*Journal of Biomolecular NMR*, **21:** 269–273, 2001.
*KLUWER/ESCOM*
© 2001 *Kluwer Academic Publishers. Printed in the Netherlands.*

# Efficient identification of amino acid types for fast protein backbone assignments

Horng D. Ou[a], Helen C. Lai[a], Zach Serber[a] & Volker Dötsch[b,*]
[a]*Graduate Group in Biophysics, and* [b]*Departments of Pharmaceutical Chemistry and Cellular & Molecular Pharmacology, University of California San Francisco, San Francisco, CA-94143, U.S.A.*

Received 10 July 2001; Accepted 20 August 2001

*Key words:* automated assignment, calmodulin, HSQC, SAM-domain, selective labeling

## Abstract

We describe a procedure that allows for very efficient identification of amino acid types in proteins by selective $^{15}$N-labeling. The usefulness of selective incorporation of $^{15}$N-labeled amino acids into proteins for the backbone assignment has been recognized for several years. However, widespread use of this method has been hindered by the need to purify each selectively labeled sample and by the relatively high cost of labeling with $^{15}$N-labeled amino acids. Here we demonstrate that purification of the selectively $^{15}$N-labeled samples is not necessary and that background-free HSQC spectra containing only the peaks of the overexpressed heterologous protein can be obtained in crude lysates from as little as 100 ml cultures, thus saving time and money. This method can be used for fast and automated backbone assignment of proteins.

The first step of a protein structure determination by NMR spectroscopy is the assignment of the protein backbone resonances. Since the introduction of heteronuclear multidimensional NMR experiments, this task has been performed by matching resonance frequencies between pairs of spectra to obtain intra- and interresidual correlations (Clore and Gronenborn, 1991; Ikura et al., 1990; Montelione and Wagner, 1990). In principle, searching for peaks with matching resonance frequencies should be amenable to computer-assisted automation, and over the past decade, several different groups have developed computer algorithms aimed at providing either fully automated or semi-automated assignment procedures (Buchler et al., 1997; Leutner et al., 1998; Lukin et al., 1997; Moseley et al., 2001; Zimmerman et al., 1997, 1994). However, the algorithms suffer from the fact that NMR spectra contain certain confounding features and artifacts. These include overlapping peaks, missing peaks, multiple resonance frequencies caused by the presence of different protein conformations, and spectral artifacts such as $T_1$-noise. While an experienced NMR spectroscopist can identify the cause of these artifacts and solve these problems, they pose an enormous challenge for computer algorithms. Recent approaches using genetic algorithms (Bartels et al., 1997) have improved the reliability of these automated assignment procedures. However, in particular for larger proteins, problems still exist. One way to improve the method is to incorporate information about the amino acid type into the assignment protocol. For some amino acids such as alanine and threonine the combination of their $^{13}$Cα and $^{13}$Cβ chemical shifts can be used for identification. However, for many other amino acid types, chemical shifts are not unique and can only be used to establish a likelihood factor that a peak in a spectrum belongs to a certain amino acid (Grzesiek et al., 1993).

This problem can be solved by two different approaches that exclusively select information about certain amino acid types. The first method relies on designing NMR pulse sequences that take advantage of the homo- and heteronuclear spin-spin coupling networks in the individual amino acids (Dötsch et al., 1996; Dötsch and Wagner, 1996; Feng et al., 1996;

[*] To whom correspondence should be addressed. E-mail: volker@picasso.ucsf.edu

Rios et al., 1996; Schubert et al., 1999, 2001). These experiments suppress coherence transfer pathways of all amino acids except for the selected ones and the resulting spectra contain only peaks of the targeted amino acids. The second approach relies on the selective incorporation of $^{15}$N-labeled amino acids into the protein (Hibler et al., 1989; Lee et al., 1995; McIntosh et al., 1990; Muchmore et al., 1989). If auxotrophic bacterial strains are used or if the overexpression conditions are carefully chosen (Lee et al., 1995), this method results in NMR spectra that contain exclusively peaks of the selected amino acids. The biggest advantage of the first method is that it calls for only one $^{13}$C/$^{15}$N doubly-labeled sample. In contrast, selective $^{15}$N-labeling requires the production of one sample per labeled amino acid type. In particular, if the purification procedure is complicated, the selective-labeling method can become very time-consuming. In addition, several of the selective $^{15}$N-labeled amino acids are quite expensive. On the other hand, some of the filter elements that have to be incorporated into the non-selective NMR pulse sequences in order to make them amino-acid type specific are relatively long and do not work very well for larger proteins. Moreover, some of the editing schemes do not completely suppress coherences of the non-selected amino acids, leading to potential breakthrough peaks that can cause considerable problems for automated assignment procedures (Dötsch et al., 1996).

