J Biomol NMR (2007) 38:151–159
DOI 10.1007/s10858-007-9157-7

ARTICLE

A systematic analysis of backbone amide assignments achieved via combinatorial selective labelling of amino acids

C. Jeremy Craven · Moza Al-Owais · Martin J. Parker

Received: 11 January 2007 / Accepted: 15 March 2007 / Published online: 26 April 2007

Abstract With the advent of high-yield cell-free expressions systems, many researchers are exploiting selective isotope labelling of amino acids to increase the efficiency and accuracy of the NMR assignment process. We developed recently a combinatorial selective labelling (CSL) method capable of yielding large numbers of residue-type and sequence-specific backbone amide assignments, which involves comparing cross-peak intensities in ^{1}H–^{15}N HSQC and 2D ^{1}H–^{15}N HNCO spectra collected for five samples containing different combinations of ^{13}C- and ^{15}N-labelled amino acids [Parker MJ, Aulton-Jones M, Hounslow A, Craven C J (2004) J Am Chem Soc 126:5020–5021]. In this paper we develop a robust method for establishing the reliability of these assignments. We have performed a detailed statistical analysis of the CSL data collected for a model system (the B1 domain of protein G from *Streptococcus*), developing a scoring method which allows the confidence in assignments to be assessed, and which enables the effects of overlap on assignment fidelity to be predicted. To further test the scoring method and also to assess the performance of CSL in relation to sample quality, we have applied the method to the CSL data collected for GFP in our previous study.

M. Al-Owais · M. J. Parker (✉)
Astbury Centre for Structural Molecular Biology, Institute of Molecular and Cellular Biology, University of Leeds, Leeds LS2 9JT, UK
e-mail: m.j.parker@leeds.ac.uk

C. Jeremy Craven
Department of Biotechnology and Molecular Biology, University of Sheffield, Sheffield S10 2TN, UK

Keywords Cell-free protein synthesis · Combinatorial selective labelling · NMR assignment · Protein G · Selective isotope labelling

Introduction

The ^{1}H–^{15}N HSQC (or TROSY) spectrum is the foundation of structural, functional, and dynamics studies of protein molecules. With the availability of efficient cell-free expression systems capable of producing milligram quantities of protein (Spirin et al. 1988; Kigawa et al. 1999; Yokoyama, 2003; Swartz et al. 2004; Jewett and Swartz, 2004; Klammt et al. 2004), a number of groups have exploited selective isotope labelling of amino acid types for the rapid assignment of ^{1}H–^{15}N HSQC cross-peaks (Kigawa et al. 1995; Yakuki et al. 1998; Guignard et al. 2002; Klammt et al. 2004; Ozawa et al. 2004; Shi et al. 2004). Recently, combinatorial selective labelling (CSL) schemes have been devised for providing large numbers of residue-type (Wu et al. 2006) and sequence-specific (Parker et al. 2004) assignments using only a limited number of samples, which can be rapidly and cost-effectively produced in parallel in commercially available cell-free systems.

The CSL method we devised recently (Parker et al. 2004) is based on the dual amino acid-selective $^{13}C/^{15}N$ labelling technique (Kainosho and Tsuji 1982; Yabuki et al. 1998), 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 are extracted on the basis of ^{13}C–^{15}N spin couplings; if an (*a*)*b* pair exists only once in the sequence then a unique cross-peak will appear in the ^{1}H–^{15}N 2D HNCO spectrum,

and the NH group of the residue in the sequence with type (*a*)*b* can be assigned unambiguously. The CSL method requires five protein samples, each containing a different combination of 16 labelled amino acid types which are individually either 100% ^{13}C,^{15}N-labelled or 50% ^{15}N-labelled (see Table 1, for example). 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 can be assigned in the four samples as there are 2^4 100%^{15}N/50%^{15}N labelling patterns (sample 1 is the fully labelled reference). For a particular cross-peak, the amino acid type of the preceding residue in the sequence is obtained by examining the presence or absence of peaks in the five 2D HNCO spectra. Therefore, all 16 × 16 possible amino acid pairs are identifiable simultaneously from these five samples. The use of 100% and 50% ^{15}N labelling (rather than 100% and 0%) is necessary to maintain the possibility of observing H_N detected cross-peaks in the HNCO spectra from all samples (Parker et al. 2004).

We demonstrated the feasibility of the CSL method using the 27 kDa, beta barrel protein GFP as a model system, with samples prepared using the Rapid Translation System (RTS) 500 *E. coli* HY kit (Roche). The poor solution characteristics of GFP, however, precluded a careful investigation of the robustness of our assignments, with respect to inherent spectral noise, the accuracy with which the amino acid isotope mixes can be made, and scrambling and dilution of amino acids in the cell-free system. In this paper we chose a protein with excellent NMR characteristics (the B1 domain of protein G from *Streptococcus*), which has allowed us to address these issues. We have performed a detailed statistical analysis of the CSL data collected for protein G, developing a scoring method that allows the confidence in assignments to be assessed, and which enables the effects of overlap on assignment fidelity to be predicted. This scoring function is further applied to the CSL data collected for GFP.

