Journal of Magnetic Resonance 195 (2008) 187–195

Contents lists available at ScienceDirect

Journal of Magnetic Resonance

journal homepage: www.elsevier.com/locate/jmr

Amino acid type identification in NMR spectra of proteins via β- and γ-carbon edited experiments

David Pantoja-Uceda, Jorge Santoro *

Instituto de Química Física Rocasolano, CSIC, Serrano 119, 28006 Madrid, Spain

ARTICLE INFO

Article history:
Received 22 July 2008
Revised 8 September 2008
Available online 17 September 2008

Keywords:
Proteins
Amino acid type editing
Assignment
Triple-resonance experiments
CBCA(CO)NH
CBCANH

ABSTRACT

In this work, we introduce a set of pulse sequences that provide amino acid type identification of the NH correlation signals of proteins. The first pulse sequence is a modification of the CBCA(CO)NH experiment that exploits spin-coupling topologies to differentiate between amino acid types. A set of eight 2D $^{1}H–^{15}N$ correlation spectra is recorded where the sign of the cross-peaks change from one spectrum to another according to the amino acid type of the preceding residue in the protein sequence. Linear combination of these eight data sets produces four subspectra. Taking also into account the sign of the correlation signals, this method allows the classification of the NH signals into six different groups, depending on the character of the preceding residue. This sequence is complemented with a (CGCBCACO)NH experiment that allows the subdivision of the largest of these groups into two smaller ones. Finally, a modification of the CBCANH experiment led to a similar classification of NH signals into six different groups, but now depending on the type of its own amino acid. The set of pulse sequences is demonstrated with two proteins of small to moderate size.

1. Introduction

For the investigation of proteins by NMR a sequence-specific assignment of the resonances is a prerequisite. The usual strategy consists on the acquisition of triple resonance experiments [1] that correlate each ^{1}H, ^{15}N frequency pair with Cα, Cβ or other frequencies of the same and the sequential amino acid and thus form a chain of spin systems that can be matched to the amino acid sequence of the protein. This powerful strategy is not devoid of problems, like degeneracy in ^{1}H and ^{15}N chemical shifts, overlapping of Cα or Cβ frequencies in the CBCA(CO)NH and CBCANH spectra, missing ^{13}C signals, total absence of some spin systems due to proton exchange or unfavorable dynamics, etc. Under these circumstances methods that add in extra data can be of great aid in the assignment process.

One way to improve the assignment of protein backbone resonances, either manually or by automatic methods is to incorporate information about amino acids types. For some amino acids, G, A, and the pair S/T, the combination of their $^{13}Cα$ and $^{13}Cβ$ chemical shifts can be used for identification [2]. However, for the rest of amino acids chemical shifts are not unique, and categorization cannot be made on this basis. Two different approaches have been used to classify the NMR signals into amino acid types. The first method relies on the selective incorporation of labeled amino acids into the protein [3–5]. The second approach employs pulse sequences that made use of the homo- and heteronuclear spin–spin coupling networks in the individual amino acids to select the amino acid types [6–15]. The biggest advantage of this second method is that only one doubly-labeled sample is needed. However, this approach is not absent of drawbacks. Some of the proposed strategies require the registering of many spectra and become very time-consuming. A second problem is the appearance of breakthrough peaks, what can cause problems, especially in automated assignment methods. Finally, some of the filters incorporated into the pulse sequences are long and the sequences do not work well for large proteins.

The 20 protein amino acid side chains form eight topology classes with respect to the number of hydrogen atoms attached to Cβ carbons and to the number and type of carbons at the γ position. Therefore, the Cβ carbons are ideal as starting point in pulse sequences for amino acid type identification. Recently, an experiment called HADAMAC, for HADamard-encoded AMino-ACid-type-editing, based on the CBCA(CO)NH experiment and exploiting the topology of the Cβ carbons has been proposed [15]. This experiment led to the classification of the $^{1}H–^{15}N$ HSQC signals into seven classes depending on the preceding amino acid type. In this paper, we present a set of experiments for amino acid typing based on similar principles. A first pulse sequence, having many points in common with HADAMAC, yields the classification of the $^{1}H–^{15}N$ correlation signals into six different classes, depending on the nature of the preceding amino acid. Main differences of our sequence from HADAMAC are the handling of signals originating from Hα magnetization and the distinction of aromatic side chains from res-

* Corresponding author. Fax: +34 91 564 24 31.
E-mail address: jsantoro@iqfr.csic.es (J. Santoro).

1090-7807/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmr.2008.09.010

idues lacking a carbon at the γ position. A second pulse sequence allows subdividing the most populated class, including seven amino acids types and coinciding with one of the HADAMAC classes, into two subclasses of 3 and 4 amino acids. Combined use of both experiments provides a classification of the signals into seven classes, having from 1 to 4 members. Finally, another pulse sequence allows a classification of the $^{1}H–^{15}N$ signals depending on the nature of its own amino acid. Altogether, this set of pulse sequences provides a very complete classification of the $^{1}H–^{15}N$ HSQC signals, and consequently of the spin systems associated with them, and can be of great aid in the sequence-specific assignment process.

2. Pulse sequences

The pulse sequence for β carbon edited (CBCACO)NH, which is shown in Fig. 1A, is based on the CBCA(CO)NH experiment [16]. In the regular experiment, the observed magnetization follows two different pathways: Hβ → Cβ → Cα → CO → N → H and

