

# The DNA binding domain of c-Myb: Over-expression and NMR characterization

P. K. Radha\*, Anup Madan\*, L. C. Padhy\*<sup>§</sup> and R. V. Hosur\*†

\*Chemical Physics Group, <sup>†</sup>Molecular Biology Group, Tata Institute of Fundamental Research, Homi Bhabha Road, Bombay 400 005, India

The DNA-binding domain of the *Drosophila melanogaster* c-Myb protein, 160 residues long, containing three conserved imperfect repeats of nearly 50 residues each has been overexpressed in *E. coli* strain BL21(DE3). The protein is expressed to the extent of atleast 20% of the total cellular protein. It has been purified by a four-step protocol developed in the laboratory. The protein has been characterized by various 2D NMR experiments. Several specific amino acids have been identified. A three-dimensional NMR spectrum has been recorded to achieve dispersion of cross-peaks into different 2D planes.

THE c-Myb protein, an oncogene product, is a nuclear protein and binds to DNA in a sequence-specific manner<sup>1-5</sup>. The DNA binding activity is mediated by a highly conserved amino-terminal region which consists of three imperfect stretches [R1-R2-R3] of approximately 50–53 amino acids, characterized by tryptophan repeats every 18/19 residues in each stretch<sup>4,6,7</sup>. It is conjectured that the tryptophan repeats play important role in the DNA interaction and recognition process. Out of the three conserved stretches, two stretches (R2-R3) have been shown to be critical for the DNA binding in the chicken c-Myb. With a view to investigate the structure-function relationship, recently Ogata *et al.*<sup>8</sup> have determined structures of a synthetic peptide containing the third stretch of the protein [R3] by NMR. The results show the presence of a helix-turn-helix motif, characteristically seen in DNA binding motifs of other DNA binding proteins<sup>9</sup>.

To further study the structure and mechanism of interaction of the entire segment of three conserved regions with DNA, we have over-produced this segment (hereafter referred to as R123) of *Drosophila melanogaster* myb (*Dm*-myb). Over-expression has been achieved in *E. coli* cells to the extent of approximately 20% of the total cellular protein using the vector system developed by Studier<sup>10</sup>. The protein has been purified and characterized by two-dimensional NMR spectroscopy. For complete three-dimensional structure determination of the protein by NMR – the protein being

among the largest ever handled by NMR in the world – we have successfully extended the NMR experiments into the third dimension. We present here the highlights of our experiments and results while the details will be published elsewhere.

## Methods

### Protein preparation

The strategy used for over-producing the conserved stretch of the *Drosophila melanogaster* c-Myb protein is shown in Figure 1. The *p*-Gem2 vector containing a full

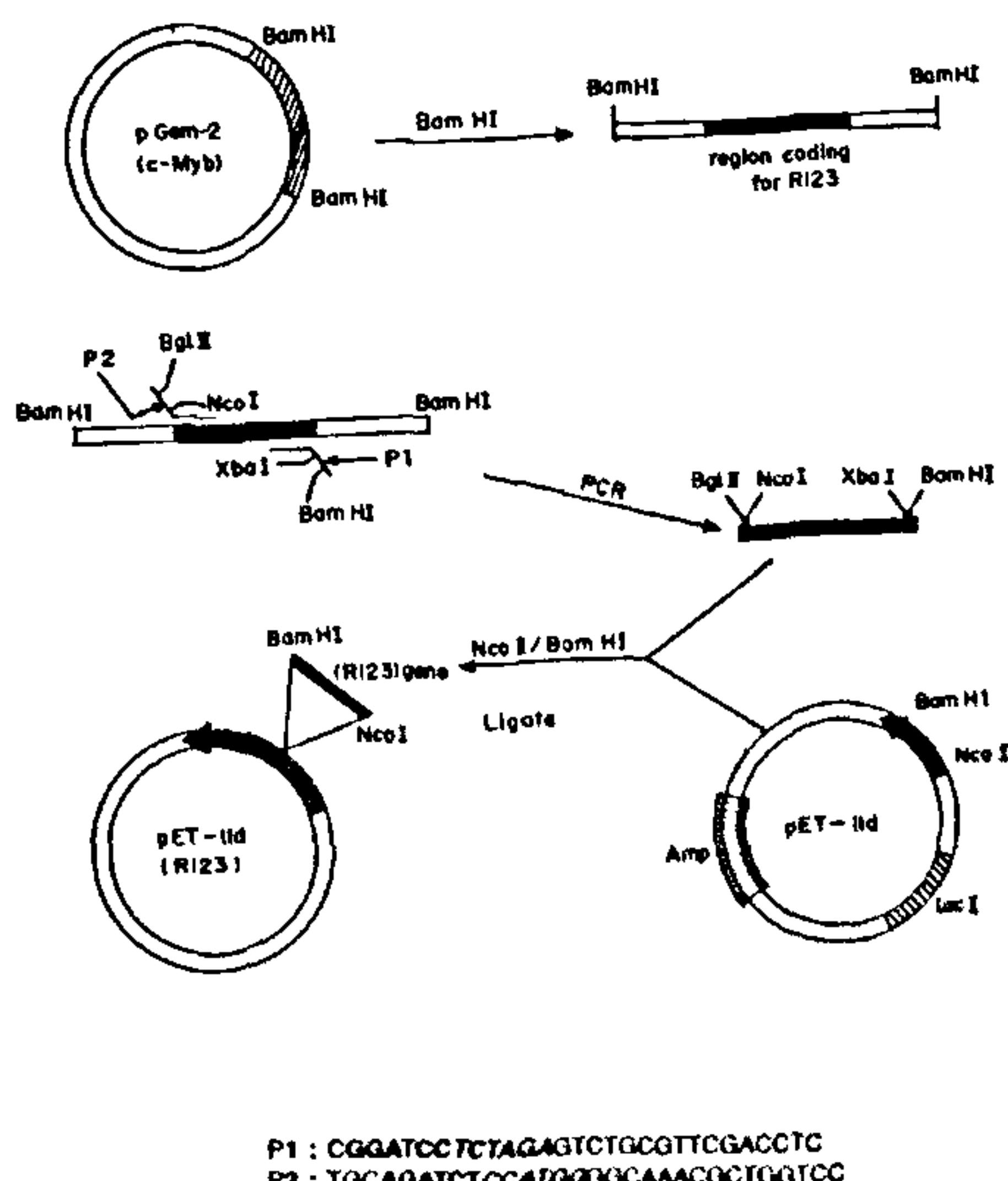


Figure 1. Strategy used to construct an expression vector pET-11d(R123) to over-express the conserved stretch of the *Dm* c-Myb protein in *E. coli*. The primer sequences used for PCR generation of DNA are also indicated.