We wanted to investigate whether the method of selective labeling with $^{15}$N-labeled amino acids could be transformed into a more efficient tool. Our approach is based on an earlier observation by Clore and Gronenborn who reported that [$^{15}$N-$^{1}$H]-HSQC spectra of proteins overexpressed in bacteria can be obtained in crude cell lysates (Gronenborn et al., 1996). Similar results were reported recently (Almeida et al., 2001). In addition, we have shown that [$^{15}$N-$^{1}$H]-HSQC spectra can be obtained with overexpressed proteins inside living E. coli cells (Serber et al., 2001; Serber et al., in press), in which only protein resonances of the overexpressed protein are visible and no resonances of endogenous bacterial proteins could be detected. The only other peaks in these spectra are peaks with a much narrower line width than the resonances of the overexpressed protein. The chemical shift range of these additional peaks is between 8.0-8.5 ppm, which in combination with the narrow line width suggests that these peaks are non-protein resonances, presumably caused by $^{15}$N-incorporation into small molecules, e.g. amino acids. In the experiments by Gronenborn and Clore these additional peaks were eliminated by buffer exchange of the crude cell lysate (Gronenborn and Clore, 1996). During our 'in-cell' NMR experiments with NmerA and calmodulin, however, we discovered that it is possible to obtain 'in-cell'-[$^{15}$N-$^{1}$H]-HSQC spectra with selectively $^{15}$N-labeled samples without any background signals (Serber et al., 2001). Based on these findings, we reasoned that we could use selective $^{15}$N-labeling as an efficient tool to obtain amino acid type identification by completely eliminating any purification steps.

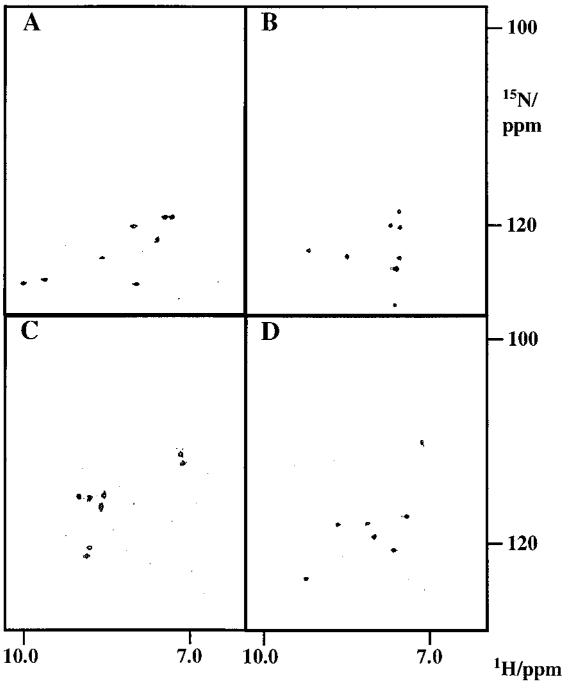

*Figure 1.* [$^{1}$H,$^{15}$N]-HSQC spectra of calmodulin, selectively $^{15}$N-labeled with different amino acids: (A) isoleucine, (B) lysine, (C) phenylalanine and (D) valine. With the exception of the peak at $^{1}$H: 7.6 ppm, $^{15}$N: 129 ppm in the lysine spectrum, all peaks represent calmodulin resonances.