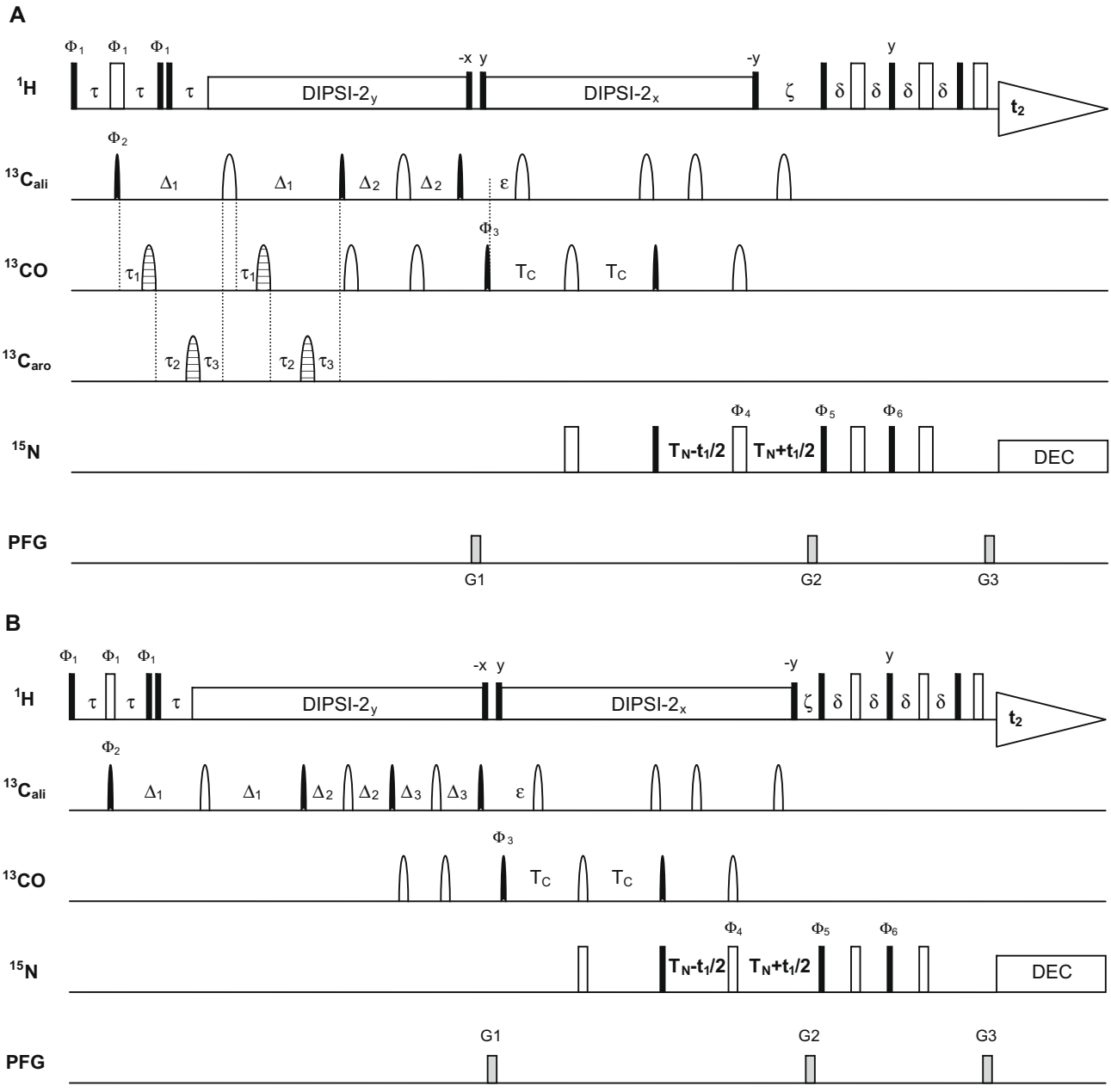

Fig. 1. Pulse sequences for sequential carbon edited HSQC experiments. All radiofrequency pulses are applied along the *x*-axis unless indicated. 90° and 180° rectangular pulses are represented by filled and unfilled bars, respectively. ^{13}C pulses have the shape of Gaussian cascades Q5 (black filled shapes) and Q3 (open shapes) with durations of 307 and 192 μs at 800 MHz, respectively. ^{13}C aliphatic pulses are centered at 41 ppm and CO pulses at 173 ppm. Pulsed field gradients G_1 to G_3 of sinusoidal shape are applied along the *z*-axis with a 1 ms length and amplitudes of 30%, 80% and 8.1% from the maximal intensity. For quadrature detection echo/antiecho data are recorded by sign alternation of the gradient G_2 with simultaneous inversion of phase ϕ_5 [17]. (A) Sequential β-carbon edited HSQC. Striped shapes correspond to Q3 Gaussian cascades of 340 μs duration (800 MHz), applied at 120 ppm ($^{13}C_{aro}$) and 180 ppm (^{13}CO). The delays are adjusted to $\Delta_1 = 9.5–10.5$ ms; $\Delta_2 = 4.5$ ms; $\tau = 3.7$ ms; $\zeta = 5.5$ ms; $\delta = 2.3$ ms; $\varepsilon = 4.4$ ms; $T_C = 11.5$ ms; $T_N = 12.4$ ms. For every t_1 value 8 repetitions of the experiment are separately recorded with the parameter settings given in Table 1 and $\phi_3 = \phi_4 = x$. The two-step phase cycle of Table 1 can be augmented by adding two further steps with phases ϕ_3, ϕ_4, and ϕ_{rec} inverted. (B) Sequential γ-carbon edited HSQC. The delays used are $\Delta_1 = 9.4$ ms; $\Delta_2 = 7.15$ ms; $\Delta_3 = 4.9$ ms; $\tau = 3.7$ ms; $\zeta = 5.5$ ms; $\delta = 2.3$ ms; $\varepsilon = 4.4$ ms; $T_C = 12.4$ ms; $T_N = 12.4$ ms. The phase cycling is $\phi_1 = 45°, 45°, 135°, 135°, 225°, 225°, 315°, 315°$; $\phi_2 = 8x, 8(-x)$; $\phi_3 = x$; $\phi_4 = x$; $\phi_5 = x, -x$; $\phi_6 = -y, y$; $\phi_{rec} = 2(x, -x, -x, x), 2(-x, x, x, -x)$. This minimal phase cycling can be extended by adding 16 further steps with phases ϕ_3, ϕ_4, and ϕ_{rec} inverted.

Hα → Cα → Cα → CO → N → H. In edited experiments, in order to obtain a clean amino-acid-typing based on the characteristics of the β carbon, the second magnetization pathway is usually suppressed, or at least strongly reduced in intensity. Several methods have been used for this purpose. The first one is to evolve the $^{13}C\alpha$–^{13}CO coupling for a time $1/2J_{C\alpha-CO}$ during the $2\Delta_1$ evolution period [6]. This method, however, also suppress the β carbons of D and N. A second possibility is to use an evolution time of $2\Delta_2 = 1/(2J_{C\alpha-C\beta})$ [15]. This approach lengthens the duration of the pulse sequence in about 5 ms, thereby producing a loss in sensitivity. Finally, the initial refocused INEPT can be replaced by a MUSIC coherence transfer scheme [10,19] selective for CH_2. This approach will suppress completely the undesired pathway for all amino acids except glycine. Unfortunately, the sequence also suppress Cβ carbons of T, A, V, and I. Our approach is different to all these methods. Instead of suppressing the undesired pathway, we retain it, but arranging the pulse sequence in such a way that it generates its own subspectrum. To this end, the pulse sequence starts with a MUSIC-like coherence transfer, but storing separately the scans in which all carbons appear with the same sign and those in which the CH and CH_3 carbons appear inverted relative to the CH_2 carbons. The sum of the two data sets restores the MUSIC-CH_2 filter and selects the Cα carbon of G and the Cβ carbons of all residues, except T, A, V, and I. On the contrary, the difference preserves the Cα carbons of all residues, except G, and the Cβ carbons of T, A, V, and I. The separation of this group into two subspectra, one for the Cα carbons and other for the Cβ ones, will be described below. At the end of the delay $2\Delta_1$, omitting for the moment the effect of the selective CO and aromatic carbon pulses applied during this period and considering only the Cβ magnetization, the product operator term that will be transferred to Cα magnetization has an intensity of

$$2B_yA_z \sin(2\pi\Delta_1 J_{C\alpha-C\beta}) \cos^{nG}(2\pi\Delta_1 J_{C\beta-C\gamma}) \quad (1)$$

where nG is the number of aliphatic γ carbons attached to the β carbon. Therefore, a value of $\Delta_1 \approx 3/(8J_{C-C})$ will yield a sign inversion for all amino acid types with a single aliphatic γ carbon (T, K, R, P, L, E, Q, and M) with respect to the rest of amino acids, while providing high transfer efficiencies for all residues. The rest of the sequence is identical to the regular CBCA(CO)NH: the magnetization is transferred to the Cα; from there it is relayed via the carbonyl carbon to the nitrogen of the following residue, N(i + 1), and finally detected on the H(i + 1) amide proton. Therefore, the sign of the ^1H–^{15}N correlation peak, negative for amino acid types with a single aliphatic γ carbon and positive for the rest, will reveal the nature of the preceding residue in the protein sequence.

