Implementation of the main chain directed assignment strategy

Computer assisted approach

Sarah J. Nelson,* Diane M. Schneider,[‡] and A. Joshua Wand[‡]

*Department of NMR and Medical Spectroscopy,[‡] Institute for Cancer Research, Fox Chase Cancer Center, Philadelphia, Pennsylvania 19111 USA

ABSTRACT A computer-assisted procedure has been developed to apply the main chain directed (MCD) assignment strategy to the analysis of ^{1}H NMR spectra of proteins. The underlying mathematical foundation of this procedure, termed MCDPAT, is presented. MCDPAT is based upon the expanded library of MCD patterns defined previously (A. J. Wand and S. J. Nelson. 1991. *Biophys. J.* 59:1101–1112), and has been evaluated with both simulated and experimental data from the protein ubiquitin. The influence of the precision, spectral variation, and inherent degeneracy upon the design of the procedure is explored. Several approaches have been taken to overcome the uncertainty introduced by these variables. These include a hierarchical approach to both primary pattern recognition and subsequent construction of MCD-defined units of secondary structure. It is shown that the MCDPAT procedure, in conjunction with automated statistically based spectral analysis, leads to the successful MCD assignment of the protein ubiquitin. The implications and limitations of this approach are discussed.

INTRODUCTION

A critical step in the elucidation of protein structures from NMR data is the assignment of resonances to particular protons. This is conventionally a time-consuming procedure, often requiring extensive interactive analysis of an extensive database of two- and, more recently, three-dimensional NMR spectra (1–7). There have been several attempts to automate parts of the analysis, especially the delineation of J-coupling networks (8–12). The main chain directed (MCD) procedure is a novel approach to making such assignments, which is particularly amenable to automation (13, 14). Its characteristic feature is that it places the major emphasis upon patterns of spatial and bonding relationships between main chain protons. In the preceding paper in this series (15), the underlying basis of the MCD method was re-examined and an expanded library of patterns was developed. These patterns were evaluated for their ability to correctly identify the structural features presented by a data base of 39 protein structures. Several patterns were found to be extremely robust to the structural variations in these proteins and are able to form the basis for a formal application of the MCD strategy. Here we address the implementation of this strategy as a well-defined, mathematical procedure. This procedure is used to perform a computer assisted study of the influence of spectral degeneracy in the primary NMR data upon the utility of the MCD assignment strategy.

The MCD procedure involves the identification of patterns of bonding and distance relationships between protons of amino acid residue amide NH-C$_\alpha$H-C$_\beta$H subspin systems (NAB sets). The ultimate aim is to achieve sequence specific main chain amide, alpha, and beta protein assignments with a minimal initial reliance upon the analysis of J-correlated spectra. The necessary bonding and distance relationships are derived from J-correlated and NOE-correlated spectroscopy, respectively. Hence the primary information which defines these relationships are cross-peaks in multidimensional spectra. As such, they are indirect because they relate frequencies rather than protons. The ability to define these relationships therefore depends upon how accurately cross-peak positions can be determined and whether individual resonance frequencies can be associated with a specific unique proton. In the preceding paper, the underlying frequencies and fidelities of the MCD patterns were examined in the context of the structural features presented by proteins (15). However, to complete an investigation of the utility of the MCD approach, it is clearly necessary to examine the effects of spectral degeneracy upon the analysis.

Dr. Nelson's present address is Department of Radiology, University of California at San Francisco, San Francisco, CA 94143.

Dr. Wand's present address is Department of Biochemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801.

Abbreviations used in this paper: A, alpha hydrogen; B, beta hydrogen; COSY, J-correlated spectroscopy; DQF, double quantum filter; MCD, main chain directed; N, amide hydrogen; NMR, nuclear magnetic resonance; NOE, nuclear Overhauser effect; NOESY, NOE-correlated spectroscopy; RCT, relayed coherence transfer.

0006-3495/91/05/1113/10 $2.00

MATHEMATICAL FORMULATION

A precise, mathematical description of the problem can be made as follows. The information obtained from J-correlated spectra contains information relating particular groups of protons to individual NAB sets. These protons will be denoted as "elements" and summarized at the set E.

$$E = \{\text{elements} - \text{members of NAB sets}\}.$$

The J-correlated data defines relationships or mappings on the set E as follows.

$$(a) \quad S{:}E \rightarrow \{1, 2, 3, \ldots . M_{\mathrm{NAB}}\}$$

$S(e)$ is the number of the NAB set to which the element e belongs.

$$(b) \quad T{:}E \rightarrow \{N, A, B\}$$

$T(\mathrm{e})$ is the type of an element, i.e. amide, alpha, or beta.

$$(c) \quad R{:}E \rightarrow F_1$$

$R(\mathrm{e})$ is the frequency assigned to element e and F_1 is the set of all frequencies of members of the NAB sets. Note that when two elements have very similar frequencies it is not always possible to distinguish them and so the mapping R is not necessarily one to one. If two elements e_1 and e_2 are such that:

$$R(\mathrm{e}_1) = R(\mathrm{e}_2)$$

and they are termed to be "degenerate."

Descriptions of elements with different characteristics can be conveniently made using simple set algebra. For example, let

$$S_1 = \{\mathrm{e} \in \mathrm{E} \text{ such that } R(\mathrm{e}) = f_o\}$$

$$S_2 = \{\mathrm{e} \in \mathrm{E} \text{ such that } T(\mathrm{e}) = N\}.$$

Then $S_1 \cap S_2$ is the set of all amide protons with frequency f_o. Members of this set are degenerate.