To test this hypothesis, we used calmodulin and expressed samples that were selectively $^{15}$N-labeled on lysine, phenylalanine, valine or isoleucine. For each experiment a 250 ml bacterial culture was grown to an optical density of 0.8 in regular LB medium, harvested by centrifugation, resuspended in 100 ml of fresh medium and induced with 0.4 mM IPTG. Lysine-labeled samples were expressed in minimal M9 medium supplemented with 220 mg/l of $^{15}$N-labeled lysine in regular BL21 bacteria for five hours at

37 °C. All other selectively labeled samples were expressed in the strain DL39 avtA::Tn5 λDE3 (LeMaster and Richards, 1988; McIntosh and Dahlquist, 1990) that is auxotrophic for Asp, Val, Ile, Leu, Phe and Tyr. The bacteria were induced at room temperature for two days in minimal M9 medium that was supplemented with 500 mg/l Ala, 400 mg/l Asp, 230 mg/l Val, Ile, Leu, 170 mg/l Tyr and 130 mg/l Phe. For each selectively labeled sample, half of the above stated amount was used for the $^{15}$N-labeled amino acid while the other amino acids were unlabeled. The cells were harvested by centrifugation and resuspended in 2 ml buffer (10 mM Hepes pH 6.3, 100 mM KCl, 12 mM $CaCl_2$, 10 mM β-mercaptoethanol, protease inhibitors (Complete-tablets, Böhringer-Mannheim)) and incubated for one hour with lysozyme. Cell lysis was completed by a freeze-thaw method during which the resuspended cells were frozen with liquid nitrogen and then thawed in water three times. We decided to lyse the bacteria instead of measuring the spectra inside living cells because our earlier work demonstrates that cell lysis reduces the line width considerably (Serber et al., 2001). In order to reduce the viscosity of the solution, DNAse I was added. Finally, the cell debris was removed by centrifugation, 500 μl of the supernatant were transferred into a 5 mm NMR tube, and 50 μl of $D_2O$ were added. Figure 1 shows all spectra obtained with the selectively labeled calmodulin samples. In all but one spectra only peaks belonging to the overexpressed protein are visible. The one exception is the lysine-labeled sample, which shows, in addition to the protein peaks, one extra peak ($^1$H: 7.6 ppm, $^{15}$N: 129 ppm) that most likely represents a metabolic product of lysine. This demonstrates that with the exception of this one peak in the lysine-spectrum, artifact-free HSQC spectra of amino-acid type selectively labeled proteins can be obtained from crude cell lysates even without buffer exchange. All spectra were measured with 64 complex points in the indirect dimension. For the lysine-labeled sample 8 scans per increment were recorded, yielding a total measurement time of 22 min. The lower expression level of the DL39 bacteria reduced the signal-to-noise ratio of the HSQC spectra measured with 8 scans. Therefore, these spectra were measured with 32 scans per increment. All expected peaks in the obtained spectra were visible with the exception of one lysine and one phenylalanine. The higher number of scans extended the measurement time to 1.5 h.

To further test the robustness of our method we applied it to one of our current structural projects in the lab, the SAM domain of the p53 homologue p63. We produced four samples, selectively $^{15}$N-labeled on lysine, phenylalanine, leucine and isoleucine following the protocol described above. The resulting HSQC spectra, shown in Figure 2, confirm that all visible peaks belong to the overexpressed protein, in this case the SAM domain, with the exception of one additional peak in the lysine spectrum representing again a metabolic product of lysine. All spectra of the SAM domain were recorded with 30 complex points in the indirect dimension. The lysine sample was measured with eight scans per increment and all other spectra with 32 scans. The total measurement time per spectrum were 10 min and 41 min, respectively.

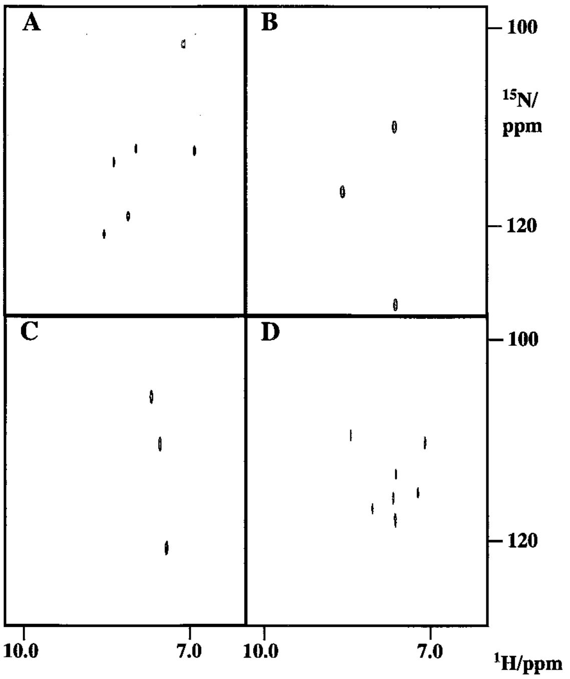

*Figure 2.* Spectra of the p63 SAM domain selectively labeled with $^{15}$N-labeled amino acids in 10 mM Hepes buffer, pH 7.0, 5 mM DTT: (A) isoleucine, (B) lysine, (C) phenylalanine and (D) leucine. All expected resonances with the exception of two leucines that are located in the unstructured N- and C-termini are present.