Table 1 Labelling scheme used for the protein G samples

AA	Sample				
	1	2	3	4	5
Asn	C	N	N	N	N
Tyr	C	N	N	N	C
Met	C	N	N	C	N
Leu	C	N	N	C	C
Phe	C	N	C	N	N
Ile	C	N	C	N	C
Gln	C	N	C	C	N
Asp	C	C	C	C	C
Ser	C	C	N	N	N
Ala	C	C	N	N	C
Val	C	C	N	C	N
Lys	C	C	N	C	C
Thr	C	C	C	N	N
Gly	C	C	C	N	C

''C'' denotes amino acids that are 100% ^{15}N/^{13}C labelled, and ''N'' denotes amino acids that are labelled 50%^{15}N/50%^{14}N (and 100% ^{12}C). Glu was not labelled due to its presence at high concentration in the RTS reconstitution buffer. Trp was not labelled due to prohibitive cost. His was not labelled owing to use of a His tag. Arg, Cys and Pro do not occur in protein G

Materials and methods

Sample preparation

The gene for the B1 domain of protein G from *Streptococcus* was amplified with forward (CGTGATTACCCATGGACACCTACAAACTGATCCTG) and reverse (GTTACCGA AGGGGGTTCTCATCATCATCATCATCATTAACCCGGGATCCGGTAAC) primers designed to add a 6 × His tag at the C-terminus, yielding the final protein sequence: MDTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTEGGSHHHHHH. The forward and reverse primers contained *Nco*I and *Xma*I restriction sites, respectively. Following digestion with the appropriate restriction enzymes (NEB) the gene was cloned into the in vitro expression plasmid pIVEX2.3d (Roche) using standard methods, and sequencing performed to confirm identity. In vitro protein expressions were performed using the RTS 500 *E. coli* HY kit (Roche), according to the manufacturer's instructions, in the RTS ProteoMaster Instrument at 30°C for 24 h. ^{15}N- and ^{13}C/^{15}N-labelled (CK Gas) and unlabelled (Sigma) amino acid stock solutions were made to a concentration of 84 mM in the reconstitution buffer supplied with the RTS 500 kit. For the CSL, 6 ml solutions containing the appropriate mixtures of labelled and unlabelled amino acids at 4.2 mM in reconstitution buffer were prepared (see Table 1). Expressed proteins were purified using HisTrap HP Columns, and desalted and buffer exchanged using PD-10 Columns (GE Healthcare) according to manufacturer's instructions. Purified proteins were concentrated to a final volume of 0.5 ml in 20 mM sodium acetate buffer and 0.01% (v/v) NaN_3, pH 4.3 using 1 kDa cutoff spin concentrators (Centricon). For NMR, D_2O was added to 10% (v/v). The protein concentrations in the CSL samples were ca. 200 μM as assessed by UV absorption at 280 nm (see below for more precise relative quantitation by NMR).

NMR spectroscopy

All spectra were acquired at 25 °C on a Bruker Avance spectrometer operating at 800 MHz. The HSQC spectra were acquired as 150 complex pairs in the indirect dimension (acquisition time = 112 ms), with 8 scans per FID, giving a total experiment time of ca. 40 min. 2D HNCO planes were acquired as 60 complex pairs in the indirect dimension (acquisition time = 22.5 ms as a constant time dimension concurrent with the $^{13}C'$–^{15}N refocusing period), with 64 scans per FID and a total experiment time of ca. 2 h. Spectra were processed into matrices of dimensions 4096 × 4096. Shifted sine bell window functions were applied in each dimension. In the direct dimension the processed acquisition time was reduced to 21 ms to broaden the lines and increase the overlap in the spectra. The assignment of Gronenborn et al. (1991) was obtained from the supplementary deposition to PDB entry 1GB1, and transferred to our construct and conditions using a 3D HNCA spectrum acquired on sample 1.

Extraction of peak intensity ratios

The intensities of peaks in the HSQC and HNCO spectra were determined in Felix, and transferred to UNIX text files for further manipulation using python scripts. In order to convert absolute intensities in the spectra into ratios the following three-step procedure was employed. First, the relative protein concentrations of the five samples were determined by comparing well-resolved upfield shifted resonances in proton 1D spectra. These concentrations were used to apply an intensity correction to the spectra from each of samples 2–5. Next, the intensities of the peaks in the HSQC spectra of samples 2–5 were individually converted to ratios by dividing by the intensity of the corresponding peak in the HSQC spectrum of sample 1. Implicitly, the ratios in sample 1 are precisely 1.0, so that any error in the measurement of the intensity for sample 1 is manifested as a uniform distortion of the ratios for all of samples 2–5. In order to handle the data from all five samples in a more even-handed manner, a final overall scaling was then applied to the five ratio measurements in order to maximise the value of the scoring function for each peak, as described in Results and discussion. The same procedure was applied for intensities measured in the HNCO spectra.

The variance of ratios due to baseline noise was calculated by first measuring the baseline noise in the spectra in regions devoid of peaks. The variance of each individually calculated ratio was then calculated using a Monte Carlo procedure. Random values from a distribution with variance equal to that of the baseline noise measured in the two corresponding spectra were added to the two experimental intensities involved in the calculation of each ratio. This procedure was repeated 100 times and the variance of the resulting distribution of ratios was calculated.

