Journal of Biomolecular NMR, 3 (1993) 185–204
ESCOM

J-Bio NMR 105

Amino acid type determination in the sequential assignment procedure of uniformly $^{13}C/^{15}N$-enriched proteins

Stephan Grzesiek and Ad Bax

Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892, U.S.A.

Received 14 October 1992
Accepted 15 November 1992

Keywords: ^{13}C chemical shifts; Constant-time; Triple resonance; Sequential assignment; 3D NMR

SUMMARY

Experiments and procedures are described that greatly alleviate the sequential assignment process of uniformly $^{13}C/^{15}N$-enriched proteins by determining the type of amino acid from experiments that correlate side chain with backbone amide resonances. A recently proposed 3D NMR experiment, CBCA(CO)NH, correlates C^{α} and C^{β} resonances to the backbone amide ^{1}H and ^{15}N resonances of the next residue (Grzesiek, S. and Bax, A. (1992) *J. Am. Chem. Soc.*, **114**, 6291–6293). An extension of this experiment is described which correlates the proton H^{β} and H^{α} resonances to the amide ^{1}H and ^{15}N resonances of the next amino acid, and a detailed product operator description is given. A simple 2D-edited constant-time HSQC experiment is described which rapidly identifies H^{β} and C^{β} resonances of aromatic or Asn/Asp residues. The extent to which combined knowledge of the C^{α} and C^{β} chemical shift values determines the amino acid type is investigated, and it is demonstrated that the combined C^{α} and C^{β} chemical shifts of three or four adjacent residues usually are sufficient for defining a unique position in the protein sequence.

INTRODUCTION

In the past few years, a large array of heteronuclear 3D and 4D experiments have been proposed to enable the sequential assignment of larger proteins that have been uniformly isotopically enriched with ^{13}C and ^{15}N (Ikura et al., 1990; Kay et al., 1990a; Powers et al., 1991; Wagner et al., 1991; Boucher et al., 1992; Clubb et al., 1992; Grzesiek and Bax, 1992a; Kay et al., 1992; Palmer et al., 1992). These techniques are all geared to provide an as unique as possible sequential correlation between the amide $^{1}H/^{15}N$, the $^{1}H^{\alpha}/^{13}C^{\alpha}$, and the ^{13}CO resonances of one residue, and the same resonances of the next residue. The most difficult step in establishing this connectivity chain is to correlate H^{N}/N with the intraresidue H^{α}/C^{α}, mainly because of the relatively small value of $^{1}J_{NC\alpha}$ and the short transverse relaxation times of C^{α} nuclei in larger

proteins (Grzesiek et al., 1992). For larger proteins, the most sensitive experiments for bridging the N–C^{α} bond are HNCA (Ikura et al., 1990; Grzesiek and Bax, 1992a) which correlates the amide $^{1}H/^{15}N$ to the intraresidue $^{13}C^{\alpha}$, and the ^{15}N-edited HOHAHA (Marion et al., 1989a) and ^{15}N separated 3D and 4D NOESY experiments (Marion et al., 1989b; Kay et al., 1990b) which correlate the amide $^{1}H/^{15}N$ to the intraresidue $^{1}H^{\alpha}$ or, in the 4D case, to $^{1}H^{\alpha}$ and $^{13}C^{\alpha}$. If for sensitivity reasons, these are the only experiments available to obtain the N-C^{α} connectivity, an ambiguity in the assignment occurs whenever $^{1}H/^{15}N$ or $^{1}H^{\alpha}/^{13}C^{\alpha}$ resonances of a given residue do not have unique chemical shifts, i.e., when the frequencies of one amide or C^{α}/H^{α} site cannot be distinguished from the amide or C^{α}/H^{α} resonances of another residue at the resolution available in the 3D/4D spectrum. In spite of a widespread misconception, none of the 4D experiments proposed to date for backbone assignments solves this problem. In fact, because the digital resolution in 4D spectra is generally lower than in 3D spectra, the ambiguity occurs even more frequently (Bax and Grzesiek, 1993). If sensitivity permits, the common case of degenerate H^{α}/C^{α} resonances can be solved by correlating the amide $^{1}H/^{15}N$ with the intraresidue ^{13}CO resonance (Clubb et al., 1992) or the intraresidue $^{13}C^{\beta}$ resonance (Grzesiek and Bax, 1992b). Very often, however, only relatively short stretches, spanning approximately half a dozen residues, can be correlated in an unambiguous manner.

The same problem of ambiguous connections between adjacent residues also arises in the conventional sequential assignment procedure applied to much smaller unlabeled proteins. Nevertheless, reliable sequential assignments can be made provided that the amino acid type for many of the connected residues can be established on the basis of homonuclear J connectivities (Wüthrich, 1986). A similar solution can be adopted for the heteronuclear sequential assignment procedure, if the H^{α}/C^{α} frequencies are sufficiently unique to permit an unambiguous connection to the side-chain resonances, which identify the residue type and which are correlated with C^{α}/H^{α} via the HCCH-COSY (Bax et al., 1990a; Kay et al., 1990c) and HCCH-TOCSY (Bax et al., 1990b; Fesik et al., 1990) experiments. As discussed previously for the protein interferon-γ (Grzesiek et al., 1992), a 31.4 kD homo-dimer rich in α-helical structure, the resonance dispersion and H^{α}/C^{α} resolution are frequently too poor for this purpose. Therefore, a more direct approach for connecting backbone and side-chain resonances is clearly desirable.

Two experiments have been proposed recently which directly correlate the amide resonances with the preceding $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ (CBCA(CO)NH, Grzesiek and Bax, 1992c) or with both the intraresidue $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ and the preceding $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ (CBCANH, Grzesiek and Bax, 1992b). The CBCA(CO)NH experiment is particularly applicable to larger proteins, because magnetization transfer occurs via large J couplings. We now describe an extension of this experiment, named HBHA(CBCACO)NH, which correlates the H^{β} and H^{α} resonances with the backbone amide of the next residue. A detailed product operator description is presented which applies to both the HBHA(CBCACO)NH and the CBCA(CO)NH experiments. In addition, a simple and sensitive editing procedure is described, which allows selective observation of the H^{β}/C^{β} correlations for aromatic or Asp/Asn residues in a constant-time HSQC spectrum (Santoro and King, 1992; van de Ven and Philippens, 1992; Vuister and Bax, 1992), further facilitating the amino acid type assignment.

The uniqueness of amino acid type identification on the basis of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts is discussed in detail. For this purpose, a database was compiled containing the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts of 600 residues from six different proteins, and a probability distribution of $^{13}C^{\alpha}$

and $^{13}C^{\beta}$ chemical shifts for the different amino acid types was derived. A program was developed which, on the basis of this probability distribution and the primary sequence of a protein, locates a set of sequentially J-correlated residues with known C^{α}/C^{β} shifts at their most probable positions in the primary sequence. It is shown that this position usually becomes unique for proteins containing up to a few hundred amino acids once there are at least three to four J-correlated C^{α}/C^{β} pairs. Using the same strategy, the information content of the primary sequence, which is responsible for this uniqueness in location, can be used to sort out ambiguities in the J-correlations between amino acid pairs.

EXPERIMENTAL

Experiments were carried out at a ^{1}H frequency of 600 MHz on a Bruker AMX-600 spectrometer, equipped with an external 150-W class A/B power amplifier for ^{13}C and operating with software version 911101.3. The 3D HBHA(CBCACO)NH spectrum was obtained from a $52^{*}(t_1) \times 32^{*}(t_2) \times 512^{*}(t_3)$ data set (where n* refers to n complex data points), with acquisition times of 13 ms (t_1), 20.0 ms (t_2) and 55.3 ms (t_3). The total measuring time was 2.5 days. Mirror-image linear prediction (Zhu and Bax, 1990) in the t_2 domain was used, and data were zero-filled to yield a $256 \times 128 \times 1024$ matrix for the absorptive part of the final 3D spectrum. The edited constant-time HSQC resulted from a $128^{*} \times 1024^{*}$ data matrix with acquisition times of 26 ms (t_1) and 53 ms (t_2). Edited data were generated by subtraction of the two halves of the difference experiment (see below). The signal in the t_1 dimension, which does not exhibit T_2 relaxation decay, was extended to 256* by mirror-image linear prediction, prior to zero-filling to 512* and Fourier transformation. The total accumulation time was 2h. All spectra shown were acquired for a sample containing uniformly $^{15}N/^{13}C$-enriched calmodulin complexed with Ca^{2+} and M13, a 26-residue peptide fragment for which it has high affinity ($K_d \sim 10^{-9}$ M) (Ikura et al., 1992). The complex ($M_r \sim 19.7$ kD) was dissolved in 90% H_2O/10% D_2O for the HBHA(CBCACO)NH experiment, and in 99.9% D_2O for the CT-HSQC experiment. All spectra were recorded at 35°C, using a sample concentration of 1.5 mM.