The other information is the set of NOE cross-peaks. This set results from a mapping, D, involving pairs of frequencies from the set of all possible frequencies F onto the set $\{0, 1\}$, where 0 implies no cross-peak and 1 implies that the relationship corresponds to a cross-peak.

$$F \times F \rightarrow \{0, 1\}.$$

If $R(\mathrm{e}_1) = f_1$, $R(\mathrm{e}_2) = f_2$ and $D(f_1, f_2) = 1$ then elements (protons) e_1 and e_2 show a nuclear Overhauser effect and are thus close in space. An ideal case for the MCD procedure is provided by the situation where R is a one to one mapping and cross-peaks are perfectly defined. In this situation, as was the case in the preceding study (15), NOE's between individual protons are unambiguously defined. Simple set algebra can then be used to identify unique MCD patterns and then fit them together into pieces of secondary structure.

In practice, R is not one to one and positions of NOE cross-peaks are accurate only to a certain tolerance $f_1 \pm \xi_1, f_2 \pm \xi_2$. The set $\{(\mathrm{e}_1, \mathrm{e}_2)\}$ of pairs for protons for which $R(\mathrm{e}_1)\epsilon[f_1 - \xi_1, f_1 + \xi_1]$ and $R(\mathrm{e}_2)\epsilon[f_2 - \xi_2, f_2 + \xi_2]$ may have several members and it is impossible to distinguish which of these pairs of protons give rise to the identified cross-peak. Hence, all members of the set must be assumed to potentially have an NOE connection. In this manner false or ambiguous patterns are introduced into the analysis and a major part of dealing with real (experimental) data is in picking out the true from the false patterns. One way of doing this is to try to construct "robust" patterns which make use of NOE connections to at least two elements in each NAB set. Another approach is to fit the patterns together and pick out only those which give rise to larger structural units. Both of these approaches are illustrated below.

IMPLEMENTATION OF THE MCD STRATEGY

The key steps of the MCD strategy have been implemented as a collection of computer programs which search for and fit together MCD patterns. These are summarized in Fig. 1 and are jointly termed MCDPAT. The modular nature of MCDPAT has provided great flexibility and allowed rapid testing of numerous alternative strategies for fitting patterns together. In this approach, the lists of NAB sets and NOE cross-peaks are first translated into a form where the relationships between individual elements (protons) can be studied. At least four basic criteria for classifying the elements are available; NAB set, frequency, possible NOE's, and whether it has been found participating in any MCD pattern or higher level structural unit. Other characteristics such as the intensity of NOE's and cross-peak lineshape are potentially of utility but have not been used in the present analysis. The NAB set definitions and NOE list are used to define the fields of a data structure as follows: for each proton, the type (N, A, or B), the other protons in its NAB set and the frequency as obtained from J-correlated spectra are first specified. For a given tolerance on peak positions, main chain protons with similar frequencies are identified and classified as being degenerate. This comprises all the NAB and frequency information required for subsequent analyses.

The NOE list gives relationships between pairs of

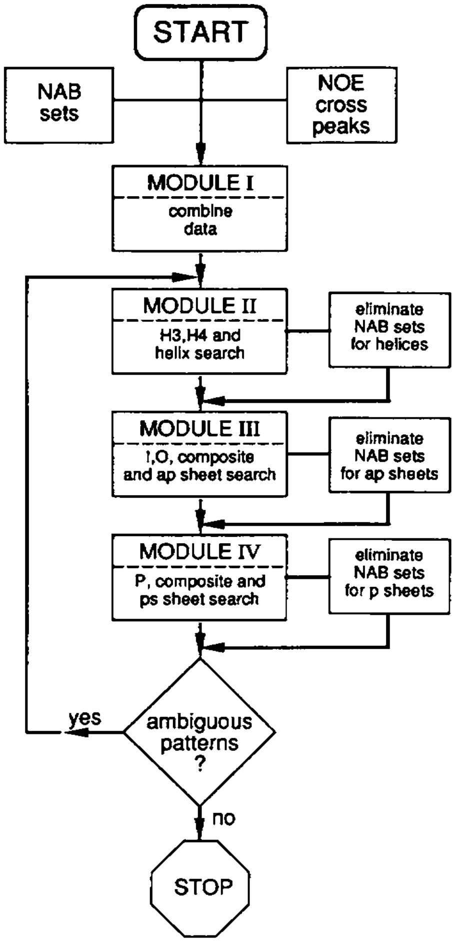
START
NAB sets
NOE cross peaks
MODULE I
combine data
MODULE II
H3,H4 and helix search
eliminate NAB sets for helices
MODULE III
I,O, composite and ap sheet search
eliminate NAB sets for ap sheets
MODULE IV
P, composite and ps sheet search
eliminate NAB sets for p sheets
yes
ambiguous patterns ?
no
STOP

FIGURE 1 Schematic outline of the MCDPAT procedure.

frequencies and tolerances representing their accuracy. All pairs of protons which have frequencies within the tolerances of an NOE cross-peak are assumed to potentially have an NOE relationship. To simplify the pattern search, it is easier to separate the NOE relationships according to whether they are N × N, N × A, N × B, A × A, A × B, B × B. At the end of this phase of the analysis, each proton is fully classified according to NAB set, frequency, and NOE relationships.

Subsequent modules of the MCDPAT procedure use this information to identify the residues which participate in helix, anti-parallel, and parallel sheet structures. The first type of structure to be considered is the helix. Triplets of NAB sets with NOE's between neighboring amides, from an amide to the beta proton in its own NAB set and from the beta proton to the neighboring amide proton on the right are classified as an H3 pattern (15). Pairs of triplets with two overlapping NAB sets are then identified to give the basic H4 unit (Fig. 2). All of these relationships are established using simple set operations: union, intersection, membership, and complementation. Depending on the number of additional N × A and A × B NOE's between these NAB sets, its status is defined according to a scale in the range 1–12 (15).