The group of amino acids with an even number of aliphatic Cγ carbons can be further subdivided by taking into account the nature of the gamma substituent. For instance, by insertion in the $2\Delta_1$ period of selective 180° pulses affecting only the CO carbons it is possible to evolve the $J_{C-C'}$ couplings. Therefore, for D and N residues the intensity of the Cβ magnetization antiphase with respect to Cα that is present at the end of the $2\Delta_1$ period in the pulse sequence will be

$$\sin(2\pi\Delta_1 J_{C\alpha-C\beta}) \cos(2\pi T_X J_{C-C'}) \quad (2)$$

where $2T_X$ is the effective evolution time of the $J_{C-C'}$ coupling, $2 * (\tau 3 + \tau 2 + p - \tau 1)$ with the arrangement of Fig. 1A. For $T_X \approx 0$ the cosine term will be positive and approximately equal to 1, while for $T_X \approx 1/J_{C-C'}$ it will be negative and close to −1. Cβ magnetization of all other residues is not affected by the selective pulses, and its behavior will remain the one described above, independently of the position of the selective pulses. Therefore, if two experiments are acquired, one with $T_X \approx 0$ and a second one with $T_X \approx \Delta_1 \approx 1/J_{C-C'}$, its difference will show signals only for the D and N residues. In contrast, the sum of the two datasets will show the signals of all residues not affected by the selective pulses. Small signals of D and N residues can also be present in this second subspectrum if T_X is not exactly $1/J_{C-C'}$. Nevertheless, these breakthrough peaks are easily recognized by comparing the intensity in both subspectra, and therefore do not hinder a correct amino acid classification. Since Cα coherence is evolving during the delay Δ_1 under the Cα–C′ scalar coupling, it is affected by the CO selective inversion pulses in a similar way as described above for the Cβ magnetization of D and N. Consequently the G residues will appear in the same subspectrum as D and N. Furthermore, this method will split the subspectrum of CH/CH_3 carbons into two subspectra, one originated by the Cβ carbons of T, A, V, and I, that are not coupled to a carbonyl carbon, and other one by the Cα carbons of all residues except G. Hence, although magnetization starting at Hα is not suppressed it will not disturb the classification of the NH signals into amino acid types. The same kind of selection described for carbons bounded to a CO can be made to differentiate the amino acids with an aromatic carbon at the γ position: F, Y, W, and H. Combining all selection schemes, the eight experiments described in Table 1 are obtained. By linear combination of the eight data sets, four amino-acid-type edited ^1H–^{15}N spectra are obtained: one corresponding to the aromatic residues, other to the D, N, and G residues, a third one including all other amino acids with a CH_2 β-carbon, and finally, one including the T, A, V, and I residues. The T, A, V, and I residues are not separated in different subspectra by this

Table 1
Delays and pulse phases used in the β-carbon edited experiments of Figs. 1A and 3

	Parameters								Encoding				
	$\tau 1$	$\tau 2$	$\tau 3$	ϕ_1	ϕ_2	ϕ_5	ϕ_6	ϕ_{Rec}	FYWH	DNG	CS, (*Long*[a])	AVI, (T)	Cα[b]
1	$\Delta 1$-C1[c]-p[d]	C1 + C2[e] − $\Delta 1$	$\Delta 1$-C2-p	45°, 315°	x	x	$-y$	+, −	+	+	+, (−)	+, (−)	−
2	$\Delta 1$-C1-p	C1+C2-$\Delta 1$	$\Delta 1$-C2-p	225°, 135°	x	x	$-y$	+, −	+	+	+, (−)	−, (+)	+
3	$\Delta 1$-C1-p	C1-p-4 μs	4 μs	45°, 315°	$-x$	x	$-y$	+, −	+	−	−, (+)	−, (+)	+
4	$\Delta 1$-C1-p	C1-p-4 μs	4 μs	225°, 135°	$-x$	x	$-y$	+, −	+	−	−, (+)	+, (−)	−
5	4 μs	C2-p-4 μs	$\Delta 1$-C2-p	45°, 315°	x	$-x$	y	+, −	−	+	−, (+)	−, (+)	−
6	4 μs	C2-p-4 μs	$\Delta 1$-C2-p	225°, 135°	x	$-x$	y	+, −	−	+	−, (+)	+, (−)	+
7	4 μs	$\Delta 1 - 2 * (p + 4\ \mu s)$	4 μs	45°, 315°	$-x$	$-x$	y	+, −	−	−	+, (−)	+, (−)	+
8	4 μs	$\Delta 1 - 2 * (p + 4\ \mu s)$	4 μs	225°, 135°	$-x$	$-x$	y	+, −	−	−	+, (−)	−, (+)	−

Eight repetitions of the experiments are performed using the parameter settings given in this table. The subspectrum corresponding to a particular amino acid type is obtained by linear combination of the eight data sets with the relative sign taken from the encoding column.

[a] The *Long* group includes E, Q, M, K, R, P, and L residues.

[b] In the CBCANH experiment the signs of the Cα column are reversed.

[c] C1 is $1/4J_{C\gamma-CO}$ = 5.32 ms [18].

[d] p is the duration of the 180° selective pulse.

[e] C2 is $1/4J_{C\gamma-Caro}$ = 5.68 ms [18].

scheme. However, since they differ by the presence of an odd, T, or even, A, V, and I, number of aliphatic carbons at the Cγ position, they appear with opposite signs. Similarly, the *Long* group of residues appears in the same subspectrum as C and S, but with opposite sign. A fifth subspectrum, corresponding to the pathway starting at Hα, can be obtained. This Cα subspectrum is of minor interest for the classification in amino acid types, since it will show signals for all amino acids except glycine. However, it can be useful to determine the phase corrections for the other subspectra, to verify the assignments to glycine, and to classify the NH signals of the side chains.

The NH_2 moieties of Asn and Gln side chains should not give signals in the spectrum, since the evolution time of the J_{NH} coupling, ζ, is adjusted to maximize the ^{15}N to ^{1}H transfer for a NH group, thereby abolishing the transfer for moieties of other multiplicity. However, proteins are usually dissolved in H_2O with a 5–10% of D_2O, so that the NHD groups of Asn and Gln side chains will be visible. For both side chains there are two pathways ending into observable magnetization: one starts at the carbon contiguous to the CO group and another starts one carbon further away from the CO. The carbon responsible of the first pathway is topologically identical in both side chains, a CH_2 group bounded to a CO and to an aliphatic carbon. Therefore, both side chains will give a signal in the DNG subspectrum that, according to the modulation with the J_{C-C} coupling during the $2\Delta_1$ period, will be negative. The second pathway, however, is different for both side chain types, and allows its classification. In the case of Asn, it starts at a CH carbon bounded to a CO group. Therefore, Asn side chains will give signals in the Cα subspectrum, with a sign opposite to that of the backbone NH's. For Gln, the starting carbon is a CH_2 neither bounded to CO nor to aromatic carbons. Consequently, signals of Gln side chains will appear in the CS + *Long* subspectrum with negative sign.