Results and discussion

HSQC and HNCO peak intensity patterns

Five samples of the B1 domain of protein G from *Streptococcus* were prepared using the combinatorial labelling scheme shown in Table 1. For each sample, a ^{1}H–^{15}N HSQC and a 2D ^{1}H–^{15}N plane of a 2D HNCO were acquired. The peak intensities measured in the HSQC and HNCO spectra of the five samples were normalised as described in Materials and methods. The resulting peak intensity ratios are shown in Figs. 1 and 2. In Fig. 1 the data from the HSQC spectra are grouped according to the amino acid type of the residue giving rise to the HSQC

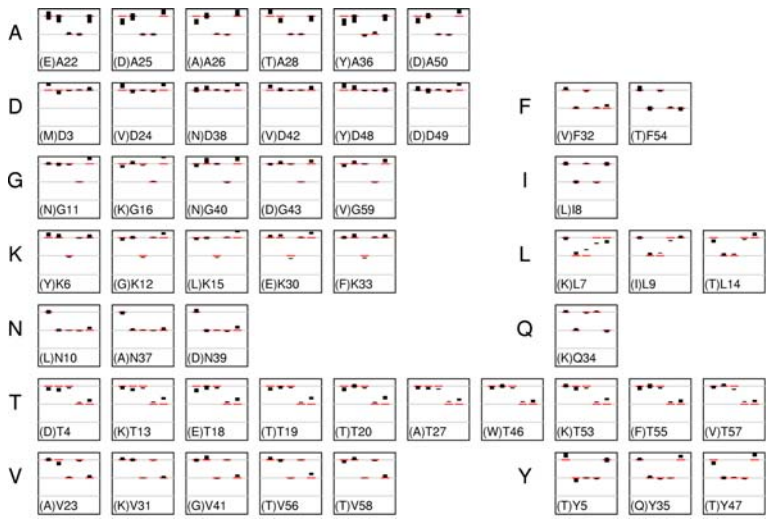

Fig. 1 Peak intensity ratios from HSQC spectra for protein G plotted for samples 1–5 (left to right respectively). The grey lines from top to bottom mark ratios of 0, 0.5 and 1.0, respectively. The observed ratios of peak intensities, scaled as described in the text, are shown by black rectangles. The expected values in each of the samples are shown by red lines. Data involving Asp have been scaled in samples 2–5 as described in the text

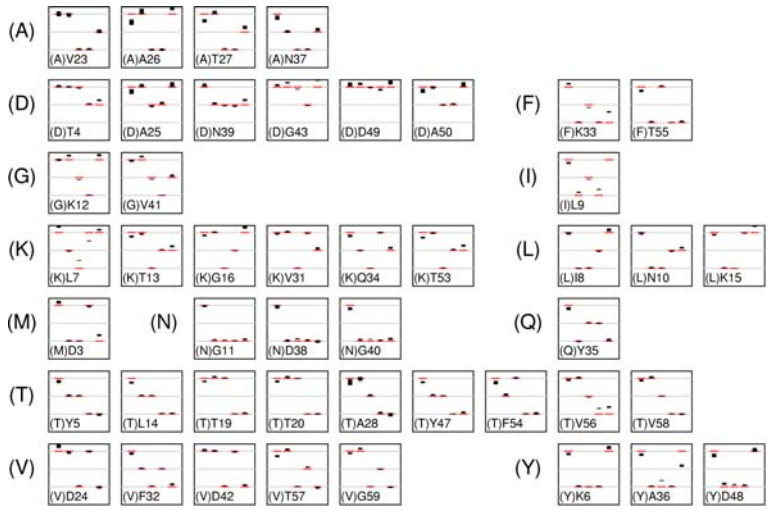

Fig. 2 Peak intensity ratios from HNCO spectra for protein G, plotted as for Fig. 1

peak. With the amino acid pair of the residue denoted (*a*)*b*, as introduced above, red horizontal lines in the figure mark ideal ratios of 1.0 (*b* labelled 100% $^{15}N/^{13}C$) or 0.5 (*b* labelled 50%^{15}N/50%^{14}N). In Fig. 2 the data from the HNCO spectra are instead grouped by the amino acid type of the preceding residue (i.e. *a*), and the red lines mark ideal ratios of 1.0 (*b* labelled 100% $^{15}N/^{13}C$, *a* labelled 100% $^{15}N/^{13}C$), 0.5 (*b* labelled 50%^{15}N/50%^{14}N, *a* labelled 100% $^{15}N/^{13}C$) and 0.0 (*a* labelled 50%^{15}N/50%^{14}N). The error bars in these two figures equate to twice the standard deviation (SD) in intensity ratios, calculated purely on the basis of measurement of the baseline noise in the spectra (see Materials and methods).

