Amino acid selective 'unlabelling' for residue-specific NMR assignments in proteins

H. S. Atreya and K. V. R. Chary*

Department of Chemical Sciences, Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400 005, India

A novel methodology for sequence-specific resonance assignments in proteins, using amino acid selective 'unlabelling' is presented. The strategy is based on selective unlabelling of amino acid residues in uniformly or fractionally ^{13}C or/and ^{15}N labeled proteins, which simplify the multi-dimensional heteronuclear NMR spectra. This aids in sequence-specific resonance assignments of both backbone and side-chain nuclei. The methodology has been demonstrated by unlabelling Lys residues in a 15 kDa calcium-binding protein from *Entamoeba histolytica* (*Eh*-CaBP).

STRUCTURE determination of large proteins (> 10 kDa) using triple resonance NMR techniques has been greatly aided by the ability to label these macromolecules with ^{13}C and ^{15}N. Whether it is a labelled or an unlabelled protein, sequence-specific resonance assignments remain an important and essential step towards its complete three-dimensional (3D) structural characterization[1]. Since the last decade, several double and triple resonance experiments have been proposed to carry out sequence-specific ^{1}H, ^{13}C and ^{15}N NMR assignments in isotope labelled proteins[2]. Despite the demonstrated utility of such techniques for the structural characterization of proteins, one encounters several problems in the resonance assignment procedure. In principle, it should be possible to walk all along the backbone of the polypeptide chain starting at the C-terminal and ending at the N-terminal of a given protein, by making use of various backbone nuclei that participate in the magnetization transfer. However, in practice, when these techniques are applied to large proteins with molecular weights in excess of 15 kDa, rapid relaxation rates of the nuclei may result in the broadening of several cross peaks, thus hampering the complete sequence-specific resonance assignments. Pro residues which lack ^{1}H^{N} further aggravate the assignment problem. This prompts one to have as many good starting points as possible along the polypeptide chain of a given protein. Ala, Gly, Ser and Thr have been the most easily identifiable amino acid residues, primarily because of their characteristic ^{13}C$^{\alpha}$ and ^{13}C$^{\beta}$ chemical shifts[3]. As evident from Figure 1 *a*, Gly (^{13}C$^{\alpha}$) always resonates up-field of 50 ppm in a region well separated from the ^{13}C$^{\alpha}$ chemical shifts of all other residues and thus helps in their identification. On the other hand, Ala (^{13}C$^{\beta}$) and Ser (^{13}C$^{\beta}$)/Thr (^{13}C$^{\beta}$) resonate less than 24 ppm and more than 58 ppm, respectively, in regions well separated from ^{13}C$^{\beta}$ chemical shifts of all other residues (Figure 1 *b*) and thus help in their unambiguous identification. This characterization is based on the complete chemical shift data of proteins available with BioMagResBank (BMRB)[4]. Further, it is interesting to note that, on an average, the percentage composition of Ala, Gly, Ser and Thr residues amounts to as much as 25% (Figure 2). Thus, in principle, it should be a straight-forward approach to complete the sequence-specific resonance assignments with these residues as starting points. In practice, however, even with four triple resonance spectra such as CBCANH (ref. 5), CBCA(CO)NH (ref. 6), HNCO (ref. 7) and HN(CA)CO (ref. 8), which provide information about H^{N}, ^{15}N, ^{13}C$^{\alpha}$, ^{13}C$^{\beta}$ and ^{13}C′ chemical shifts, the automatic resonance assignment success rate turns out to be less than 50% (Table 1). Besides Ala, Gly, Ser and Thr residues, if one can identify some

