REFINEMENT AND AUTOMATION OF THE MAIN CHAIN DIRECTED ASSIGNMENT PROCEDURE FOR THE ANALYSIS OF 2-D ^{1}H SPECTRA OF PROTEINS

Sarah J. Nelson,[†] Diane M. Schneider[#] Deena L. DiStefano[#]
and A. Joshua Wand[#] Department of NMR & Medical Spectroscopy[†],
Institute for Cancer Research[#]
Fox Chase Cancer Center, Philadelphia, PA 19111

INTRODUCTION

The elucidation of protein structures from NMR data requires the interpretation of large multi-dimensional spectra. These data are subjected to extensive processing and analysis in order to reveal the information of interest. J-correlated and distance correlated relationships between pairs of protons are most often derived from a variety of 2-D (or 3-D) homonuclear spectra (1-4). More recently labelling with ^{13}C or ^{15}N in conjunction with heteronuclear spectroscopy has provided an alternative or complementary route for distinguishing and identifying resonances (5,6). All relationships between pairs of nuclei are represented by crosspeaks in 2-D or 3-D spectra. The first task in analyzing such data is therefore the identification of crosspeak positions in individual spectra. The cornerstone of any automated or computer assisted procedure for structure determination is to use the computer to produce a list of crosspeak positions for each spectrum being considered.

The second stage of the analysis is to assign spectral frequencies to particular protons in the molecule. One approach to this problem is the Sequential Assignment Procedure (7-9), which places initial emphasis on analyzing the side chain spin systems of each amino acid. The amino acid sequence is used to position short runs of spin systems within the protein. Secondary structure is inferred from the presence of NOEs indicating short distances between main chain protons from different residues in the molecule. A second approach is the Main Chain Directed assignment procedure (10,11), which places more initial emphasis on the main chain NH-C_αH-C_βH subspin systems. Characteristic patterns of short distances between these protons are used to identify different types of secondary structures. The secondary structural units are then placed within the sequence and remaining protons assigned by a reduced side chain analysis.

Clearly, as the size of proteins studied increases, it will become necessary to make use of all available information in making assignments. It seems likely therefore that, in future studies, relationships between side chain and main chain protons will need to be considered in parallel rather than sequentially. To assist in these analyses, computer aided pattern searches are already becoming critical parts of the analysis. Bodenhausen et al (12) have presented attempts to automatically identify side chain spin systems directly from J-correlated spectra. We report here upon recent advances in the refinement and automation of the analysis of main chain spin systems. These include the use of an automated peak-picking algorithm (PIQABLE2) and the derivation of strategies for identifying helix, anti-parallel and parallel sheet structures using only relationships between main chain protons (MCDPAT). Our studies have been guided by simulations based upon crystal structure data and involved examination of NMR data derived from several different proteins. The detailed statistical basis is being reported elsewhere (13,14) and thus we will concentrate here on the overall procedure and its practical implementation.

PEAK-PICKING AND PIQABLE2

The PIQABLE algorithm was originally developed for analysis of 1-D low signal to noise spectra. It automatically separates slowly varying baseline from statistically significant peaks and random noise of constant variance (15,16). The basic algorithm has natural extensions to both 2-D and 3-D data, a discussion of its underlying assumptions being presented elsewhere (19). The 2-D version of the algorithm (PIQABLE2) can be used to separate univariate ridges from bivariate peaks