To evaluate the usefulness of these spectra for automatic assignment procedures, we have compared the chemical shifts of the peaks in the amino-acid selective labeled spectra with the corresponding chemical shifts in the [$^{15}$N,$^1$H]-HSQC spectra of the purified proteins. A very good agreement was obtained for most of the resonances with an average deviation of 0.015 ppm in the proton dimension and 0.05 ppm in the $^{15}$N-dimension for the SAM domain and 0.02 ppm and 0.07 ppm in the calmodulin spectra. The slightly higher deviations in the calmodulin spectra as compared to the spectra of the SAM domain are most likely caused by differences in the buffer composition. For two peaks in the SAM domain spectra, however, much larger changes occurred that reached 0.05 ppm in the proton dimension and 0.16 ppm in the $^{15}$N dimension. The corresponding values reached 0.07 ppm and 0.3 ppm for four peaks in the calmodulin spectra. Despite these larger deviations, most of these peaks could be unambiguously assigned in the spectra of the purified protein because these peaks are located in a relatively empty spectral region. On the other hand, some of the peaks which showed a very good chemical shift agreement could not be unambiguously assigned because of peaks overlap with other resonances. In total, 5 out of 8 Ile, all Lys, 5 out of 8 Phe and 5 out of 7 Val could be unambiguously assigned in the calmodulin spectra and 5 out of 6 Ile, 1 out of 2 Lys, all Phe and all Leu in the SAM domain spectra.

As demonstrated, no artifact peaks (with the exception of one peak in the lysine spectrum) are detectable in these spectra. The robustness of the method makes selective $^{15}$N-labeling very attractive as the basis for automated assignment. Incorporation of definite amino acid type identifications as opposed to likelihoods will make computer-based automated assignment protocols more reliable. The ability to measure artifact free HSQC spectra in crude cell lysates without buffer exchange makes the method also very efficient and abolishes the argument that selective $^{15}$N-labeling is too time-consuming to be practical. Since the amount of labeled amino acid that is used for these experiments is very small, the costs of the experiment is reduced at the same time.

However, crucial for the success of this method is a good overexpression level of the protein. For poorly expressed proteins, the crude cell lysate of a larger culture would have to be concentrated and part of the high-throughput character of the method would be lost. The expression level of the DL39 strain is significantly lower than the expression level of BL21 bacteria. Other groups have shown that BL21 cells can be used for amino acid selective labeling if the expression protocol is carefully adjusted (Lee et al., 1995). In our experiments we have used (with the exception of lysine) auxotrophic strains because we were concerned about potential background peaks caused by cross-labeling of other types of amino acids. Such cross-labeling would severely limit the usefulness of this approach for automated assignment procedures. The method could be even more efficient if auxotrophic strains with a higher expression level than the DL39 strain were used.

Another critical parameter is how closely the buffer conditions in the crude lysate resemble the buffer in the purified sample. If the crude cell lysate contains components that interact with the overexpressed protein, substantial differences in chemical shift can occur that could make an unambiguous assignment impossible. Our results have demonstrated that, however, even with excellent correspondence between the chemical shifts not all resonances can be unambiguously assigned due to peak overlap. Nonetheless, we were able to assign a majority of the peaks and a reduced number of resonances for which the amino acid type is known will still provide valuable information for an automatic assignment program. Moreover, in the case of peak overlap, a computer algorithm could narrow down the possible assignments since it can rely on the fact that one of the (for example) two overlapping peaks has to be a certain amino acid type.

In conclusion, we have demonstrated that artifact free HSQC spectra of selectively labeled proteins can be obtained in crude bacterial lysates without buffer exchange. Our current procedure allows us to obtain spectra of at least 5–6 samples within one day. We believe that this number is sufficient to make fully automated backbone assignment protocols with high reliability and efficiency possible.