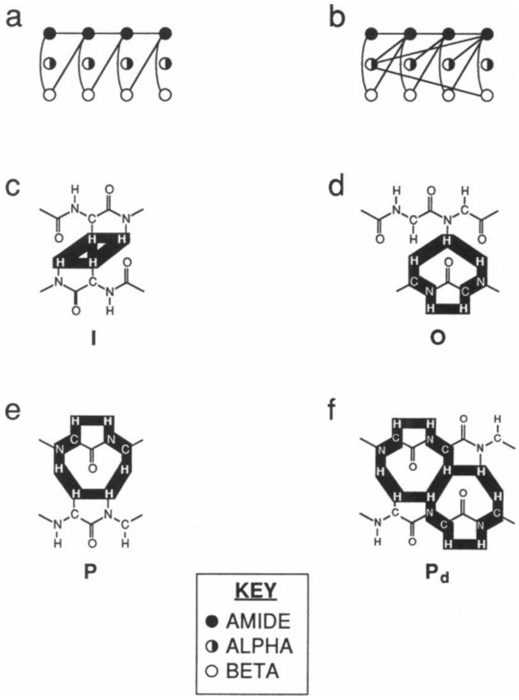

FIGURE 2 Basic MCD patterns which are used extensively in the MCDPAT procedure: (*a*) helix H4-1, (*b*) helix H4-12, (*c*) antiparallel sheet inner loop I, (*d*) outer loop O, (*e*) parallel sheet single loop P, and (*f*) parallel sheet double loop P_d.

To construct extensive pieces of helix, it is necessary to identify a convenient startpoint or "seed" and to build out from there. The most robust helix patterns are H4-12 units (15). Studies with ideal data indicate that they are frequent and show high fidelity for inter-proton distances of up to 4.2 Å. Hence they are the first seeds to be examined. Any other H4 units which overlap by three NAB sets with the seed are added on, followed by H4's which in turn overlap with them. The process is repeated until the longest possible helix is constructed. A new seed is then considered. If all the H4-12 units have been consumed, the H4 with next highest NOE status becomes the new seed. After all consistent pieces of helix are identified, the analysis stops.

The antiparallel sheet structures are considered next. The two major units for building these structures are inner and outer loops. The inner loop is identified as follows (see Fig. 2). Each element i is first tested to see that it is of type "A," is not part of an existing structure and has an NOE to an alpha element j from a different

NAB set. The two sets S_1 and S_2 are then defined.

S_1 = {elements of type "N" with an NOE to element i}

S_2 = {elements of type "N" with an NOE to element j}.

$S_3 = S_1 \cap S_2$ is then the set of all possible amides with an NOE to both i and j. If S_3 has less than two members there is no I pattern involving elements i and j. Alternatively, if S_3 has exactly two members, a single unambiguous I pattern is defined. Finally, if S_3 has more than two members, a number of possible I's can be defined. In the latter case, further criteria such as participation in composite patterns must be used to distinguish which if any, of the I's are real. Again, these relationships are specified by defining appropriate sets and testing for set union, intersection, membership, and complement.

Each inner loop is potentially surrounded by six outer loops (see Fig. 3), which may in turn be attached to other inner loops. Generally, sheets may involve two or more strands, each typically being 3–6 residues long (17). There is likely to be only one inner loop which is totally surrounded by outer loops. The seeds used for antiparallel sheets are OIO_h units which are classified according to whether they participate in higher order patterns $OOIOO_h$ and $OOIOO_{vh}$ (see Fig. 3). Consider first an $OOIOO_h$ pattern. If this has OO_h in common with another $OOIOO_h$ then we have an extensive two-strand region of antiparallel sheet. Alternatively it may overlap by an O unit with only an OIO_h or IO_h. In the latter case the two NAB sets on the exposed side of the I may be ambiguous. If this type of seed is extended as far as possible to each side, a maximum piece of two strand antiparallel sheet is defined. After having built up the largest possible set of overlapping two-strand patterns, vertical connections to additional strands are considered to give three-strand or four-strand units involving overlaps with IOI_v, OIO_v, and OO_v patterns. In this manner, several pieces of antiparallel sheet can be identified. Once inconsistencies are eliminated, remaining patterns can be tested to see if they participate in the sheet units.

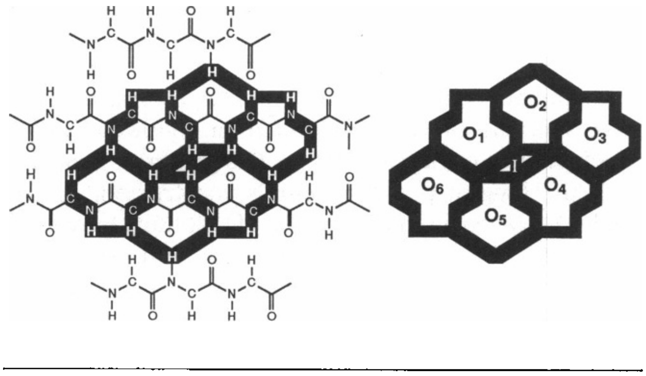

FIGURE 3 Connections between inner and outer loop patterns which are used to construct antiparallel sheets. Composite patterns are as follows: IO_h-IO_1, IO_4; OIO_h-O_1IO_4; $OOIOO_h$-$O_6O_1IO_4O_3$; OO_h-O_1O_6, O_4O_3; OO_v-O_6O_5, O_4O_5, O_1O_2, O_2O_3; IO_v-IO_2, IO_5; OIO_v-O_2IO_5; $OOiOO_{vh}$-$O_1O_6IO_2O_4$, $O_1O_6IO_5O_4$, $O_3O_4IO_5O_1$, $O_3O_4IO_2O_1$.

The third type of structure to be identified is the parallel sheet. The basic single pattern (P) comprises amide-alpha NOE's and J-correlated relationships between four NAB sets (see Fig. 2). Because only four cross-residue set NOE's are involved, it has a low intrinsic fidelity (15) and is very sensitive to degeneracy. This is the reason for considering it after helices and antiparallel sheets have been defined. The basic two-strand seed for a parallel sheet is a pair of intersecting P_d patterns, composite pattern PP_d (see Fig. 4). PP_d units which have a common P_d can be fitted together to form an extended two-strand chain. Once the longest possible unit is defined, vertical connections can be examined using either antiparallel sheet strands (OPI or POI units) or other parallel sheet strands (P_h or P_v units).