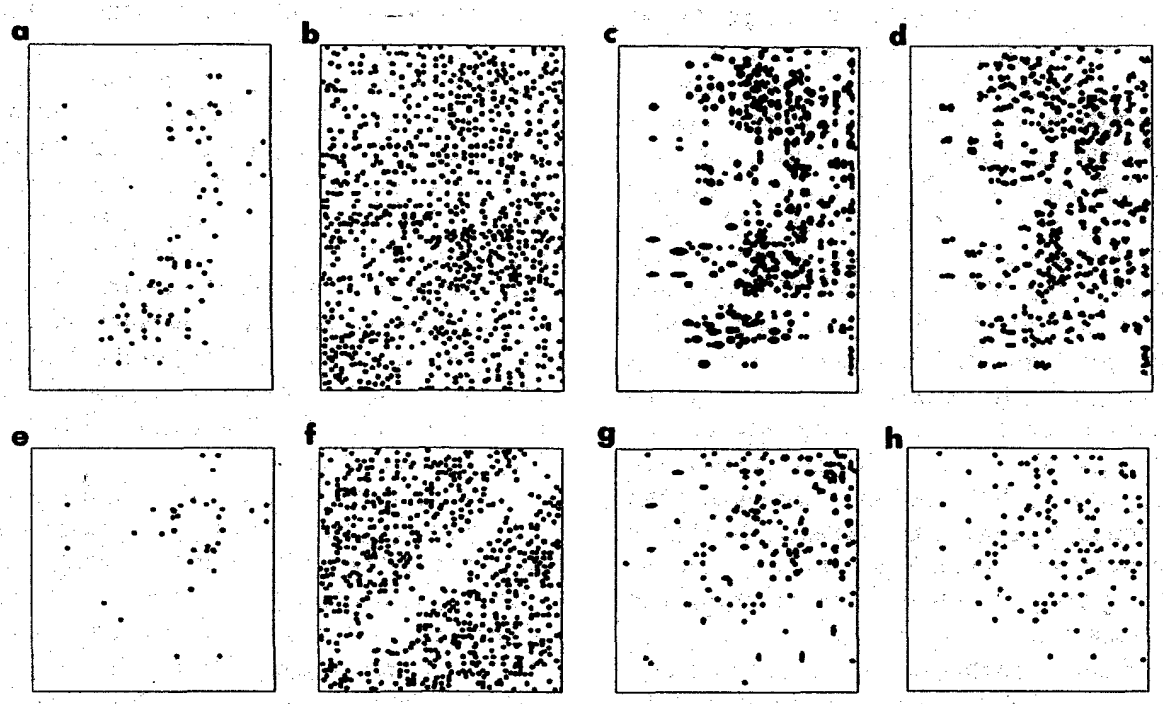

Figure 1: Comparison of peak picking procedures applied to amide-alpha (a-d) and amide-amide (e-h) region of a 500MHz NOESY spectrum of ubiquitin: a)and e) simulations based on main chain proton distances from crystal structure, b) and f) local maximum search, c) and g) manual peak detection, d) and e) PIQABLE2 analysis.

and noise which has slowly changing variance in one or both coordinate directions. As large regions of 2-D NMR spectra display these characteristics, we have used this algorithm to obtain lists of NOE crosspeak positions. Of particular interest is the accuracy of peak position estimates, whether the algorithm could reliably identify crosspeaks between main chain amide, alpha and beta protons and how the limited precision in these estimates imparts subsequent assignment strategies. Human ubiquitin was used as the first test system for the combination of PIQABLE2 with the Main Chain Directed (MCD) assignment procedure. Crosspeak sets were obtained by the following three methods:

(i) visual peak-picking of a NOESY spectrum obtained at 500MHz

(ii) searching for local maxima in the 500MHz spectrum, followed by eliminating those maxima which occurred on only one side of the diagonal and

(iii) PIQABLE2 applied to the 500MHz NOESY spectrum.

A crosspeak set was also generated on the basis of the crystal structure (18). For these data, each main chain proton was assigned its actual frequency (11,19). A crosspeak was assumed to exist when the crystal structure predicted a distance between two protons of less than 4.2A. The relevant MCD patterns were predicted by applying the MCDPAT procedure to the simulated crystal structure data assuming zero error in crosspeak position. Schematics of typical peak distributions obtained in this manner are shown in Figure 1. In order to visualize the peaks, the radii of the circles representing peaks are exaggerated from the peak picking tolerance ($\pm$ 6 datapoints). For the manual peaks, some tolerances were relatively

large and, to give a reasonable representation, the values used were either the same as for the automated peaks or twice the estimated tolerance, depending on which was the larger.