\*For correspondence

length c-Myb clone was the gift of Rashmi Sood. PCR was used to simultaneously amplify the fragment coding for R123 and generate start and stop codons at appropriate places. Primer P2 contained BglII & NcoI site overlapping with ATG start codon and primer P1 contained an XbaI site overlapping with TAG stop codon & BamHI site. The reaction was carried out in a thermal cycler for 20 cycles of denaturation at 95°C, annealing and extension at 75°C for 45 seconds. The success of PCR reaction was checked on 3% NuSieve agarose gel. The reaction product was fused between NcoI and BamHI sites of pET-11d. The ligation mixture was transformed into *E. coli* BL21(DE3) using standard procedures<sup>11</sup> and the transformants were selected for ampicillin resistance. The plasmid DNAs from transformants were further screened by various enzyme digestions for the presence of the insert and the recombinants were identified. The recombinants were grown at 37°C overnight in a medium containing 10 g of Bactotryptone, 10 g of NaCl and 5 g of Bactoyeast extract, pH 7.0 (LB medium) and 100 µg/ml of ampicillin. The culture was diluted and allowed to grow just before induction. Expression was induced by addition of IPTG to 25 µM when the O.D.<sub>600</sub> reached around 0.7. Cells were harvested after 3 hours by centrifugation at 5000 rpm for 10 minutes at 4°C using GSA rotor. The protein was found to come in both the soluble form and as an inclusion body. In order to solubilize the insoluble portion (inclusion bodies) of the protein as well, 8 M urea was included in the lysis buffer. The cells from every litre of culture were suspended in 20 ml of lysis buffer (8 M urea, 20 mM Tris.HCl pH 8.0, 2 mM EDTA, 1 mM PMSF), sonicated (1 min × 5 at 40% maximum intensity) using a Branson sonifier and centrifuged at 10,000 rpm for 10 minutes using SS34 rotor in Sorvall RC5B centrifuge to remove the cell debris. The supernatant contained the R123 protein along with other soluble cellular proteins and DNA.

#### *Purification of the recombinant protein and refolding*

DNA and lipids were removed from the cell-free extract by adding PEI (polyethyleneimine) to a concentration of 0.03%. The precipitate was removed by centrifugation at 15,000 rpm for 10 minutes using SS34 rotor. From the supernatant the proteins were precipitated by adding ammonium sulphate solution to a final concentration of 50% saturation and the precipitate was pelleted by centrifugation at 15,000 rpm for 10 minutes using SS34 rotor. The pellet was redissolved in 20 ml of the lysis buffer. Sample was loaded on DE-52 (Whatman) column (5.0 × 45 cm) pre-equilibrated with lysis buffer and various fractions of 4 ml were collected. The elution of the protein was monitored by measuring O.D.<sub>280</sub>.

Fractions containing the proteins were analysed on 15% SDS-PAGE gel electrophoresis and were visualized by Coomassie Brilliant Blue R250 staining. Fractions containing R123 (based upon the mobility of proteins in 15% SDS-PAGE gel) were pooled. The pH of the pooled fractions was adjusted to 6.5 by MES and loaded on SP-sephadex column (2.5 × 15 cm) pre-equilibrated with buffer A (8 M urea, 20 mM Na-MES pH 6.5, 0.15 M NaCl). The column was washed extensively with buffer A and developed with a linear gradient of NaCl (0.15–1.0 M) in buffer A. Fractions of 2.0 ml were collected and analysed on 15% SDS-PAGE gel electrophoresis. Active fractions were pooled and found to be of approximately 99% purity as shown in Figure 2, lane 6.

For refolding, the protein was diluted to 200 µg/ml with 20 mM Tris.HCl pH 7.0, 1.0 mM EDTA, 10 mM DTT buffer and gently rocked at 4°C for 24 hours. The urea concentration in this solution was 0.1 M. The protein was concentrated by ultrafiltration on an Amicon YM10 membrane and dialyzed against three changes of water to remove the remaining urea and DTT. The dialyzed sample was lyophilized.

The protein samples for NMR were prepared by dissolving the protein in 0.5 ml solution of either 10 mM sodium acetate buffer, or 10 mM Na-phosphate buffer. The experiments were performed in the pH ranges of 1.5–7.5. The sample was in 90% <sup>1</sup>H<sub>2</sub>O/10% <sup>2</sup>H<sub>2</sub>O for the experiments in water. For experiments in <sup>2</sup>H<sub>2</sub>O, the NMR sample was in 99.99% <sup>2</sup>H<sub>2</sub>O and was prepared by repeated lyophilization from 99.8% <sup>2</sup>H<sub>2</sub>O. The lyophilized protein was finally dissolved in 99.99% <sup>2</sup>H<sub>2</sub>O. The protein concentration in each of the samples was approximately 6 mM.

#### *NMR data acquisition and processing*

NMR spectra were recorded on a Bruker AMX-500 spectrometer, operating at 500 MHz for <sup>1</sup>H. All spectra were recorded at 289 K. Two-dimensional 2QF COSY<sup>12</sup>  $\omega_1$ -scaled 2QF-COSY<sup>13</sup>, Clean TOCSY<sup>14</sup> double quantum<sup>15, 16</sup>, NOESY<sup>17</sup> and three-dimensional NOESY-TOCSY<sup>18</sup> spectra for <sup>1</sup>H<sub>2</sub>O sample were acquired using the standard methods. The Clean TOCSY spectra for <sup>1</sup>H<sub>2</sub>O sample were acquired with spin-lock periods of 38 and 61 ms and the NOESY with a mixing time of 60 ms. The 3D NOESY-TOCSY spectrum was obtained with a NOE mixing time of 100 ms and spin lock period of 60 ms. Spectra recorded for the <sup>2</sup>H<sub>2</sub>O sample were 3QF COSY<sup>16</sup> and TOCSY with spin lock time of 33 ms.

Spectra were collected in phase-sensitive mode, using time-proportional phase incrementation (TPPI<sup>19</sup>). 2048 points were collected and averaged with 96 or 128 scans in the  $t_2$  dimension. In  $t_1$  dimension 450–580 points were acquired. Spectral widths of 5208.33 Hz for the

$^1\text{H}_2\text{O}$  sample and 4032.26 Hz for the  $^2\text{H}_2\text{O}$  sample were used. The 3D NOESY-TOCSY spectra were recorded with 512 points in the  $t_3$  dimension averaged to 16 scans, 64 in the  $t_2$  and 64 in the  $t_1$  dimension.

All the spectra were processed on Aspect-X32 data station using the UXNMR software. Before Fourier transformation, the time domain data were apodized with phase-shifted sine square functions. The 2D data were processed with 2048 real points in both the dimensions. For 3D with  $512 \times 64 \times 64$  acquired points in the three dimensions linear prediction was used to extend the data to 128 points along  $t_2$  and  $t_1$  dimensions. Automated baseline corrections were performed on all the processed data. Chemical shifts are expressed with respect to HDO whose position was calibrated with respect to sodium 3-(trimethylsilyl) [2,2,3,3- $^2\text{H}_4$ ]propionate([ $^2\text{H}$ ] TSP).

