J|A|C|S
COMMUNICATIONS

Published on Web 04/01/2004

A Combinatorial Selective Labeling Method for the Assignment of Backbone Amide NMR Resonances

Martin J. Parker,*,† Marc Aulton-Jones,† Andrea M. Hounslow, and C. Jeremy Craven*

Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield S10 2TN, U.K.

Received November 14, 2003; E-mail: m.j.parker@leeds.ac.uk; c.j.craven@shef.ac.uk

One of the unique features of NMR spectroscopy is that it can provide a set of residue-specific probes with which to study protein−ligand interactions.[1] The prime source for these probes is the $^{1}H-^{15}N$ correlation spectrum, ligand interaction sites being inferred from perturbations of $^{1}H-^{15}N$ cross-peaks upon titration of ligand into the sample. Clearly, this requires the assignment of cross-peaks to specific residues, which remains a significant challenge despite many technological developments. Recently, several groups have presented methods using selective amino acid labeling to accelerate the identification of ligand binding sites;[2−4] however, these techniques only probe a very limited number of residues. Here we present a scheme for the efficient assignment of a much larger number of $^{1}H-^{15}N$ cross-peaks simultaneously using only five selectively labeled samples that can be rapidly and cost-effectively produced in parallel in a commercially available in vitro translation system. We term the method combinatorial selective labeling (CSL).

The CSL method is based upon the dual amino acid-selective $^{13}C/^{15}N$ labeling technique,[5,6] in which the carbons of one amino acid type *a* are labeled with ^{13}C, and the amide nitrogens of another amino acid type *b* are labeled with ^{15}N. If an (*a*)*b* pair exists only once in the protein sequence then a single cross-peak will appear in the $^{1}H-^{15}N$ 2D HNCO spectrum, and the NH group of the residue type *b* can be unambiguously assigned. In a simple approach, the identification of each different (*a*)*b* pair would require a separate sample, demanding a prohibitively large total number of such samples. Our novel approach is to use a much smaller number of samples produced with different combinations of labeled amino acids, using the resulting patterns of cross-peak intensities across these samples to differentiate each (*a*)*b* pair.

The experiments that we present here required the production of five protein samples. Each sample contains a different combination of 16 labeled amino acid types, which are individually either 100% $^{13}C/^{15}N$ labeled or $50\%^{15}N/50\%^{14}N$ labeled. The labeling scheme is shown in Figure 1. For each sample, two NMR spectra are acquired: a $^{1}H-^{15}N$ HSQC spectrum and a $^{1}H-^{15}N$ 2D HNCO spectrum. Following normalization of the spectra, comparison of the relative peak intensities in the HSQC spectra establishes the amino acid type of each peak. Arginine residues, for example, would have a peak intensity pattern in samples 2−5 of $^{1}/_{2}$:1:$^{1}/_{2}$:$^{1}/_{2}$, respectively. The 16 amino acid types chosen here can be assigned in these four samples as there are 2^4 (= 16) such patterns. For a particular cross-peak, the amino acid type of the preceding residue in the sequence is established by examining the presence or absence of peaks in the five 2D HNCO spectra. For instance, if the preceding residue is a threonine, then the HNCO cross-peaks will show the pattern absent:absent:present:absent. Therefore, all 16 × 16 possible amino acid pairs are identifiable simultaneously from these five samples (15 × 16 if proline is included in the set, as in Figure 1).

Type	Sample				
	1	2	3	4	5
G	●●	◑	◑	◑	◑
V	●●	◑	◑	◑	●●
T	●●	◑	◑	●●	◑
S	●●	◑	◑	●●	●●
R	●●	◑	●●	◑	◑
Q	●●	◑	●●	◑	●●
N	●●	◑	●●	●●	◑
M	●●	◑	●●	●●	●●
L	●●	●●	◑	◑	◑
K	●●	●●	◑	◑	●●
I	●●	●●	◑	●●	◑
F	●●	●●	◑	●●	●●
D	●●	●●	●●	◑	◑
C	●●	●●	●●	◑	●●
A	●●	●●	●●	●●	◑
P	●	●	●	●	●

Figure 1. The labeling pattern for the CSL method. Red and blue filled circles denote 100% ^{13}C and ^{15}N labeling. Blue half-filled circles denote samples in which a 50:50 mix of ^{15}N and ^{14}N amino acids is used. The 50% "background" ^{15}N labeling is necessary to allow the observation of an HNCO cross-peak to be independent of the nature of the *b* residue in an (*a*)*b* pair. The particular choice of these 16 amino acids was made partly to simplify the development of the method and also due to problems of supply and solubility of the amino acids histidine, tryptophan, and tyrosine. Glutamate cannot be labeled in the standard RTS system due to the presence of a large excess of unlabeled glutamate in the buffer. Sample 1 is necessary to provide reference intensities for comparison with spectra from the other samples. The full 20 amino acids could be accommodated by either introducing a sixth sample or by using the same labeling pattern for pairs (or more) of uncommon amino acids. The use of an in vitro translation system alleviates problems of label scrambling[6,7] and allows the rapid and cost-effective production of the five samples required, it being possible to produce and purify five his-tagged samples in the system in 2 days.