From the crystal structure, 116 crosspeaks between main chain protons were involved in MCD patterns. Of these, 9 would be overlapping due to degeneracy. The local maximum search of the 500 MHz spectrum found over 30000 peaks with 4056 symmetric maxima. Only 85 of these corresponded to the crosspeaks defining MCD patterns. The missing crosspeaks defined mainly relationships between protons in the sheet structures, especially those close to the solvent track. The spatial distribution of local maxima in the spectrum appeared to be stochastic, showing no obvious pattern (see Figure 1). This suggests that a large fraction of the peaks were due to random noise.

Manual peak-picking from the upper diagonal of an expanded contour plot of the spectrum gave 847 crosspeaks. In this case, the beta-beta and alpha-beta regions were not analyzed thoroughly because they were thought to have minimal influence on identifying MCD patterns. To ensure that overlapping peaks were not missed, the tolerance on peak position was set by the size of the lowest contour around the identified peak and ranged from $\pm$ 1 to $\pm$ 6 data points (2.8Hz/point). It was thought that these tolerances would be able to be relaxed but, in practice, this was not possible without losing critical peaks. When these tolerances were used to compare peaks found with those predicted from the crystal structure, 95 of the 116 MCD peaks were identified. Of the 21 missing, 8 were between alpha-beta protons and would not cause MCD patterns to be missed. The remaining 13 crosspeaks were either at the ends of sheets or involved crosspeaks close to the solvent.

The PIQABLE2 algorithm requires a definition of baseline, which define the relative rates of change of these three components in both dimensions. For this analysis a range of parameters were used. In each case, approximately 2000 crosspeaks are found, with between 90 and 99 of the 116 MCD crosspeaks being identified. In general, the same MCD crosspeaks were found as in the manual analysis, with the addition of some of the missing 8 alpha-beta crosspeaks. For some parameter values, crosspeaks close to the spectrum diagonal were missed and, as with the other peak-picking procedures, crosspeaks near the solvent track were sometimes not identified. Again, there is a clear spatial pattern of PIQABLE2 crosspeaks (see Figure 1), suggesting that most of them were real. In this case the peak position tolerances were found to be between $\pm$1 and $\pm$2 data points. This is consistent with the accuracy of positions found in the analysis of 1-D spectra using the PIQABLE algorithm.

Overall, both manual and PIQABLE2 peak picking found the majority of crosspeaks needed for defining MCD patterns in ubiquitin. Missing crosspeaks were due mainly to difficulties near the solvent track or for protons in residues near the ends of sheets. The latter is caused by larger distances between protons in residues spanning the transitions from classical secondary structure. The local maximum search found a larger total number of crosspeaks and less MCD crosspeaks. This is clearly an inferior procedure and apparently found a large number of spurious peaks. The improved accuracy of the PIQABLE2 analysis can be attributed to (1) its ability to remove the influence of ridges, (2) that it made use of additional local smoothing to refine peak position estimates and (3) the difficulties associated with visually estimating maximum positions from a contour plot. It is possible that the accuracy of manual estimates could be improved by picking from a grey level image of the spectrum. Even in this case, it seems unlikely that significant and consistent improvements over the PIQABLE2 accuracy would be made. Hence, for both speed and reliability, PIQABLE2 would appear to be the best of the procedures studied.

THE MCDPAT PROCEDURE

The basic information which is used for the MCDPAT procedure is a list of the frequencies of main chain amide, alpha and beta protons for each residue (NAB set) and a list of crosspeaks obtained from the NOESY. The NAB sets are

identified from J-correlated spectra (COSY, RELAYED COSY and TOCSY) by relatively simple rules. In our current analyses, this step has been performed by hand. In future, we expect to automate it by making use of PIQABLE2's peak-picking abilities. The MCDPAT procedure requires a set NOESY crosspeaks characterized by both a position and estimated position tolerance so that they can be mapped to specific pairs of NAB protons. The key to easy automation of the search for MCD patterns is the way in which these two types of information are organized. Our approach can be summarized as follows. Let the main chain protons of interest be defined by a set E of objects under study. Information about individual protons or relationships between protons can be viewed as mappings on the set E. Joint properties are then defined by simple set algebra on the inverse images of these mapping. For example, suppose
S_1 = {protons with frequencies in the range (f_1,f_2)} and S_2 = {alpha protons}. Then the intersection of S_1 and S_2 describes the alpha protons with frequencies in the range (f_1,f_2). Note that the frequencies in the NAB list depend upon the peak-picking in J-correlated spectra and it thus is not always be possible to associate each crosspeak in the NOESY with a unique pair of protons. This is partly due to the tolerances on peak positions and partly to the occurrence of degeneracy. If there is a crosspeak at (f_1,f_2) in a 2-D spectrum then it is assumed that there is a short distance relationship between all pairs of protons in the range $(f_1-\varepsilon_1,f_1+\varepsilon_1)$ and protons with frequencies in the range $(f_2-\varepsilon_2,f_2+\varepsilon_2)$, where ε_1 and ε_2 are the peak-picking tolerances. Clearly, the smaller the values of ε_1 and ε_2, the less the chance of including false relationships.