## Results

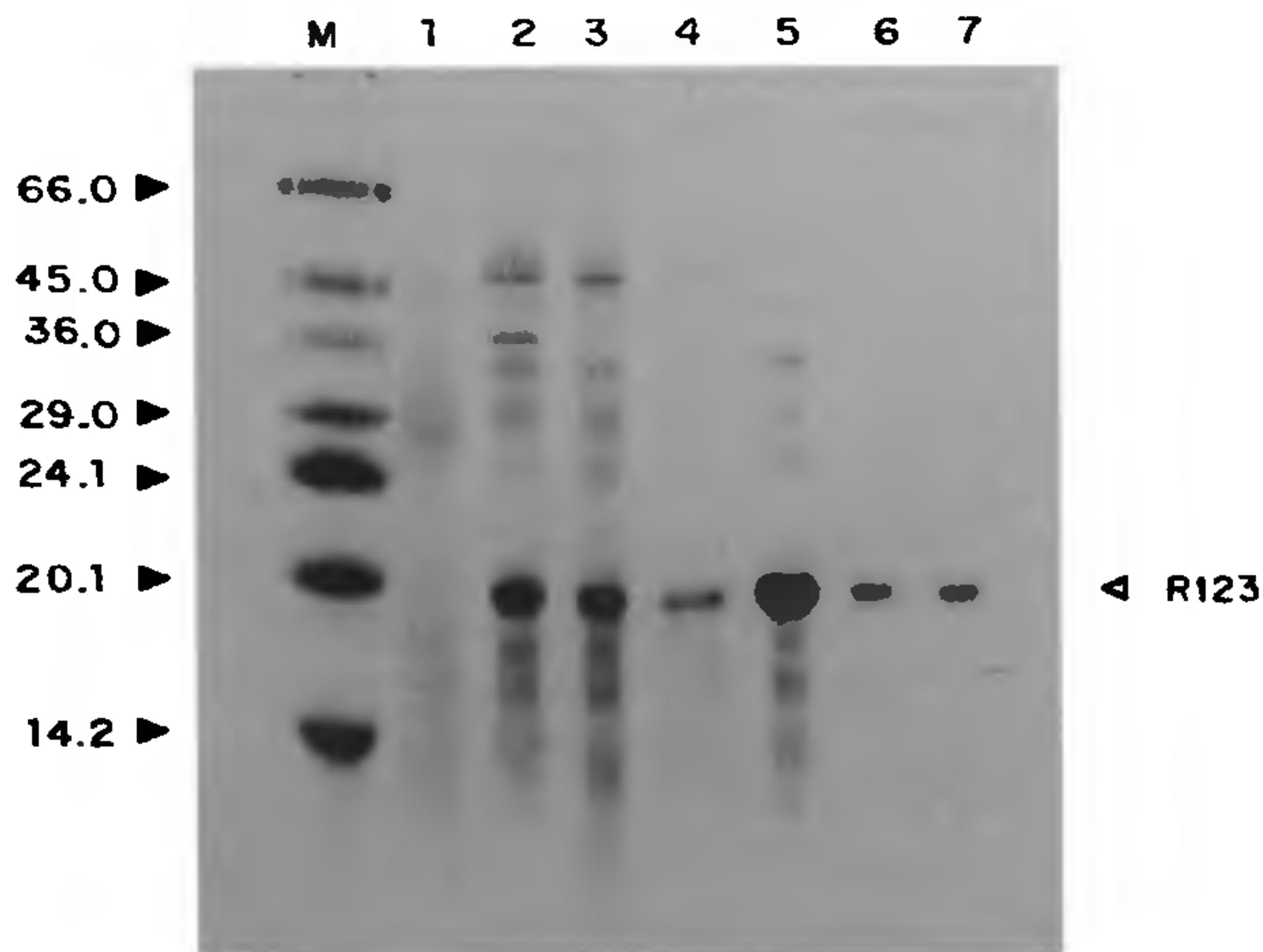
### Overexpression and purification of the protein

For over-producing the DNA binding domain of the *Drosophila melanogaster* c-Myb protein in *E. coli*, our strategy was to make use of bacteriophage T7 expression system developed by Studier *et al.*<sup>24</sup> Figure 2

shows the results at various stages of protein purification. The presence of the protein with an apparent molecular weight of about 19.5 kDa (indicative of R123) revealed that *E. coli* BL21(DE3) cells transformed with pET-11d(R123) plasmid expressed R123 protein on induction with IPTG (Figure 2; inset: lane 2).

Polyethyleneimine (PEI) causes precipitation of nucleic acids and acidic proteins at low ionic strength by forming charge neutralization complexes and cross bridges between the complexes leading to flocculation. By using PEI to a final concentration of 0.03% not only DNA gets removed but some of the acidic proteins also get precipitated (Figure 2; lane 3). PEI from the supernatant was removed by precipitating the proteins by ammonium sulphate precipitation. It was found that at 50% saturation of ammonium sulphate all the protein of interest gets precipitated. The precipitate was found to be soluble only in buffer containing 8 M urea.

Further purification of the R123 protein was achieved by ion-exchange chromatography. The anion exchange column DE-52 (Whatman) retained the acidic proteins which were not precipitated by PEI. The protein of interest came entirely in flow. This partial purification of the protein using DE-52 column results in significant enrichment of the protein (Figure 2; inset: lane 5). Final purification of the protein was accomplished by cation



**Figure 2.** 15% SDS-polyacrylamide gel analysis showing entire cell protein and purification of R123 from *E. coli*, BL21(DE3)/pET-11d(R123). Lane 1: BL21(DE3)/pET-11d(R123) whole cell lysate before induction. Lane 2: BL21(DE3)/pET-11d(R123) whole cell lysate prepared after 3 hours of induction. Lane 3: after PEI precipitation of the crude lysate. Lane 4: after 50% ammonium sulphate saturation, the precipitate is redissolved in the lysis buffer and analysed. Lane 5: DE-52 cellulose (Whatman) pool containing the active fractions. Lane 6: SP-sephadex pool containing the active fractions. Lane 7: after refolding the protein. Molecular weight markers are indicated with arrows and expressed in kDa.

exchanger SP-sephadex column. It was found that on equilibrating the column with buffer containing 0.15 M NaCl R123 was retained in the column and eluted at 0.83% of NaCl on developing the column with linear NaCl gradient (0.15–1.0). The peak corresponding to R123 showed a single band on 15% SDS-PAGE gel electrophoresis at approximately 19.5 kDa (Figure 2; inset: lane 6). SDS-PAGE gel electrophoresis of the refolded protein showed a single band indicative of no degradation products during refolding (Figure 2; inset: lane 7).

The purified and refolded R123 protein has been tested for its DNA-binding properties. It has been observed that protein binds to DNA tightly and strongly discriminates between cognate target and non-cognate target DNA sequences, assuring us about the retention of biological activity in the truncated myb protein.

### 2D NMR characterization of the protein

NMR has emerged as an invaluable tool for detailed characterization of biological macromolecules in aqueous solutions. At the lowest level, the behaviour of protein molecules under different experimental conditions such as pH, ionic strengths, concentration and temperature stability can be studied. At a somewhat higher level, specific residues in a large protein can be identified which can serve as monitors for interaction studies with substrates. At a still higher level, three-dimensional structures of the molecules can be determined and finally the dynamism can be studied at the highest level of experimental sophistication and data analysis.

In the following, we will first establish a correlation between the amino acid composition of the R123 protein and the number of peaks in different 2D spectra. This will not only provide complementary support – though preliminary – to the successful production of the desired protein, but will also throw light on the chemical shift dispersion among the various types of protons in the protein, a feature which is at the heart of feasibility of 3D structure determination. Next, we will try to identify some amino acid types in the spectra so that they can be monitored for interaction studies with its specific DNA. Table 1 shows the amino acid sequence and composition of the R123 protein.

Two-dimensional *J*-correlated spectra in  $^2\text{H}_2\text{O}$  solution provide fingerprints of the number of residues and to some extent also of the amino acid composition of a well-folded protein. Each residue except prolines and glycines produces one NH-C<sup>a</sup>H cross-peak while each glycine produces two peaks and proline gives rise to none. Thus the number of peaks expected for a given protein can be calculated from the knowledge of its amino acid composition. These NH-C<sup>a</sup>H peaks appear in a distinct region of the 2D NMR spectrum which is consequently referred to as the 'fingerprint' region of the spectrum. Thus, analysis of the fingerprint region of 2D NMR spectra is the first step in NMR characterization of a given protein. Often the number of peaks observed will be slightly less than the expected number. This is because (i) depending upon the lability of the protein, the peaks belonging to the terminal residues may not be seen due to fast exchange of the amide protons with the solvent protons and (ii) due to large linewidths of some resonances, the net intensities and

**Table 1** Amino-acid sequence and composition of R123 protein.

#### Amino acid sequence:

MGKRW SKSEDVLLKQLVETHGENWEIIGPHFK  
DRLEQQVQQQRWAKVLNP ELIKGPWT RDED DMV  
IKLVRNFGPKKWT LIA RYLNGRIGKQC RERWH  
NHLPNPKKTAWTEKEDEIIYQAHLELGNQWA  
KIAKRLPGRTDNAIKNHWNSTMRRKYDVERRL.

