J|A|C|S COMMUNICATIONS

Published on Web 06/17/2009

Automated Assignment in Selectively Methyl-Labeled Proteins

Yingqi Xu,[†,‡,§] Minhao Liu,[†,‡,§] Peter J. Simpson,[†,‡,§] Rivka Isaacson,[†,‡,§] Ernesto Cota,[†,‡,§] Jan Marchant,[†,‡,§] Daiwen Yang,[||] Xiaodong Zhang,[†,§] Paul Freemont,[†,§] and Stephen Matthews*[,†,‡,§]

Division of Molecular Biosciences, Cross Faculty NMR Centre, Centre for Structural Biology, Imperial College London, South Kensington, London, SW7 2AZ, U.K., and Department of Biological Sciences, National University of Singapore, 14 Science Drive 4, Singapore 117543

Received March 16, 2009; E-mail: s.j.matthews@imperial.ac.uk

Although NMR spectroscopy is usually employed for structural studies of relatively small proteins, the advent of deuteration and transverse relaxation optimized spectroscopy (TROSY)[1] techniques has pushed the molecular weight limit for backbone assignment to ~100 kDa.[2,3] For structure determination, more nuclear Overhauser effects (NOEs) can be collected using selective protonation of methyl groups,[4,5] stereoarray isotope labeling (SAIL),[6] or fully protonated samples,[7] which are supplemented with other measurements such as residual dipolar couplings.[8] In addition to its role in structure determination, selective methyl protonation also provides excellent probes for monitoring interactions and dynamics. High quality spectra can be recorded in very large systems (up to ~1 MDa),[9] due to the proton multiplicity, favorable relaxation properties, and methyl-TROSY[10] effects. Methods are available for labeling isoleucine (δ1), leucine, valine,[4,5] alanine,[11,12] and methionine[13] methyl groups efficiently.

In very large systems, it is not possible to obtain backbone assignments; therefore alternative strategies are required for the assignment of methyl resonances. One successful approach includes splitting the system into smaller fragments[9,13] that are within the range of routine NMR experiments and transferring the assignments to the larger system. Mutation of specific methyl-containing residues[9,13,14] can also guide the assignment of some residues as can the use of NOE correlations in combination with available crystal structures.[9,13] These methods are often time-consuming as several mutants and fragments may have to be tried, and chemical shift changes due to mutation or truncation often complicate interpretation.

Here we propose a fully automated method that can reliably assign selectively labeled methyl groups without resorting to data on mutants or smaller fragments. The input is a crystal structure, which is available for most, if not all, cases of reported NMR studies of large systems,[9,13,15] together with a few NMR spectra, namely a 2D $^{1}H-^{13}C$ HMQC (TROSY) presenting the peaks to be assigned, an H_mC_mC experiment correlating the methyl resonances with the directly bonded ^{13}C (distinguishing between valine and leucine methyl peaks and which arise from the same residue), and a 3D CCH-NOESY giving the NOE network among methyl groups. All these experiments have been shown to be effective for systems with molecular weight >300 kDa.[9] In short, we predict chemical shifts (using SHIFTS[16] or our own program for large and oligomeric proteins for ^{1}H and SHIFTX[17] for ^{13}C) and NOE correlations from the crystal structure and score the comparison with experimental NMR data. The ranked scores provide the initial assignment from which an automated assignment-swapping protocol bootstraps the final assignment.

We first developed and tested the plausibility of this method using simulated data on 60 proteins for which both crystal structures and chemical shift assignments are available (Table S1). For each protein, NOESY spectra were simulated according to the crystal structure and experimental chemical shifts with various distance cutoff settings, increasing levels of absent NOEs, and structural differences between solution and crystal forms. For each methyl peak, the related NOE peaks are separated according to the amino acid type of the donor methyl. To see whether a peak p could be assigned to a methyl r, we calculate a score (S) using eq S1, which comprise two terms that describe the match between experiment and prediction for chemical shifts and NOEs, respectively (see Supporting Information for further details). The score is maximal when the numbers of predicted and experimental NOEs are equivalent and the chemical shifts match. In principle, the higher the score, the more likely that peak (p) should be assigned to a particular residue (r). A best-first procedure is applied to derive the assignment automatically: from all the possible (p, r) pairs for a protein, the one with the highest score is examined and the peak p is assigned to the methyl r accordingly; then all the pairs involving these p or r are removed, and further assignments are made iteratively. For an assignment made between p and r, if there is no other choice for either p or r with a higher or close (within 0.2) score, it is regarded as reliable.