Although, as discussed below, the error estimates based solely on baseline noise appear to underestimate the noise in the data, for the vast majority of residues, the pattern of ratios conforms to that expected based on the labelling pattern in Table 1. There are, however, residues for which the pattern is less clear-cut (Leu 7, for example). In order to apply the CSL method to proteins of unknown assignment it is necessary both to understand (as far as is possible) the reasons for deviations from the expected patterns, and to establish a methodology for assessing the reliability of a particular assignment.

Assignment scoring function

In outline, our strategy was as follows. We first assumed that we did not know the assignment of the peaks in the acquired spectra. For a peak observed in sample 1 we then used a weighted sum of squares (χ^2) to define how well the observed pattern of intensities for the corresponding peaks across all five samples matched the pattern expected for the 16×16 possible (*a*)*b* pairs. Our test was not restricted to amino acid pairs that actually occur in the protein G sequence, in order that the statistics obtained should be applicable to larger proteins. A peak was assigned to the amino acid pair giving the lowest χ^2, and a confidence score for this assignment was determined as the ratio of this χ^2 value to that obtained for the amino acid pair yielding the second lowest χ^2 value.

We first defined separate contributions to χ^2 based on the HSQC spectra (where only the amino acid type of *b* determines the peak intensity), and on the HNCO spectra (where the amino acid types of both *a* and *b* determine the peak intensity), and then combined these values.

In principle the contribution to χ^2 from the HSQC spectra, $\chi^2_{\mathrm{HSQC}}(i,b)$, if a peak position *i* is provisionally assigned to amino acid type *b* (*b* = A, D, F, ..., Y), could be defined as:

$$\chi^2_{\mathrm{HSQC}}(i,b) = (x_{i,2} - r_{b,2})^2/\sigma^2_{i,2} + (x_{i,3} - r_{b,3})^2/\sigma^2_{i,3} + (x_{i,4} - r_{b,4})^2/\sigma^2_{i,4} + (x_{i,5} - r_{b,5})^2/\sigma^2_{i,5} \quad (1)$$

where $x_{i,k}$ is the observed ratio for the peak *i* in the HSQC spectrum of sample *k*, defined as the intensity for peak *i* in sample *k* divided by the intensity for peak *i* in sample 1; $r_{b,k}$ is the expected ratio for a peak belonging to an amino acid of type *b* in sample *k*; and $\sigma_{i,k}$ is the standard error in the ratio for the peak *i* in sample *k*. For each sample the SD of baseline noise in the spectrum was measured and used to estimate $\sigma_{i,k}$ for each peak (see Materials and methods).

However, with the definition of Eq. 1, any measurement error in the intensity of sample 1 will uniformly elevate or depress the values of $x_{i,2-5}$. In order to prevent this over-emphasis of the observed intensity in sample 1, we introduced a factor α, which scales each observed ratio $x_{i,k}$ such that $\alpha\, x_{i,k}$ is the ratio that would have been calculated if it were possible to measure the intensity in sample 1 with zero error. α was estimated as the value that minimised

χ^2_{HSQC} (i,b) for each particular i and b, with χ^2_{HSQC} (i,b) defined as:

$$\chi^2_{\mathrm{HSQC}}(i,b) = (\alpha-1)^2/\sigma^2_{i,1} + (\alpha x_{i,2} - r_{b,2})^2/\sigma^2_{i,2} + (\alpha x_{i,3} - r_{b,3})^2/\sigma^2_{i,3} + (\alpha x_{i,4} - r_{b,4})^2/\sigma^2_{i,4} + (\alpha x_{i,5} - r_{b,5})^2/\sigma^2_{i,5} \quad (2)$$

The first term in this expression was introduced to include the deviation of the observed ''ratio'' for sample 1 from the idealised value.

The contribution to χ^2 from the HNCO spectra, $\chi^2_{\mathrm{HNCO}}(i,a,b)$, was defined similarly to equation 2, except with the possibility that the trial assignment can be made to $(a)b$ = (A)A, (A)D, (A)F, ..., (G)G) and with $x_{i,k}$ being the observed ratio in the appropriate HNCO spectrum. The combined χ^2 value was defined as:

$$\chi^2(i,a,b) = \chi^2_{\mathrm{HSQC}}(i,b) + \chi^2_{\mathrm{HNCO}}(i,a,b) \quad (3)$$

A peak at position i was assigned to the $a(b)$ pair yielding the lowest $\chi^2(i,a,b)$. The score for this assignment was then obtained by dividing this lowest $\chi^2(i,a,b)$ value by $\chi^2(i,a,b)$ for the $a(b)$ pair yielding the second lowest $\chi^2(i,a,b)$ value. A low score is thus indicative of a high discrimination between the two top-scoring assignments, and the score tends to one as the χ^2 values for the two best assignments become similar.

