma$H	1150	43.3/21.7	26.0/9.2	9.5/2.3
C$^\beta$H*-C$^\gamma$H	75	1.3/0.0	0.0/0.0	0.0/0.0
C$^\beta$H-C$^\gamma$H*	425	5.2/11.8	0.2/0.0	0.0/0.0
C$^\beta$H*-C$^\gamma$H*	25	0.0/0.0	0.0/0.0	0.0/0.0
H-H	4975	27.6/8.1	13.5/2.7	4.7/0.5
H*-H	1400	3.8/3.9	0.1/0.3	0.0/0.0
H*-H*	75	0.0/0.0	0.0/0.0	0.0/0.0
HR-HR	175	0.0/0.0	0.0/0.0	0.0/0.0
H2-H2	1575	9.2/0.8	0.4/0.1	0.0/0.0

[a] In this table, '*' represents the degenerate protons. H-H represents the three-bond couplings among all nondegenerate protons except those in aromatic rings, H*-H those between the degenerate protons and the nondegenerate protons, H*-H* those among the degenerate protons, HR-HR those among the protons in the aromatic rings, and H2-H2 represents the two-bond couplings. The numbers to the left of each slash are the frequency of the violations of each class of coupling constant in the BPTI(7, hr, hn) problems, whereas those to the right come from the BPTI(7, lr, ln) problems.

28–45% of the coupling constants were recovered to within 1.0 Hz. The line widths also ran 9 to 27 Hz rms too wide, indicating that many of the calculated cross peaks were 'trying' to overlap with the target cross peaks by broadening themselves. Interestingly, by using a starting line width of 50 instead of 7 Hz, these figures could be improved to 65–90% of the chemical shifts recovered to within 0.01 ppm, and 67–92% of the coupling constants recovered to within 1.0 Hz. The reason is probably that when large initial line widths are used, sufficient overlap is present between the starting and target spectra to pull many of the cross peaks in the right direction.

CONCLUSIONS

We have shown that a simple procedure for fitting homonuclear 2D COSY spectra by spectra simulated in the frequency domain holds considerable promise as a means of estimating proton–proton coupling constants in at least small proteins (as well as, by inference, many other types of molecules), while simultaneously improving on the precision of the given chemical shifts and effective line widths. Provided that the errors in the assigned chemical shifts are not much larger than the line widths, the local-minimum problem is surprisingly mild and the coupling constants can be accurately and reliably recovered with essentially no a priori assumptions on their values. Similarly, provided that the signal resolution is at least ca. 100 FIDs per ppm and the digital resolution is greater than ca. 200 points per ppm in both dimensions, the ill-conditioning problem is not serious and the quality of the results is remarkably insensitive to random noise in the FIDs. The procedure described in this paper could certainly be extended to heteronuclear experiments, and perhaps even higher dimensional spectra, although the computer time required would be substantially greater.

In addition to using a reasonably high digital and signal resolution, in applying this method to experimental data it appears advisable to process the data in such a way that the line shapes are easy to approximate by a simple funtion. We would recommend a Gaussian, not only because it is simple but also because it should alleviate peak overlap problems (Bodenhausen et al., 1977; Pearson, 1987). Because we have used a symmetric Gaussian to model the peaks, approximately equal numbers of points in t_1 and t_2 would also be advisable, although using separate line widths in ω_1 and ω_2 should pose no serious problems. In addition, we would recommend eliminating the diagonal region and water signal of the spectrum from the sum of squares being minimized, primarily to avoid having the fit dominated by the large but heavily overlapped peaks that are present there. This will also improve the computational efficiency of the procedure, as will eliminating any parts of the spectrum that are lacking cross peaks. Finally, we would recommend starting the refinement from an underestimate of the line widths, save possibly for peaks (e.g. amide protons) that may have significant chemical shift errors. Further details concerning how the data should be processed for best results with the procedure, and results with actual experimental data, may be found in our companion paper (Yang et al., 1994).

In order to obtain highly accurate coupling constants in all cases, e.g. methylene protons, it will probably be necessary to take account of strong coupling. This is best done by diagonalizing the scalar-coupling Hamiltonian directly (Widmer and Wüthrich, 1986), particularly since the analytic gradients of the transitions and intensities are now available (Maalouf, 1993). We expect, however, that the peak matrix obtained by the relatively simple fitting procedure described in this paper will tend to be more accurate than the actual coupling constants, because the sum of squares being minimized depends on the peak contours directly, whereas the coupling constants also depend upon the assumed line shape as well as the weak-coupling approximation. Once the COSY peak matrix is available, a 'reduced' peak matrix containing the digitized contours of the *multiplets* in the 1D spectrum is easily calculated from it, which can then be used as an aid in the analysis and simulation of many other kinds of 2D NMR spectra, particularly NOESY (Denk et al., 1986; Havel et al., 1994). Indeed, the procedure described in this paper can be viewed as a first step towards the development of a general method for decomposing experimental 2D NMR spectra into a product of matrices of the form $\mathbf{S} = \mathbf{PCQ}^T$. Like eigenvalue decompositions in linear algebra, the availability of such a decomposition of a 2D NMR spectrum makes it much easier to analyze, and thus these decompositions constitute an elegant and powerful approach to solving many problems in NMR spectroscopy.

