A combinatorial selective labelling method for the assignment of backbone NMR resonances

Marc Aulton-Jones and Martin J. Parker

Introduction
NMR spectroscopy is an enormously powerful technique with which to study a wide range of biochemical problems in solution. One of its unique features is that it can provide a set of residue specific probes with which to study protein interactions. The prime source for these probes is the $^1$H-$^{15}$N HSQC spectrum. Interaction sites are inferred from perturbations of crosspeaks upon titration of ligand or unlabelled protein into the sample. These are then mapped on to the experimental structure. The method is particularly suited for studying weakly interacting systems, and has led to an active area called ‘SARS by NMR’, where lead pharmaceutical compounds with affinities as low as 10mM can be screened rapidly for interaction with the target protein.

Application of the method obviously demands the assignment of crosspeaks to specific residues in the protein. Despite many technological developments, this remains a significant challenge, especially for large systems. Recently, several groups have presented methods using selective amino acid labelling to accelerate the identification of protein interaction sites. These techniques only probe a very limited number of residues however. We have developed a Combinatorial Selective Labelling (CSL) method for the efficient assignment of the majority of HSQC crosspeaks using a small number of samples that can be rapidly and cost effectively produced in parallel in a commercially available cell-free system.

The CSL method
The CSL method is based upon the dual amino acid-selective $^{13}$C/$^{15}$N labelling technique, which utilises protein samples in which the main chain carbonyl carbons of one amino acid type (A) are labelled with $^{13}$C, and the amide nitrogens of another amino acid type (B) are labelled with $^{15}$N. The NMR signals of the amino acid residues that possess a $^{13}$CO-$^{15}$N linkage can be extracted on the basis of the $^{13}$C-$^{15}$N spin coupling, i.e. if an (A)B pair exists only once in the sequence then a unique crosspeak will appear in the $^1$H-$^{15}$N 2D HNCO spectrum, and the NH group of the residue type B can be unambiguously assigned. To date these selective labelling methods have only been applied to identify single peaks in $^1$H-$^{15}$N correlation spectra. In a simple approach, the identification of each different AB 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 labelled with different combinations of amino acids, using the resulting patterns of crosspeak intensities across these samples to differentiate each AB pair.

The CSL method requires five protein samples, each containing a different combination of 16 labelled amino acid types (see Fig. 1 (a)). For each sample, a $^1$H-$^{15}$N HSQC spectrum and a $^1$H-$^{15}$N 2D plane of an HNCO spectrum are acquired. Comparison of the relative peak intensities in the HSQC spectra yields the amino acid type of each peak. The 16 amino acid types chosen here can be assigned in the 4 samples as there are $2^4$ (= 16) such patterns (sample 1 is the reference). For a particular crosspeak, one can tell the amino acid type of the preceding residue in the sequence by examining the presence or absence of peaks in the five 2D HNCO spectra. Therefore, all 16x16 possible amino acid pairs are identifiable simultaneously from these five samples. 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 have demonstrated the feasibility of the method using the 27 kDa, beta barrel protein GFP as a model system. The samples were labelled in the pattern depicted in Fig. 1 (a), using the Rapid Translation System (RTS) 500 E. coli HY kit. The HSQC spectrum of GFP with all 16 of the chosen set of amino acids fully $^{13}$C/$^{15}$N labelled (sample 1) is shown in Fig. 1 (b). The majority of the resolved peaks in the HSQC/HNCO spectra display intensity patterns conforming to those expected for the published assignments (green peaks in Fig. 1 (b)), demonstrating that the method is practicable.

![Fig. 1: LHS (a) CSL labelling method. Red and blue filled circles denote 100% $^{13}$C and $^{15}$N labelling. Blue half filled circles denote samples in which a 50:50 mix of $^{15}$N and $^{14}$N labelled amino acids is used. RHS (b) $^1$H-$^{15}$N HSQC spectrum of GFP labelled according to sample 1 in (a). The peaks coloured in green show 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.](image)

Traditional assignment 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. The assignment of a particular crosspeak in our method to a particular (A)B pair depends solely on information about that single crosspeak, which makes our method much less sensitive to crosspeak overlap or missing crosspeaks. Furthermore, and crucially, our method uses two of the most rapid/sensitive NMR experiments, which make it applicable to proteins suffering from poor solubility, tumbling or sample lifetime 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.

**Further studies**

We are currently developing and refining the CSL method. In particular, we are combining it with protein ligation technology to increase the completeness of the assignments and to reduce spectral complexity. Together with recent developments in NMR (specifically the ‘TROSY’ method), this will greatly increase the range of applicability. These developments are best achieved by exploring their application to challenging and medically important problems. We are applying the methods to: (i) the interaction of bactericidal/permeability increasing protein with lipopolysaccharide (LPS; endotoxin), an important part of the innate immune response to bacteria, and (ii) the interactions among those members of the Bcl-2 family of proteins involved in apoptosis; potential targets for cancer therapeutics.
experiments will identify those residues comprising the protein interaction sites in these systems, providing important insights on their biochemistry and aiding in the design and development of drugs.

**Collaborators**
Dr C. Jeremy Craven, University of Sheffield.

**References**

**Funding**
This work was funded by the BBSRC. M.J.P. is a BBSRC David Phillips and University Research Fellow.