The MCDPAT procedure is therefore a very simple, easy to follow series of set operations and is extremely flexible. If there are still ambiguities after helix, antiparallel sheet and parallel sheet patterns have been fitted together, the procedure is iterated. We will illustrate MCDPAT with the results of a study on the effect of degeneracy on the frequency and fidelity of different patterns and on the ability to identify secondary structures.

INFLUENCE OF DEGENERACY ON MCD PATTERN FREQUENCY AND FIDELITY

Studies of MCD patterns in data simulated from crystal structures provides a good indication of the usefulness of different patterns for the case of ideal data (15). As a first step toward applying MCDPAT to experimental data, we added degeneracy to the ideal data. Frequencies were assigned to individual protons from a crystal

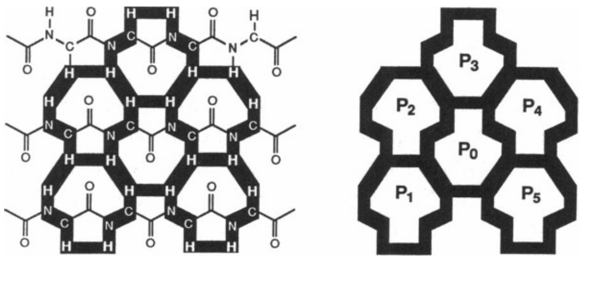

FIGURE 4 Connections between single loop patterns which are used to construct parallel sheets. Composite patterns are as follows: P_d-P_0P_5, P_3P_4, P_2P_3, P_1P_0; P_v-P_3P_0, P_1P_2, P_4P_5; P_h-P_2P_0, P_4P_0; PP_d-; $P_1P_0P_5$, $P_2P_3P_4$.

structure based upon the empirical distributions for amide, alpha, and beta hydrogen resonance frequencies (14) and the NAB set information was constructed. NOE's were then determined by searching for protons which were within a particular cut-off distance of each other and using the corresponding NAB frequencies to construct a list of "cross-peaks." The effect of degeneracy upon the MCD analysis could then be studied as a function of the tolerance (precision) of the frequency positions. The high resolution crystal structure of human ubiquitin (16) was used for these studies. The ^{1}H NMR spectrum of ubiquitin has been manually assigned using a rudimentary version of the MCD strategy (18). In the following we discuss our results obtained with simulated ubiquitin in detail.

NOE's

Table 1 shows the effect of increasing degeneracy on the numbers of NOE's found between members of the NAB sets for ubiquitin. Spectra were simulated using a resolution of 2.8 Hz/point which corresponds to those spectra used for the manual analysis (18). As expected, the numbers of apparent NOE connections which are found increases rapidly with the tolerance. For a tolerance of ±1.0 data points approximately 50% of the apparent NOE connections determined were false and for ±2.0 data points this rose to almost 75%. The situation was less severe for $N \times N$ and $N \times A$ NOE's but a particular problem for $A \times A$, $A \times B$, and $B \times B$ combinations as expected from the empirical-chemical shift distributions (14).

Basic MCD patterns

These ambiguities are reflected in the numbers of basic MCD patterns which are found. Examples are seen in Table 2. For the helix patterns both the H3 and the total number of H4 units increase rapidly with tolerance. However, the H4 patterns with NOE status 5 are more robust and those with NOE status 9 (or above) are extremely robust. Clearly the longer range periodic (i, $i + 3$) NOE connections present in H4-5 and H4-9 patterns significantly reduce the probability of detecting a false pattern.

For the basic antiparallel sheet patterns, (I, O, H), the effect of degeneracy is particularly severe (Table 2). Without being able to make use of some additional information, the chances of separating true from false patterns appears bleak. The reason for the difficulties with the inner loop subunit I are made clear by examining the pattern more closely (see Fig. 2). All the NOE connections involve just one member of each NAB set. Hence, if any of these particular elements are degenerate, false patterns are defined. Unlike the helix patterns there are apparently no additional NOE's which can be used reliably to make this pattern robust.

The problem is not as intractable as it appears, because more robust patterns can be found by forming composite patterns of overlapping I and O units. These had already been found to have higher fidelity for ideal, nondegenerate data. Table 3 compares the effect of degeneracy on the fidelity of the I, IO_h, OIO_h, OOH, IO_v, OO_v patterns as a function of tolerance. Clearly the composite patterns become more robust with the addition of each overlapping O unit.

The parallel sheet basic pattern P and its associated composite patterns are relatively unreliable, even for ideal data. Hence, it is the last type of MCD pattern to be examined and requires first the elimination of relevant helix and antiparallel sheet options. In spite of this, confirmatory information involving composite patterns (either P_d, P_h, or P_v or in combination with antiparallel

TABLE 1 **Simulation of the effect of degeneracy on numbers of NAB set interactions for human ubiquitin**

	Tolerance (data points)[*]				
Number of NOE's detected[†]	0	0.5	1.0	1.5	2.0
$N \times N$	74	84	116	137	149
$N \times A$	247	296	416	506	684
$N \times B$	174	248	386	463	613
$A \times A$	25	39	56	80	139
$A \times B$	110	165	263	350	536
$B \times B$	22	61	131	195	310
Total	652	891	1374	1731	2431

[*] For the spectra appearing in reference 19, this corresponds to 2.8 Hz/pt.

[†] Using a NOE detection limit of 4.2 Å.