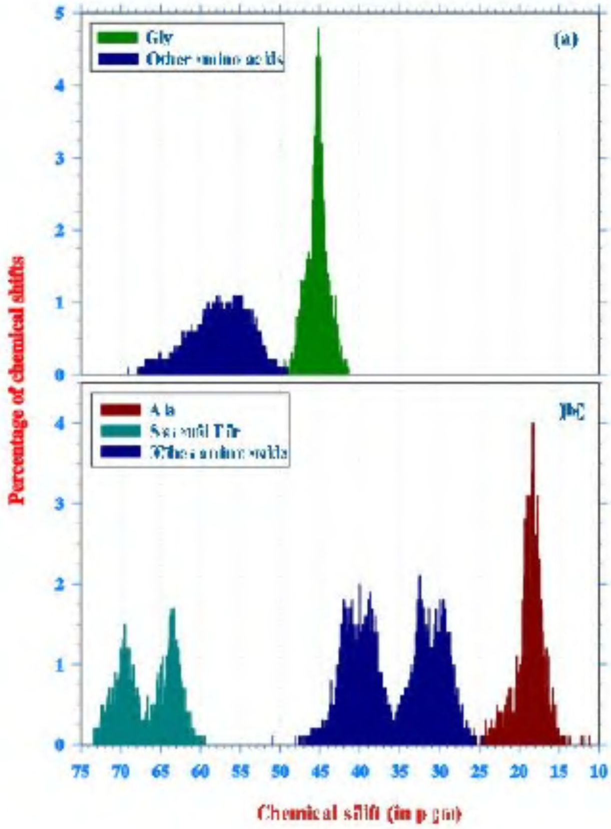

Figure 1. Distribution of (***a***) ^{13}C$^{\alpha}$ and (***b***) ^{13}C$^{\beta}$ chemical shifts of various amino acid residues using the complete chemical shift data of proteins derived from BMRB (http://www.bmrb.wisc.edu). The histograms depict the percentage of amino acids having a particular chemical shift within a range of 0.1 ppm. The total number of chemical shifts analysed in the case of ^{13}C$^{\alpha}$ and ^{13}C$^{\beta}$ spins was ~ 25,000 and ~ 21,000, respectively.

*For correspondence. (e-mail: chary@tifr.res.in)

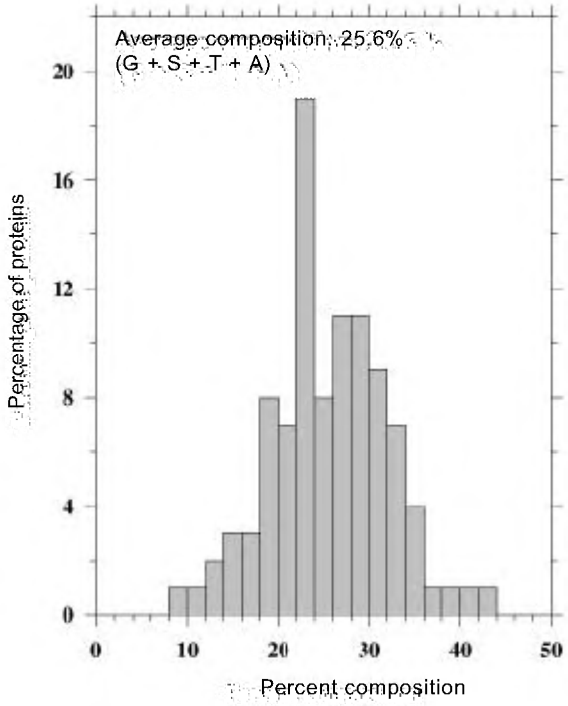

Figure 2. Total percentage composition of Gly, Ser, Thr and Ala residues taken together vs the percentage of proteins. Primary sequences of 100 proteins chosen randomly from BMRB and ranging from 50 to 370 amino acid residues in length were analysed.

Table 1. Percentage of sequence-specific resonance assignments obtained in *Eh*-CaBP with different number of identified amino acid residues

Identified amino acid residues	Percentage of residues assigned* (out of 134 residues)	Percentage of correct assignments	Percentage of correct assignments (out of 134 residues)
G, S/T, A	63.1	80.1	50.5
G, S, T, A	63.4	83.4	52.9
G, S, T, A, K	81.3	84.1	68.4
G, S, T, A, K, L	89.1	89.4	79.7

*Assignments were obtained in the case of *Eh*-CaBP, using an in-house developed algorithm (a modified version of TATAPRO[3]).

more amino acid residues, it has been observed in the case of a 15 kDa calcium-binding protein from *Entamoeba histolytica* (*Eh*-CaBP) using an in-house developed algorithm (modified version of TATAPRO[3]) that, the success rate increases reasonably and the percentage of incorrect assignments reduces proportionately (Table 1). Thus, there is a need for an unambiguous identification of as many peaks as those belonging to specific amino acid residues, other than those which are easily identifiable in various triple resonance spectra. In this direction, amino acid-specific labelling has been used by several researchers[9,10]. In such a procedure, a specific amino acid type in a protein is selectively labelled by feeding the host micro-organism with the desired isotopically labeled (^{15}N or/and ^{13}C) amino acid, while supplying the rest of the amino acids in the unlabelled form[9,10]. Such a specific amino acid residue labelling approach results in direct sequence-specific resonance assignment of the nuclei belonging to that particular amino acid residue, if it occurs only once in the protein primary sequence. If the labelled amino acid residue occurs more than once in the primary sequence, the assignment would then be residue-specific and would provide alternative starting points in sequence-specific resonance assignments as discussed earlier. This methodology, which has been used in selective labelling of several proteins, becomes prohibitively expensive when more than one amino acid residue has to be labelled.