During the MCDPAT procedure, the following information for each main chain proton is kept:
(1) NAB set membership
(2) type (amide, alpha or beta)
(3) frequency from the NAB list
(4) protons with which there may be an NOE
(5) patterns or secondary structural units that the proton has currently been found to participate in.

With appropriate data structures, the MCDPAT procedure is implemented as a relatively straightforward set of logical rules. Searching for patterns and fitting the patterns together uses set algebra operations (e.g. union, intersection, membership, complementation) as basic tools. This gives a very flexible and powerful basis for studying complex relationships between main chain protons. These concepts can readily be extended to include other protons (e.g. side chain spin systems).

The MCDPAT procedure first searches for basic helix H4 patterns (see Figure 2). These are ordered according to the number of characteristic NOEs they exhibit (between 10 and 15). Pieces of helical structure are generated by fitting the overlapping H4 units together, beginning with the H4's of highest NOE status (seeds). If an inconsistency or degeneracy is encountered, a new seed is considered. When all the H4 units have been considered, the pieces of helix are sorted and the elements of NAB sets corresponding to unambiguous pieces of helix are eliminated from further consideration. In this way false relationships, generated during mapping of NAB frequencies onto the set of NOE crosspeaks are eliminated.

The next basic MCD patterns to be found are for anti-parallel sheets (Figure 2); inner loops (I), outer loops (O) and hybrid loops (H). Fitting them together is more complex due to the multiple patterns and the potentially multi-strand nature of sheets. A second level of composite patterns between I and O subunits are first identified (IO_h, IO_v, OIO_h, where the 'h' refers to horizontal patterns connecting only two strands and the 'v' to vertical patterns connecting more than two strands). The OIO_h patterns are the seeds for building bigger pieces of sheets. They are extended out both horizontally and vertically as far as possible using overlapping composite patterns until an ambiguity is reached. At this stage any relevant hybrids can be fitted to the sheets. When all possible combinations have been made, the anti-parallel sheet structures are sorted and unambiguous ones elimated from further search.