#### Amino acid composition:

Alanine	7	Leucine	14
Arginine	15	Lysine	17
Asparagine	11	Methionine	3
Aspartate	8	Phenylalanine	2
Cysteine	1	Proline	6
Glutamine	8	Serine	3
Glutamate	13	Threonine	7
Glycine	9	Tryptophan	9
Histidine	6	Tyrosine	3
Isoleucine	11	Valine	7

Total number of amino acid residues = 160

peak heights may get reduced resulting in poor signal-to-noise ratio. A larger than the expected number of peaks could imply some compositional or conformational heterogeneity of the sample.

In the R123 protein, the fingerprint region of the *J*-correlated spectra is expected to contain 163 peaks. We have used a variety of *J*-correlated techniques namely double quantum filtered (DQF) COSY, shift-scaled DQF-COSY, double quantum (DQ) spectroscopy and clean-TOCSY to check for the consistency of results. However, it is important to note that excepting clean-TOCSY spectra, all others have positive and negative components in the cross-peak fine structures which will tend to cancel each other in the event of large linewidths of the protons. Clean-TOCSY has all positive components and consequently we were able to count the

maximum number of peaks (154) in the fingerprint region of this spectrum (Figure 3 *a*).

Clean-TOCSY spectrum also provides correlations between the NH-C<sup>β</sup>H, C<sup>γ</sup>H etc. protons for each amino acid residue and thus can be used to identify spin systems of the various residues in the spectrum (Figure 3 *b*). This information can be cross-checked with the C<sup>α</sup>H-C<sup>β</sup>H, C<sup>γ</sup>H etc. cross-peaks present in the clean-TOCSY spectra in <sup>2</sup>H<sub>2</sub>O solutions or with other C<sup>α</sup>H-C<sup>β</sup>H, C<sup>β</sup>H-C<sup>γ</sup>H etc. cross-peaks present in the *J*-correlated spectra such as DQF-COSY, E-COSY spectra in <sup>2</sup>H<sub>2</sub>O solutions. We have carried out such an exercise and more than 100 NH-C<sup>α</sup>H-(C<sup>β</sup>H<sub>1</sub>, C<sup>β</sup>H<sub>2</sub>) spin systems could be identified from the spectra. In this context, the E-COSY spectrum is also helpful in distinguishing non-exchangeable AMX spin systems expected for the residues Ser, Asn, Asp, Trp, Tyr, His, Phe, Cys from the spin systems of long side chains; the AMX systems produce well-resolved (+, -) patterns in the cross-peaks. R123 has 37 AMX spin systems and we have been able to identify a large number of these systems.

The above analysis has indicated that under the experimental conditions of 16°C and pH 4.2, there is a fairly good chemical shift dispersion among the NH, C<sup>α</sup>H and C<sup>β</sup>H protons of the protein molecule, so much so that nearly 90% of the expected peaks could be resolved. This is indeed suggestive of a well-folded structure for the molecule and is encouraging for attempting 3D structure determination by NMR.

The next step is to identify residue specific peaks in the spectra. In this sense, Trp residues show great promise, and they are of special interest in the present case because of their hypothesized involvement in specific DNA binding. The NH proton in the indole ring of the side chain of Tryptophan residue resonates at a distinctly low-field position (> 10 ppm) in the spectrum; consequently, the cross-peak between NH (ring) and C<sup>7</sup>H (ring) protons in the residue appears in a distinct region and can be easily identified. Figure 4 *a* shows a portion of a NOESY spectrum of the R123 protein from which 9 peaks of Trp can be easily counted, which is the expected number for the protein. Here we assume that all the peaks are intra-residue indole peaks; they are expected to be strong due to their fixed short distance. Many of these correlations are also seen in TOCSY and DQ spectra (data not shown).

Triple quantum-filtered (TQF) COSY allows unique identification of certain spin systems. For a response to appear in a TQF COSY spectrum, it is necessary that at least three spins are *J*-coupled to each other. While many amino acid spin systems satisfy this condition, the NH-(C<sup>α</sup>H<sub>1</sub>, C<sup>α</sup>H<sub>2</sub>) of Gly and C<sup>β</sup>H<sub>3</sub>-(C<sup>γ</sup>H<sub>1</sub>, C<sup>γ</sup>H<sub>2</sub>) of Ile are of special significance. If a TQF-COSY spectrum is recorded in <sup>1</sup>H<sub>2</sub>O, the fingerprint region of the spectrum would contain cross-peaks belonging exclusively to glycines. Similarly, among the methyl containing residues, only the isoleucines have the above coupling