Amino acids belonging to the *Long* group can be further classified by using a (CGCBCACO)NH experiment. This experiment, shown in Fig. 1B, is similar to the (CBCACO)NH one, but extended by a relay step. Several pathways end in observable magnetization and have intensities:

$$\cos(2\pi J_{C-C}\Delta_1)\cos(2\pi J_{C-C}\Delta_2)\cos(2\pi J_{C-C}\Delta_3) \tag{3a}$$

$$\sin(2\pi J_{C-C}\Delta_1)\cos(2\pi J_{C-C}\Delta_2)\cos^{nG}(2\pi J_{C-C}\Delta_2)\sin(2\pi J_{C-C}\Delta_3) \tag{3b}$$

$$\cos(2\pi J_{C-C}\Delta_1)\cos^{nG}(2\pi J_{C-C}\Delta_1)\sin(2\pi J_{C-C}\Delta_2) \times \cos^{nG}(2\pi J_{C-C}\Delta_2)\sin(2\pi J_{C-C}\Delta_3) \tag{3c}$$

$$\sin(2\pi J_{C-C}\Delta_1)\cos^{nG}(2\pi J_{C-C}\Delta_1)\sin(2\pi J_{C-C}\Delta_2)\cos(2\pi J_{C-C}\Delta_3) \tag{3d}$$

$$\sin(2\pi J_{C-C}\Delta_1)\cos^{nD}(2\pi J_{C-C}\Delta_1)\sin^2(2\pi J_{C-C}\Delta_2) \times \cos^{nG-1}(2\pi J_{C-C}\Delta_2)\sin(2\pi J_{C-C}\Delta_3) \tag{3e}$$

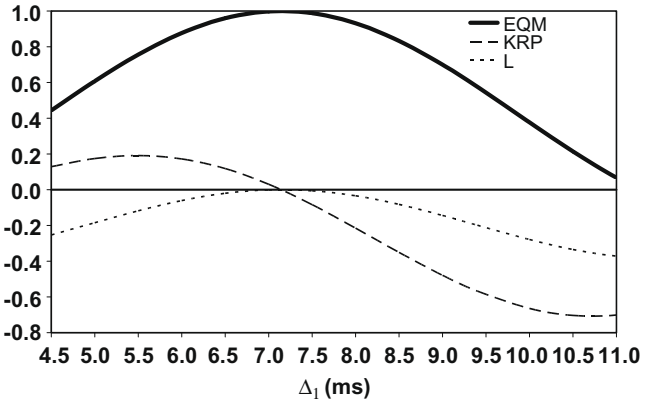

Fig. 2. Transfer amplitudes as function of the delay Δ_1 for amino acids with one gamma carbon in the γ-carbon edited experiment of Fig. 1B. Parameters used: J_{CC} = 35 Hz, Δ_2 = 7.15 ms; Δ_3 = 14.3 ms − Δ_1.

Terms (3a) and (3b) correspond to magnetization of Cα during the $2\Delta_1$ period, terms (3c) and (3d) to magnetization of Cβ, and term (3e) to magnetization of Cγ. Of these five terms only the last one can give information on the nature of the Cγ carbon, so that the pulse sequence should eliminate or reduce the other four. Terms corresponding to Cα magnetization are eliminated, except for glycine residues, by starting the pulse sequence with a MUSIC-CH_2 coherence transfer [10]. The sum of terms (3c) and (3d) can be eliminated for amino acids lacking a Cγ carbon by tuning the delays so that $\Delta_1 + \Delta_3 = 1/(2J_{C-C})$. Finally, to maximize term (3e) for residues with one Cγ carbon a value of $\Delta_2 \approx 1/4J_{C-C}$ should be used. Under these conditions, the intensities expected for the different amino acids of the *Long* group as a function of Δ_1 is shown in Fig. 2. As can be seen, for values of Δ_1 ranging from 8.5 to 10.5 ms signals are of sufficient intensity and can be subdivided in two groups according to their sign, positive for E, Q, and M, and negative for K, R, P, and L. G will also appear in this spectrum, giving a strong and positive signal. The rest of amino acids should give no signal. However, deviations of the carbon–carbon scalar coupling constants from the standard value of 35 Hz can lead to positive or negative breakthrough signals. These breakthrough signals do not hinder the correct classification of residues, since only signals catalogued as *Long* or as DNG in the β-carbon edited experiment should be analyzed here.

The same kind of β-carbon edition described above can be used to modify the CBCANH experiment [20], yielding to the sequence shown in Fig. 3. The analysis of the results of this experiment demands some considerations. In this experiment, the edited Cβ magnetization is transferred to the Cα carbon; from there it is transferred via $^{1}J_{C\alpha N}$ and $^{2}J_{C\alpha N}$ couplings to the own and to the sequential nitrogens, N(i) and N(i + 1), and finally to the H(i) and H(i + 1) amide protons. Hence, every NH will appear in two subspectra, one revealing the nature of their own side chain, and a second one corresponding to the character of the sequential side chain. This second signal will coincide with the one observed in the edited (CBCACO)NH experiment. Usually $^{1}J_{C\alpha N}$ is larger than $^{2}J_{C\alpha N}$ and the intraresidual signal is more intense than the sequential one. However, this is not always the case. Therefore, if a ^{1}H–^{15}N correlation peak appears just in one subspectrum, its presence can be attributed to the intraresidual pathway only in the case that its classification does not coincide with the classification obtained in the edited (CBCACO)NH experiment for the sequential sidechain.

Similarly to the inclusion of the β-carbon edition, a γ-carbon edition could be included in the CBCANH experiment. Nevertheless, this experiment would be of no utility. Differently to the β-carbon edited experiment, the γ-carbon edition do not produce different subspectra. Hence, every NH observed will always exhibit contributions of its own side chain and of the sequential one. These contributions do not correspond exclusively to pathways starting at Cγ, but include also pathways starting at Cβ that are no suppressed due to deviations of the scalar couplings from the standard value of 35 Hz. Consequently the sign of the signal observed cannot be related unambiguously with the topology of the intraresidual side chain. Therefore, this putative pulse sequence has not been considered.

3. Results and discussion

The above set of pulse sequences has been tested with samples of uniformly ^{15}N and ^{13}C labeled ubiquitin (1.7 mM, 76 amino