If a pair appears *n* times in the sequence, then *n* peaks will appear in these spectra with the same intensity pattern, and the assignment will be *n*-fold degenerate.

We tested the method on a truncated version of the cycle3 version of green fluorescent protein (GFP) from *Aequorea victoria*,[8] a 27 kDa protein. The samples were labeled in the pattern depicted in Figure 1, using the rapid translation system 500 *Escherichia coli* HY kit (Roche Diagnostics Ltd). The HSQC spectrum of GFP with all 16 of the chosen set of amino acids fully $^{13}C/^{15}N$ labeled (sample 1) is shown in Figure 2A. Two cross-peaks are shown in detail in Figure 2B. The HSQC intensities correspond to residue types F (peak *x*) and L (peak *y*) (see Figure 1), and the HNCO patterns correspond to preceding residue types S and I, respectively. Therefore, the two cross-peaks can be assigned to amino acid pairs (S)F and (I)L, respectively. The amino acid pair SF occurs only once in the sequence, and therefore peak *x* can be assigned to F100. The amino acid pair IL occurs twice in the sequence, and thus peak *y* is assigned to either L15 or L137. The published assignment[8] confirms these results.

This protein does not display ideal NMR characteristics. Under the sample conditions used here, we estimate the correlation time to be ca. 21 ns, and furthermore, the protein displays regions of

† Present address: Astbury Centre for Structural Molecular Biology, School of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT, U.K.

10.1021/ja039601r CCC: $27.50 © 2004 American Chemical Society

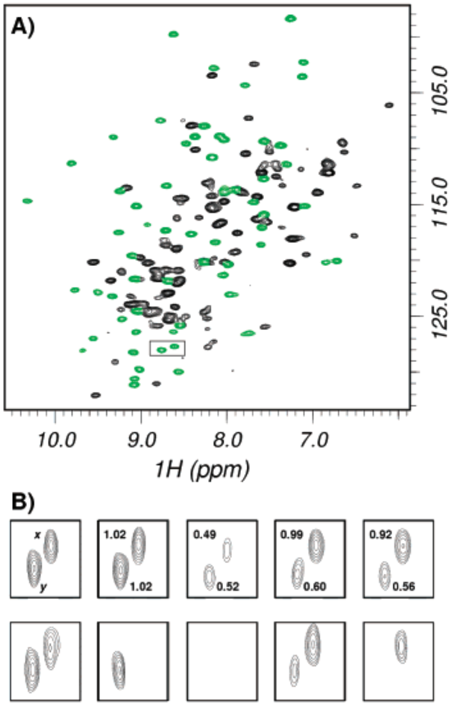

Figure 2. (A) $^{1}H-^{15}N$ HSQC spectrum of GFP labeled according to sample 1 of Figure 1. The peaks colored in green showed the correct pattern of intensities, based on cutoffs relative to sample 1 of 0.75 and 0.25 for the HSQC and HNCO spectra, respectively. The majority of the remaining peaks are either severely overlapped, weak, or are side-chain amide peaks. The sample concentration was ca. 100 uM (ca. 1 mg yield from a 10 mL reaction). The total experiment time for each HSQC spectrum was 20 h, and for each HNCO spectrum the total experiment time was 40 h. The spectra were acquired on a 500 MHz Bruker Avance spectrometer, equipped with a cryoprobe. The sample conditions were as described in ref 8. (B) Detail from the five HSQC spectra (upper five panels; from left to right, samples 1−5) and five HNCO spectra (lower five panels). The region shown corresponds to the rectangular region marked in (A). The intensities (relative to sample 1) are marked for the two peaks *x* and *y* in the HSQC spectra for samples 2−5.

missing and weakened resonances[8] due to conformational exchange broadening. In this demonstration of the feasibility of the method, 61 residues could be assigned to the correct (*a*)*b* pair. It is anticipated that in proteins with better NMR characteristics (giving improved sensitivity and resolution), and with refinement of amino acid formulations, that the assignment rate would be greatly increased.