TABLE 2 **Simulation of the effect of degeneracy on numbers of basic MCD patterns for ubiquitin**

Pattern	Number found[*] Tolerance (data points)[†]				
	0	0.5	1.0	1.5	2.0
H3	23	25	40	65	88
all H4	18	18	34	77	108
H4-5	11	11	16	23	24
H4-9	8	8	8	9	11
I	9	32	50	80	345
O	26	36	59	77	134
H	1	7	22	39	326
P	20	49	105	411	504
P_d	7	10	41	442	942

[*] Using a NOE detection limit of 4.2 Å.

[†] For the spectra appearing in reference 19, this corresponds to 2.8 Hz/pt.

TABLE 3 **Simulation of the effect of degeneracy on numbers of composite antiparallel sheet patterns for ubiquitin**

Pattern	Number found* Tolerance (data points)‡				
	0	0.5	1.0	1.5	2.0
I	9	32	50	80	345
IO_h	11	16	23	37	165
OIO_h	3	3	4	5	13
OO_h	4	5	7	9	37
IO_v	8	13	19	32	183
IOI_v	3	6	14	14	405
OIO_v	0	0	0	2	20
OO_v	1	1	1	1	1

*Using a NOE detection limit of 4.2 Å.

‡For the spectra appearing in reference 19, this corresponds to 2.8 Hz/pt.

sheet options) often revolves the ambiguities introduced by degeneracy.

Secondary structure

By constructing units of secondary structure beginning with the most robust MCD patterns as seeds, many false patterns are eliminated. This is particularly true for the helix subunits. Table 4 shows the helices which are unambiguously defined for the simulated ubiquitin data as a function of tolerance. Because of the large number of cross-residue NOE's which comprise the H4 patterns, the two helix structures are unambiguous right up to +/−2.0 data points tolerance with only a single false H4 unit at +/−1.5 and +/−2.0 tolerance. Hence, as long as all the NOE's are identified, the procedure for finding helices is very reliable and robust to degeneracy.

The search for antiparallel sheets is more complex because of the two-dimensional nature of the structure. For ubiquitin, there are two antiparallel sheets found in the ideal data: a double strand unit which is seven residues long (APS-I) and a triple strand unit which is five residues at the widest part (APS-II). As the tolerance increases, the ends of the structures become more ambiguous until, at +/−2.0 data points, the three strand unit cannot be clearly separated (see Table 5). There is one false four residue sheet which appears at +/−1.0 data point.

The situation is even worse for parallel sheets. When the residues participating in helix structures are eliminated, the situation still becomes uncertain at +/−1.5 data point. For ubiquitin, a two-strand, five residue long piece of parallel sheet connects the two pieces of antiparallel sheet. If one searches first for parallel structures which are connected to antiparallel units, this parallel sheet is still identified. This is because the two-dimensional nature of the composite structure adds robustness by requiring extra cross-residue NOE relationships. The same would be true for a multistrand parallel sheet structure.

TABLE 4 **Helices defined for simulated ubiquitin data as a function of tolerance**

Tolerance (data points)‡	Residues forming helix*		
	I	II	Others
0	23→33	56→59	—
0.5	23→33	56→59	—
1.0	23→33	56→59	—
1.5	23→27, 26→33	56→59	{52, 43, 50, 67}
2.0	23→27, 26→33	56→59	{52, 43, 50, 67}

*Using a NOE detection limit of 4.2 Å.

‡For the spectra appearing in reference 19, this corresponds to 2.8 Hz/pt.

APPLICATION TO EXPERIMENTAL DATA

To examine the application of MCDPAT to experimental data, results from a previous manual analysis of ubiquitin spectra were used. The majority of NAB sets

TABLE 5 **Sheets defined for simulated ubiquitin data as a function of tolerance***

Tolerance (data points)‡	APS-I	APS-II	Others	APS
0	2→7	69→72		3→7
	17←12	45←41		65→69
		48→50		
0.5	2→7	69→72		3→7
	17←12	45←41		65→69
		48→50		
1.0	2→7	69→71	55,76	4→7
	17←12	44←42	20,22	66→69
		49→50		
1.5	3→7	69→70	55,76	Options found were ambiguous
	16←12	44←42	20,22	
		49→50		
2.0	3→7	Options found were ambiguous		
	16←12			

*Using a NOE detection limit of 4.2 Å.

‡For the spectra appearing in reference 19, this corresponds to 2.8 Hz/pt.

had been identified from the 500 MHz COSY, RCT COSY, and TOCSY spectra. These provided the definition of NAB protons. NOE peak positions were obtained either by manual "peak-picking" of the NOESY spectrum or automatic peak analysis using the PIQABLE2 algorithm (19, 20). Experimental conditions and basic processing of the raw spectral data were as reported previously (18). The computer program FT-NMR (Hare Research, Inc., Woodinville, WA) was used for the manual analysis. In this case, peaks were taken to be at the center of elliptical contours with tolerances defined by the dimensions of the ellipse at the lowest contour level. For the PIQABLE2 analysis, peak positions were determined from the local maxima of baseline subtracted data within the regions designated as containing statistically significant peak regions.

Where possible, the frequencies of the NAB set amide, alpha, and beta protons were recalibrated to be consistent with the NOE frequencies using intra-NAB set cross-peak positions. Having constructed appropriate NAB set and NOE cross-peak files, the MCDPAT procedure was applied. A range of different peak position tolerances were used in an attempt to determine the underlying accuracy of the peak picking procedures.

For manual "peak-picking" 847 cross-peaks were identified. The regions of the spectrum analyzed were from the upper diagonal and excluded the alpha-beta and beta-beta region. Estimated tolerances on peak positions ranged from ±1 to ±6 data points (2.8 Hz/point). The tolerance factors considered for the MCD-PAT procedures ranged from 0.1 to 0.5. In practice, to identify most of the helix subunits indicated from the simulated ubiquitin data, the largest tolerance factor was required. If this was used and only H4 units with high NOE status (≥ 9 for seeds and ≥ 5 to extend from seeds) were considered, the complete α helix from residues 23 to 33 was defined. The short 3_{10} helix (residues 56 → 59) was present, but had a low NOE status and could not be distinguished from false H4 patterns. The degeneracy was relatively high with a total of 375 H4 units, 83 with status greater than or equal to 5, and 4 with status greater than or equal to 9.