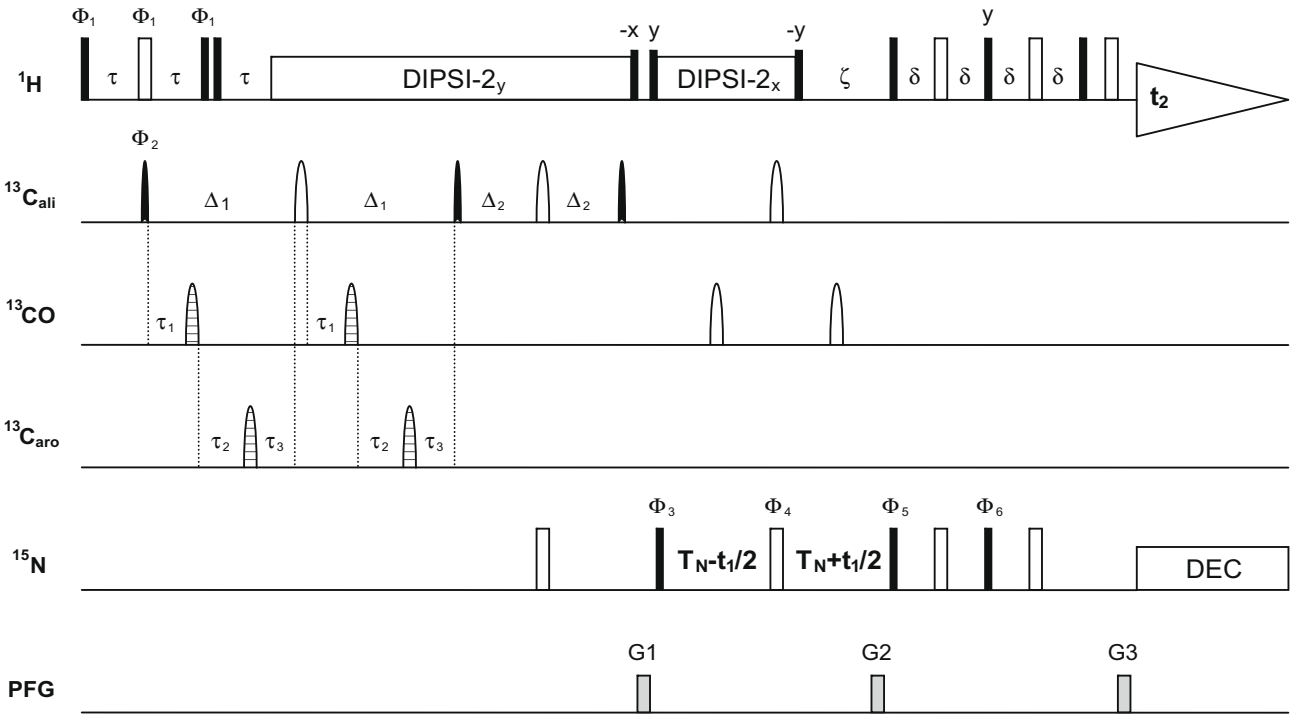

Fig. 3. Pulse sequence for the intraresidual β-carbon edited HSQC experiment. All radiofrequency pulses are applied along the *x*-axis unless indicated. 90° and 180° rectangular pulses are represented by filled and unfilled bars, respectively. ^{13}C pulses have the shape of Gaussian cascades Q5 (black filled shapes) and Q3 (open shapes) with durations of 307 and 192 μs at 800 MHz, respectively. ^{13}C aliphatic pulses are centered at 41 ppm and CO pulses at 173 ppm. Striped shapes correspond to Q3 Gaussian cascades of 340 μs duration (800 MHz), applied at 120 ppm ($^{13}C_{aro}$) and 180 ppm (^{13}CO). The delays are adjusted to $\Delta_1 = 9.5–10.5$ ms; $\Delta_2 = 8.5$ ms; $\tau = 3.7$ ms; $\zeta = 5.5$ ms; $\delta = 2.3$ ms; $T_N = 12.4$ ms. For every t_1 value 8 repetitions of the experiment are separately recorded with the parameter settings given in Table 1 and $\phi_3 = \phi_4 = x$. Pulsed field gradients G_1 to G_3 of sinusoidal shape are applied along the *z*-axis with a 1 ms length and amplitudes of 30%, 80% and 8.1% from the maximal intensity. For quadrature detection echo/antiecho data are recorded by sign alternation of the gradient G_2 with simultaneous inversion of phase ϕ_5 [17].

acids) and of one of the domains of the protein Pub1p (0.2 mM, 101 amino acids). All the NMR experiments were performed at 298 K on a Bruker AV 800 spectrometer equipped with a TCI cryoprobe. The ubiquitin edited spectra were acquired using 24 × 1024 complex data points with spectral widths of 3243.7 Hz (^{15}N) and 9615.4 Hz (^{1}H) and 16 scans (2 scans per data set in the β-carbon edited experiments). The duration of each experiment was 17 min. For the Pub1p edited spectra an acquisition data matrix of 55 × 1024 complex points was used with spectral widths of 2838.3 (^{15}N) and 10416.7 (^{1}H) and 128 scans (16 scans per data set in the β-carbon edited experiments), resulting in an accumulation time of 5 h per experiment.

Fig. 4 shows the regular $^{1}H–^{15}N$ HSQC spectrum of ubiquitin together with the subspectra generated with the sequential β-carbon edited experiment. The regular HSQC shows all expected NH peaks, except those of E24 and G53. In the sequential edited spectrum, all peaks observed in the HSQC spectrum are present at least in one subspectrum. Relative intensity of the NH peaks is different to what is observed in the HSQC. This fact is due to the different attenuation by relaxation that each signal suffers during the edited sequence and to the different number and values of the J_{CC} coupling constants of the residue preceding the observed amide. Particularly noticeable is the case of the C-terminal NH, G76, reflecting the characteristics of G75, that shows a much higher intensity than the other NH signals. Conversely, signals of T7, S57 and E51 are much weaker than average.

The FHWY subspectrum shows the four expected peaks (F4–V5, F45–A46, Y59–N60 and H68–L69) together with some extra peaks of much lower intensity. This kind of residual peaks appears in all subspectra, due to variations in scalar coupling constants and to pulse imperfections. It should be pointed out that residual peaks of some residues can be of the same intensity as real peaks of other residues and can easily be confused if the subspectra are analyzed separately. However, residual peaks of a NH are always much less intense than the real peak detected in the correct subspectrum. In the two proteins studied, the intensities of these artifacts do not exceed 12% of the intensity of the correct peak. Therefore, a joint analysis of all subspectra easily allows distinguishing residual peaks from real ones. In the following, we will refer only to "correct peaks", that is to the peak appearing in the subspectrum with the highest intensity at a given ^{1}H, ^{15}N coordinates. In the DNG subspectrum, 11 positive peaks out of the 12 expected from the ubiquitin sequence are detected. The missing peak corresponds to one of the NH's not observed in the regular $^{1}H–^{15}N$ HSQC (D52–G53). This subspectrum shows also some sharp and negative peaks arising from the side chain NHD groups. This difference in sign, positive for backbone NH's and negative for the side chain ones, hinders the erroneous assignment of a peak to the wrong category. The TAVI subspectrum can be subdivided into two subspectra by representing separately the negative and positive peaks. In the negative subspectrum, the seven peaks corresponding to the seven threonines present in ubiquitin are observed. The positive subspectrum shows 11 peaks, while 12 are expected. As in the case of the DNG subspectrum, the lost peak (I23–E24) corresponds to a NH not observed in the regular HSQC. Finally, the CSEQMKRPL group is the most numerous, showing 37 signals. The positive subspectrum displays three signals, corresponding to the NH's following the three serines present in ubiquitin, together with a residual peak of remarkable intensity. This residual peak corresponds to the G76 NH and, in spite of having an intensity comparable to the correct "CS peaks", it has only 2% of the intensity of the correct G76 NH peaks appearing in the DNG subspectrum. The negative subspec-