DESCRIPTION OF THE PULSE SCHEMES

HBHA(CBCACO)NH

The pulse scheme used in the HBHA(CBCACO)NH experiment is given in Fig. 1. This experiment is conceptually quite analogous to the CBCA(CO)NH (Grzesiek and Bax, 1992c) experiment. However, because no detailed description of the CBCA(CO)NH experiment has been previously given, and because some novel ideas are incorporated in the HBHA(CBCACO)NH pulse scheme, a detailed product operator formalism analysis of the magnetization transfer is presented. The purpose of the experiment is to correlate the H^{α} and H^{β} resonances of residue i, with the amide ^{1}H and ^{15}N of residue i + 1. Briefly, the experiment works as follows. After the H^{α} and H^{β} transverse magnetization, created by the initial ^{1}H 90° pulse, evolves during a semi-constant-time evolution period (see below) between time points a and b in Fig. 1, the magnetization is transferred to the attached $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ nuclei. During the time 2γ, a major fraction of C^{β} magnetization is relayed to C^{α} while a fraction of magnetization remains on C^{α}. Subsequently, the C^{α} magnetization is relayed via the carbonyl to the amide of the next residue, where after a

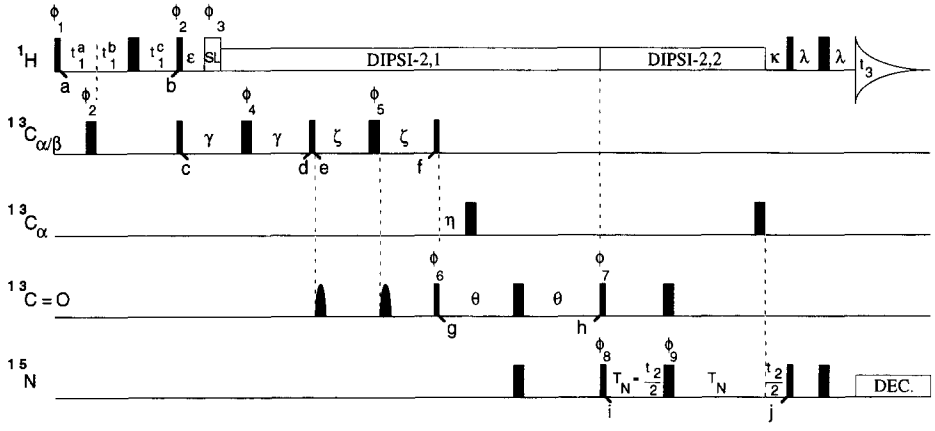

Fig. 1. Pulse scheme of the HBHA(CO)NH experiment. Narrow and wide pulses correspond to 90° and 180° flip angles, respectively. Pulses for which the phase is not indicated are applied along the x axis. The ^{1}H carrier is set to the H_2O frequency for the first part of the pulse sequence, up to time h, and is switched to the center of the amide region (8.4 ppm) thereafter. Thus, the frequency of the ^{1}H broad-band decoupling (with a 5-kHz RF field) is switched at this point, and the two modes of decoupling, using a coherent DIPSI-2 scheme (Shaka et al., 1988), are marked DIPSI-2,1 and DIPSI-2,2. ^{15}N decoupling is accomplished using WALTZ-16 modulation with a 1.5-kHz RF field. Carbonyl pulses have a shaped amplitude profile, corresponding to the center lobe of a sinx/x function and a duration of 202 μs. The carrier for the $C^{\alpha/\beta}$ pulses is positioned at 46 ppm, for the C^{α} pulses at 56 ppm, and for the carbonyl pulses at 177 ppm. The power of the 90° and 180° $^{13}C_{\alpha/\beta}$ pulses, as well as the 180° $^{13}C^{\alpha}$ pulses is adjusted such that they do not excite the ^{13}CO nuclei (i.e., 4.43, 11.4, 10.6 kHz RF field for 150.9 MHz ^{13}C frequency, respectively). Phase cycling is as follows: $\phi_1 = y$; $\phi_2 = x, -x$; $\phi_3 = y$; $\phi_4 = 8(x),8(y),8(-x),8(-y)$; $\phi_5 = 4(x),4(-x)$; $\phi_6 = 2(x),2(-x)$; $\phi_7 = 48.5°$ (Bloch–Siegert phase error compensation); $\phi_8 = 4(x),4(-x)$; $\phi_9 = 8(x),8(-x)$;Acq. = x,2(−x),x,−x,2(x),2(−x),2(x),−x,x,2(−x),x. Quadrature in the t_1 and t_2 domains is obtained by changing the phases ϕ_1 and ϕ_8, respectively, in the usual States-TPPI manner (Marion et al., 1989c). Delay durations are: ε = 2.1 ms; SL = 1.0 ms; γ = 3.1 ms; ζ = 3.7 ms; η = 4.7 ms; θ = 11.4 ms; T_N = 11.2 ms; κ = 5.4 ms; λ = 2.25 ms. The initial delays for the 'semi-constant-time' proton evolution period (see text) were set to t_1^a = 1.5 ms; t_1^b = 0; t_1^c = 1.5 ms. Increments for the delays were set according to Eq. 3 for a total acquisition time of 13 ms.

constant-time ^{15}N evolution period, the magnetization of the amide proton is detected during the time t_3.

The relevant terms of the magnetization transfer pathway are expressed in terms of the product operator formalism (Ernst et al., 1987). For clarity, relaxation terms are not included, constant multiplicative factors are omitted, and only terms that result in observable magnetization during the detection period, t_3, are retained. The spin operators used are $\mathbf{H}^{\alpha}$, $\mathbf{H}^{\beta}$, $\mathbf{C}^{\alpha}$, $\mathbf{C}^{\beta}$, and $\mathbf{C}'$ for the ^{13}C spins of the i-th amino acid as well as $\mathbf{N}$ and $\mathbf{H}^{N}$ for the amide ^{15}N and proton spin of residue i + 1. The resonance offset of a nucleus, X, is denoted δ_X, whereas the coupling constant between adjacent nuclei, X and Y, is denoted J_{XY}. The sequence will be described assuming RF phases of the pulses corresponding to the first step of the phase cycle given in the legend to Fig. 1.

The t_1 evolution period, during which H^{α} and H^{β} magnetization evolves, is of a 'semi-constant-time' nature. During this period, between time points a and b in Fig. 1, H^{α} and H^{β} spins must become antiphase with respect to their directly attached ^{13}C nuclei, which requires a time of $1/(2J_{CH})$, i.e. ~ 3.5 ms. However, because transverse relaxation of H^{α} and H^{β} is relatively fast ($T_2(H^{\alpha})$ ~ 6 ms for interferon-γ), it is desirable to also use this 3.5-ms period for frequency labeling. In fact, to minimize relaxation losses, a slightly shorter value of 3 ms is used for the

initial duration of this period, i.e. for $t_1 = 0$. When the length of the evolution period increases, the dephasing period is also gradually increased to $1/(2J_{CH})$.

Between time points a and b, the ^1H chemical shift evolution is active for a time

$$t_1 = t_1^a + t_1^b - t_1^c \tag{1a}$$

and J_{CH} evolution is effective for a time t_{1J} given by:

$$t_{1J} = t_1^a - t_1^b + t_1^c \tag{1b}$$

An initial condition of $t_{1J} = 3$ ms and $t_1 = 0$ for the first data point can be realized by setting:

$$t_1^a(1) = t_1^c(1) = 1.5 \text{ ms}; \; t_1^b(1) = 0 \tag{2a}$$

The total acquisition time AT_1 is defined by the last t_1 increment, N, as:

$$AT_1 = t_1^a(N) + t_1^b(N) - t_1^c(N) \tag{2b}$$

Usually, AT_1 is longer than 3.5 ms, and the individual components can be set such that

$$t_1^c(N) = 0; \; t_{1J}(N) = t_1^a(N) - t_1^b(N) = 3.5 \text{ ms}; \tag{2c}$$

This requires the following increments for the three components of t_1:

$$\Delta t_1^a = [(AT_1 + 3.5 \text{ ms})/2 - 1.5 \text{ ms}]/(N-1) \tag{3a}$$

$$\Delta t_1^b = (AT_1 - 3.5 \text{ ms})/(2(N-1)) \tag{3b}$$

$$\Delta t_1^c = -(1.5 \text{ ms})/(N-1) \tag{3c}$$

and the spectral width in the t_1 dimension, SW_1, is given by:

$$SW_1 = (\Delta t_1^a + \Delta t_1^b - \Delta t_1^c)^{-1} \tag{4}$$

The procedure described above incorporates the INEPT transfer time into the t_1 evolution period, thereby reducing the apparent relaxation rate in the t_1 dimension by a factor of $(AT_1 - t_{1J}(1))/AT_1$. This amounts to a factor 10/13 for the calmodulin data shown, and to a factor 5/8 for the protein interferon-γ (data not shown). This procedure also scales the size of homonuclear ^1H–^1H J splitting by the same factor, causing an additional small increase in resolution. Note that this procedure is applicable to many different heteronuclear shift correlation experiments, including the older ^{13}C detected ones, without increasing the number of RF pulses.