In this communication, an alternate and inexpensive methodology for amino acid residue-specific resonance assignments in proteins is described. The strategy is based on selective unlabelling of amino acid residues in uniformly or fractionally ^{13}C or/and ^{15}N labelled proteins, which simplify the multi-dimensional heteronuclear NMR spectra. This aids in sequence-specific resonance assignments of both backbone and side-chain nuclei. In this approach, a particular amino acid in a protein is 'unlabelled' by feeding the host micro-organism with ^{15}N labelled ammonium chloride or/and ^{13}C labelled glucose as the sole source of nitrogen and carbon, respectively, along with the desired amino acid to be assigned, in unlabelled form. This renders the desired amino acid residue in the protein unlabelled, as a result of which cross peaks due to these residues are not observed in any of the double and triple resonance spectra. A comparison of such a spectrum with that of a control spectrum involving a uniformly ^{15}N or/and ^{13}C labelled protein then enables one to distinguish peaks, and hence the corresponding chemical shifts of nuclei, belonging to the unlabelled amino acid residues. The methodology is demonstrated by the identification of peaks arising from all the Lys residues present in *Eh*-CaBP.

Recombinant proteins generally are enriched isotopically with ^{13}C and ^{15}N by growing the microbial host in a M9 salt medium, supplemented with a ^{15}N labelled ammonium chloride as the sole source of nitrogen or/and ^{13}C labelled glucose as the sole source of carbon[9,10]. *Eh*-CaBP was cloned and over-expressed in *E. coli* BL21(DE3) strain containing pET-3c expression system, the protocol for which has been described earlier[11,12]. Uniformly ^{15}N labelled *Eh*-CaBP was produced using a minimal medium for the bacterial growth having the following composition: M9 salts[13] supplemented with 0.250 g of $MgSO_4.2H_2O$, 0.015 g of $CaCl_2$ and containing 1.0 g of $^{15}NH_4Cl$ per litre of culture as sole source of ^{15}N and 4.0 g/l of ^{12}C-D-glucose as the sole source of carbon. Cells were induced at mid-log phase (O.D. ~ 0.60) with IPTG and grown for 4 h. In order to