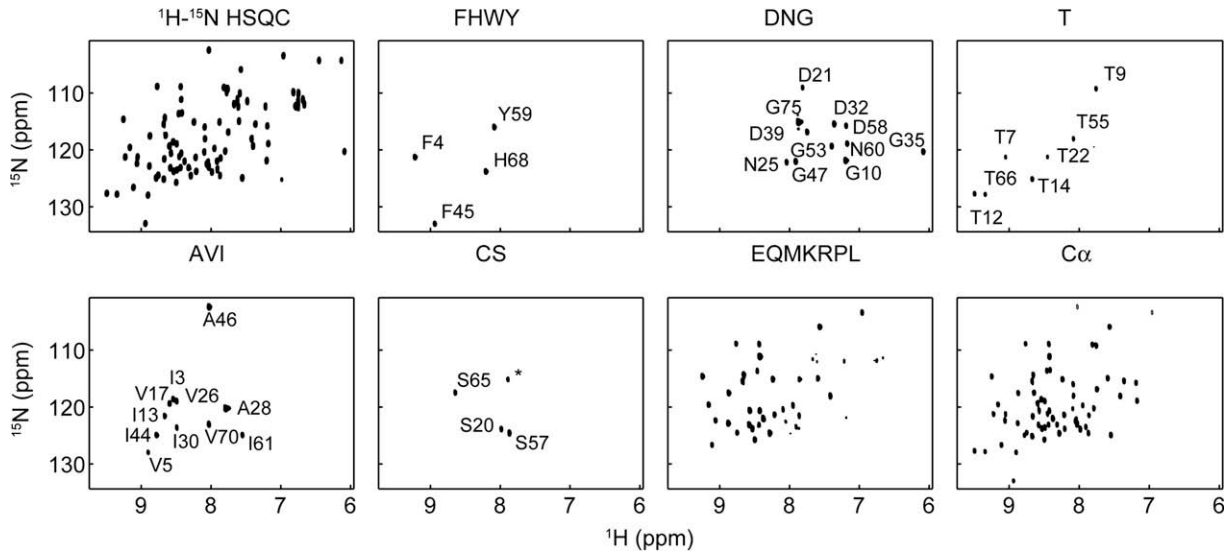

Fig. 4. Regular ^{1}H–^{15}N HSQC and sequential β-carbon edited ubiquitin spectra. The cross-peaks are labeled according to the previous residue in the ubiquitin sequence. Only positive contours are shown for FHWY, DNG, and Cα subspectra. T and AVI, as well as EQMKRPL and CS, correspond to negative and positive contours of the same subspectrum. The signal with an asterisk in the CS subspectrum corresponds to a residual peak of G76.

trum shows the expected number of backbone NH peaks, 34, together with some NHD signals of side chains. In summary the 70 backbone NH peaks observed in the HSQC spectrum of ubiquitin can be classified into one of the six groups. For the Pub1p domain the results obtained were quite similar, and the analysis of the β-carbon edited experiment allowed the unambiguous classification of 86 NH's from the 89 observed in the HSQC spectrum. Two of the NH's for which no classification was possible show signals of low intensity in the HSQC spectrum and do not give any signal in the edited spectra. The other unclassified NH corresponds to a signal that partially overlaps with other NH of the same type and of greater intensity, thus difficulting the classification.

In both proteins the amino-acid-type group corresponding to amino acids with a single aliphatic gamma carbon (E, Q, M, K, R, P, and L) is very copious (34 NH's in the case of ubiquitin and 30 for Pub1p). This group can be subdivided into two smaller groups by using the γ-carbon edited experiment. Fig. 5a shows an example of the application to the classification of the ubiquitin NH peaks. As expected, NH peaks of residues following K, R, P or L appear negative in γ-carbon edited spectrum, while NH peaks of residues that follow E, Q or M appear positive. The γ-carbon edited experiment allowed the assignment of 20 KRPL and 12 EQM peaks for ubiquitin and of 16 KRPL peaks and 9 EQM peaks for Pub1p. Although the main utility of the γ-carbon edited experiment is the subdivision of the EQMKRPL group, it can be also used to analyze the DNG subspectrum. As explained in the description of the pulse sequence, NH's following G residues will give a positive and intense signal in this kind of spectrum, while NH's following aromatic, D, N, C or S residues can give positive, zero, or negative signals, depending on the value of the $J_{C\alpha-C\beta}$ coupling constant. Therefore, signals classified as DNG in the β-carbon edited experiment and appearing negatives in the γ-carbon edited experiment cannot be G. An example of the application of this rule is shown in Fig. 5b. From the five DNG signals appearing in the DNG subspectrum excerpt, two appear negatives in the γ-edited spectrum, thereby being classified as DN, one does not appear at all, most likely being also DN, and two are positive and of high intensity, probably being G. This proposal can be confirmed by using the Cα subspectrum, since this subspectrum shows NH signals for all residues except those following glycine (see Fig. 5c). The use of two criteria for the separation of DN and G, sign in the γ-carbon edited experiment and presence/absence in the Cα subspectrum, make the process very robust.

The β-carbon edited experiment not only allows the classification of the backbone NH's into one of six groups but also the distinction of N and Q side chain NHD resonances. Fig. 5d gives an example of application to the Pub1p domain. As was explained above, the side chains NHD signals will appear in two subspectra. The DNG subspectrum will show signals of both kind of side chains, but the EQMKRPL subspectrum will show signals only for the Q side chains, while in the Cα subspectrum only the N side chains will appear. This fact has allowed the classification of all side chains, except one, in the case of ubiquitin and of 9, from the 11 existing, in the domain of Pub1p. The three side chains that cannot be classified do not give any signal in any of the three subspectra, probably because of relaxation during the pulse sequence.

The classification of the backbone NH resonances can be further extended by using the β-carbon edited CBCANH experiment of Fig. 3. The subspectra produced by this experiment show signals for NH's for which either their own side chain or the sequential one is of the selected type. Signals corresponding to sequential neighbors appear usually with weak intensity, or are even absent, and coincide with the signals observed in the same subspectrum type in the β-carbon edited CBCACONH experiment. Therefore, intra-residual signals are easily recognized. Fig. 6 shows some of the subspectra obtained for the ubiquitin sample, together with corresponding subspectra of the β-carbon edited CBCACONH experiment. A clean, although not perfect, separation of cross-peaks from different types is observed. As in the case of the CBCACONH experiment residual peaks are of much lower intensity, usually less than 10%, than that of the correct correlation peak. Therefore, the subspectrum with the highest intensity, once the sequential peak is excluded from the comparison, yields the classification of the amino acid. This classification is unambiguous in most cases. However, in cases in which only one peak is observed and its type coincides with the one of the sequential amino acid, the categorization of the own amino acid remains ambiguous, since two different situations can lead to this result. First, the sequential and the own amino acids are of the same type. Second, the intraresidual peak is attenuated by relaxation during the pulse sequence beyond detection and only the sequential peak is observed. Although the first situation is the most probably, it is recommended not assign