Assignment scores for protein G

The scores for all peaks present in sample 1 were calculated and are plotted in rank order of increasing discrimination (i.e. of *decreasing* score) in Fig. 3. Apart from the three cases with the worst scores (Leu 7, Ala 36 and Lys 33), all the assignments are correct. The cross-peaks for Leu 7 and Leu 9 are heavily overlapped, and Leu 9 yields the fifth worst score. Likewise the cross-peaks for Ala 36 and Lys 33 are heavily overlapped, and yield the second and third worst scores, respectively. In addition to spectral overlap, other issues arising in the samples were: incomplete deformylation of the N-terminal Met (giving rise to duplicate peaks); dilution of labelled Ala by metabolites; and deamidation of Asn to Asp. Incomplete deformylation in proteins expressed in *E. coli* lysates has also been observed by Torizawa et al. (2004). They circumvented this problem by making use of a cleavable N-terminal tag, which also appeared to increase yield. Dilution of labelled Ala was suspected due to the low peak intensities observed for Ala residues in the HSQC spectra, as noted by Shi et al. (2004). We confirmed this by recording a 2D ^{1}H–^{1}H TOCSY spectrum of sample 1 with no X-nucleus decoupling, in which singlet peaks were observed at the centres of the X-coupled H_N–H_α and H_N–H_β quartets (dilution ~50%; data not shown). Addition of 2 mM amino-oxy-acetate, which inhibits conversion of pyruvate to Ala by a pyridoxal-dependent transaminase (Lopukhov et al. 2002), did not affect the level of dilution, in contrast to the observations of Shi et al. (2004). Our observation of Asn deamidation to Asp is opposite to the transamidation of Asp to Asn observed by Ozawa et al. (2004). 1D proton spectra recorded for the Asp and Asn stock solutions confirmed that the deamidation must occur after addition of the amino acids to the cell lysate. Further developments in in vitro protein expression are necessary to overcome these problems. However, with due account for deamidation of Asn (as discussed below), the scoring method is sufficiently robust not to be affected by these relatively small effects.

With the labelling pattern in Table 1, the deamidation of Asn (which is labelled 50%^{15}N/50%^{14}N in samples 2–5), leads to a reduction in samples 2–5 of the intensities of cross-peaks associated with Asp (which is labelled 100% ^{15}N/^{13}C in samples 2–5). Reductions occur in the intensities of HSQC and HNCO cross-peaks where the residue b in $(a)b$ is Asp, or in the intensities of HNCO cross-peaks where the residue a is Asp. If x% of Asn is converted to Asp, then the reduction in the ^{13}C content of Asp is $x/(100 + x)$%, whilst the reduction in the ^{15}N content is $(x/2)/(100 + x)$%. From inspection of the ratios observed in samples 2–5 for cross-peaks involving Asp, it appears that this effect is constant and that approximately 30% of Asn is converted to Asp. This observation is supported by the intensity of Asp cross-peaks that appeared in the HSQC

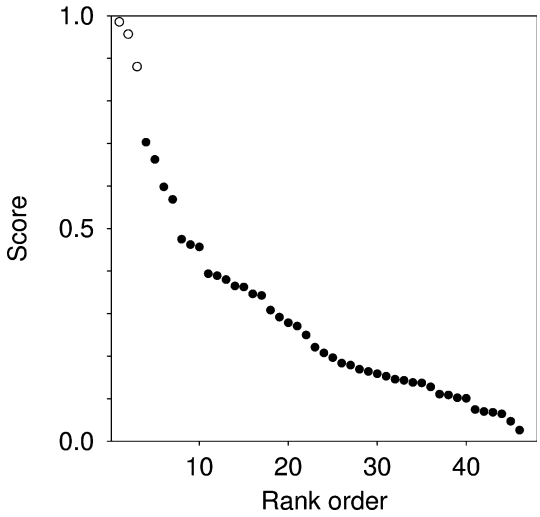

Fig. 3 Score plot for the data from protein G. The residues are plotted in the order of decreasing score (i.e. increasing discrimination). Filled circles are for residues that are assigned to the correct $a(b)$ pair, whereas open circles are residues that are incorrectly assigned

spectrum of a sample in which the only added labelled amino acids were Asn and Ala (data not shown). In Figs. 1 and 2, appropriate correction factors have been applied to cross-peaks involving Asp. In the scoring scheme, an alternative approach was successfully used, which did not require prior knowledge of the assignment of cross-peaks. The observed ratios were not adjusted; instead the *expected pattern* for the test pair (*a*)*b* was adjusted if it involved Asp.

Residues in pairs (*a*)*b* where *a* is not labelled (Trp and Glu for protein G with our labelling scheme) give rise to an HSQC cross-peak but to no HNCO cross-peak in all five samples. The absence of an HNCO cross-peak in sample 1 precludes the calculation of peak ratios, and therefore such cross-peaks cannot be assigned by the method described above. There are four such residues in protein G ((E)A22, (E)K30, (E)T18 and (W)T46), and the data for these are included in Fig. 1 but not in Fig. 2. For these four residues a score calculated solely on the basis of $\chi^2_{\mathrm{HSQC}}(i,b)$ correctly assigns the amino acid type.

Effect of overlap and noise on assignment confidence

In the scoring analysis above (and also in Figs. 1, 2), the values for $\sigma_{i,k}$ were derived solely from measurements of the baseline noise in the HSQC and HNCO spectra. It is clear from Figs. 1 and 2 however that there are many instances where the measured ratio deviates from the expected ratio by more than twice $\sigma_{i,k}$. For normally distributed data one expects a much smaller number of such exceptions (less than 5%). The unbiased estimate for the SD in the differences between the measured and ideal ratios is approximately 0.07, whereas the SD expected from baseline noise alone is approximately 0.02. This latter value was also observed in repeated measurements on one sample. The greater SD of the data must arise from a combination of the many factors and processes involved in the preparation of five separate physical samples and the measurements thereon.