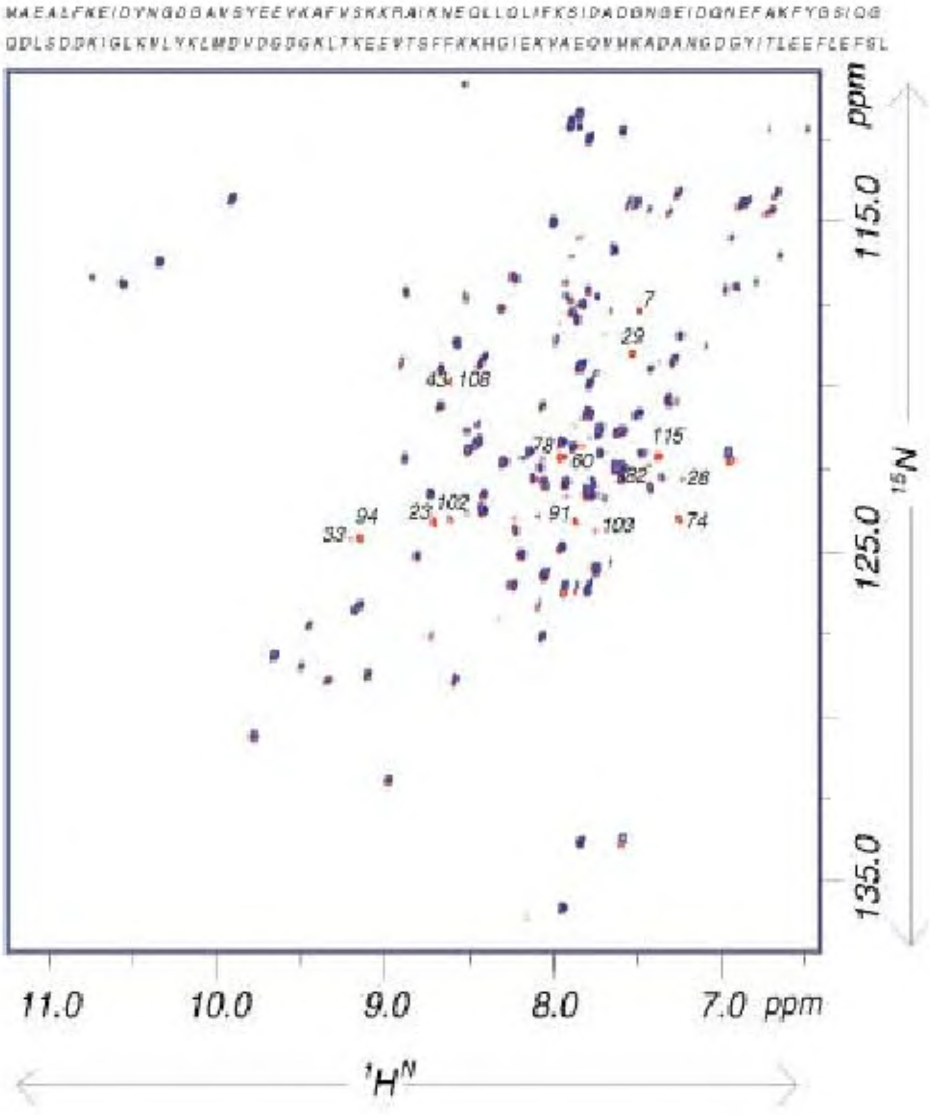

Figure 3. Superimposed 2D [^{15}N-^{1}H] HSQC spectra of uniformly ^{15}N labelled (red colour) and Lys unlabelled–^{15}N labelled *Eh*-CaBP (blue colour) recorded in a mixed solvent of 90% H_2O and 10% ^2H_2O at 35°C and pH = 6.5. Experimental parameters were as follows: τ is 5 ms, recycle delay 1 s, 64 scans/t_1 increment. Time domain data points were 100 and 4096 along t_1 and t_2, respectively. The ^{1}H carrier frequency was kept at the water resonance (4.68 ppm) and ^{15}N carrier frequency was at the centre of the amide nitrogen region (123.8 ppm). The data were multiplied with a sine bell window function shifted by $\pi/3$ and a Gaussian resolution enchancement window function along the t_1 and t_2 axes, respectively, and zero-filled to 2048 and 4096 data points along t_1 and t_2 axes, respectively, prior to 2D-FT. The digital resolution along the ω_1 and ω_2 axes corresponds to 0.83 Hz/pt and 1.95 Hz/pt, respectively. Cross peaks corresponding to 16 Lys residues, which are absent in the Lys-unlabelled spectrum clearly show up. The primary sequence of the protein is displayed on the top.

unlabel all the Lys residues in *Eh*-CaBP, the protein was over-expressed using the same medium described above along with the unlabelled lysine to a final composition of 0.5 g/l. Further, 0.5 g of unlabelled lysine was added at the time of induction along with IPTG. *Eh*-CaBP was expressed to the extent of ~30% of the total cell proteins. The purity of the protein was checked by SDS-PAGE. The yield of uniformly labelled *Eh*-CaBP was ~60 mg of purified protein per litre of culture. Expectedly, the yield turned out to be more in the medium containing lysine. This methodology of selective unlabelling thus requires unlabelled amino acids and $^{15}NH_4Cl$, as against labelled amino acids for selective labelling which is highly expensive.

NMR experiments were carried out on a Varian Unity + 600 NMR spectrometer equipped with a pulsed-field-gradient unit and triple resonance probe with actively shielded Z-gradients, operating at ^{1}H frequency of 600.051 MHz. 2D [^{15}N-^{1}H] HSQC (ref. 14) measurements were performed with a sample of 0.6 ml of

1 mM *Eh*-CaBP in 30 mM $CaCl_2$ and 50 mM deuterated TRIS buffer, pH = 6.5 and temperature of 35°C, in a mixed solvent of 90% H_2O and 10% ^2H_2O. Data transformation and processing were done on Silicon Graphics workstation (R10000-based Indigo II Solid Impact Graphics) using the FELIX97 software (Microsoft Inc, USA). Other experimental conditions used in recording the spectra are described in the caption of Figure 3.

Figure 3 shows the superimposed 2D [^{15}N-^{1}H] HSQC spectrum of uniformly ^{15}N labelled *Eh*-CaBP and Lys unlabelled-^{15}N labelled *Eh*-CaBP. As evident, the cross peaks that are present in the control experiment (red colour) but not in Lys-unlabelled experiment (blue colour), correspond to Lys residues and satisfy our previous resonance assignments[12]. Thus, all cross peaks belonging to the 16 Lys residues could be identified unambigously. However, it may not be always true that such unlabelling helps in a straightforward residue-specific assignment. In the event of simultaneous degeneracy in the ^{15}N and H^N chemical shifts, it is impossible to decipher the absence or presence of a peak. In such a situation, one can record a 3D HNHA (ref. 15) or ^{15}N-edited 3D TOCSY/NOESY (ref. 16) to resolve the ambiguity.