This method is qualitatively different from the traditional methods based on HNCA-type experiments that link spins and map them onto the primary sequence. Such methods are sensitive to the completeness of the data for all residues, since incorrect or incomplete information about one residue can confound assignment of another residue. In contrast, the assignment of a particular cross-peak in our method to a particular (*a*)*b* pair depends solely on information about that single cross-peak. (Cases of absolute resonance overlap would be detectable by the deviation of the relative peak intensities in the HSQC spectrum from 0.5 or 1.) Furthermore, our method uses two of the most sensitive NMR experiments ($^{1}H-^{15}N$ HSQC and $^{1}H-^{15}N$ 2D HNCO), which makes it applicable to proteins suffering from poor solubility or tumbling characteristics. The analysis of the data in our technique is also much less demanding and could be carried out by someone with limited NMR experience. Thus, the method opens up opportunities for applying NMR to systems that give less than ideal spectra, to studying groups of related proteins rapidly in parallel, and to making NMR more accessible to non-NMR specialists. All the elements of the technique from protein production, NMR data collection, through data analysis are ideal for robotic and computational automation.

A clear disadvantage of the method is that only partial assignments are obtained. For instance, in GFP 43% of residues are in unique (*a*)*b* pairs, 35% are in (*a*)*b* pairs that occur twice in the protein sequence, 14% are in (*a*)*b* pairs that occur three times in the protein sequence, and 8% exhibit higher degeneracy. While incomplete, these assignments would still provide a very large number of residue-specific probes, which will in general be randomly distributed in the protein. In many applications, an incomplete assignment will be quite adequate, for instance, to resolve the choice between different models. Reese and Dötsch[4] showed recently that binding information could be obtained by studying the patterns of chemical shift perturbations in a number of single amino acid-type labeled samples. By searching for a region with an appropriate amino acid composition they were able to identify a binding site, which they proposed could be verified using dual selective amino acid labeling to identify a particular single amino acid. Our technique provides much more detailed information with a comparable number of samples.

The method that we have presented here could be developed in various ways. In cases of severe spectral overlap the number of amino acids that are ^{15}N labeled could be reduced, while retaining the full pattern of ^{13}C labeling. This would reduce the number of potentially overlapping cross-peaks in the $^{1}H-^{15}N$ planes, without compromising the ability to identify the nature of the preceding residue. Combining the method with segmental isotope labeling[9] would both decrease the spectral complexity and improve the number of unambiguous assignments. The availability of deuterated ^{13}C and ^{15}N amino acids would also greatly increase the range of applicability of this method by making use of TROSY[10] and saturation transfer[11] methods.

Acknowledgment. The work was funded by the BBSRC, U.K. M.J.P. is a BBSRC David Phillips Fellow and University Research Fellow (University of Leeds). We acknowledge the provision of FELIX software from Accelrys. We thank Paul Gillingham and Mark Wells for assistance with the NMR spectroscopy and Sophie Jackson for a clone of GFP.

References

(1) Zuiderweg, E. R. P. *Biochemistry* **2002**, *41*, 1−7.

(2) Weigelt, J.; Dongen, M. v.; Uppenberg, J.; Schultz, J.; Wikstrom, M. *J. Am. Chem. Soc.* **2002**, *124*, 2446−2447.

(3) Guignard, L.; Ozawa, K.; Pursglove, S. E.; Otting, G.; Dixon, N. E. *FEBS Lett.* **2002**, *524*, 159−162.

(4) Reese, M. L.; Dötsch, V. *J. Am. Chem. Soc.* **2003**, *125*, 14250−14251.

(5) Kainosho, M.; Tsuji, T. *Biochemistry* **1982**, *21*, 6273−6279.

(6) Yabuki, T.; Kigawa, T.; Dohmae, N.; Takio, K.; Terada, T.; Ito, Y.; Laue, E. D.; Cooper, J. A.; Kainosho, M.; Yokoyama, S. *J. Biomol. NMR* **1998**, *11*, 295−306.

(7) Zartler, E. R.; Hanson, J.; Jones, B. E.; Kline, A. D.; Martin, G.; Mo, H.; Shapiro, M. J.; Wang, R.; Wu, H.; Yan, J. *J. Am. Chem. Soc.* **2003**, *125*, 10941−10946.

(8) Khan, F.; Stott, K.; Jackson, S. *J. Biomol. NMR* **2003**, *26*, 281−282.

(9) Yamazaki, T.; Otomo, T.; Oda, N.; Kyogoku, Y.; Uegaki, K.; Ito, N.; Ishino, Y.; Nakamura, H. *J. Am. Chem. Soc.* **1998**, *120*, 5591−5592.

(10) Pervushin, K.; Riek, R.; Wider, G.; Wüthrich, K. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 12366−12371.

(11) Takahashi, H.; Nakanishi, T.; Kami, K.; Arata, Y.; Shimada, I. *Nat. Struct. Biol.* **2000**, *7*, 220−223.

JA039601R