At the end of the semi-constant-time evolution period, H^β and H^α magnetizations are antiphase with respect to their directly attached C^α and C^β spins, and the pair of 90° ($^1H/^{13}C$) pulses applied

at time b transfers the ^1H magnetization that is perpendicular to the ϕ_2 axis into antiphase ^{13}C magnetization (time c):

$$-H_z^{\alpha}C_y^{\alpha} \sin(\pi J_{CH}\, t_{1J})\cos(2\pi\delta_{H\alpha}t_1) \tag{5a}$$

$$-H_z^{\beta}C_y^{\beta} \sin(\pi J_{CH}t_{1J})\cos(2\pi\delta_{H\beta}t_1) \tag{5b}$$

Refocusing of this antiphase magnetization follows a different time dependence for methine, methylene, and methyl carbons. A compromise value of 2.1 ms (Burum and Ernst, 1980) was used for this refocusing delay, ε, in the scheme shown in Fig. 1. At the end of this refocusing delay, a water purge pulse, labeled SL, is applied and ^1H decoupling is switched on. Beginning at time point c and continuing to time point d, the transverse carbon magnetization dephases due to homonuclear J coupling with adjacent carbons. The effect of J coupling between aliphatic and carbonyl carbons, however, is eliminated by adjusting the power of the $180°_{\phi 4}$ pulse in such a way that it does not excite the carbonyl ^{13}C nuclei. By temporarily neglecting the sine and cosine terms of Eq. 5 and assuming complete refocusing of the antiphase carbon magnetization during ε, the two spin operators of Eq. 5 evolve from c to d as:

$$H_z^{\alpha}C_y^{\alpha} \rightarrow -C_x^{\alpha}\cos^m(2\pi\gamma J_{C\alpha C\beta}) \tag{6a}$$

$$H_z^{\beta}C_y^{\beta} \rightarrow -C_y^{\beta}C_z^{\alpha}\sin(2\pi\gamma J_{C\alpha C\beta})\cos^n(2\pi\gamma J_{C\beta C\gamma}) \tag{6b}$$

where the exponent m is zero for glycine and 1 for all other residues. The exponent n is 2 for valine and isoleucine, 0 for alanine, aspartic acid, asparagine, cysteine and serine, and 1 for all other residues. The $C^{\alpha/\beta}\ 90°_x$ pulse at point d converts the spin operators of expression 6 into

$$C_x^{\alpha} \rightarrow C_x^{\alpha} \tag{7a}$$

$$C_y^{\beta}C_z^{\alpha} \rightarrow -C_z^{\beta}C_y^{\alpha} \tag{7b}$$

Between time points e and f, these terms dephase due to J_{CC} coupling to their adjecent carbons. Note that during this interval, the coupling from the $^{13}C^{\alpha}$ to the carbonyl ^{13}C nucleus is active, as a selective 180° carbonyl pulse is applied at the same time as the 180° $C^{\alpha/\beta}$ pulse. The spurious phase shift of C^{α} magnetization caused by the Bloch–Siegert effect associated with the carbonyl pulse (Vuister and Bax, 1992; McCoy and Mueller, 1992) is compensated for by the application of the 'dummy' shaped ^{13}CO pulse at the beginning of the transfer interval, 2ζ. The relevant evolution of the operators in Eq. 7 between time points e and f is given by:

$$C_x^{\alpha} \rightarrow C_y^{\alpha}C_z'\cos^m(2\pi\zeta J_{C\alpha C\beta})\sin(2\pi\zeta J_{C\alpha C'}) \tag{8a}$$

$$C_z^{\beta}C_y^{\alpha} \rightarrow -C_y^{\alpha}C_z'\sin(2\pi\zeta J_{C\alpha C\beta})\sin(2\pi\zeta J_{C\alpha C'}) \tag{8b}$$

The pair of $^{13}C^{\alpha/\beta}$ and C′ 90° pulses at time f converts these terms according to:

$$\mathbf{C}_y^{\alpha}\mathbf{C}_z' \rightarrow -\mathbf{C}_z^{\alpha}\mathbf{C}_y' \tag{9}$$

During the next transfer interval, 2θ, between time points g and h, antiphase C′ magnetization refocuses for a time period, 2η, and the transverse carbonyl ^{13}C dephases with respect to the neighboring nitrogen for total time 2θ:

$$\mathbf{C}_z^{\alpha}\mathbf{C}_y' \rightarrow -\mathbf{C}_y'\mathbf{N}_z \sin(2\pi\eta J_{C\alpha C'}) \sin(2\pi\theta J_{C'N}) \tag{10}$$

Again, the $^{13}C^{\alpha}$ pulse in the first part of this transfer period, which serves to introduce the C^{α} to C′ coupling for a duration of 2η, causes a change in the ^{13}CO resonance frequency due to the Bloch–Siegert effect. The concomitant change in the ^{13}CO phase in this case is compensated for by an adjustment of the phase ϕ_7 of the last carbonyl 90° pulse. At time h, the simultaneous 90° $^{13}C'$ and ^{15}N 90° pulses transfer the antiphase C′ magnetization into antiphase N magnetization:

$$\mathbf{C}_y'\mathbf{N}_z \rightarrow -\mathbf{C}_z'\mathbf{N}_y \tag{11}$$

During the subsequent ^{15}N constant-time period of total duration $2T_N$, between time points i and j, the $\mathbf{C}_z'\mathbf{N}_y$ transverse ^{15}N magnetization rephases with respect to its carbonyl coupling partner. The effect of the $^{13}C^{\alpha}$–^{15}N J coupling is eliminated during the first $2T_N - t_2$ fraction of the ^{15}N evolution period by the 180° ^{15}N pulse and by the fact that the 180° C′ pulse is adjusted such that it does not excite the C^{α} spins. During the last fraction, t_2, of the constant-time ^{15}N evolution period, the 180° $^{13}C^{\alpha}$ pulse decouples the $^{15}N/^{13}C^{\alpha}$ interaction. During the delay, κ, ^1H decoupling is switched off and ^{15}N magnetization becomes antiphase with respect to its attached proton spin. These effects are all summarized by:

$$\mathbf{C}_z'\mathbf{N}_y \rightarrow -\mathbf{N}_y\mathbf{H}_z^N \sin(2\pi T_N J_{C'N}) \sin(\pi\kappa J_{NH}) \cos(2\pi\delta_N t_2) \tag{12}$$

$^1J_{NH}$ is quite uniform in proteins and typically falls in the 92–94 Hz range. With $\kappa = 5.4$ ms, the $\sin(\pi\kappa J_{NH})$ term in Eq. 12 is very close to one and therefore can be safely omitted.

The final reverse INEPT, applied following time j, transforms $\mathbf{N}_y\mathbf{H}_z^N$ into $-\mathbf{H}_x^N$, which is observed during the detection period, t_3. As is clear from the above, for a given amide proton, the HBHA(CBCACO)NH experiment yields correlations to the H^{α} and H^{β} of the preceding residue. Omitting relaxation losses, and neglecting pulse imperfections and incomplete de- and rephasing during the first INEPT and the final reverse INEPT transfers, the amplitudes of the two cross peaks to H^{α} and H^{β} are obtained by combining the trigonometric terms of Eqs. 5–12:

$$\begin{aligned}\mathbf{H}_z^{\alpha} \rightarrow \mathbf{H}_x^N &\cos^m(2\pi\gamma J_{C\alpha C\beta}) \cos^m(2\pi\zeta J_{C\alpha C\beta}) \sin(2\pi\zeta J_{C\alpha C'}) \times \\ &\sin(2\pi\eta J_{C\alpha C'}) \sin(2\pi\theta J_{C'N}) \sin(2\pi T_N J_{C'N}) \cos(2\pi\delta_{H\alpha} t_1) \cos(2\pi\delta_N t_2)\end{aligned} \tag{13a}$$

$$\begin{aligned}\mathbf{H}_z^{\beta} \rightarrow \mathbf{H}_x^N &\cos^n(2\pi\gamma J_{C\beta C\gamma}) \sin(2\pi\gamma J_{C\alpha C\beta}) \sin(2\pi\zeta J_{C\alpha C\beta}) \sin(2\pi\zeta J_{C\alpha C'}) \times \\ &\sin(2\pi\eta J_{C\alpha C'}) \sin(2\pi\theta J_{C'N}) \sin(2\pi T_N J_{C'N}) \cos(2\pi\delta_{H\beta} t_1) \cos(2\pi\delta_N t_2)\end{aligned} \tag{13b}$$

Although not immediately clear from Eq. 13, it should be borne in mind that, for two non-

equivalent H^{β} methylene protons, the correlations have approximately half the intensity of that of a H^{β} methine proton. This additional factor of ~ 0.5 results from the fact that while rephasing of C^{β} magnetization occurs during the delay, ε, for the coupling to one of the H^{β} protons, simultaneous dephasing occurs due to coupling to the second H^{β} proton (Burum and Ernst, 1980).

CBCA(CO)NH

The previously published CBCA(CO)NH pulse scheme (Grzesiek and Bax, 1992c), which correlates C^{β} and C^{α} resonances with the amide of the next residue, is very similar to that of the HBHA(CBCACO)NH experiment; only the initial part of the pulse scheme is slightly different. In the CBCA(CO)NH experiment, the semi-constant-time evolution period of the HBHA(CBCACO)NH experiment is replaced by a regular INEPT transfer, and the dephasing/rephasing period, 2γ, is replaced by a constant-time evolution period (Fig. 2A). The CBCA(CO)NH experiment can also be performed without the first INEPT transfer (Fig. 2B), using presaturation of the ^{1}H resonances, resulting in a heteronuclear NOE enhancement of the ^{13}C magnetization. Comparison of the INEPT-enhanced versus the NOE-enhanced version of the CBCA(CO)NH experiment gives an approximately 1.5 times lower signal-to-noise ratio. However, significantly better suppression of the H_2O resonance is obtained with the NOE-enhanced version because all ^{1}H magnetization is effectively destroyed by the ^{1}H saturation. As in our case