Tests with simulated data show that greater than 90% of assignments made for alanine β and isoleucine δ1 sites are correct (Tables S2, S3). Furthermore, ~60% of assignments are regarded as reliable with 99% accurate assignments. Leucine and valine methyls can also be assigned with similar reliability. As expected performance deteriorates in cases where mismatches occur between cutoff distances for predicted and experimental NOEs (Table S2), high numbers of NOEs are absent (Table S4), or significant structural differences exist between crystal and solution states (Table S5). Nevertheless, greater than 50% correct assignments are obtained overall and for the reliably assigned resonances at least 90% correctness can be expected. Preliminary assignments based on the scoring algorithm provide the start point for an automated routine in which assignments were systematically swapped and reevaluated. Changes that result in an increase in the score (see Supporting Information for further details), i.e., result in an increase in the number of expected NOE peaks, according to the current assignments and structure, or a decrease in the difference between predicted and observed chemical shifts were preserved. The process is continued without manual intervention until no further changes are observed. We determined its effectiveness on simulated data for ILV-labeled MBP at a 7.0 Å cutoff with 30% of the NOEs randomly removed. After a few rounds of automated swapping a plateau in the number of expected NOEs was reached (434 compared to 440 for a totally correct assignment), and 106 correct methyl group assignments were derived out of 122. The incorrect assignments possess very few NOEs and common chemical shifts. It is worth noting that intermethyl distances are readily observed over 8 Å in deuterated proteins and increasing the cutoff to 8.5 Å gives a 100% correct assignment for MBP.

We next applied our automated method to a genuine experimental data set, namely the 300 kDa, ILV-labeled proteasome ($\alpha_7\alpha_7$) for which excellent spectra have been recorded by the Kay group.[9] The crystal

[†] Division of Molecular Biosciences, Imperial College London.
[‡] Cross Faculty NMR Centre, Imperial College London.
[§] Centre for Structural Biology, Imperial College London.
[||] National University of Singapore.

10.1021/ja9020233 CCC: $40.75 © 2009 American Chemical Society

structure for the complex of $\alpha_7\beta_7\beta_7\alpha_7$ and 11S (1YAU) was initially used to predict chemical shifts and NOE peaks. To provide an accurate estimate of the NOE cutoff distances, scores were calculated over a range of values; Figure S1 demonstrates that for MBP the highest score corresponds to the value closest to the actual cutoff. For the proteasome, the highest score was observed for cutoff values between 8.5 and 9.2 Å (Figure S1); at this stage a total of 49 out of 93 ILV methyl assignments were correct (without considering stereospecific assignments for leucine and valine). Assignment swapping runs were carried out at 0.1 Å intervals between 8.5 and 9.2 Å. For each run the number of expected NOEs and correct assignments increased dramatically; i.e., 8.9 Å generates a total of 578 expected NOEs (up from 316) and increases the number of correct assignments from 49 to 87 (out of 93). Although the performance was similar across the cutoff range, several of the incorrect assignments were different suggesting that the true minimum had not been reached. Assignments from all eight runs were subsequently ranked according to numbers of expected, unexpected, and absent NOEs together with the chemical shift prediction (Table S6); the highest ranked assignment for each peak was chosen which gave 92 correct methyl groups assignments and one error (Figure 1A).

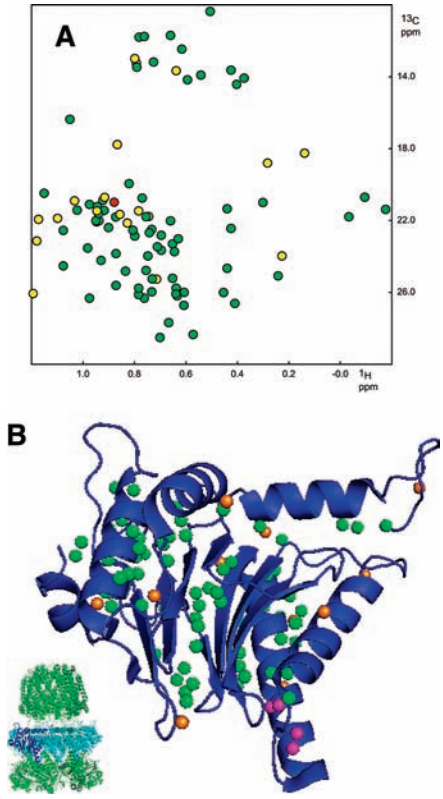

Figure 1. (A) Schematic diagram of the 1H−^{13}C HMQC spectrum for ILV-labeled proteasome $\alpha_7\alpha_7$. Green (reliable, Table S6) and yellow peaks are correctly assigned automatically; the red peak is wrongly assigned. (B) Ribbon representation of proteasome α monomer with correctly assigned methyl groups shown as spheres (green − reliable; orange − remaining methyl groups; pink − L106, V107). L106, V107 lie on one face of the α_7 ring; NMR spectra were recorded on $\alpha_7\alpha_7$, while the crystal structure used for NOE and chemical shift predictions represents the 11S-proteasome ($\alpha_7\beta_7$) complex, in which these residues lie at the $\alpha_7\beta_7$ interface (inset - position of the α subunit shown in dark blue).