Finally, what is the effect of amino acid metabolism in *E. coli* on such unlabelling? Biosynthesis of amino acids in bacteria is known to be regulated at the level of enzymatic activity and at the level of gene expression. Transaminase catalysed nitrogen exchange leads to isotopic dilution and mis-incorporation of the label at undesired sites[9,10]. This is true even for the unlabelling strategy outlined in this communication. Therefore, the conversion of unlabelled amino acid(s), via the various metabolic pathways, to other amino acid(s) is undesirable. In the case of prototrophic *E. coli* strains, those amino acids which do not metabolize to other amino acids (e.g. Lys, Arg, Met, Pro and Cys) and which are simultaneously abundant in the protein can be chosen for unlabelling. On the other hand, mis-incorporation of unlabelled amino acids to their metabolic derivative can be prevented by using appropriate *E. coli* genetic backgrounds where such mis-incorporations are curtailed.

The approach outlined here can thus be used for residue-specific assignments in proteins, which form an important input for complete sequence-specific resonance assignments. Moreover, this is the best method for simultaneous unlabelling of specific amino acids in a protein, while fractionally labelling (15% ^{13}C labelling)[17] the rest. Such procedure can be used to simplify the spectra of large proteins and thus enable stereospecific resonance assignments[17] of methyl groups in Val and Leu residues (to be published elsewhere). Efforts are further on to simplify ^{13}C/^{15}N edited 3D TOCSY and 3D NOESY spectra using such unlabelling strategies to derive more structural restraints leading to high precision protein structures.

1. Wüthrich, K., in *NMR of Proteins and Nucleic Acids*, Wiley Pulishers, New York, 1986, pp. 1–292.
2. Bax, A. and Grzesiek, S., *Acc. Chem. Res.*, 1993, **26**, 131–138.
3. Atreya, H. S., Sahu, S. C., Chary, K. V. R. and Govil, G., *J. Biomol. NMR*, 2000, **17**, 125–136.
4. Seavey, B. R., Farr, E. A., Westler, W. M. and Markley, J. L., *ibid.*, 1991, **1**, 217–236.
5. Wittekand, M. and Mueller, L., *J. Magn. Reson.*, 1993, **B101**, 201–205.
6. Grzesiek, S. and Bax, A., *ibid.*, 1992, **96**, 432–440.
7. Clubb, R. T., Thanabal, V. and Wagner, G., *J. Biomol. NMR*, 1992, **2**, 203–210.
8. Kay, L. E., Ikura, M., Tschudin, R. and Bax, A., *J. Magn. Reson.*, 1990, **89**, 496–514.
9. Muchmore, D. C., McIntosh, L. P., Russell, C. B., Anderson, D. E. and Dahlquist, F. W., *Methods Enzymol.*, 1989, **B177**, 44–73.
10. McIntosh, L. P. and Dahlquist, F. W., *Q. Rev. Biophys.*, 1990, **23**, 1–38.
11. Prasad, J., Bhattacharya, S. and Bhattacharya, A., *Cell. Mol. Biol. Res.*, 1993, **39**, 167–175.
12. Sahu, S. C., Atreya, H. S., Chauhan, S., Bhattacharya, A., Chary, K. V. R. and Govil, G., *J. Biomol. NMR*, 1999, **14**, 93–94.
13. Sambrook, J., Fritsch, E. F. and Maniatis, T., in *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989, vols 1–3, 2nd ed.
14. Muller, L., *J. Am. Chem. Soc.*, 1979, **101**, 4481–4484.
15. Vuister, G. W. and Bax, A., *ibid.*, 1993, **115**, 7772–7777.
16. Marion, D., Kay, L. E., Sparks, S. W., Torchia, D. A. and Bax, A., *ibid.*, 1989, **111**, 1515–1517.
17. Neri, D., Szyperski, T., Otting, G., Senn, H. and Wüthrich, K., *Biochemistry*, 1989, **28**, 7510.

ACKNOWLEDGEMENTS. The facilities provided by the National Facility for High Field NMR, supported by Department of Science and Technology, Department of Biotechnology, Council of Scientific and Industrial Research, and Tata Institute of Fundamental Research, Mumbai, are gratefully acknowledged. We thank Prof. Alok Bhattacharya, Jawaharlal Nehru University, New Delhi for providing us the *Eh*-CaBP expression plasmid and Prof. B. J. Rao, TIFR, Mumbai for his suggestions. This communication is dedicated to Prof. G. Govil on his 60th Birthday.

Received 18 April 2000; accepted 17 June 2000