To assess the influence of noise on assignment correctness, we simulated data that corresponded to the ideal ratio patterns expected for the residues in protein G with random noise added at various levels, corresponding to SDs ranging from 0.07 to 0.2. The resulting score plots obtained from the simulated data are shown in Fig. 4. As expected, as the noise level increases, the overall score values get worse, and also the number of incorrect assignments increases. However, a cutoff score value of approximately 0.4 can be defined, below which no incorrect assignments are made.

The plot of scores for the experimental data (Fig. 3) is similar to that obtained from the simulated data with a SD of 0.07 (Fig. 4), i.e. close to the value inferred from the differences in the measured and ideal ratios. However, the score increases markedly for the seven lowest scoring cross-peaks. As noted above, several of these are overlapped. To investigate the issue of overlap further we simulated datasets with varying levels of overlap. For each degree of overlap 1,000 test assignments were performed. For each test a random (*a*)*b* pair was selected and its ideal ratio pattern calculated. Then a second random (*a*)*b* pair was selected and its pattern (scaled by a factor defined as the degree of overlap) was added to the first pattern. This was then normalised so that the ratio observed in sample 1 was 1.0. Random noise with a SD of 0.07 was then added. The resulting pattern of ratios was then scored, as described above. Once scored, the tests were divided into two classes:

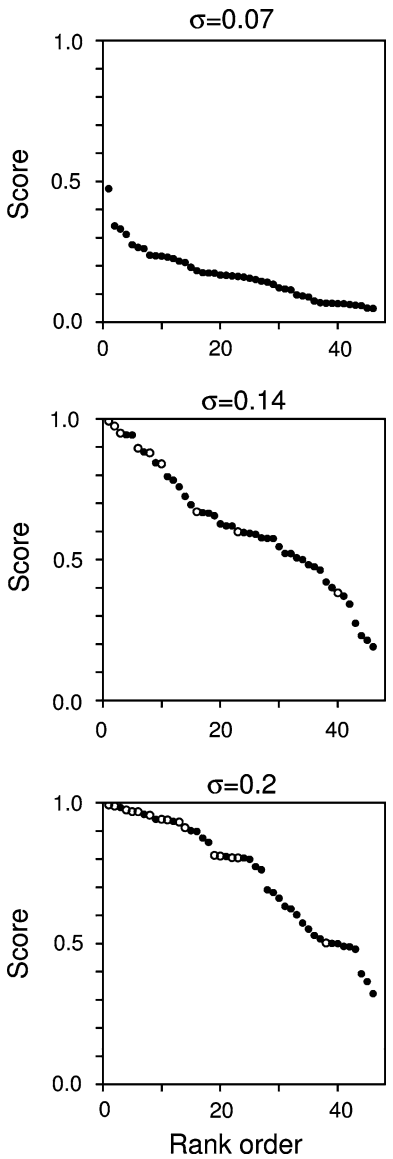

Fig. 4 Simulated score plots for protein G, created with varying levels of normally distributed noise. The data are plotted as for Fig. 3

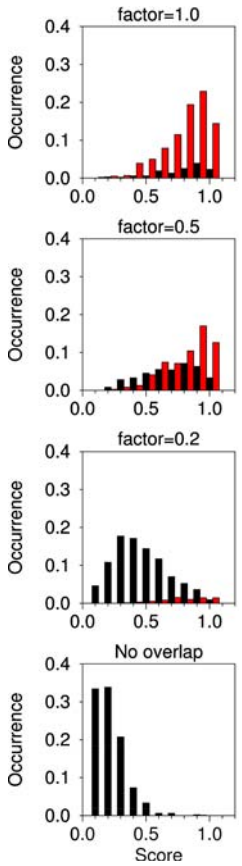

Fig. 5 The effect of overlap on peak scoring and correctness of assignments. The 1,000 trial residue pairs considered for each degree of overlap were grouped into bins corresponding to intervals of score of 0.1, and according to whether the assignment was made correctly (black bars) or incorrectly (red bars), as defined in the text

those where the assignment was correct (i.e. corresponds to the major component of the cross-peak for values of overlap <1, or corresponds to either component for values of overlap of 1) and those where the assignment was incorrect. Histograms of the classified scores are plotted in Fig. 5. Although overlap does cause incorrect assignments, the overwhelming majority occur with scores above the previously introduced cutoff of 0.4. Overlap also increases the scores of correctly assigned peaks. The elevated score values calculated for the overlapped cross-peaks in the protein G data are in accordance with these general observations.