For the antiparallel sheet patterns, all except one NOE was present to complete the two strand unit 2 → 7, 12 → 17 and the three strand unit 48 → 50, 42 → 45, 68 → 71. The missing NOE cross-peak corresponded to the amide proton of residue 16 and the alpha proton of residue 15, and would have a position very close to the solvent. Although the patterns were present, even with the elimination of the residues from the helix, it was impossible to unambiguously define these pieces of secondary structure because of the high degeneracy: 393 inner loops (five corresponding to the simulated data) and 381 outer loops. Clearly, without additional information, the manual method of peak picking described here was too inaccurate.

There was good reason to expect that the automatic PIQABLE2 algorithm would perform significantly better. Simulations with low signal-to-noise one-dimensional spectra had been able to identify peak positions to within ±1 to ±2 data points and, in addition, the two-dimensional algorithm is capable of removing the effect of baseline components corresponding to ridge-like artifacts. A range of different peak detection and baseline parameters were used for the PIQABLE2 analysis. As in the manual case, a range of different tolerances were used for applying the MCDPAT procedure (±0.5 to ±2 data points). Many fewer false MCD patterns were identified than in the manual case. For example, there were 81 total H4's with a tolerance of ±1.5 data points and 26 H4's with NOE status greater than or equal to 5. Apart from a single missing NOE at residue 27, the entire α-helix 23 → 33 could be identified using this tolerance and the 3_{10} helix could be pinned down to either residues 56 → 59, or 56, 63, 58, 59.

The basic cores of the antiparallel sheet units 2 → 6, 13 → 17 and 42 → 45, 68 → 71 could be unambiguously defined from the PIQABLE2 data. Three NOE's present in the manual analysis were missing at the ±1.5 tolerance used here. These comprised one alpha-alpha and two alpha-amide cross-peaks which were close to the solvent. There were cross-peaks within ±3 data points of two of the missing NOE's, but the other was absent. This suggests that the cross-peaks close to the solvent are not identified as accurately or reliably with PIQABLE2 as cross-peaks in the rest of the spectrum. From Fig. 5 it is clear that the noise level is higher there and could cause difficulty in identifying cross-peaks. If partial MCD patterns (i.e., with a single NOE missing) are used to build from the basic core region, all the antiparallel sheet units are unambiguously defined.

To examine the benefits of improved spectral resolution, a 600-MHz NOESY spectrum of ubiquitin was also analyzed. Peaks were once more detected using PIQABLE2 and an NOE list obtained. The NAB set frequencies from the 500- to 600-MHz data were remapped using the intra-NAB set NOE's. MCDPAT was then implemented using the PIQABLE2 peak list and modified NAB set frequencies. Almost identical results were obtained as with the 500-MHz data for the helices; the α-helix being fully identified, the 3_{10} helix restricted to either residues 56 → 59 or 56, 57, 58, 61. For the antiparallel sheets, slightly different NOE's were determined, but 2 alpha–alpha NOE cross-peaks were missed. If an IO_h pattern was allowed without the