## Acknowledgements

We are grateful to Wesley Straub for providing the calmodulin plasmid. Financial support was provided by NIH (GM56531-02), the UCSF Innovations in Basic Sciences Award and the Sandler Family supporting foundation. H.D.O., H.C.L., and Z.S. were supported by an NIH Training Grant (GM08284).

## References

Almeida, F.C.L., Amorim, G.C., Moreau, V.H., Sousa, V.O., Creazola, A.T., Americo, T.A., Pais, A.P.N., Leite, A., Netto, L.E.S., Giordano, R.J. and Valente, A.P. (2001) *J. Magn. Reson.*, **148**, 142–146.

Bartels, C., Güntert, P., Billeter, M. and Wüthrich, K. (1997) *J. Comput. Chem.*, **18**, 139–149.

Buchler, N.E.G., Zuiderweg, E.R.P., Wang, H. and Goldstein, R.A. (1997) *J. Magn. Reson.*, **125**, 34–42.

Clore, G.M. and Gronenborn, A.M. (1991) *Science*, **252**, 1390–1399.

Dötsch, V., Oswald, R.E. and Wagner, G. (1996) *J. Magn. Reson.*, **110**, 304–308.

Dötsch, V. and Wagner, G. (1996) *J. Magn. Reson. B*, **111**, 310–313.

Feng, W., Rios, C.B. and Montelione, G.T. (1996) *J. Biomol. NMR*, **8**, 98–104.

Gronenborn, A.M. and Clore, G.M. (1996) *Protein Sci.*, **5**, 174–177.

Grzesiek, S. and Bax, A. (1993) *J. Biomol. NMR*, **3**, 185–204.

Hibler, D.W., Harpold, L., Dell'Acqua, M., Pourmotabbed, T., Gerlt, J.A., Wilbe, J.A. and Bolton, P.H. (1989) *Meth. Enzymol.*, **177**, 74–86.

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

Lee, K.M., Androphy, E.J. and Baleja, J.D. (1995) *J. Biomol. NMR*, **5**, 93–96.

LeMaster, D. and Richards, F.M. (1988) *Biochemistry*, **27**, 142–150.

Leutner, M., Gschwind, R.M., Liermann, J., Schwarz, C., Gemmecker, G. and Kessler, H. (1998) *J. Biomol. NMR*, **11**, 31–43.

Lukin, J.A., Grove, A.P., Talukdar, S.N. and Ho, C. (1997) *J. Biomol. NMR*, **9**, 151–166.

McIntosh, L.P. and Dahlquist, F.W. (1990) *Quart. Rev. Biophys.*, **23**, 1–38.

Montelione, G.T. and Wagner, G. (1990) *J. Magn. Reson.*, **87**, 183–188.

Moseley, H.N.B., Monleon, D. and Montelione, G.T. (2001) *Meth. Enzymol.*, **339**.

Muchmore, D.C., McIntosh, L.P., Russell, C.B., Anderson, D.E. and Dahlquist, F.W. (1989) *Meth. Enzymol.*, **177**, 44–73.

Rios, C.B., Feng, W., Tashiro, M., Shang, Z. and Montelione, G.T. (1996) *J. Biomol. NMR*, **8**, 345–350.

Schubert, M., Oschkinat, H. and Schmieder, P. (2001) *J. Magn. Reson.*, **148**, 61–72.

Schubert, M., Smalla, M., Schmieder, P. and Oschkinat, H. (1999) *J. Magn. Reson.*, **141**, 34–43.

Serber, Z., Keatinge-Clay, A.T., Ledwidge, R., Kelly, A.E., Miller, S.M. and Dötsch, V. (2001) *J. Am. Chem. Soc.* **123**, 2446–2447.

Serber, Z., Ledwidge, R., Miller, S.M. and Dötsch, V. (2001) *J. Am. Chem. Soc.*, in press.

Zimmerman, D.E., Kulikowski, C.A., Feng, W., Tashiro, M., Powers, R. and Montelione, G.T. (1997) *J. Mol. Biol.*, **269**, 592–610.

Zimmerman, D.E., Kulikowski, C.A., Wang, L.L., Lyons, B.A. and Montelione, G.T. (1994) *J. Biomol. NMR*, **3**, 241–256.