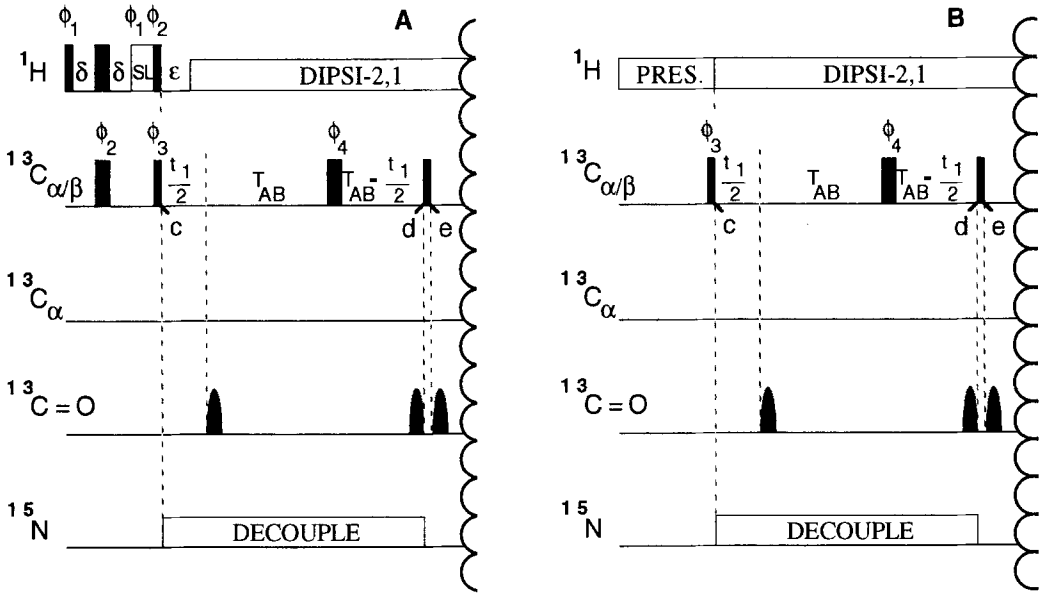

Fig. 2. Pulse scheme of the CBCA(ON)NH experiment. Only the first part, which is different from the pulse scheme of the HBHA(CO)NH experiment (Fig. 1), is shown. Besides the steps listed below, all delays and phase cycling steps are identical to those in Fig. 1. Delays are: δ = 1.5 ms, T_{AB} = 3.3 ms. A: scheme with INEPT transfer from protons to attached C^{α}/C^{β} carbons. The H_2O resonance is suppressed by a 1.8-ms purge pulse, SL. Phase cycling is as follows: ϕ_1 = y; ϕ_2 = x,−x; ϕ_3 = x. Quadrature in the t_1 domain is obtained by changing the phase ϕ_3 in the usual States-TPPI manner (Marion et al., 1989c). B: scheme with heteronuclear NOE enhancement of the ^{13}C magnetization by proton presaturation. Phase cycling is as follows: ϕ_3 = y,−y. Quadrature in the t_1 domain is obtained as in Fig. 2A.

the H_2O suppression obtained with the INEPT-enhanced scheme is adequate, and because the residual H_2O signal does not interfere with that of the directly observed amide proton signals, we prefer to use the INEPT-enhanced version of the experiment.

Edited HSQC

We recently demonstrated the use of a constant-time HSQC ^1H-^{13}C correlation scheme that increases the ^{13}C resolution of the 2D ^1H–^{13}C correlation spectrum of uniformly ^{13}C-enriched proteins (Vuister and Bax, 1992). Such a 2D spectrum can be obtained quite rapidly and frequently resolves many of the ^1H–^{13}C correlations. Here we describe an equally sensitive analog of this CT-HSQC experiment, which selects only correlations to carbons that are directly coupled to either an aromatic or to a carbonyl carbon, and thereby provides a rapid means for verifying the residue type assignment for aromatic and Asn/Asp residues.

The pulse scheme of the edited CT-HSQC experiment is shown in Fig. 3. The ^1H and $^{13}C^{\alpha/\beta}$ pulses are identical to those in the original timing diagram (Vuister and Bax, 1992), and the total time, 2T, between the end of the initial INEPT transfer (time a) and the start of the reverse INEPT transfer (time b) is adjusted to $N/J_{C\alpha C\beta}$ (N = 1,2). The edited CT-HSQC scheme is a difference experiment, where in one case the modulation due to J coupling between aliphatic and carbonyl or aromatic carbons evolves for a time $1/J_{C\beta C\gamma}$, and in the other the effect of J coupling during the constant-time evolution period is removed, resulting in signals of opposite phase. These two 'halves' of the difference experiment are briefly discussed below.

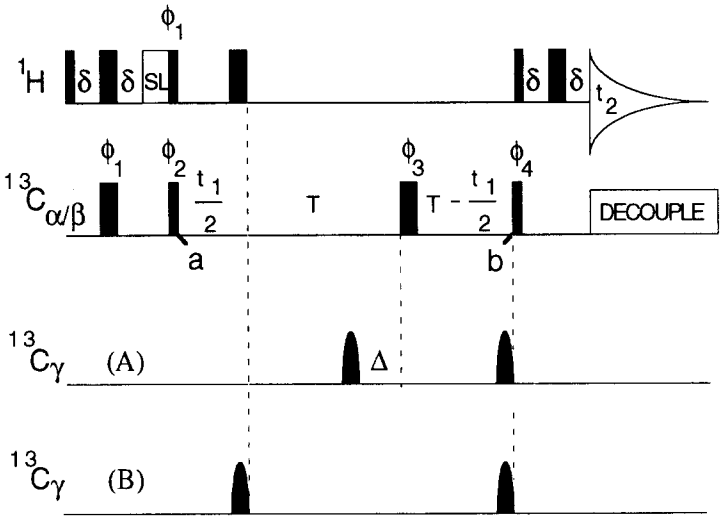

Fig. 3. Pulse scheme of the edited constant-time HSQC experiment. The ^1H carrier is set to the H_2O frequency. ^{13}C decoupling is accomplished using WALTZ-16 modulation with a 4.5-kHz RF field. $^{13}C^{\gamma}$ pulses have a shaped amplitude profile, corresponding to the center lobe of a sinx/x function and a duration of 294 μs. The carrier for the $C^{\alpha/\beta}$ pulses is positioned at 43 ppm, and for the C^{γ} pulses at 133 ppm for editing the aromatic residues and at 177 ppm for selecting the resonances adjacent to carbonyls. The power of the 90° and 180° $^{13}C^{\alpha/\beta}$ pulses, except the ϕ_3 pulse, corresponds to 18 kHz, whereas the ϕ_3 pulse is applied using a 7.8-kHz (for aromatic editing) or an 11.7-kHz (for carbonyl editing) RF field. Phase cycling is as follows: ϕ_1 = y,−y; ϕ_2 = x; ϕ_3 = 2(x),2(y),2(−x),2(−y); ϕ_4 = 8(x),8(−x); Acq. = x,2(−x),x,−x,2(x),2(−x),2(x),−x,x,2(−x),x. Quadrature in the t_1 domain is obtained by changing the phase ϕ_2 in the usual States-TPPI manner (Marion et al., 1989c). Delay durations are: δ = 1.7 ms, SL = 0.5 ms, 2T = 27 ms, and Δ = 3 ms.

All pulses except the $180^\circ_{\phi 3}$ ^{13}C pulse are applied with high power in a non-selective manner. The power and duration of the $180^\circ_{\phi 3}$ pulse are adjusted to cause minimal excitation of the aromatic or carbonyl C^γ resonances. If the difference in frequency between the center of the aromatic (or carbonyl) C^γ region and the position of the $C^{\alpha/\beta}$ carrier (43 ppm in our experiments) equals W hertz, this is accomplished by setting the RF field strength of the 180° pulse to $W/\sqrt{3}$ hertz, resulting in a 180° pulse width on resonance given by $\sqrt{3}/(2W)$ s. In one 'half' of the difference experiment, the selective 180° pulses indicated on trace (A) in Fig. 3 are applied. Neglecting the effect of one-bond $J_{C\alpha C\beta}$ coupling at the end of the constant-time evolution period (time b), which is refocused if 2T is adjusted to $N/J_{C\alpha C\beta}$, transverse ^{13}C magnetization dephases due to J coupling with C^γ for a time $t_1/2 + T - \Delta$ up to the first selective 180° pulse. After this pulse, the sign of the $J_{C\beta C\gamma}$ interaction is reversed for a period, Δ, after which J dephasing is inverted once more by the $180^\circ_{\phi 3}$ pulse. Effectively, $J_{C\beta C\gamma}$ dephasing occurs for a total time, T_J, given by:

$$T_J = (t_1/2 + T - \Delta) - \Delta\ (T - t_1/2) = 2T - 2\Delta. \tag{14a}$$

The duration of Δ is adjusted such that

$$(2T - 2\Delta) = 1/(J_{C\beta C\gamma}) \tag{14b}$$

and

$$2T = N/J_{C\alpha C\beta},\ N = 1,2 \tag{14c}$$

where $J_{C\alpha C\beta}$ is the typical one-bond aliphatic carbon–carbon coupling (~ 34 Hz). Couplings between C^β and adjacent aromatic and carbonyl carbons are significantly larger (~ 43–55 Hz). Thus, for $2T = 1/J_{C\alpha C\beta}$ (~ 27 ms), Δ is set to ~ 3 ms.

In the second half of the difference experiment, the selective pulses indicated on trace (B) in Fig. 3 are applied. In this case, C^β–C^γ coupling during the first fraction, t_1, of the constant-time evolution period is eliminated by the first selective 180° (C^γ) pulse. The effect of this coupling during the remainder of the constant-time evolution period is eliminated by the $180^\circ_{\phi 3}$ pulse. Signals for aliphatic carbons coupled to the aromatic (or carbonyl) carbons will therefore have an opposite sign in the two halves of the experiment, whereas signals for other carbons will be identical in both halves of the experiment. FIDs corresponding to both halves of the difference experiment are recorded in an interleaved manner, and subtraced from one another before processing.