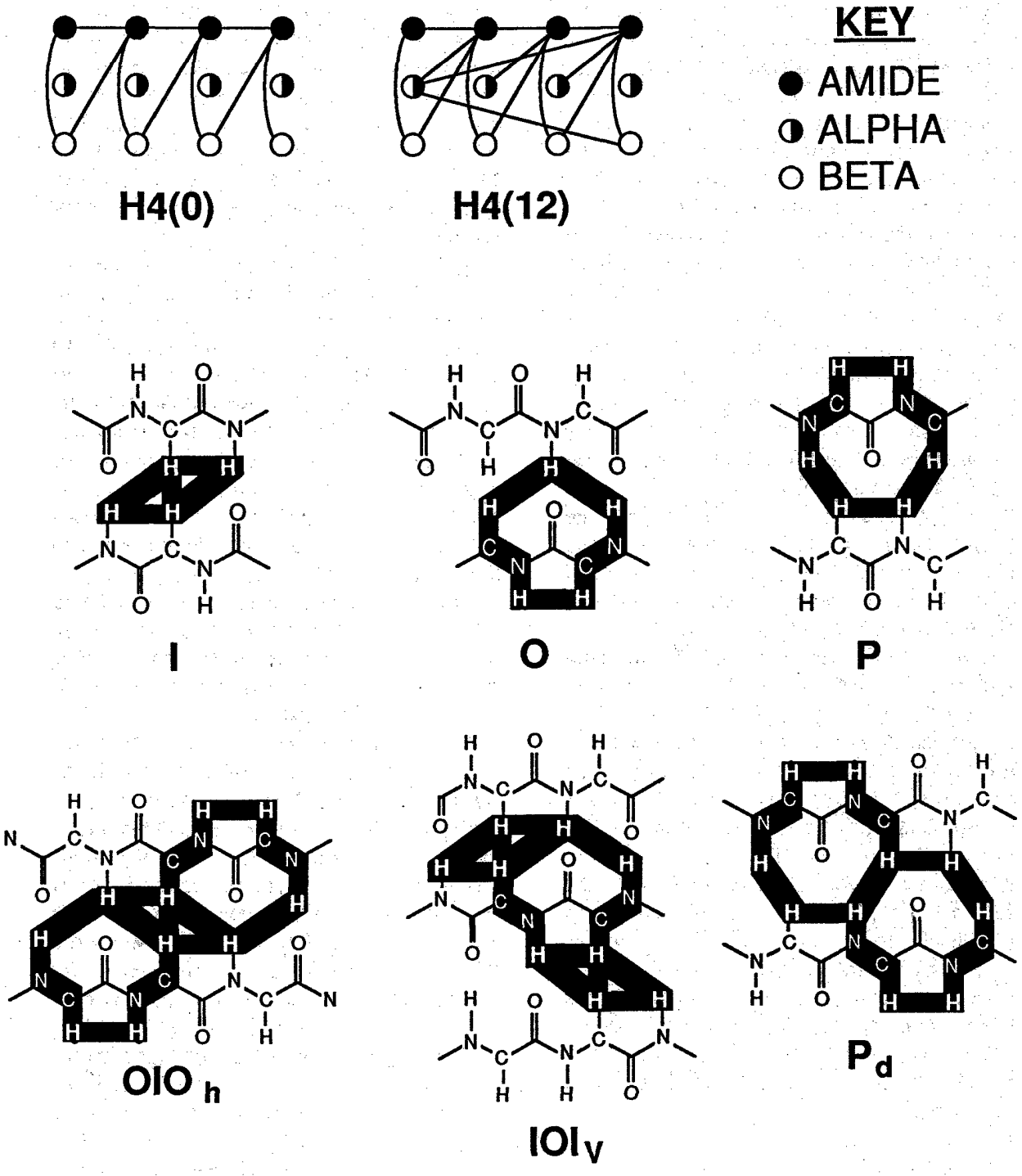

Figure 2: Examples of MCD patterns.

Third comes the study of single loop (P) parallel sheet patterns (Figure 2) and their combination with each other. Again, multiple strand structures are possible. When these structures have been identified they are checked to see if they are joined to pieces of anti-parallel sheets. Finally comes a reconciliation phase where ambiguities are addressed. Any outstanding basic or composite patterns are examined to see if their role has been better defined. This may lead to an iteration of the procedure to build new pieces of structure. If there are remaining ambiguities, they must be resolved by using other criteria such as side chain analysis of the residues involved.

APPLICATION OF MCDPAT

The MCDPAT procedure has been used for both theoretical and experimental studies. A first area where it has been useful is in finding which particular patterns are valuable for the MCD analysis. The original patterns considered where based upon empirical observations (10). With MCDPAT we were able to perform a more extensive analysis of the frequency and fidelity of different patterns. Inter-proton distances were obtained from high resolution crystal stucture data (11) and used to generate dummy "ideal" datasets for MCDPAT. A library of 39 different proteins were considered with between 26 and 287 residues. The data were ideal in the sense that there was no degeneracy (perfect peak-picking and no overlapping frequencies) and that all possible short distances were included (no missing peaks).