ACKNOWLEDGEMENTS

We thank Jeff Hoch, Sven Hyberts, Andrzej Krezel, George Maalouf, Igor Najfeld, Peter Schmieder, Alan Stern and Gerhard Wagner for many useful discussions. This work was supported by NIH Grants GM-38221 and GM-47467, and by a grant from BIOSYM Technologies, Inc. The computational facilities used were paid for in part by NIH Grant RRO-7452, and by a grant from the W.M. Keck Foundation.

REFERENCES

Berndt, K.D., Güntert, P., Orbons, L.P.M. and Wüthrich, K. (1992) *J. Mol. Biol.*, **227**, 757–775.

Bodenhausen, G., Freeman, R., Niedermeyer, R. and Turner, D.L. (1977) *J. Magn. Reson.*, **26**, 133–164.

De Marco, A., Llinás, M. and Wüthrich, K. (1978) *Biopolymers*, **17**, 2727–2742.

Denk, W., Baumann, R. and Wagner, G. (1986) *J. Magn. Reson.*, **67**, 386–390.

Ernst, R.R., Bodenhausen, G. and Wokaun, A. (1987) *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*, Oxford Science Publishers, Oxford.

Griesinger, C., Sørensen, O.W. and Ernst, R.R. (1986) *J. Phys. Chem.*, **85**, 6837–6852.

Havel, T.F. (1991) *Prog. Biophys. Mol. Biol.*, **56**, 43–78.

Havel, T.F., Najfeld, I. and Yang, J.-X. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 7962–7966.

Jones, J.A., Grainger, D.S., Hore, P.J. and Daniell, G.J. (1993) *J. Magn. Reson. Ser. A*, **101**, 162–169.

Karplus, M. (1959) *J. Chem. Phys.*, **30**, 11–15.

Kessler, H., Müller, A. and Oschkinat, H. (1985) *Magn. Reson. Chem.*, **23**, 844–852.

Kim, Y. and Prestegard, J.H. (1989) *J. Magn. Reson.*, **84**, 9–13.

Ludvigsen, S., Andersen, K.V. and Poulsen, F.M. (1991) *J. Mol. Biol.*, **217**, 731–736.

Maalouf, G.J. (1993) Poster presented at the 'Macromolecules, Genes and Computers' conference, Waterville Valley, NH.

Mádi, Z.L. and Ernst, R.R. (1988) *J. Magn. Reson.*, **79**, 513–527.

Neuhaus, D., Wagner, G., Vasák, M., Kägi, J.H.R. and Wüthrich, K. (1985) *Eur. J. Biochem.*, **151**, 257–273.

Oschkinat, H., Pastore, A., Pfändler, P. and Bodenhausen, G. (1986) *J. Magn. Reson.*, **69**, 559–566.

Otting, G., Widmer, H., Wagner, G. and Wüthrich, K. (1986) *J. Magn. Reson.*, **66**, 187–193.

Pardi, A., Billeter, M. and Wüthrich, K. (1984) *J. Mol. Biol.*, **180**, 741–751.

Pearson, G.A. (1987) *J. Magn. Reson.*, **74**, 541–545.

Ponder, J.W. and Richards, F.M. (1987) *J. Mol. Biol.*, **193**, 775–791.

Smith, L.J., Sutcliffe, M.J., Redfield, C. and Dobson, C.M. (1991) *Biochemistry*, **30**, 986–996.

Szyperski, T., Güntert, P., Otting, G. and Wüthrich, K. (1992) *J. Magn. Reson.*, **99**, 552–560.

Titman, J.J. and Keeler, J. (1990) *J. Magn. Reson.*, **89**, 640–646.

Wagner, G., Braun, W., Havel, T.F., Schaumann, T., Gō, N. and Wüthrich, K. (1987) *J. Mol. Biol.*, **196**, 611–639.

Widmer, H. and Wüthrich, K. (1986) *J. Magn. Reson.*, **70**, 270–279.

Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York, NY.

Yang, J.-X., Krezel, A., Schmieder, P., Wagner, G. and Havel, T.F. (1994) *J. Biomol. NMR*, **4**, 827–844.