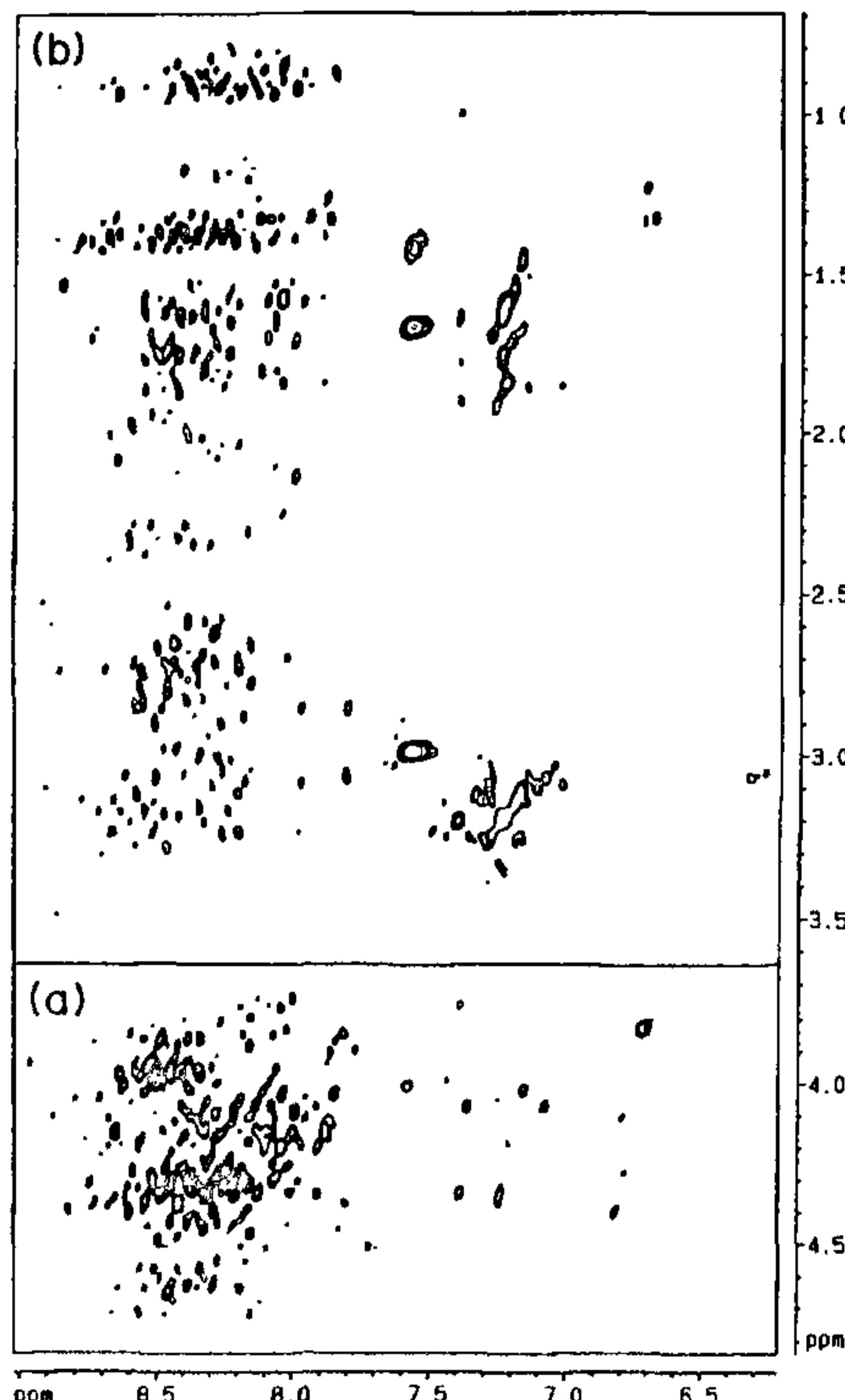
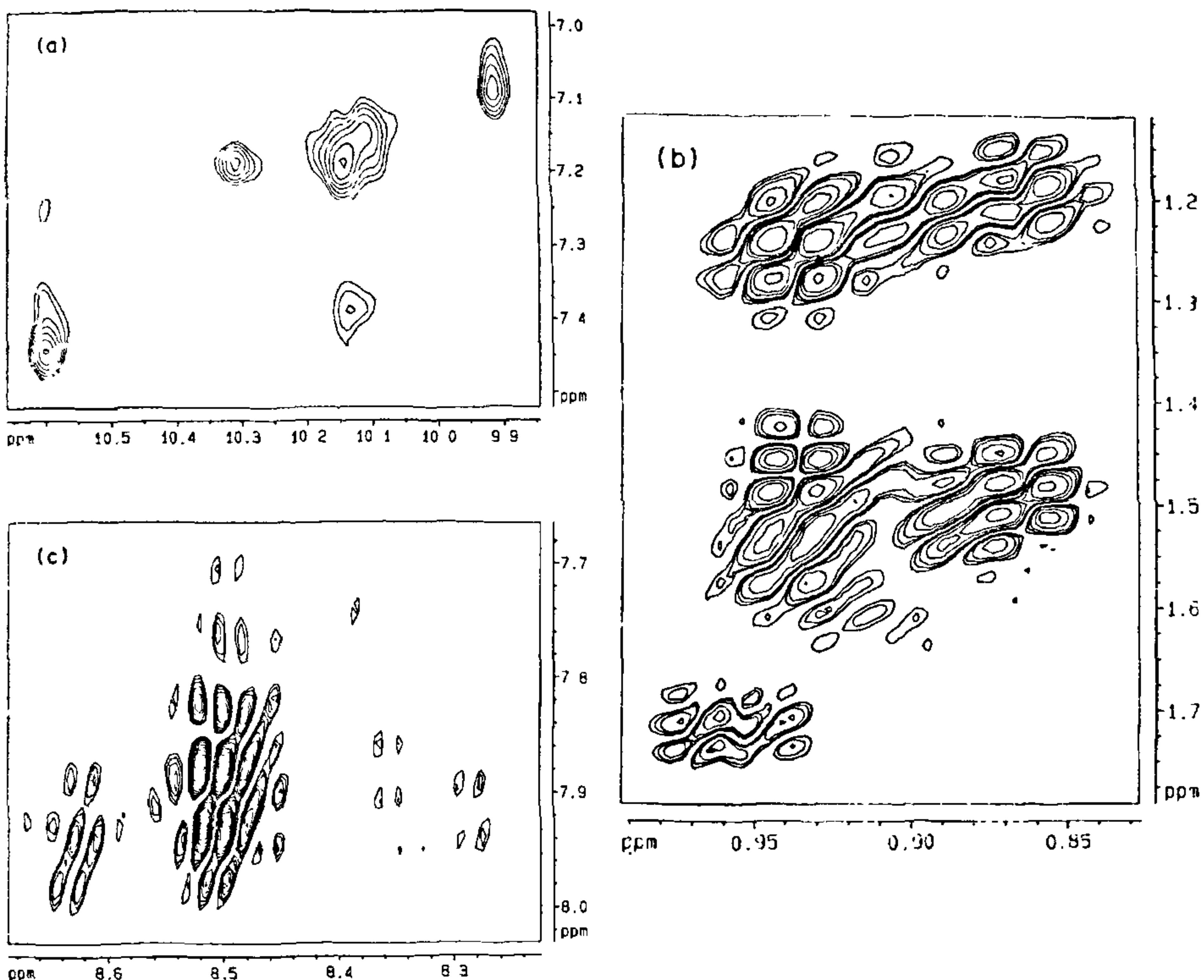


Figure 3. NH region of the 500 MHz clean TOCSY spectrum of the R123 protein in 90% <sup>1</sup>H<sub>2</sub>O/10% <sup>2</sup>H<sub>2</sub>O, at 289 K. The spectrum was acquired with a spin lock period of 60 ms *a*, Fingerprint region containing 154 NH-C<sup>α</sup>H peaks *b*, Region containing cross-peaks between NH and the rest of the aliphatic protons



**Figure 4.** Portions of various 500 MHz 2D spectra recorded for R123 protein at 289 K to identify the specific spin systems. *a*, Portion of the NOESY spectrum ( $t = 60$  ms) of the protein in 90%  $^1\text{H}_2\text{O}/10\% ^2\text{H}_2\text{O}$  showing peaks due to indole NH for the 9 Tryptophan residues. *b*, Portion of the 3QF-COSY spectrum of the protein in  $^2\text{H}_2\text{O}$  containing the  $\text{C}^{\alpha}\text{H}_1\text{-C}^{\alpha}\text{H}_1$  and  $\text{C}^{\alpha}\text{H}_2\text{-C}^{\alpha}\text{H}_2$  cross-peaks for the isoleucine residues. Both positive and negative levels have been plotted. *c*, Portion of the double quantum spectrum of the protein in 90%  $^1\text{H}_2\text{O}/10\% ^2\text{H}_2\text{O}$  showing the remote peaks at  $\omega_1 = \text{NH}$ ,  $\omega_1 = \text{H}_{\alpha 1} + \text{H}_{\alpha 2}$ , for the glycine residues.