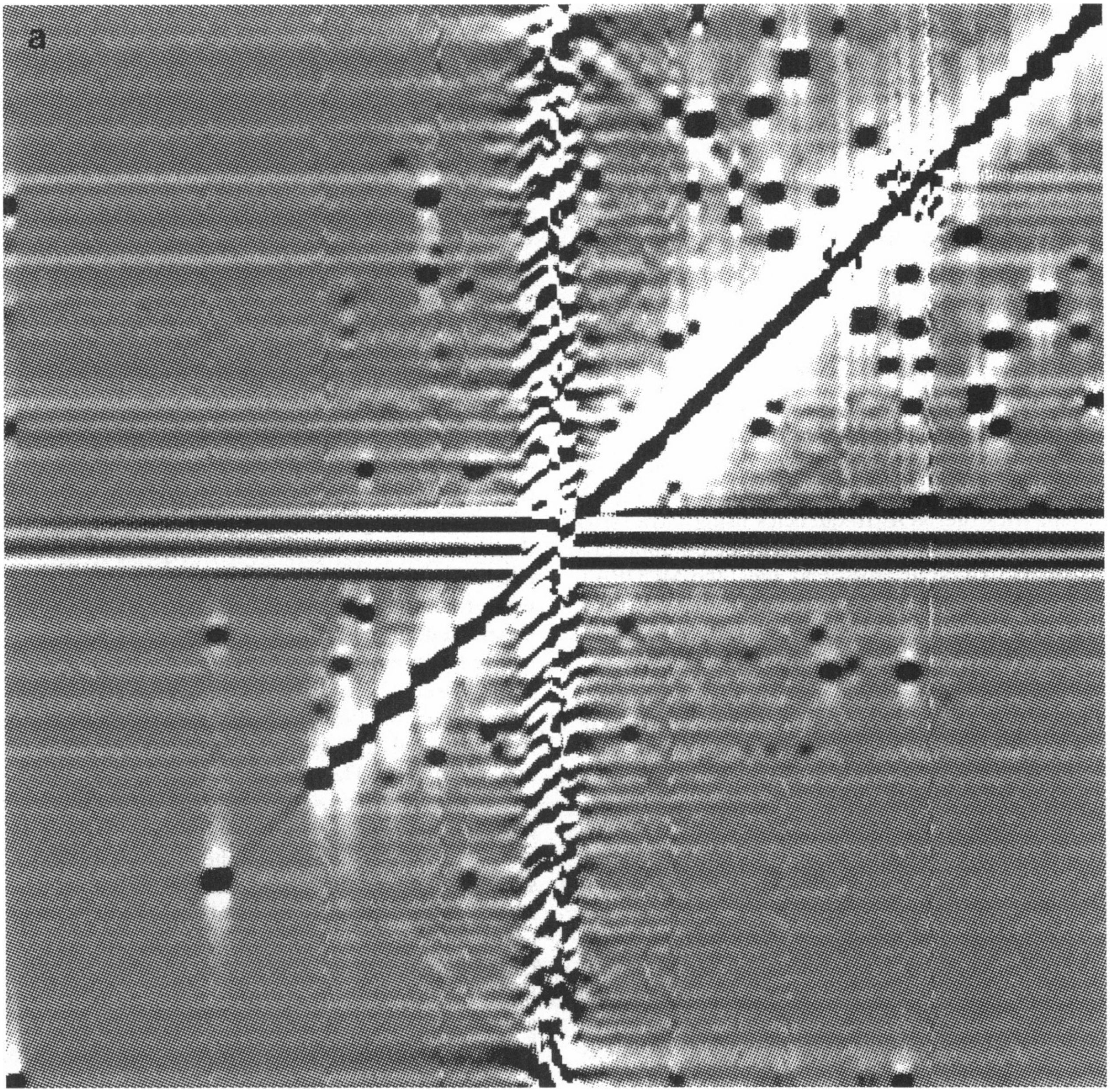

FIGURE 5 Solvent region of 500 MHz spectrum plotted as a grey level map. (*a*) Alpha–alpha regions.

alpha–alpha NOE, the complete antiparallel sheet units were unambiguously defined. Thus, for both the 500- and the 600-MHz data, there were a small number of missing NOE's corresponding to alpha–alpha or amide–alpha resonances close to the solvent. For application of MCDPAT to experimental data, it is therefore important to optimize solvent suppression and pay careful attention to identifying peaks in that region of the spectrum.

CONCLUSIONS

A procedure for identifying MCD patterns and fitting them together has been developed and implemented as the computer assisted procedure MCDPAT. Based upon studies of the crystal structures of human ubiquitin ribonuclease A and T_4 lysozyme, this works perfectly for ideal data (15). When the accuracy of peak detection is

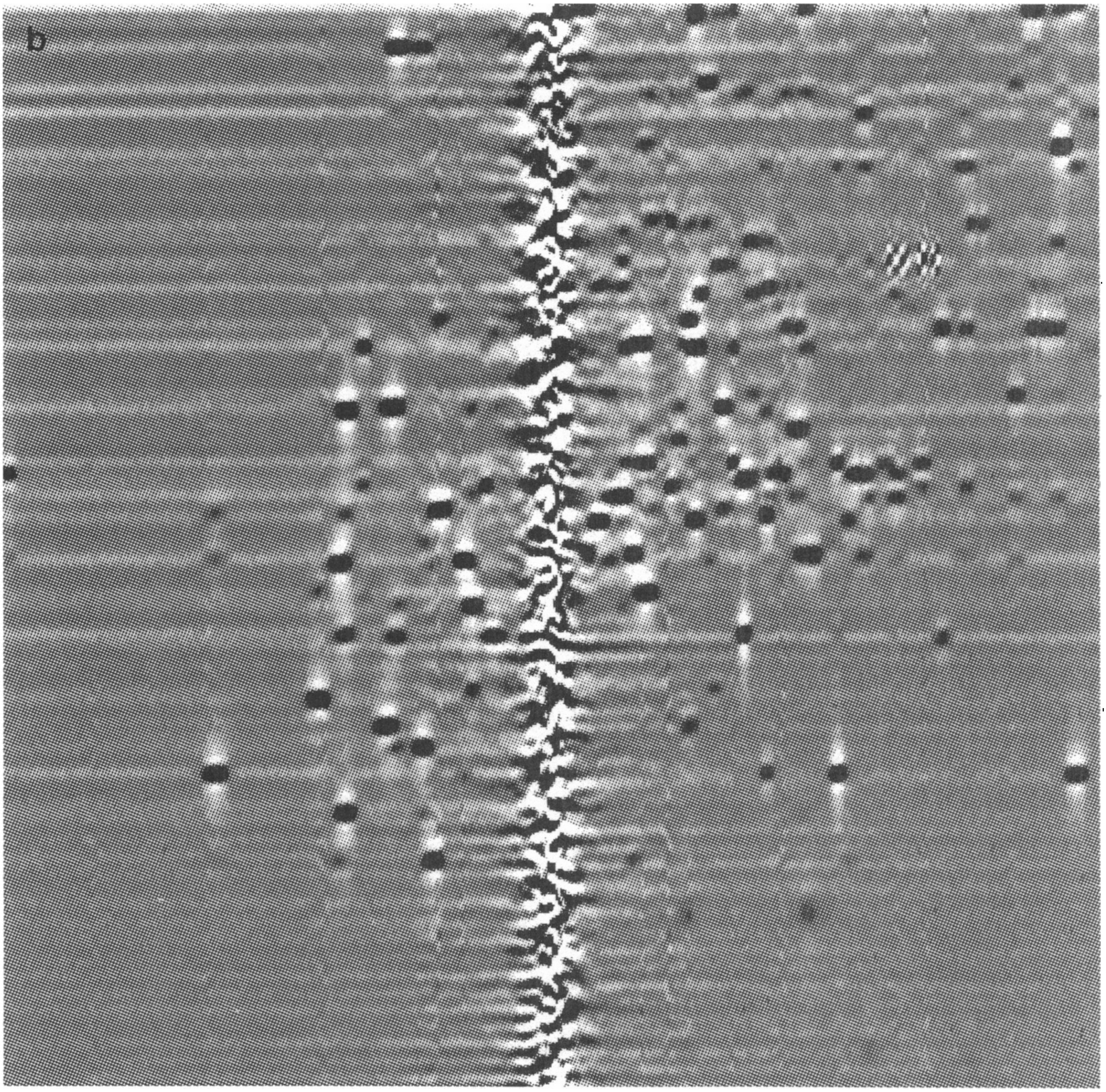

FIGURE 5 (Continued) (*b*) Amide–alpha.

between ±1 and ±2 data points (equivalent here to 2.8 Hz/point), the analysis reduces to a unique or a small number of options which we have been able to resolve on simulated degenerate data. Thus the essential part of the MCD strategy is reliable and robust to degeneracy within the tolerances given.