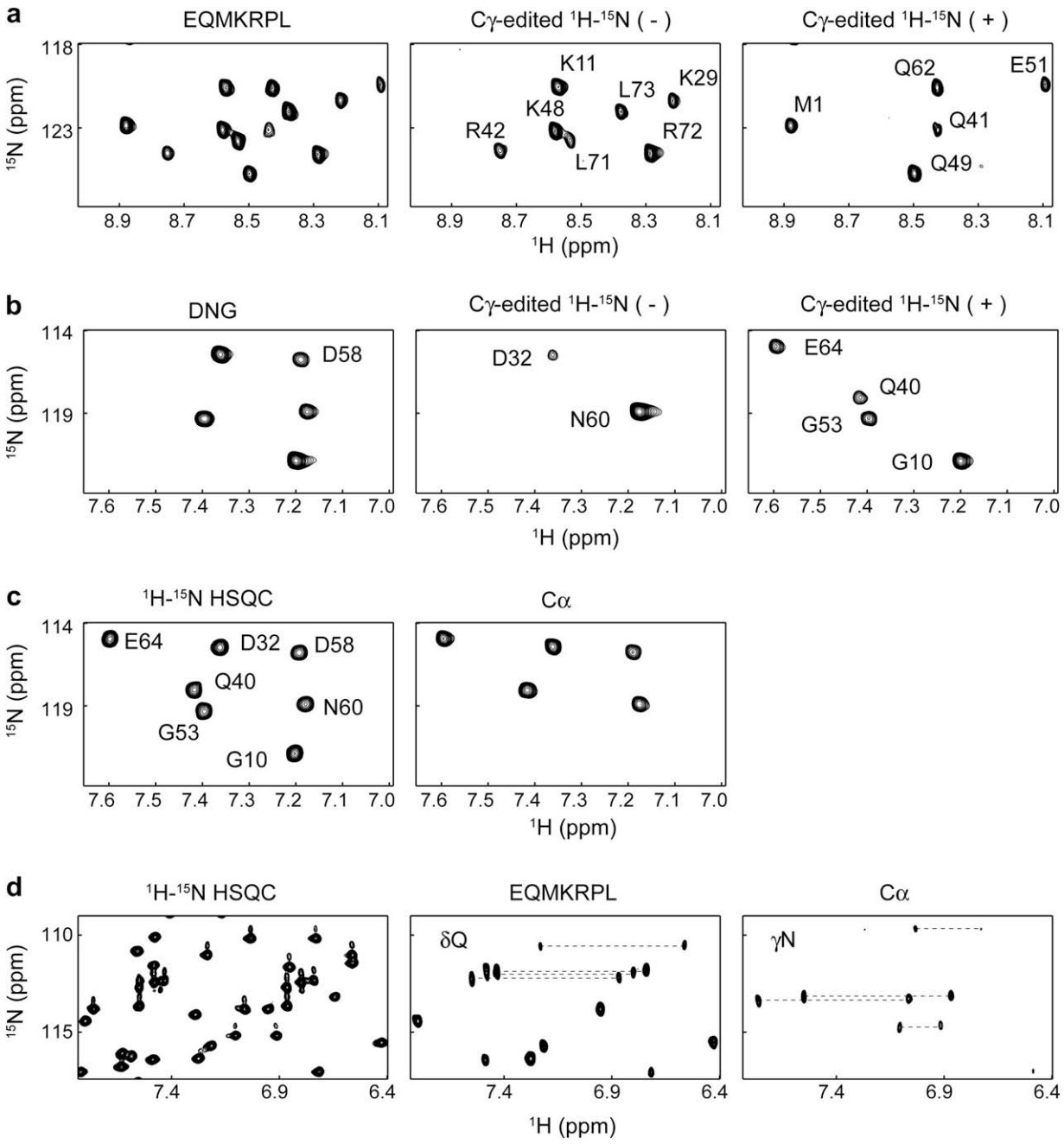

Fig. 5. (a) Excerpt of the EQMKRPL subspectrum of ubiquitin (left), and negative (middle) and positive (right) contours of the γ-carbon edited spectrum of the same spectral region; (b) excerpt of the DNG subspectrum of ubiquitin (left), and negative (middle) and positive (right) contours of the γ-carbon edited spectrum of the same spectral region; (c) $^{1}H-^{15}N$ HSQC (left), and Cα subspectrum (right) corresponding to the same spectral region shown in (b); and (d) $^{1}H-^{15}N$ HSQC (left), EQMKRPL subspectrum (middle), and Cα subspectrum (right) of the side chains spectral region of Pub1p.

the amino acid type in this situations. This is especially true in automated assignment procedures, since the use of wrong information would mislead the automated assignment program. Even using this cautious strategy 55 of the 70 NH's observed in the HSQC spectrum could be classified in ubiquitin and 60 of 86 in Pub1p.

4. Conclusions

Our sequential β-carbon edited pulse sequence has many points in common with the recently proposed HADAMAC experiment [15]. Nevertheless, there are also some significant differences. First, the separation of CH/CH_3 from CH_2 is based on DEPT in the HADAMAC experiment and on POMMIE [10,21] in our sequence. The POMMIE procedure is less sensitive than DEPT to pulse imperfections and offset effects [19,21], and should produce a cleaner separation of subspectra. A second and more important difference refers to the handling of the signals originating from Cα magnetization. While HADAMAC suppresses this pathway by using a Δ_2 value of $1/(4J_{C\alpha-C\beta})$, our method produces an exclusive subspectrum for this pathway, and therefore permits the use of shorter Δ_2 settings, what results in less attenuation of the magnetization by relaxation. Further, instead of using Cβ selective inversion pulses to obtain separate subspectra for S and T residues, we use aromatic carbon inversion pulses to obtain a subspectrum for aromatic residues. The Cβ selective pulses are difficult to implement,