This analysis showed conclusively that H4 patterns were present in almost all regions defined as having a helical secondary structure. In addition, their fidelity increased with NOE status and, for cut-off detection distances of 4.2 Å, their NOE status was usually the highest possible (status 12). This high frequency and fidelity provided the motivation for putting them first in the MCDPAT procedure. At low cut-off distances (less than 3.6A), the hybrid subunits dominated for anti-parallel sheet patterns. When the cut-off distance was increased to 4.2A, the inner and outer loops increased in frequency and had higher fidelity. Nevertheless, the inner loops had maximum fidelity only 80% and the outer loops 90%. There are two ways of inproving the fidelity of anti-parallel sheet patterns. The first is by eliminating already identified helix patterns and the second is to form composite patterns. For example, OIO_h patterns have a fidelity of 98% and OIO_v patterns of 100%.

The situation for parallel sheet single loops is much worse and the basic fidelity is as low as 10%. In this case, however, there is a considerable overlap with helix and anti-parallel sheet patterns. Eliminating these before searching for single loops first does leave behind a subset of patterns with relatively higher fidelity. When the full MCDPAT procedure was applied to data for ubiquitin, ribonuclease A and T4 lysozyme, the secondary structures identified corresponded very closely to the crystal structures. The only deviations were within the residues at the end of structural units or within regions where the crystal structure information was ambiguous (13).

To investigate the effect of spectral degeneracy on the frequency of different patterns we chose to study ubiquitin in some detail: from the NMR and crystal structures it contained the three types of secondary structures detected by MCDPAT (18,19). Each main chain proton was given a frequency by sampling at random from empirical distributions for the diferent types of protons. Using the crystal structure to define NOE relationships between protons within 4.2A and assuming a range of tolerances $\pm$ .5 to $\pm$ 2 data point on peak positions, we simulated various levels of degeneracy. The extent of degeneracy had a large effect on the number of NOEs which were identified. As the tolerance increases to $\pm$ 2, approximately 75% of the identified NOEs are false. There are variations depending on the particular types of protons concerned, but even in the best case (amide-amide) 50% of NOEs are false at $\pm$2 tolerance. Similar large increases are found in the total numbers of basic MCD patterns found. However, some patterns are relatively robust to degeneracy. These are in particular the H4 patterns with high NOE status (greater than or equal to 9). Composite anti-parallel sheet

patterns are more robust than single patterns but above ±1 tolerance they exhibit a large number of false patterns. The reason for this is clear if one examines composite inner loop patterns. Unlike the helix patterns, the inner loop involves relationships between single protons from each NAB set. Hence, any degeneracy in the relevant proton will define a false pattern. there are six possible outer loops which overlap with an inner loop. Each overlapping loop requires an NOE to a second proton of one of the NAB sets. Appropriate combinations of four outer loops will therefore place strong limitations on the effect of degeneracy. For example, the existence of an $OOIOO_h$ pattern will confer robustness to the OIO_h subpattern. Using this type of argument, OIO_h patterns can be classified according to their participation in higher level patterns. This order gives a robustness criterion for ordering the seeds in searching for anti-parallel sheet structures. Similar arguments can be applied to single loop patterns which form part of PPP_h composite patterns.

For simulated ubiquitin at zero tolerance two helices are identified in residues 23-33 and 56-59. Up to ±1 tolerance these are still uniquely defined by MCDPAT. At ±1.5 and ±2 tolerance an additional false H4 is found. Anti-parallel sheets are also identified with a single false inner loop up to ±1.5 tolerance. At 2.0 tolerance, many false patterns are found. In the case of parallel sheets, the ambiguity is too severe to identify pieces of sheet even at ±0.5 tolerance without using overlaps between parallel and anti-parallel sheet patterns. This is partly because the parallel sheet in ubiquitin conprises a relatively short run of two strands linking two pieces of anti-parallel sheet.