Assignment scores for GFP

In a previous report (Parker et al. 2004) we applied CSL to a 229-residue truncated version of the cycle3 version of GFP from *Aequorea victoria* (Khan et al. 2003). This protein has rather poor NMR characteristics. We determined the correlation time for this construct to be greater than 20 ns, and the protein displays two regions of missing and weakened resonances due to conformational exchange broadening (Khan et al. 2003). With the ca. 100 μM samples that were available to us in our initial study we were not able to obtain sufficient signal-to-noise in triple resonance experiments to elucidate the assignments of all overlapped clusters under our experimental conditions. We were unable therefore to determine assignments for all cross-peaks. There are 168 assignments available for the residue types that were labelled (His, Tyr, Trp and Glu were not labelled) (Khan et al. 2003). We were able to pick 'blind' 133 cross-peaks. The remainder were too weak to detect, or were involved in peak overlaps. Of these 133 cross-peaks, 91 were sufficiently well-resolved to pick and identify based on the previous assignment. Furthermore, despite long acquisition times with a cryoprobe, the signal-to-noise of the cross-peaks were low, corresponding to SDs between observed and ideal ratios of 0.14 for the HSQC data, and 0.47 for the HNCO data. For all these reasons, the GFP CSL data were not suitable for developing a scoring method. Nevertheless, to further test the scoring function developed above using protein G, and to assess the performance of the CSL method in relation to sample quality, we have applied it to the GFP CSL data.

The plot of scores obtained from the GFP CSL data is shown in Fig. 6. The general form of the plot is very similar to that simulated using a SD of 0.2 in Fig. 4, which is broadly in line with that measured from the observed and ideal ratios. Consequently, relatively few cross-peaks have scores below the 0.4 cutoff. For the cross-peaks that do have scores below the 0.4 cutoff, one is mis-assigned. Unlike the case of the small protein G (where we deliberately chose to

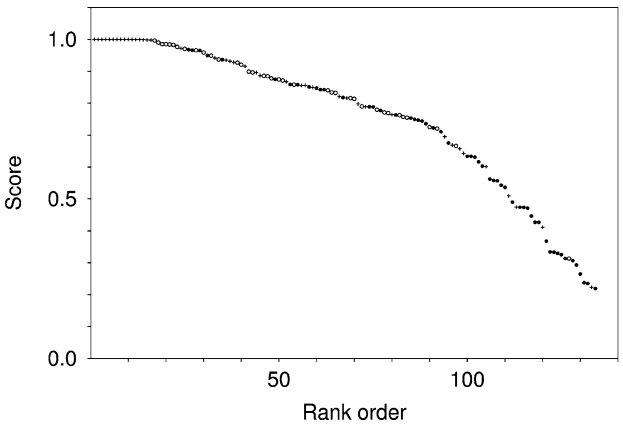

Fig. 6 Score plot for the data from GFP. The residues are plotted in the order of decreasing score (i.e. increasing discrimination). Filled circles are for residues that are assigned to the correct *a*(*b*) pair, whereas open circles are residues that are incorrectly assigned. Black crosses are for cross-peaks for which we do not have a definite assignment, i.e. cross-peaks in the HSQC spectrum of the fully labelled sample not sufficiently well-resolved to identify based on the published assignment (Khan et al. 2003)

ignore sequence information as discussed above), for the more challenging case of the large GFP protein we can use an additional filter based on knowledge of the possible *a*(*b*) pairs in the GFP sequence. Based on this knowledge, the mis-assignment can be eliminated as the cross-peak is assigned to an *a*(*b*) pair not in the GFP sequence. Therefore even for this data for GFP, which as noted is far from ideal, the scoring method appears to be adequately robust.

Conclusions

Large numbers of residue-type and sequence-specific backbone amide assignments can be obtained using a small number of selectively labelled samples designed to achieve maximal information content (Shi et al. 2004; Parker et al. 2004; Wu et al. 2006). On their own these partial assignments provide useful sets of probes for identifying ligand binding sites for proteins of known structure via chemical shift mapping (Foster et al. 1998; Takahashi et al. 2000; Reese and Dötsch, 2003). For NMR structure calculations, selective labelling techniques have been shown to increase the efficiency and accuracy of the assignment process, when combined with 3D heteronuclear data (Shi et al. 2004; Trbovic et al. 2005). In our CSL method, and the triple labelling method adopted by Shi et al. (2004), assignments are based on comparing fractional HSQC and HNCO cross-peak intensities across the samples. For these methods to be applicable to large proteins displaying complex spectra with weak and overlapped cross-peaks, a means of establishing the reliability of assignments is essential.

We have presented here a rigorous and systematic analysis of a complex labelling scheme, and used it to assess the influence of noise (arising from a combination of spectral noise and sample variation) and overlap on assignment confidence. The analysis shows that when noise and/or overlap lead to incorrect assignments they only do so with a correspondingly poor value for the score. Therefore our scoring method provides a robust means for determining the reliability of assignments using CSL with samples prepared with the RTS system that should be generally applicable. Analogous scoring approaches could be adopted for other labelling schemes, and it is important to stress that, when using alternative expression systems, any scrambling issues should be properly characterised. As the analysis of the GFP data attests, using concentrated protein samples with high signal-to-noise is essential for achieving a high contingent of CSL assignments. Further improvements in cell-free expression technology are imperative in this respect.