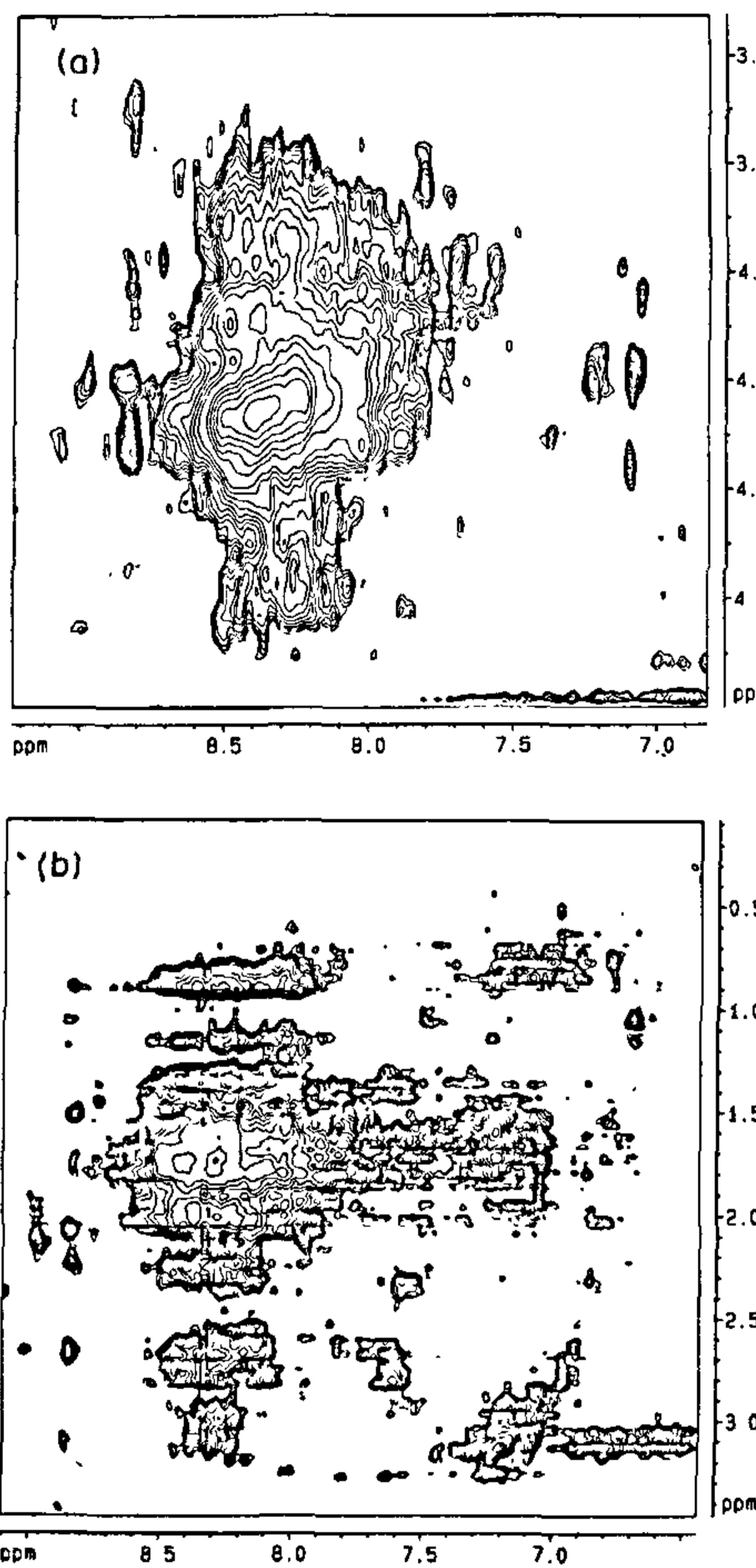
pattern and only these peaks will be present in the  $\text{CH}_3\text{-C}^{\alpha}\text{H}$  cross-peak region of the spectrum. Figure 4 *b* shows the above Ile cross-peaks in the TQF COSY spectrum recorded in  $^2\text{H}_2\text{O}$ . We were able to count here 11 peaks considering the (+---+) patterns along both  $\omega_2$  and  $\omega_1$  axes in each cross-peak. It is difficult to say at this stage whether the  $\text{C}^{\alpha}\text{H}_1$  and  $\text{C}^{\alpha}\text{H}_2$  protons are equivalent or otherwise for each of the residues. Experiments to locate the glycines from  $^1\text{H}_2\text{O}$  spectra were not successful, possibly due to the large linewidth of these protons compared to the  $\text{NH-C}^{\alpha}\text{H}_1$ ,  $\text{C}^{\alpha}\text{H}_2$  coupling constants.

All the nine glycines spin systems were however identified from double quantum (DQ) spectrum in  $^1\text{H}_2\text{O}$  solution. Here again the unique feature of glycines is that, only these produce 'remote' peaks at the chemical

shift positions  $\omega_2 = \text{NH}$  and  $\omega_1 = \text{C}^{\alpha}\text{H}_1 + \text{C}^{\alpha}\text{H}_2$ . Figure 4 *c* shows the portion of the DQ spectrum containing the remote peaks of the glycine residues. The  $\text{C}^{\alpha}\text{H}_1\text{-C}^{\alpha}\text{H}_2$  chemical shift value in conjunction with direct peaks in DQ spectrum and the range of  $\text{C}^{\alpha}\text{H}$  chemical shifts seen in the fingerprint region of the clean-TOCSY spectrum, allow identification of the individual Gly  $\text{C}^{\alpha}\text{H}$  chemical shifts. In most cases the two  $\text{C}^{\alpha}\text{H}$  protons have been found to be nearly equivalent.

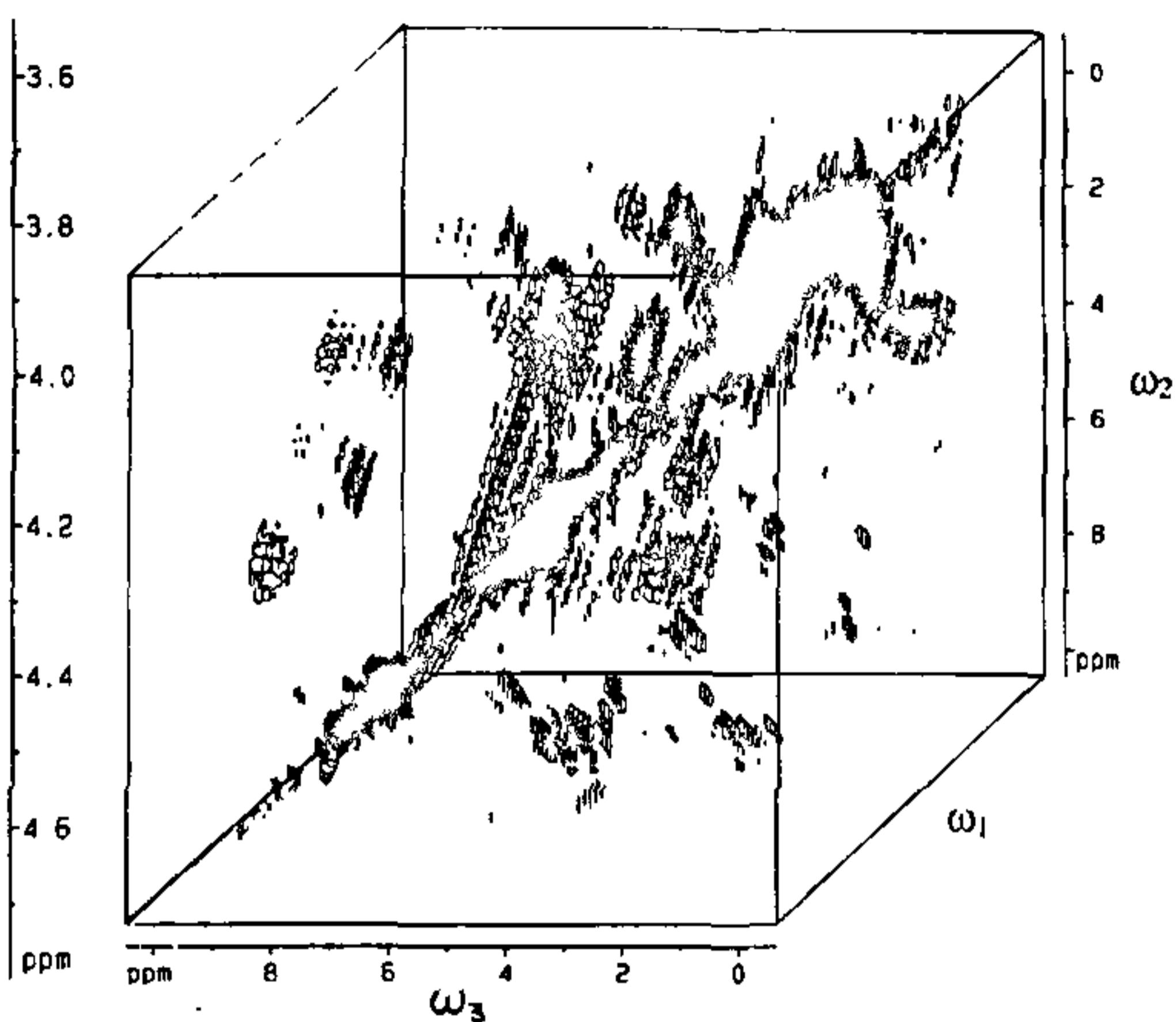
#### Extension into the third dimension

The next level of characterization of the protein, namely structure determination, requires complete sequence-