To allow the use of mirror-image linear prediction (Zhu and Bax, 1990) in the t_1 dimension of the edited HSQC spectrum, and to eliminate the need for phase correction in this dimension, a 'Bloch–Siegert compensation pulse', discussed above, is applied to the aromatic (or carbonyl) resonances just before time b.

RESULTS AND DISCUSSION

Experimental results

The experiments discussed above were applied to the protein calmodulin (CaM) (148 residues), complexed with the M13 peptide, which comprises the binding site of skeletal muscle myosin

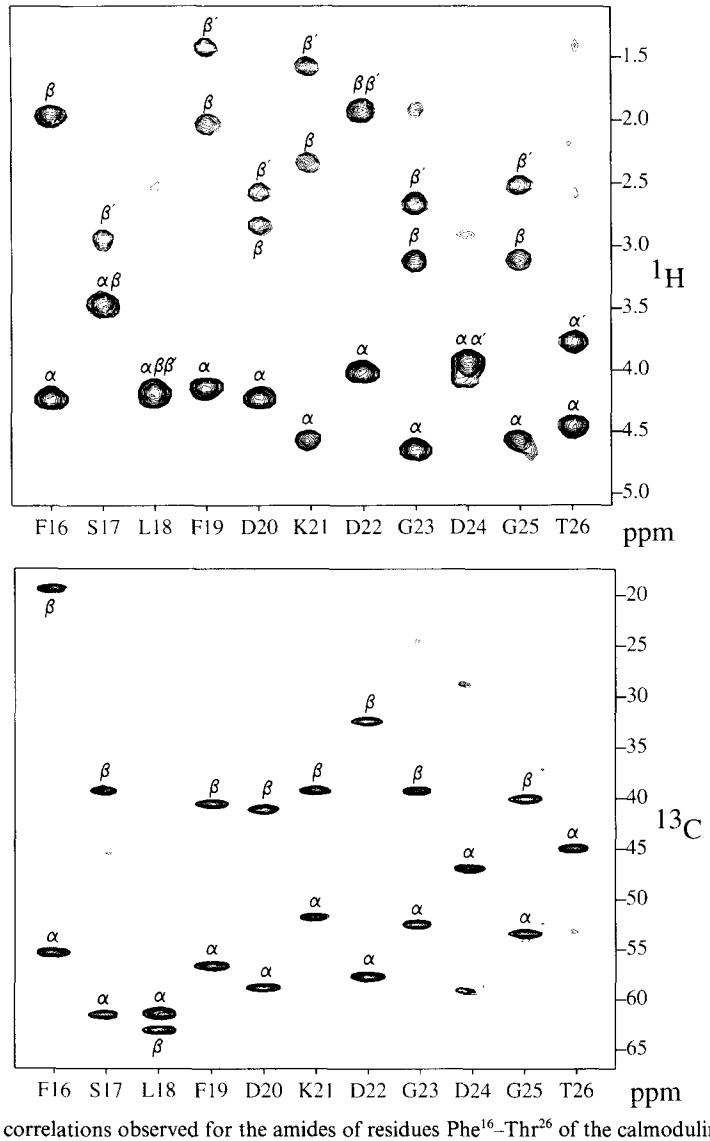

Fig. 4. Strip plot of the correlations observed for the amides of residues Phe^{16}–Thr^{26} of the calmodulin-peptide complex. Each amide correlates (A) with the H^{α} and H^{β} of the preceding residue or (B) with the corresponding C^{α} and C^{β} frequencies. Resonances which are not marked by α or β correspond to correlations to amide ^{1}H–^{15}N pairs that are close in frequency to the one for which the strip has been selected.

light-chain kinase. The structure for this globular complex has recently been solved by NMR (Ikura et al., 1992) and independently by X-ray crystallography (Meador et al., 1992).

Figure 4 shows strips for 11 adjacent residues, displaying (A) the H^{β} and H^{α} resonances of the residue preceding the observed amide in the HBHA(CBCACO)NH spectrum, and (B) the corresponding C^{β} and C^{α} resonances observed in the CBCA(CO)NH spectrum. Each of the 3D spectra required ~ 2.5 days of data acquisition. This data acquisition time is the minimum needed to obtain the required digital resolution and to execute a 32-step phase cycle needed for artifact

suppression. Considering that the signal-to-noise ratio is significantly larger than needed, the use of pulsed field gradients may considerably shorten the measuring time for these experiments by reducing the need for phase cycling, provided that sample concentrations are at least ~ 1 mM, and line widths are comparable to or narrower than for the CaM-M13 complex.

Together with the data obtained from the CBCANH experiment, which correlates both the intraresidue C^{β} and C^{α} and the C^{β} and C^{α} of the preceding residue to the backbone amide, these methods provide a very convenient and efficient path for sequential assignment (Grzesiek and Bax, 1992b). In the case of the CaM-M13 complex, the assignment was made previously with a more elaborate set of triple resonance experiments (Ikura et al., 1991), and the present experiments were used only to verify the original assignments. It was found that for two pairs of residues (Glu^{7} and Glu^{14} as well as Glu^{83} and Glu^{84}), the backbone amides and H^{α} and C^{α} resonances were interchanged. The fact that the C^{β} resonances of Glu^{6} and Lys^{13} (and of Glu^{83} and Glu^{84}) are significantly different allowed for the unambiguous reassignment of these residues.

When combined, the CBCA(CO)NH and HBHA(CBCACO)NH experiments make it possible to obtain unambiguous assignments of H^{β}, H^{α}, C^{β}, and C^{α} to the amide $^{15}N/^{1}H$ frequencies without relying on spectral editing by the strongly overlapping H^{α}/C^{α} resonances. It is important to note that all necessary experiments are carried out on a single sample, dissolved in H_2O. The remainder of the side-chain resonances can then be connected to the assigned resonances using HCCH-COSY-type (Kay et al., 1990c; Bax et al., 1990a) and HCCH-TOCSY-type (Fesik et al., 1990; Bax et al., 1990b) techniques, using a protein sample dissolved in D_2O. Previously, the main problem when making this connection was the degeneracy or near-degeneracy of C^{α}–H^{α} correla-

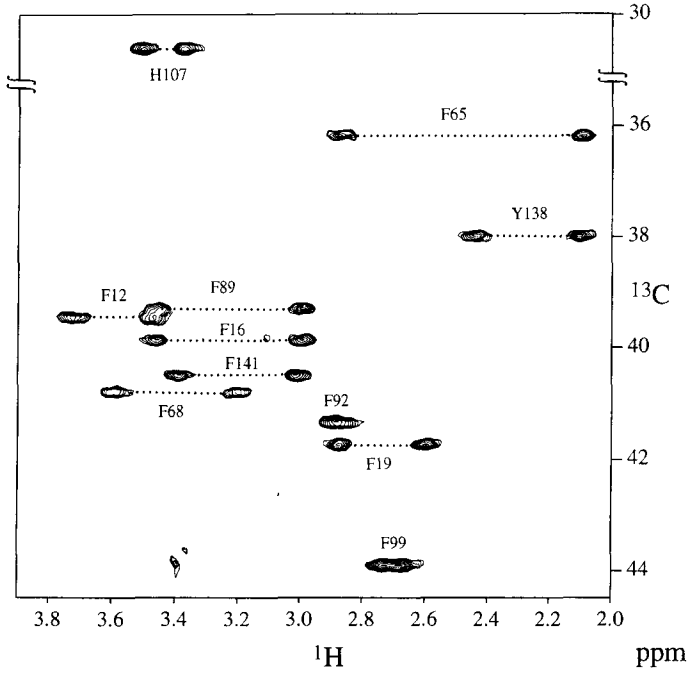

Fig. 5. 2D-edited HSQC of the calmodulin-peptide complex showing only H^{β} to C^{β} correlations of aromatic side-chain resonances.

tions and the frequently observed small differences in H^{α} and C^{α} shifts between the H_2O and D_2O samples. With the development of CBCA(CO)NH and HBHA(CBCACO)NH experiments, these problems are largely alleviated by the additional information contained in the C^{β} and H^{β} chemical shifts.

Figure 5 shows the spectrum obtained with the edited CT-HSQC sequence of Fig. 3. This spectrum readily separates the H^{β} resonances of the Phe, Tyr, and His residues from other aliphatic signals in a very crowded region of the 2D ^{1}H–^{13}C correlation spectrum. Similarly, the Asp and Asn C^{β}–H^{β} and the Glu and Gln C^{γ}–H^{γ} resonances (together with C^{α}–H^{α} of all residues) are selected by changing the frequency of the selective pulses in Fig. 3 to the carbonyl region (data not shown).

The use of C^{α} and C^{β} chemical shifts for identifying the amino acid type

In conventional 2D NMR of small proteins, the amino acid type is commonly identified by the J coupling network of the side-chain protons and their characteristic chemical shifts (Wüthrich, 1986). As was pointed out by Oh et al. (1988), ^{13}C–^{13}C coupling networks together with the characteristic ^{13}C chemical shifts are even more useful for establishing the type of amino acid. However, with the experiments discussed above, only the C^{α} and C^{β} chemical shifts (together with H^{α} and H^{β}) are generally available for this purpose during the early stages of the assignment process. It is therefore interesting to consider to what degree the C^{α} and C^{β} chemical shifts define the type of amino acid.