Acknowledgements This work was supported by a research grant from the Biotechnological and Biological Sciences Research Council (BBSRC), UK. MJP was supported by a BBSRC David Phillips fellowship and a University of Leeds research fellowship.

References

Foster MP, Wuttke DS, Clemens KR, Jahnke W, Radhakrishnan I, Tennant L, Reymond M, Chung J, Wright PE (1998) Chemical shift as a probe of molecular interfaces: NMR studies of DNA binding by the three amino-terminal zinc finger domains from transcription factor IIIA. J Biomol NMR 12:51–71

Gronenborn AM, Filpula DR, Essig NZ, Achari A, Whitlow M, Wingfield PT, Clore GM (1991) A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. Science 253:657–661

Guignard L, Ozawa K, Pursglove SE, Otting G, Dixon NE (2002) NMR analysis of in vitro-synthesized proteins without purification: a high-throughput approach. FEBS Lett 524:159–162

Jewett MC, Swartz JR (2004) Mimicking the *Escherichia coli* cytoplasmic environment activates long-lived and efficient cell-free protein synthesis. Biotech Bioeng 86:19–26

Khan F, Stott K, Jackson S (2003) Letter to the editor: ^{1}H, ^{15}N and ^{13}C backbone assignment of the Green Fluorescent Protein (GFP). J Biomol NMR 26:281–282

Kainosho M, Tsuji T (1982) Assignment of the three methionyl carbonyl carbon resonances in *Streptomyces* subtilisin inhibitor by a carbon-13 and nitrogen-15 double-labeling technique: a new strategy for structural studies of proteins in solution. Biochemistry 21:6273–6279

Kigawa T, Muto Y, Yokoyama S (1995) Cell-free synthesis and amino acid-selective stable isotope labeling of proteins for NMR analysis. J Biomol NMR 6:129–134

Kigawa T, Yabuki T, Yoshida Y, Tsutsui M, Ito Y, Shibata T, Yokoyama S (1999) Cell-free production and stable-isotope labeling of milligram quantities of proteins. FEBS Lett 442:15–19

Klammt C, Löhr F, Schäfer B, Haase W, Dötsch V, Rütterjans H, Glaubitz C, Bernhard F (2004) High level cell-free expression and specific labeling of integral membrane proteins. Eur J Biochem 271:568–580

Lopukhov LV, Ponomareva AA, Yagodina LO (2002) Inhibition of bacterial pyridoxal-depending enzymes by (aminooxy)-acetic acid improves selective ^{15}N Isotope labeling of bacterially expressed protein. Biotechniques 32:1248–1250

Ozawa K, Headlam MJ, Schaeffer PM, Henderson BR, Dixon NE, Otting G (2004) Optimization of an *Escherichia coli* system for cell-free synthesis of selectively ^{15}N-labelled proteins for rapid analysis by NMR spectroscopy. Eur J Biochem 271:4084–4093

Parker MJ, Aulton-Jones M, Hounslow A, Craven CJ (2004) A combinatorial selective labelling method for the assignment of backbone NMR resonances. J Am Chem Soc 126:5020–5021

Reese ML, Dötsch V (2003) Fast mapping of protein–protein interfaces by NMR spectroscopy. J Am Chem Soc 125:14250–14251

Shi J, Pelton JG, Cho HS, Wemmer DE (2004) Protein signal assignments using specific labeling and cell-free synthesis. J Biomol NMR 28:235–247

Spirin AS, Baranov VI, Ryabova LA, Orodov SY, Alakhov YB (1988) A continuous cell-free translation system capable of producing polypeptides in high yield. Science 242:1162–1164

Swartz JR, Jewett MC, Woodrow KA (2004) Cell-free protein synthesis with prokaryotic combined transcription-translation. Methods Mol Biol 267:169–182

Takahashi H, Nakanshi T, Kami K, Arata Y, Shimada I (2000) A novel NMR method for determining the interfaces of large protein–protein complexes. Nat Struct Biol 7:220–223

Torizawa T, Shimizu M, Taoka M, Miyano H, Kainosho M (2004) Efficient production of isotopically labeled proteins by cell-free synthesis: a practical protocol. J Biomol NMR 30:311–325

Trbovic N, Klammt C, Koglin A, Löhr F, Bernhard F, Dötsch V (2005) Efficient strategy for the rapid backbone assignment of membrane proteins. J Am Chem Soc 127:13504–13505

Wu PSC, Ozawa K, Jergic S, Su X-C, Dixon NE, Otting G (2006) Amino-acid type identification in ^{15}N-HSQC spectra by combinatorial selective ^{15}N-labelling. J Biomol NMR 34:13–21

Yabuki T, Kigawa T, Dohmae N, Takio K, Terada T, Ho Y, Laue ED, Cooper JA, Kainosho M, Yokoyama S (1998) Dual amino acid-selective and site-directed stable-isotope labeling of the human c-Ha-Ras protein by cell-free synthesis. J Biomol NMR 11: 295–306

Yokoyama S (2003) Protein expression systems for structural genomics and proteomics. Curr Opin Chem Biol 7:39–43