The wrong assignment ranks in the bottom five (Table S6) and, together with other poorly ranked assignments, has no NOE correlations and therefore cannot be assigned reliably unless its chemical shift is unique. It should be possible to correct wrong valine and leucine assignments by checking the chemical shift of the directly bonded ^{13}C nuclei. Furthermore, labeling other methyl sites, such as alanine, would provide more methyl−methyl NOEs and improve the performance further (for AILV-labeled MBP at 7 Å with 30% of NOEs removed now produces only two errors). Alanine also benefits from the shortest side chain among methyl-containing amino acids, and therefore $^{13}C_\alpha$ and $^{13}C_\beta$ chemical shifts can be readily predicted and measured in very large systems[11,12] Some correctly assigned methyl groups, such as L106 and V107, have unexpected NOE correlations, which presumably reflects subtle structural differences between solution and crystalline states (Figure 1B). Despite this, we demonstrate a new automated procedure able to rapidly assign the majority of methyl groups in very large proteins (Figure 1), without recourse to mutagenesis, truncated fragments, or manual analysis. As our method relies on structure-based chemical shift and NOE prediction, it is anticipated to work best for methyl groups in well-ordered regions of a protein and thus would complement mutagenesis approaches more suitable for relatively flexible regions. Although we envisage that our method will benefit NMR studies on very large multimeric proteins, well-dispersed methyl TROSY spectra have been demonstrated on the 723 residue malate synthase G,[4] we may therefore expect that our approach would be applicable to highly complex samples, i.e., with >200 methyl groups. It would also be particularly suitable for membrane protein systems where the lipid/detergent environment adds significantly to the apparent molecular weight; in tests 34 out of 36 methyl groups in OmpX[18] can be assigned correctly at 7 Å with 30% of the NOE data removed. Although the two errors arise from leucine residues with no NOEs, they can be assigned correctly at the 8 Å cutoff. As obstacles in resonance assignment are removed, we expect a significant growth in NMR studies of interaction and dynamics in large protein systems.

Acknowledgment. We are extremely grateful to Prof. Kay, Dr. Sprangers, and Dr. Velyvis for sharing their data. Financial support is from the Wellcome Trust (079819) and BBSRC (G004668) to S.M. and the BMRC, Singapore (R154000272305) to D.Y.

Supporting Information Available: Scoring function, tables containing results from simulated data and the ranking of proteasome assignments. The software for methyl assignment prediction from X-ray structures (MAP-XS) is available as downloadable scripts or can be run on the form-based server: http://nmr.bc.ic.ac.uk/map-xs/. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Pervushin, K.; Riek, R.; Wider, G.; Wuthrich, K. *Proc. Natl. Acad Sci. U.S.A.* **1997**, *94*, 12366–71.
(2) Salzmann, M.; Pervushin, K.; Wider, G.; Senn, H.; Wuthrich, K. *J. Am. Chem. Soc.* **2000**, *122*, 7543–7548.
(3) Fiaux, J.; Bertelsen, E. B.; Horwich, A. L.; Wuthrich, K. *Nature* **2002**, *418*, 207–11.
(4) Tugarinov, V.; Choy, W. Y.; Orekhov, V. Y.; Kay, L. E. *Proc. Natl. Acad Sci. U.S.A.* **2005**, *102*, 622–7.
(5) Tugarinov, V.; Kanelis, V.; Kay, L. E. *Nat. Protoc.* **2006**, *1*, 749–54.
(6) Kainosho, M.; Torizawa, T.; Iwashita, Y.; Terauchi, T.; Ono, A. M.; Guntert, P. *Nature* **2006**, *440*, 52–7.
(7) Xu, Y. Q.; Zheng, Y.; Fan, J. S.; Yang, D. W. *Nature Methods* **2006**, *3*, 931–7.
(8) Tjandra, N.; Bax, A. *Science* **1997**, *278*, 1111–4.
(9) Sprangers, R.; Kay, L. E. *Nature* **2007**, *445*, 618–22.
(10) Tugarinov, V.; Hwang, P. M.; Ollerenshaw, J. E.; Kay, L. E. *J. Am. Chem. Soc.* **2003**, *125*, 10420–8.
(11) Isaacson, R. L.; Simpson, P. J.; Liu, M.; Cota, E.; Zhang, X.; Freemont, P.; Matthews, S. *J. Am. Chem. Soc.* **2007**, *129*, 15428–9.
(12) Ayala, I.; Sounier, R.; Use, N.; Gans, P.; Boisbouvier, J. *J. Biomol. NMR* **2009**, *43*, 111–9.
(13) Gelis, I.; Bonvin, A.; Keramisanou, D.; Koukaki, M.; Gouridis, G.; Karamarou, S.; Economou, A.; Kalodimos, C. G. *Cell* **2007**, *131*, 756–69.
(14) Sprangers, R.; Gribun, A.; Hwang, P. M.; Houry, W. A.; Kay, L. E. *Proc. Natl. Acad Sci. U.S.A.* **2005**, *102*, 16678–83.
(15) Velyvis, A.; Yang, Y. R.; Schachman, H. K.; Kay, L. E. *Proc. Natl. Acad Sci. U.S.A.* **2007**, *104*, 8815–20.
(16) Osapay, K.; Case, D. A. *J. Am. Chem. Soc.* **1991**, *113*, 9436–9444.
(17) Neal, S.; Nip, A. M.; Zhang, H. Y.; Wishart, D. S. *J. Biomol. NMR* **2003**, *26*, 215–240.
(18) Hilty, C.; Fernandez, C.; Wider, G.; Wuthrich, K. *J. Biomol. NMR* **2002**, *23*, 289–301.

JA9020233