Figure 6 shows a plot of the random coil $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts values at neutral pH for all natural amino acids with a C^{β} carbon. The difference between the observed chemical shift and its random coil value is commonly referred to as the secondary shift. The sign and magnitude of the secondary shift depend on the protein structure (Wishart et al, 1991; Spera and Bax, 1991). For example, C^{α} resonances show a downfield shift of ~ 3 ppm in α-helical regions and a smaller upfield shift in β-sheets. C^{β} resonances on average are close to their random coil values in α-helices and shift downfield in β-sheets. In order to use C^{α} and C^{β} shifts for the identification of the side-chain type, these deviations must be taken into account. We therefore calculated these secondary shifts for 600 amino acids in six different proteins for which nearly complete C^{α} and C^{β} chemical shift assignments were available and which all were referenced in the same way. These proteins are calmodulin (Ikura et al., 1990), interleukin-1β (Clore et al., 1990), staphylococcal nuclease (SNase) (D. Torchia, personal communication), basic pancreatic trypsin inhibitor (BPTI) (Wagner and Brühweiler, 1986; Hansen, 1991), III^{Glc} (Pelton et al., 1991), and the RNA binding domain of the human hnRNP C proteins (Wittekind et al., 1992). These proteins include a representative mix of α-helix and β-sheet, and the secondary chemical shift (i.e., the deviation from the random coil chemical shift) should be representative for proteins of unknown structure.

A probability distribution, $p_{\Delta}(\Delta\delta_{C\alpha},\Delta\delta_{C\beta})$, of the secondary C^{α} and C^{β} shift values, $\Delta\delta_{C\alpha}$ and $\Delta\delta_{C\beta}$, was then derived from the 600 secondary shift pairs $(\Delta\delta^{i}_{C\alpha}, \Delta\delta^{i}_{C\beta})$ by convolution with Gaussian functions:

$$p_{\Delta}(\Delta\delta_{C\alpha},\Delta\delta_{C\beta}) = N_{P\Delta} \times \Sigma_i \exp\{-[(\Delta\delta_{C\alpha} - \Delta\delta^{i}_{C\alpha})^2 + (\Delta\delta_{C\beta} - \Delta\delta^{i}_{C\beta})^2]/\sigma^2\} \quad (15)$$

with $N_{P\Delta} = 1/(\sigma\pi N)$, and Σ_i extending over all N amino acids in the data base. In order to obtain a smooth P_{Δ} surface, the value of σ was arbitrarily set to 1 ppm. The $p_{\Delta}(\Delta\delta_{C\alpha},\Delta\delta_{C\beta})$ distribution

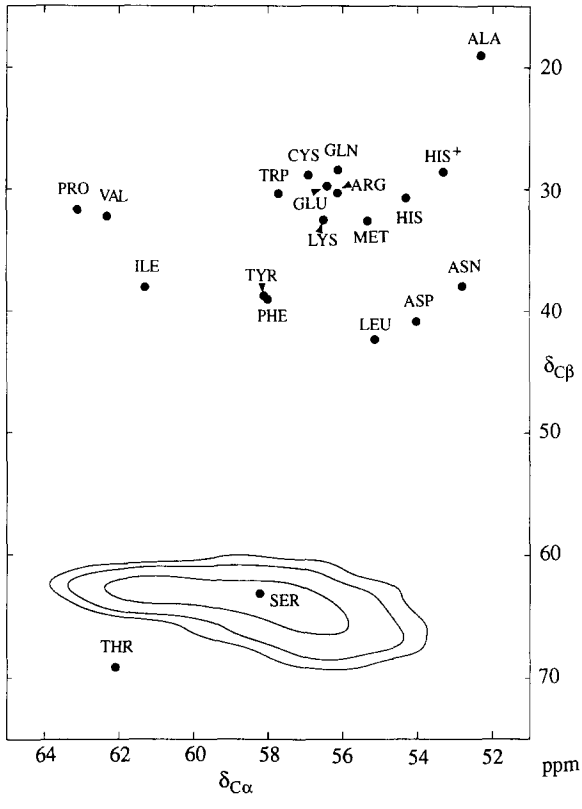

Fig. 6. Plot of the random coil chemical shift for $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ (solid dots) of the 19 natural amino acids which contain β-carbons. With the exception of histidine, the chemical shift values are taken from Spera and Bax (1991) and apply to a pH range of 5.8 to 7.4. For histidine, values from Howarth and Lilley (1978) were included (after adjusting for referencing relative to TSP) corresponding to the protonated (HIS^+) and unprotonated (HIS) state of the imidazole ring. Superimposed onto the serine random coil shift is a probability distribution of the deviation from the random coil values of 600 amino acids (see text). The contour lines are shown for probabilities of 0.95, 0.90, and 0.60 of finding the secondary shift of those 600 amino acids added to the random coil value of serine within the three contours.

is relatively broad compared to the dispersion of the random coil chemical shift values of the natural amino acids. For clarity, in Fig. 6 the shape of this distribution is only displayed for serine, and in the derivation of this shape it was assumed that it is the same for all amino acids. Contour lines shown in Fig. 6 show the regions that contain 95, 90 and 60% of the secondary chemical shifts of the 600 amino acids added to the random coil value of serine. Thus, for a serine residue in a protein of unknown structure, there is a 95% chance that the C^{α} and C^{β} chemical shifts fall within the outer contour line. Clearly, with the exception of glycine (not shown), serine, threonine, and alanine, the C^{α} and C^{β} chemical shifts are insufficiently distinct to make unique residue type assignments. However, even in the region where the contour lines, if drawn for all amino acids, would overlap extensively (i.e., for C^{β} shifts near ~ 30 ppm), a difference in probability can be calculated from a given combination of C^{α}/C^{β} chemical shifts.

The power to distinguish between different amino acids on the basis of the C^{α} and C^{β} chemical

shifts can be quantified more precisely by calculating the overlap integrals between the shift distributions for pairs of amino acids. Again, assuming identical distributions of secondary chemical shifts for all amino acids, the probability for a pair of C^{α} and C^{β} chemical shifts $(\delta_{C\alpha},\delta_{C\beta})$ to belong to an amino acid of type i, is given by:

$$p^{i}(\delta_{C\alpha},\delta_{C\beta}) = N_p \times p_{\Delta}(\delta_{C\alpha} - \delta^{i}_{C\alpha,RC},\delta_{C\beta} - \delta^{i}_{C\beta,RC}) \tag{16}$$

where N_p is a normalization constant, chosen such that $\sum_i p^{i}(\delta_{C\alpha},\delta_{C\beta}) = 1$, and $\delta^{i}_{C\alpha,RC}$ and $\delta^{i}_{C\beta,RC}$ refer to the random coil chemical shift values of residue type i. Thus, p^{i} is the probability that the $\delta_{C\alpha}$ and $\delta_{C\beta}$ shifts correspond to residue type i, disregarding the number of residues of type i in the primary sequence.

The overlap integral, O(i,j), between the chemical shift distributions of the amino acids i and j is a second important quantity, defined by:

$$O(i,j) = N_O \times \int\int p^{i}(\delta_{C\alpha},\delta_{C\beta}) \times p^{j}(\delta_{C\alpha},\delta_{C\beta})\, d\delta_{C\alpha} d\delta_{C\beta} \tag{17}$$

where the integral extends over the entire C^{α} and C^{β} chemical shift range. N_O is again a normalization constant defined by $\sum_j O(i,j) = 1$. O(i,j) is the probability that, given the C^{α} and C^{β} chemical shifts of a certain amino acid type i, it is also possible to assign to it an amino acid type j. The values of the overlap integrals (in percent) thus determined are listed in Table 1, where the entries in each row have been ordered in the direction of decreasing probability for amino acid type j. The first column contains the values for O(i,i), which represents the uniqueness of the chemical shifts of residue type i. Table 1 also lists the mean number of choices, $< N_i >$, for amino acid type i, defined as the sum $\sum_j O(i,j) \times j$, where the O(i,j) values are ordered according to $O(i,j) \geq O(i,j+1)$. The 95% number of choices, $N_{i,95}$, is also listed, representing the minimum number of probabilities, O(i,j), that must be summed up to yield $\sum_j O(i,j) \geq \sum 0.95$, and only those probabilities are shown in each row. As mentioned above, alanine, serine, and threonine can be identified uniquely with almost 100% probability. However, asparagine, aspartate, histidine^{+}, isoleucine, leucine, phenylalanine, proline, tyrosine, and valine also yield on average only two to three choices, although the N_{95} for a number of these residues can be as large as eight. Considering all amino acids except glycine, the average value for $<N_i>$ is 2.9 and the average value for $N_{i,95}$ is 6.4. The fact that glycine is identified uniquely by its C^{α} shift and by the absence of a C^{β} resonance improves the situation by slightly decreasing these averages.