**Figure 5.** Portions of the NOESY spectrum (mixing time  $t_1 = 60$  ms) of the R123 protein in 90%  $^1\text{H}_2\text{O}/10\% ^2\text{H}_2\text{O}$ , at 289 K. *a*, Fingerprint region; *b*, region containing the cross-peaks between NH and aliphatic protons other than  $\text{C}'\text{H}$ .

specific assignment of all the protons and quantification of the interactions between the assigned protons. Following established procedures, which have worked very well in many small proteins, the 2D NOESY experiment constitutes a crucial technique in this respect. Cross-peaks in NOESY spectra provide a



**Figure 6.** 3D NOESY-TOCSY spectrum of the R123 protein in 90%  $^1\text{H}_2\text{O}/10\% ^2\text{H}_2\text{O}$ , at 289 K. The spectrum was acquired in a period of 1 day only with 512 points along  $t_3$  and 64 points each along  $t_2$  and  $t_1$  directions.

fingerprint of the structure, and these are used for resonance assignments as well. However, with increasing sizes of the proteins, the overlap problems become more and more severe and individual cross-peaks cannot be discerned. This is true in the present case of R123 as well as is shown in Figure 5. Then, it is absolutely essential to add new dimensions to NMR experiments, either by providing additional nuclei (heteronuclei such as  $^{15}\text{N}$  or  $^{13}\text{C}$ ) and/or exploiting different types of  $^1\text{H}$ - $^1\text{H}$  interactions to spread the cross-peaks onto different planes. In this sense, 3D NOESY-TOCSY is a useful experiment and allows NOE and J-coupling correlations to be used along orthogonal axes in the 3D spectrum. We have successfully recorded a 3D NOESY-TOCSY spectrum of the R123 protein in an affordable short period of one day. This spectrum shown in Figure 6, has NOE information in the  $\omega_1$ - $\omega_3$  planes and TOCSY information in the  $\omega_2$ - $\omega_3$  plane. The peaks which are useful for sequential assignments are the equivalents of the 'fingerprint' region in 2D spectra. For specific assignments, the individual planes need to be plotted separately and analysed. We do indeed observe that there are fewer peaks in each plane, clearly establishing the effective enhancement of dispersion of the cross-peaks.

Detailed analysis of all the NMR spectra is in progress.

2 Sakura, H., Kanci-Ishii, C., Nagase, T., Nakagoshi, H., Gonda, T. J. and Ishii, S., *Proc Natl Acad Sci USA*, 1989, **86**, 5758-5762.

3 Biedenkapp, H., Borgmeyer, U., Sippel, A. E. and Klempnauer, K.-H., *Nature*, 1988, **335**, 835-837.

4 Howe, K. M., Reakes, C. F. L. and Watson, R. J., *EMBO J.*, 1990, **9**, 161-169.

5 Oehler, T., Hannelore, A., Biedenkapp, H. and Klempnauer, K.-H., *Nucl Acids Res*, 1990, **18**, 1703-1710.

6 Klempnauer, K.-H. and Sippel, A. E., *EMBO J.*, 1987, **6**, 2719-2725.

7 Ibanez, C. E., Gracia, A., Strober-Grasser, U. and Lipsick, J. S., *J Virol*, 1988, **62**, 4398-4402.

8 Ogata, K., Hojo, H., Aimoto, S., Nakai, T., Nakamura, A. S., Ishii, S. and Nishimura, Y., *Proc Natl Acad Sci USA*, 1992, **89**, 6428-6432.

9 Frampton, J., Leutz, A., Gibson, T. J. and Graf, T., *Nature*, 1989, **342**, 134.

10 Studier, F. W., Rosenberg, A. H., Dunn, J. J. and Duddendroff, J. W., *Methods Enzymol*, 1990, **185**, 60-89.

11 Sambrook, J., Fritsch, E. F. and Maniatis, T., *Molecular Cloning - a Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

12 Rance, M., Sorensen, O. W., Bodenhausen, G., Wagner, G., Ernst, R. R. and Wuthrich, K., *Biochem Biophys Res Commun*, 1983, **117**, 479-485.

13 Hosur, R. V., Ravikumar, M. and Sheth, A., *J Magn Reson*, 1985, **65**, 375-381.

14 Griesinger, C., Otting, G., Wuthrich, K. and Ernst, R. R., *J Am Chem Soc*, 1988, **110**, 7870-7872.

15 Braunschweiler, L., Bodenhausen, G. and Ernst, R. R., *Mol Phys*, 1983, **48**, 533-560.

16 Rance, M. and Wright, P. E., *J Magn. Reson*, 1986, **66**, 372-378.

17 Jeener, J., Meier, B. H., Bachmann, P. and Ernst, R. R., *J Chem Phys*, 1979, **71**, 4546-4553.

18 Griesinger, C., Sorenson, O. W. and Ernst, R. R., *J Magn Reson*, 1989, **84**, 14-63.

19 Drobny, C., Pines, A., Sinton, S., Weitekamp, D. and Wemmer, D., *Symp Faraday Soc.*, 1979, **13**, 49-55.

ACKNOWLEDGEMENTS We thank William Studier for making available the plasmid pET-11d, Rashmi Sood and P. N. Bhavsar for discussions. The facilities provided by the 500 MHz FTNMR National facility, supported by the Department of Science and Technology, Government of India, are gratefully acknowledged. LCP thanks the Lady Tata Memorial Trust for their support.