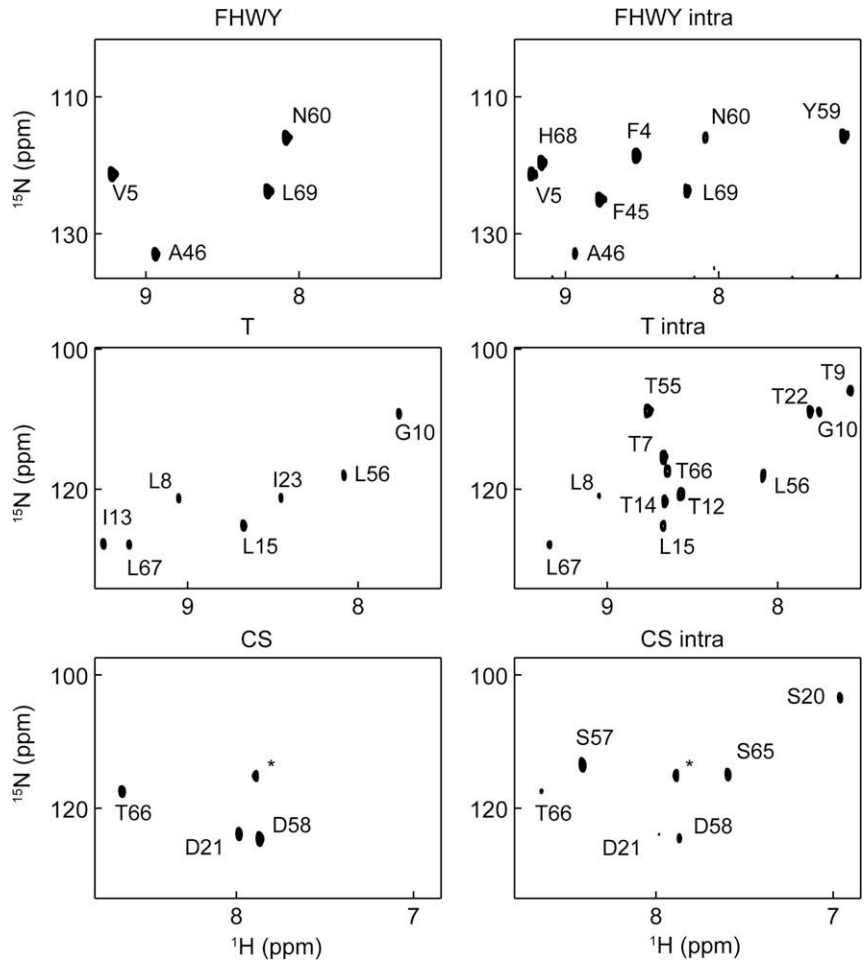

Fig. 6. Examples of sequential β-carbon edited subspectra (left) and intraresidual β-carbon edited subspectra (right) of ubiquitin. NH cross-peaks are labeled here according to its own residue in the ubiquitin sequence. The signal with an asterisk in both CS subspectra corresponds to a residual peak of G76.

since they must invert the Cβ magnetization of S and T without affecting the Cβ magnetization of other residue types, which can resonate very near. The aromatic selective pulses, however, do not suffer this problem. Finally, HADAMAC includes a manipulation of the Cα magnetization to distinguish $C\alpha H_2$ from CαH moieties, thereby producing a subspectrum for G residues. Although this manipulation could also be included in our method, it will complicate the pulse sequence unnecessarily, since G residues are easily recognized by their particular Cα chemical shift and the lack of Cβ.

Even with all these differences the amino-acid-type groups obtained with both pulse sequences are very similar. In both cases a large group, which consists of the seven amino acids with a single aliphatic γ carbon, is obtained. We have shown that this group can be subdivided into two subgroups of similar size by using a sequential γ-carbon edited pulse sequence. Using these two experiments seven amino-acid-type groups, having 1–4 elements, are differentiated. This reduces the assignment ambiguity of the ^{1}H–^{15}N HSQC peaks from 20 possible amino acid types of the preceding residue to an average of 3.

The labeling can be further extended by using an intraresidual β-carbon edited experiment. This experiment produces the same six amino-acid-type groups obtained with the sequential experiment. Unfortunately, in this case, no subdivision of the *Long* group can be obtained. Besides, the classification of the intraresidual amino acid is possible only if its type does not coincide with the one of the sequential amino acid. Therefore, only approximately 75% of the intraresidual amino acids can be unambiguously classified.

The three carbon-edited ^{1}H–^{15}N HSQC experiments proposed offer the amino acid type identification of two sequential residues, i and $i-1$, and thus a sequential pair of amino acid types. With seven types of the sequential residue and six types of the own residue, 42 different sequential pairs can occur in proteins. However, not all these pairs will be present in a particular protein, and more important, some of the pairs will appear only once. So, the ubiquitin sequence contains 10 unique pairs out of the 27 present, and Pub1p 15 unique pairs out of 35. Therefore, the experiments allow the assignment of some of the NH peaks without using any additional information.

The experiments reported here have been performed as 2D, but they can also be implemented as 3D, by adding an additional ^{13}C time evolution period to resolve overlapping ^{1}H–^{15}N correlations in a third spectral dimension. The third dimension can correspond either to CO or to Cβ/Cα. In this second case not only a separation of the overlapping signals is obtained, but also information on the Cβ and Cα chemical shifts, thus allowing the sequential assignment of the protein.

In conclusion, we have designed a set of pulse sequences that allow a clean amino acid type identification using a limited number of experiments. This amino acid classification will have a very positive impact in the assignment of protein backbone resonances, making automatic methods that include this kind of information more efficient and robust. The proposed strategy, however, is not devoid of problems. Main drawback of the proposed pulse sequences is the long delays needed to obtain the amino acid classi-

fication, what results in an important attenuation of the signal by relaxation. Therefore, the pulse sequences will be useful only for proteins of small to moderate size. Pulse sequences in Bruker language, and a program to perform the linear combination to separate the subspectra can be obtained from the page http://rmn.iqfr.csic.es.

Acknowledgments

This work was supported by Project BFU2005-01855 from the Spanish Ministerio de Educación y Ciencia. D.P-U. was supported by a "Juan de la Cierva" contract from the Ministerio de Educación y Ciencia. We thank Dr. Perez-Cañadillas for useful discussions and for the loan of the Pub1p sample.

References

[1] M. Sattler, J. Schleucher, C. Griesinger, Prog. Nucl. Mag. Res. Sp. 34 (1999) 93–158.
[2] S. Grzesiek, A. Bax, J. Biomol. NMR 3 (1993) 185–204.
[3] K.M. Lee, E.J. Androphy, J.D. Baleja, J. Biomol. NMR 5 (1995) 93–96.
[4] D.C. Muchmore, L.P. McIntosh, C.B. Russell, D.E. Anderson, F.W. Dahlquist, Meth. Enzymol. 177 (1989) 44–73.
[5] L.P. McIntosh, F.W. Dahlquist, Quart. Rev. Biophys. 23 (1990) 1–38.
[6] V. Dötsch, R.E. Oswald, G. Wagner, J. Magn. Reson. 110 (1996) 304–308.
[7] V. Dötsch, G. Wagner, J. Magn. Reson. 111B (1996) 310–313.
[8] W. Feng, C.B. Rios, G.T. Montelione, J. Biomol. NMR 8 (1996) 98–104.
[9] C.B. Rios, W. Feng, M. Tashiro, Z. Shang, G.T. Montelione, J. Biomol. NMR 8 (1996) 345–350.
[10] M. Schubert, M. Smalla, P. Schmieder, H. Oschkinat, J. Magn. Reson. 141 (1999) 34–43.
[11] M. Schubert, H. Oschkinat, P. Schmieder, J. Magn. Reson. 148 (2001) 61–72.
[12] M. Schubert, H. Oschkinat, P. Schmieder, J. Biomol. NMR 20 (2001) 379–384.
[13] M. Schubert, H. Oschkinat, P. Schmieder, J. Magn. Reson. 153 (2001) 186–192.
[14] M. Schubert, D. Labudde, D. Leitner, H. Oschkinat, P. Schmieder, J. Biomol. NMR 31 (2005) 115–127.
[15] E. Lescop, R. Rasia, B. Brutscher, J. Am. Chem. Soc. 130 (2008) 5014–5015.
[16] S. Grzesiek, A. Bax, J. Am. Chem. Soc. 114 (1992) 6291–6293.
[17] L.E. Kay, P. Keifer, T. Saarinen, J. Am. Chem. Soc. 114 (1992) 10663–10665.
[18] F. Löhr, C. Pérez, R. Köhler, H. Rüterjans, J.M. Schmidt, J. Biomol. NMR 18 (2000) 13–22.
[19] P. Schmieder, M. Leidert, M. Kelly, H. Oschkinat, J. Magn. Reson. 131 (1998) 199–202.
[20] S. Grzesiek, A. Bax, J. Magn. Reson. 99 (1992) 201–207.
[21] J.M. Bulsing, W.M. Brooks, J. Field, D.M. Doddrell, J. Magn. Reson. 56 (1984) 167–173.