Although at first sight the average values for $<N_i>$ and $N_{i,95}$ appear too large for making individual amino acid type assignments, unique type assignments frequently can be made if the resonances for a stretch of amino acids, linked by J connectivities, are compared with the primary sequence of the protein. Consider, for example, a protein of N amino acids and a given stretch of J-connected amino acids with length S. Then there are 20^{S} possible amino acid sequences for a stretch of this length. However, in the whole protein only $N - S + 1$ stretches of length S exist. Therefore, the probability of finding an arbitrary stretch in the protein is only $(N - S + 1)/20^{S}$. If for every amino acid in the stretch, the C^{α} and C^{β} shifts are known, and the mean number of choices for side chain types is C (C = 2.9 in Table 1), then there are roughly C^{S} different stretches of length S that are compatible with the NMR data. On average, out of these C^{S} possibilities, only

TABLE 1
OVERLAP INTEGRALS BETWEEN THE C^{α} AND C^{β} CHEMICAL SHIFT DISTRIBUTIONS FOR THE NATURAL AMINO ACIDS[a]

TYPE											<N>	N_{95}
	ALA											
ALA	99.91										1.00	1
	SER											
SER	99.78										1.00	1
	THR											
THR	99.78										1.00	1
	ASP	LEU	PHE	TYR	ASN							
ASP	37.22	19.51	16.22	15.46	8.24						2.51	5
	LEU	ASP	PHE	TYR	ILE							
LEU	41.75	21.88	15.19	14.01	5.15						2.26	5
	PHE	TYR	ILE	ASP	LEU							
PHE	28.48	28.28	15.65	12.41	10.36						2.67	5
	PRO	VAL	LYS	TRP	ILE							
PRO	44.94	41.57	3.20	2.77	2.58						2.02	5
	ILE	TYR	PHE	LEU	VAL	ASP						
ILE	41.13	23.02	22.61	5.07	3.06	2.41					2.24	6
	TYR	PHE	ILE	ASP	LEU	ASN						
TYR	28.63	28.43	16.02	11.89	9.61	2.67					2.67	6
	HIS^{+}	GLN	HIS	CYS	GLU	ARG	TRP					
HIS^{+}	33.59	18.82	12.35	11.75	9.46	7.37	3.23				2.98	7
	VAL	PRO	LYS	TRP	ILE	TYR	MET					
VAL	41.93	38.78	4.99	3.28	3.12	1.92	1.72				2.27	7
	ARG	GLU	TRP	HIS	CYS	GLN	MET	LYS				
ARG	16.40	15.47	12.67	12.63	12.53	9.46	8.90	7.37			4.20	8
	ASN	ASP	MET	LYS	TYR	PHE	LEU	HIS				
ASN	55.94	12.38	7.13	6.57	5.21	4.95	2.34	1.50			2.57	8
	TRP	ARG	GLU	CYS	LYS	MET	HIS	GLN				
TRP	18.81	14.53	13.63	11.40	10.91	9.92	9.33	6.51			4.16	8
	GLN	CYS	GLU	ARG	HIS^{+}	HIS	TRP	MET				
GLN	20.10	17.11	14.39	11.59	11.26	11.15	6.96	4.26			3.79	8
	CYS	GLU	GLN	ARG	TRP	HIS	HIS^{+}	MET	LYS			
CYS	17.64	15.92	15.02	13.47	10.69	10.32	6.17	5.49	4.72		3.99	9
	GLU	ARG	CYS	TRP	HIS	GLN	MET	LYS	HIS^{+}			
GLU	16.35	15.42	14.76	11.85	11.77	11.71	7.05	5.79	4.60		4.13	9
	HIS	ARG	GLU	CYS	GLN	TRP	MET	HIS^{+}	LYS			
HIS	19.04	14.66	13.70	11.14	10.56	9.45	8.32	7.00	5.43		4.14	9
	MET	LYS	ARG	TRP	HIS	GLU	CYS	GLN	ASN			
MET	21.73	18.75	11.79	11.46	9.50	9.37	6.77	4.61	2.77		3.90	9
	LYS	MET	TRP	ARG	GLU	HIS	CYS	GLN	VAL	ASN		
LYS	23.11	19.94	13.40	10.38	8.18	6.59	6.18	3.41	2.75	2.71	3.88	10

[a] Probabilities are listed in % that, given the C^{α} and C^{β} frequency distributions of a certain amino acid type i, one could also assign to it an amino acid type j. The entries in the rows have been ordered according to decreasing probabilities, and only the most likely are listed, so that the sum of their probabilities is at least 95%. Also listed is the mean number, <N>, and the 95% number, N_{95}, of the choices as defined in the text.

a fraction $(N - S + 1)/20^S$ is compatible with the protein primary sequence. We define this fraction as:

$$F = C^S(N - S + 1)/20^S \tag{18}$$

Whenever F drops below 1, on average there will be only a single choice to position this stretch in the protein primary sequence. A critical stretch length S_C can be defined, such that $F(S_C) = 1$. Usually, the condition $N \gg S_C - 1$ applies, and S_C can be approximated by:

$$S_C \approx \ln(N)/\ln(20/C) \tag{19}$$

For a typical case, such as interferon-γ (Grzesiek et al., 1992) with N = 134, S_C is approximately 2.5, and even for the hypothetical case of N = 1000, S_C only rises to a value of 3.6. Therefore one expects that stretches of more than two or three amino acids with given C^α and C^β frequencies can often be located uniquely in the primary sequence. This is illustrated for the stretch comprising residues Asn[11], Leu[12], Lys[13], Lys[14] in interferon-γ (labeled I, II, III, IV in Table 2), where the C^α and C^β frequencies are known, and the interresidue connections between the spin systems are established by J-coupled backbone assignment experiments. The individual probabilities $p^i(\delta_{C\alpha},\delta_{C\beta})$ for the C^α/C^β combinations to correspond to amino acid type i are listed in Table 2. For each of the four residues, $N_{i,95} \geq 5$; i.e., the chemical shifts are not particularly unique. However, if the individual probabilities are combined in a stretch, the situation improves dramatically. We define the probability $p_S(k)$ that a stretch of sequentially connected C^α and C^β frequency pairs $(\delta_{C\alpha}(i),\delta_{C\beta}(i))$ of length S, with i = 1, 2, ..., S, begins at position k in the primary sequence by:

$$p_S(k) = N_{PS} \times \prod_{i=1}^{S} p^{aa(k+i-1)}(\delta_{C\alpha}(i),\delta_{C\beta}(i)) \tag{20}$$

where aa(n) gives the type of the n-th amino acid in the primary sequence and the normalization constant, N_{PS}, is chosen such that $\sum_{k=1}^{N-S+1} p_S(k) = 1$. The five most probable locations (if their probability exceeds 1%) for stretches of length 2, 3, and 4 comprising the J-linked spin systems of the amino acids I-IV, listed above, are presented in Table 3. Clearly, stretches of length two still yield several locations with significant probabilities. For example, the two-residue fragment comprising residues II and III corresponds most likely to a Leu-Lys dipeptide for which there are

TABLE 2
AMINO ACID TYPE PROBABILITIES FOR THE C^α AND C^β CHEMICAL SHIFT PAIRS OF RESIDUES 11–14[a] OF INTERFERON-γ

Spin system		$\delta_{C\alpha}$	$\delta_{C\beta}$	ARG[b]	ASN	ASP	ILE	LEU	LYS	MET	PHE	PRO	TRP	TYR	VAL
I	(Asn[11])	56.40	37.98	0.0	53.7	3.9	1.3	0.9	2.4	0.1	18.0	0.0	0.2	19.0	0.4
II	(Leu[12])	58.25	41.94	0.0	0.0	10.2	11.6	56.0	0.0	0.0	11.3	0.0	0.0	11.0	0.0
III	(Lys[13])	60.74	32.51	0.8	0.0	0.0	0.1	0.0	28.1	5.9	0.0	25.3	4.0	0.0	35.3
IV	(Lys[14])	59.38	32.35	3.7	0.0	0.0	0.0	0.0	43.6	28.4	0.0	4.0	5.8	0.0	9.0

[a] Labeled as spin systems I, II, III, and IV.
[b] Only amino acid types that have a probability of at least 2% for one of the four spin systems are listed.

two occurrences in the amino acid sequence, and with lower probability to Ile-Val, Tyr-Val, Asp-Val and a number of other (not listed) dipeptides. If the stretch I-II-III is considered (Table 3), the unique Asn^{11}-Leu^{12}-Lys^{13} tripeptide is identified with a probability of 77%. When the stretch of all four J-linked residues is considered, the only stretch of residues compatible with the C^{α}/C^{β} shifts of residues I-IV is the tetrapeptide Asn^{11}-Leu^{12}-Lys^{13}-Lys^{14}, with a probability of 99.98%. The computer program which calculates the probabilities in the manner described above is available at no cost via electronic mail (see below). In the example discussed above, the C^{α} and C^{β} frequencies fall in the most crowded part of the C^{α} and C^{β} plane, and therefore the number of possible amino acid type choices is higher than average. For many other stretches in interferon-γ with more unique C^{α} and C^{β} frequencies, the location of the stretch becomes unique to more than 95% for a stretch length of 3 or less.