Received 27 April 1993, revised accepted 10 December 1993

## Crystallization of Azadirachtin-A

T. R. Govindachari<sup>†</sup>, Geetha Gopalakrishnan,  
R. Raghunathan and S. S. Rajan<sup>†</sup>

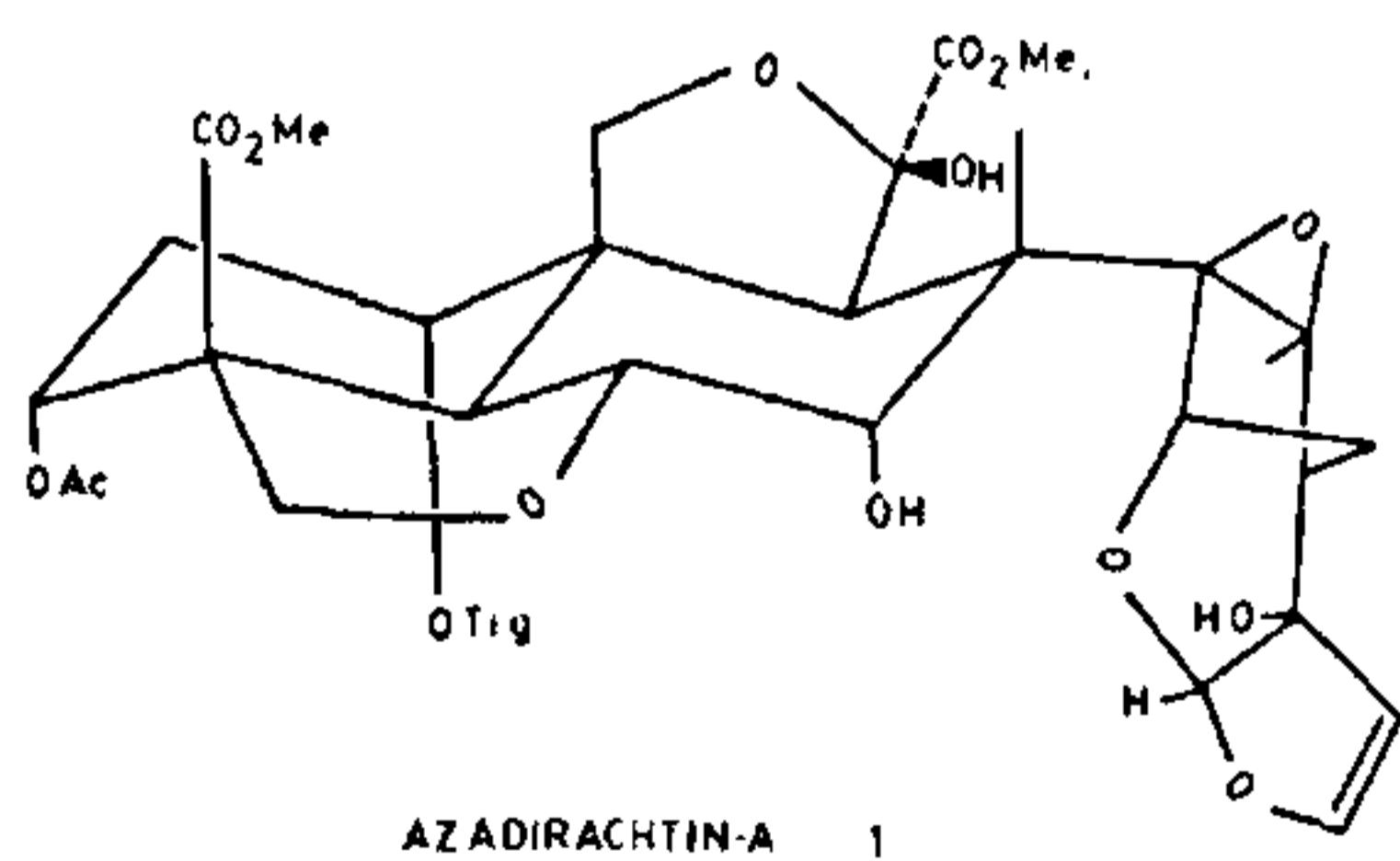
Centre for Agrochemical Research, SPIC Science Foundation, 110,  
Mount Road, Guindy, Madras 600 032, India

<sup>†</sup>Department of Biophysics and Crystallography, Madras University,  
Guindy Campus, Madras 600 025, India

**Azadirachtin-A has been obtained in a crystalline state for the first time. The crystal parameters have been measured by X-ray diffraction. The physical characteristics of crystalline Azadirachtin-A are also recorded.**

AZADIRACHTIN-A is the most abundant of the various azadirachtins A-K, present in the kernels of *Azadirachta indica* A. Juss, along with numerous triterpenoids, mostly meliacins. It has become justly famous because of its remarkable biological activity producing ecdysis inhibition in over two hundred species of insects at a level of 1 ppm and antifeedant activity at 10-20 ppm. This finding has stimulated extensive work not only on *A. indica* (neem)<sup>1</sup> but has also given a boost to the study of many other plants in the hope of obtaining environmentally-friendly agents for the control of insect pests.

Azadirachtin-A was first isolated by Butterworth and Morgan<sup>2</sup> in 1968, who established the molecular formula C<sub>35</sub>H<sub>44</sub>O<sub>16</sub> and delineated important structural features. However, it took nearly twenty years and work in many laboratories for arriving at the correct structure (1) principally by NMR studies, which have been summarized<sup>3,4</sup>. The correct structure was arrived at independently by Kraus *et al.*<sup>5</sup> and Ley and co-workers<sup>6</sup> and the latter also carried out an X-ray crystal structure analysis of detigloyl-22,23-dihydroazadirachtin<sup>7</sup>. It has been reported that Azadirachtin-A is a microcrystalline substance from which crystals suitable for X-ray studies could not be obtained and hence the compulsion to carry out the studies on a derivative.



<sup>†</sup>TRG dedicates this paper to his teacher Prof. B. B. Dey, who loved crystals greatly

In an effort in this laboratory to obtain Azadirachtin-A of the highest attainable purity, conditions for obtaining the substance in a crystalline form have been developed. Azadirachtin-A obtained by the improved procedure developed<sup>8</sup> in this laboratory (this material will be referred to as HPM) was employed in these studies. When HPM was dissolved in the minimum quantity of boiling carbon tetrachloride<sup>9</sup> and set aside overnight, a microcrystalline powder was obtained (Figure 1a). However, when an excess of carbon tetrachloride was employed, approximately 1.2 ml per mg of HPM and the solution set aside overnight, sheaves of long needles arranged in circles were obtained (Figure 1d). After drying at 110° in *vacuo* (0.1 mm) for 8 h, the crystals had m.p. 159-160°.  $[\alpha]_D^{24.8} = -60^\circ$  (c = 0.5, CHCl<sub>3</sub>). The crystals gave <sup>1</sup>H-NMR spectrum identical with spectra reported in earlier studies. The hplc chromatogram was also identical with those of samples (microcrystalline) determined to be 98% pure by NMR spectrum. Unfortunately the crystals were found to occlude carbon tetrachloride in varying amounts in different batches, despite prolonged drying in *vacuo* at 110°. The crystals cannot be termed 'pure'.

However, crystallization of HPM from ethyl acetate-hexane gave crystals which were again sheaves of long needles, arranged in circular clusters. It was possible to pick out several crystals for X-ray studies. One such crystal on examination by X-ray diffraction gave the following data: crystal dimension 0.5 × 0.05 × 0.1 mm belongs to tetragonal system, space group P4<sub>1</sub> with cell parameters  $a = b = 20.960(5)$  Å and  $c = 9.145(4)$  Å. Using these crystals, satisfactory diffraction patterns were obtained and the structure of Azadirachtin-A has been determined. A full paper furnishing the X-ray data will be published soon.

These crystals (obtained by crystallization from ethyl acetate) after drying at 30° in *vacuo* (0.1 mm) had m.p. 174° and  $[\alpha]_D^{24.8} = -71.4^\circ$  (c = 0.21, CHCl<sub>3</sub>). In hplc analysis using a RP-8 column, with acetonitrile-water (28:72) as mobile phase<sup>10</sup> and peak detection, at 215 nm, no impurities could be detected. In retrospect, it appears fortunate that Azadirachtin-A was not obtained in a crystalline condition earlier, as otherwise the outstanding and extensive NMR studies for structure elucidation would not have been undertaken. It would have been merely a case of rationalizing NMR data for a structure solved by X-ray crystallography, as is the case with numerous complex natural products.

It may also be added that Azadirachtin-H isolated in this laboratory<sup>8</sup> and whose structure was elucidated by NMR<sup>11</sup> gave crystals which are tetragonal in space group 14 with cell parameters  $a = b = 29.145(3)$  Å and  $c = 7.695(1)$  Å. The structure has been determined by X-ray diffraction and the results will be published.