These simulations were very valuable in defining robust patterns and specifying that cut-off tolerances should ideally be ±1 to ±1.5 data points. This is substantially in the range achieved by PIQABLE2 but is well below the manual peak-picking tolerances (see above). We therefore went on to investigate the ability to define structural units in experimental NMR data for which the NOESY had been peak-picked by PIQABLE2. Our predictions concerning the effect of degeneracy were confirmed. One notable difference was that, particularly for helices, protons within a structure tended to have a higher degeneracy than unrelated protons. This meant that many of the false patterns linked NAB sets within the correct structure but with incorrect orientation. Such inconsistencies were easier to resolve than totally false patterns. Both of the helices from the simulated data were found, but the shorter 3_{10} helix had a problem distinguishing between residues 57 and 63. When the residues participating in the helices were removed, the anti-parallel sheet patterns reduced to two families of composite patterns. By building out to the maximum extent possible residues 3-7 and 13-17 are seen to participate in a sheet together with 42 → 45 and 68 → 71. It was not possible to extend further or find all the parallel sheet structure because of missing crosspeaks close to the solvent track. If we extend with patterns complete except for such crosspeaks, all but a few NOEs at the ends of the sheet are present. This highlights the practical difficulty of missing information. We are currently trying two approaches to solving this problem. Firstly, we are making a more precise definition of partial patterns and secondly, we are studying 600MHz data which has higher resolution and is less affected by the residual solvent track. For larger proteins, it is likely that both types of approaches will be required.

CONCLUSIONS

The combination of PIQABLE2 and MCDPAT are potentially very powerful for studying relationships between main chain protons and identifying secondary structural units based upon NMR data. Remaining difficulties are associated with peak-picking near the solvent track which particularly affcets the definition of sheet structures. This seems to be a feature of both manual and automatic peak-picking procedures. Obtaining better peak resolution by using a higher field strength, together with improving solvent supression or collecting additional spectra at different temperatures in order to move the relevant crosspeaks further away from the solvent will reduce the impact of this technical problem. Anticipated improvements to MCDPAT will

include the use of PIQABLE2 to peak-pick J-correlated spectra and automatic definition of NAB sets. Further refinements will be based upon experience with experimental data for a other proteins. The general approach which we have developed to analysis of multiple relationships between protons may also be of value in studying side chain spin systems or analyzing 3-D proton spectra.

ACKNOWLEDGMENTS: This work was supported by NSF Grant #DIR89-04066 (SJN), NIH Research Grant GM35490 (AJW), by an NIH Postdoctoral Fellowship GM12574 (DMS), by the Pew Memorial Trust and by NIH Grants CA-06927 and QR-05539

REFERENCES

1. Wuthrich K. NMR of Proteins and Nucleic Acids (Wiley, New York 1986).
2. Ernst R.R., Bodenhausen G. and A. Wokaun. Principles of Nuclear Magnetic Resonance in One and Two Dimensions (Oxford University Press, Oxford, 1987).
3. Griesinger C., Sorensen W. and R.R. Ernst. J. Magn. Reson. 73, 574 (1987).
4. Viuster G.W., Boelens R. and R. Kaptein. J. Magn. Reson. 80, 176 (1988).
5. Bax A. and M.A. Weiss. J. Magn. Reson. 71, 571 (1987). 6. Grigley R.H., Redfield A.G., Loomis R.E. and F.W. Dahlquist. Biochemistry 24, 817 (1985).
7. Wuthrich K., Wides G., Wagner G. and W. Braun. J. Mol. Biol. 155, 311 (1982).
8. Billiter M., Braun W. and K. Wuthrich. J. Mol. Biol. 155, 321 (1982).
9. Wagner G. and K. Wuthrich. J. Mol. Biol. 155, 347 (1982).
10. Englander S.W. and A.J. Wand. Biochemistry 26, 5953 (1987).
11. Wand A.J. and S.J. Nelson. Trans.ACA 24, 131 (1988).
12. Pfandler P. and G. Bodenhausen. J. Magn. Reson. 79, 99 (1988).
13. Wand A.J. and S.J. Nelson, in preparation.
14. Nelson S.J. and Schneider D.M. and A.J. Wand, in preparation.
15. Nelson S.J. and T.R. Brown, J. Magn. Reson. 75, 229 (1987).
16. Nelson S.J. and T.R. Brown, J. Magn. Reson. 84, 95 (1989).
17. Nelson S.J. and T.R. Brown. Bull. Magn. Reson. (1990).
18. Vijay-Kumar S., Bugg C.E. and W.J. Cook. J. Mol. Biol. 194, 531 (1987).
19. DiStefano D.L. and A.J. Wand. Biochemistry 26, 7272 (1987). 1987