As mentioned above, a degeneracy in linking adjacent residues via J coupling occurs if the amide $^1H/^{15}N$ or C^{α}/C^{β} frequencies are not unique. Even in these cases, the procedure described above is applicable in a straightforward manner. If, for example, there is D-fold degeneracy of each $^1H/^{15}N$ chemical shift pair, there are $C^S D^{S-1}$ instead of C^S stretches with defined amino acid type possible starting from the first amino acid spin system. In this case, Eq. 19 for the critical stretch length has to be replaced by:

$$S_C = \ln(N/D)/\ln(20/(CD)) \tag{21}$$

TABLE 3
PROBABILITY OF PRIMARY SEQUENCE POSITIONS FOR STRETCHES OF J-CONNECTED RESIDUES, LISTED IN TABLE 2[a]

Strands of length 2:

I + II		II + III		III +IV	
43.8	Asn^{11} –Leu^{12}	26.3	Leu^{12}–Lys^{13}	17.8	Val^{80} –Lys^{81}
14.7	Phe^{30} –Leu^{31}	26.3	Leu^{34}–Lys^{35}	17.8	Val^{6} –Lys^{7}
8.8	Asn^{60} –Phe^{61}	6.8	Ile^{50} –Val^{51}	14.2	Lys^{13} –Lys^{14}
8.6	Asn^{98} –Tyr^{99}	6.5	Tyr^{5} –Val^{6}	14.2	Lys^{87} –Lys^{88}
3.2	Asp^{103} –Leu^{104}	6.0	Asp^{22}–Val^{23}	14.2	Lys^{88} –Lys^{89}

Strands of length 3:

I + II + III		II + III + IV	
76.7	Asn^{11}–Leu^{12}–Lys^{13}	79.2	Leu^{12}–Lys^{13}–Lys^{14}
15.5	Asn^{60}–Phe^{61}–Lys^{62}	19.5	Tyr^{5}–Val^{6}–Lys^{7}
5.5	Tyr^{54}–Phe^{55}–Lys^{56}		

Strands of length 4:

I + II + III + IV	
100.0	Asn^{11}–Leu^{12}–Lys^{13}–Lys^{14}

[a] The probability for locating stretches of length 2, 3, and 4, which can be generated from spin systems I–IV of Table 2, in the primary sequence of interferon-γ are calculated according to Eq. 20. The positions in the primary sequence have been ordered according to decreasing probabilities. The five most probable positions are listed, if their probability exceeds 1%.

As this expression depends only logarithmically on D, for a case of two-fold degeneracy in every amino acid connection, the critical stretch length for the example of interferon-γ becomes only 3.4. Therefore even with extensive degeneracy, the procedure is able to sort out the correct stretch among all the degenerate ones and to assign to it the correct position in the primary sequence, provided the stretch is longer than, on average, 3.4 amino acids.

The discussion presented above assumes that all 20 amino acid types are present with comparable abundance in the protein primary sequence. The fact that, in practice, some types of amino acids are much more abundant than others does not dramatically affect the value of the critical stretch length, S_C (Eq. 19). This can be qualitatively understood by considering an example where only 10 amino acid types are present in the protein primary sequence. In this case, the factor 20 in Eq. 19 must be replaced by 10, increasing S_C by ~56%. However, with fewer types of amino acids available, the average number of choices, C in Eq. 19, also decreases. This causes the effective increase in S_C to be much less than 56%.

CONCLUSIONS

The experiments and procedures described in this paper provide a rapid approach for making complete backbone and C^β/H^β assignments in proteins uniformly enriched with ^{13}C and ^{15}N. The ability to J correlate backbone amides to the H^β/C^β resonances assists with residue type identification and is a useful step in the sequential assignment process. Simple edited CT-HSQC experiments allow identification of aromatic residues and Asn and Asp residues. Once the probability that a J-linked spin system corresponds to the different amino acid types has been calculated, a simple computer program can be used to calculate the probability of locating a stretch of J-linked spin systems in the primary sequence of the protein. The development of these NMR methods and analysis procedures provides an important step towards fully automating the resonance assignment process in isotopically uniformly enriched proteins.

ACKNOWLEDGEMENTS

We thank Dennis Torchia and Jeffrey Pelton for making the assignments of SNase and III^{Glc} available to us in a computer-readable form, and Frank Delaglio, Dennis Torchia, David Live, and Jacob Anglister for valuable discussions. This work was supported by the AIDS Targeted Anti-Viral Program of the Office of the Director of the National Institutes of Health.

APPENDIX

Software available

Source code (in C) for two algorithms described above can be obtained, using the FTP protocol, from our program data base. The first program calculates the probability for a C^α/C^β chemical shift pair to correspond to a given amino acid type (Table 2). With a given stretch of J-linked C^α/C^β pairs and the primary sequence, the second program calculates the probabilities of all possible positions in the primary sequence for this stretch (Table 3).

REFERENCES

Bax, A, Clore, G.M., Driscoll, P.C., Gronenborn, A.M., Ikura, M. and Kay, L.E. (1990a) *J. Magn. Reson.*, **87**, 620–627.

Bax, A, Clore, G.M. and Gronenborn, A.M. (1990b) *J. Magn. Reson.*, **88**, 425–431.

Bax, A. and Grzesiek, S. (1993) *Acc. Chem. Res.*, in press.

Boucher, W., Laue, E.D., Campbell-Burk, S. and Domaille, P.J. (1992) *J. Am. Chem. Soc.*, **114**, 2262–2264.

Burum, D.P. and Ernst, R.R. (1980) *J. Magn. Reson.*, **39**, 163–168.

Clore, G.M., Bax, A., Driscoll, P.C., Wingfield, P.T. and Gronenborn, A.M. (1990) *Biochemistry*, **29**, 8172–8184.

Clubb, R.T., Thanabal, V. and Wagner, G. (1992) *J. Magn. Reson.*, **97**, 213–217.

Ernst, R.R., Bodenhausen, G. and Wokaun, A. (1987) Principles of Magnetic Resonance in One and Two Dimensions, Claredon Press, Oxford pp. 25–29.

Fesik, S.W., Eaton, H.L., Olejniczak, E.T., Zuiderweg, E.R.P., McIntosh, L.P. and Dahlquist, F.W. (1990) *J. Am. Chem. Soc.*, **112**, 886–888.

Grzesiek, S. and Bax, A. (1992a) *J. Magn. Reson.*, **96**, 432–440.

Grzesiek, S. and Bax, A. (1992b) *J. Magn. Reson.*, **99**, 201–207.

Grzesiek, S. and Bax, A. (1992c) *J. Am. Chem. Soc.*, **114**, 6291–6293.

Grzesiek, S., Döbeli, H., Gentz, R., Garotta, G., Labhardt, A.M. and Bax, A. (1992) *Biochemistry*, **31**, 8180–8190.

Hansen, P.E. (1991) *Biochemistry*, **30**, 10457–10466.

Howarth, O.W. and Lilley, D.M.J. (1978) *Prog. NMR Spectrosc.*, **12**, 1–40.

Ikura, M., Kay, L.E. and Bax, A. (1990) *Biochemistry*, **29**, 4659–4667.

Ikura, M., Kay, L.E., Krinks, M. and Bax, A. (1991) *Biochemistry*, **30**, 5498–5504.

Ikura, M., Clore, G.M., Gronenborn, A.M., Zhu, G., Klee, C.B. and Bax, A. (1992) *Science* **256**, 632–638.

Kay, L.E., Ikura, M., Tschudin, R. and Bax, A. (1990a) *J. Magn. Reson.*, **89**, 496–514.

Kay, L.E., Clore, G.M., Bax, A. and Gronenborn, A.M. (1990b) *Science*, **249**, 411–414.

Kay, L.E., Ikura, M. and Bax, A. (1990c) *J. Am. Chem. Soc.*, **112**, 888–889.

Kay, L.E., Wittekind, M., McCoy, M.A., Friedrichs, M.S. and Mueller, L. (1992) *J. Magn. Reson.*, **98**, 443–450.

Marion, D., Driscoll, P.C., Kay, L.E., Wingfield, P.T., Bax, A., Gronenborn, A.M. and Clore, G.M. (1989a) *Biochemistry*, **28**, 6150–6156.

Marion, D., Kay, L.E., Sparks, S.W., Torchia, D.A. and Bax, A. (1989b) *J. Am. Chem. Soc.*, **111**, 1515–1517.

Marion, D., Ikura, M., Tschudin, R. and Bax, A. (1989c) *J. Mag. Res.*, **85**, 393–399.

McCoy, M.A. and Mueller, L. (1992) *J. Magn. Reson.*, **99**, 18–36.

Meador, W.E., Means, A.R. and Quiocho, F.A. (1992) *Science*, **257**, 1251–1255.

Oh, B.H., Westler, W.M., Darba, P. and Markley, J.L. (1988) *Science*, **240**, 908–911.

Palmer III, A.G., Fairbrother, W.J., Cavanagh, J., Wright, P.E. and Rance, M. (1992) *J. Biomol. NMR*, **2**, 103–108.

Pelton, J.G., Torchia, D.A., Meadow, N.D., Wong, C.-Y. and Roseman, S. (1991) *Biochemistry*, **30**, 10043–10057.

Powers, R., Gronenborn, A.M., Clore, G.M. and Bax, A. (1991) *J. Magn. Reson.*, **94**, 209–213.

Santoro, J. and King, G.C. (1992) *J. Magn. Reson.*, **97**, 202–207.

Shaka, A.J., Lee, C.J. and Pines, A. (1988) *J. Magn. Reson.*, **77**, 274–293.

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

van de Ven, F.J.M. and Philippens, M.E.P. (1992) *J. Magn. Reson.*, **97**, 637–644.

Vuister, G.W. and Bax, A. (1992) *J. Magn. Reson.*, **98**, 428–435.

Wagner, G. and Brühweiler, D. (1986) *Biochemistry*, **25**, 5839–5843.

Wagner, G., Schneider, P. and Thanabal, V. (1991) *J. Magn. Reson.*, **93**, 436–440.

Wishart, D.S., Sykes, B.D. and Richards, F.M. (1991) *J. Mol. Biol.*, **222**, 311–333.

Wittekind, M., Görlach, M., Friedrichs, M., Dreyfuss, G. and Mueller, L. (1992) *Biochemistry*, **31**, 6254–6265.

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

Zhu, G. and Bax, A. (1990) *J. Magn. Reson.*, **90**, 405–410.