The successful application of MCDPAT to experimental data depends upon three factors: being able to identify NAB sets from J-correlated spectra, accurately peak-picking NOE cross-peaks, and being able to match the NAB frequencies and NOE frequencies adequately. That NAB sets can be succesfully identified from J-correlated spectra has been confirmed by previous manual analyses of ubiquitin (20). Peak-picking cross-peaks from NOESY spectra has proved more difficult. The manual method that we used for ubiquitin is too inaccurate, producing many ambiguous antiparallel sheet patterns because of degeneracy. Analysis with PIQABLE2 was excellent in all regions of the spectrum apart from close to the solvent (with ±30 data points). From examination of the spectrum, this is not too surprising, as the noise level there is relatively high. However, despite a small number of missing amide–alpha and alpha–alpha cross-peaks, all the helix structures and the core of the two pieces of antiparallel sheet structure for ubiquitin were directly identified using PIQABLE2 cross-peak lists for both 500- and 600-MHz NOESY spectra. If we allowed partial MCD patterns it was possible to extend and complete the antiparallel sheet structures. However, the use of spectra acquired in D_20 to identify alpha–alpha cross-peaks generally makes this step unnecessary. Finally, the successful mapping of the NAB frequencies from the 500-MHz spectra into frequencies from the 600-MHz spectrum, *via* intra-NAB set NOE's, demonstrates the robustness of the procedure to variations in experimental conditions.

In summary, these studies have shown that the combination of PIQABLE2 and MCDPAT successfully identifies units of secondary structure from cross-peaks in two-dimensional ^{1}H spectra. This procedure can be substantially automated and uses only relationships between main chain amide, alpha, and beta protons. Having identified a number of neighboring residues, the analysis of side chain spin systems is considerably simplified and assignment of resonances can proceed much more rapidly.

We thank Deena L. Di Stefano for technical assistance and helpful discussion.

This work was supported by National Institutes of Health (NIH) research grants DK-39806 and GM-35940 (Dr. Wand), National Science Foundation (NSF) research grant DIR89-04066 (Dr. Nelson), NIH grants CA-06927 and RR-05539, the Pew Memorial Trust, by Instrumentation grants RR-2499 (NIH) and DMB84-13986 (NSF), by the Fanny Ripple Foundation and by a gift from Marie Z. Montrose Cole, and by an appropriation from the Commonwealth of Pennsylvania awarded to the Institute for Cancer Research. D.M.S. is the recipient of an NIH postdoctoral fellowship (GM-12594).

Received for publication 18 July 1990 and in final form 10 January 1991.

REFERENCES

1. Wüthrich, K. 1986. NMR of Proteins and Nucleic Acids. Wiley, New York.
2. Ernst, R. R., G. Bodenhausen, and A. Wokaun. 1987. Principles of Nuclear Magnetic Resonance in One and Two Dimensions. Oxford University Press, Oxford.
3. Vuister, G. W., R. Boelens, and R. Kaptein. 1988. Nonselective three-dimensional NMR spectroscopy. The 3D NOE-HOHAHA experiment. *J. Magn. Reson.* 80:176–185.
4. Griesinger, C., W. Sørensen, and R. R. Ernst. 1987. A practical approach to three-dimensional NMR spectroscopy. *J. Magn. Reson.* 73:574–579.
5. Wüthrich, K., G. Wider, G. Wagner, and W. Braun. 1982. Sequential resonance assignments as a basis for determination of spatial protein structures by high resolution proton nuclear magnetic resonance. *J. Mol. Biol.* 155:311–319.
6. Billeter, M., W. Braun, and K. Wüthrich. 1982. Sequential resonance assignments in protein ^{1}H nuclear magnetic resonance spectra: computation of sterically allowed proton-proton distances and statistical analysis of proton-proton distances in single crystal protein conformations. *J. Mol. Biol.* 155:321–346.
7. Wagner, G., and K. Wüthrich. 1982. Sequential resonance assignments in protein ^{1}H nuclear magnetic resonance spectra. *J. Mol. Biol.* 155:347–366.
8. Pfändler, P., and G. Bodenhausen. 1988. Topological classification of fragments of coupling networks and multiplet patterns in two-dimensional NMR spectra. *J. Magn. Reson.* 79:99–123.
9. Meier, B. U., Z. L. Mádi, and R. R. Ernst. 1987. Computer analysis of nuclear spin systems based on local symmetry in 2D spectra. *J. Magn. Reson.* 74:565–573.
10. Pfändler, P., and G. Bodenhausen. 1986. Automated analysis of two-dimensional NMR spectra of mixtures by pattern recognition. *J. Magn. Reson.* 70:71–78.
11. Mádi, Z., B. U. Meier, and R. R. Ernst. 1987. Detection of cross peaks in two-dimensional NMR by cluster analysis. *J. Magn. Reson.* 72:584–590.
12. Novic, M., U. Eggenberger, and G. Bodenhausen. 1991. *J. Magn. Reson.* In press.
13. Englander, S. W., and A. J. Wand. 1987. Main chain directed strategy for the assignment of ^{1}H NMR spectra of proteins. *Biochemistry.* 26:5953–5958.
14. Wand, A. J., and S. J. Nelson. 1988. Refinement and automation of the main chain assignment of ^{1}H spectra of proteins. *In* NMR and X-Ray Crystallography: Interferences and Challenges. M.C. Etter, editor. AIP, New York. 131–144.
15. Wand, A. J., and S. J. Nelson. 1990. Refinement of the main chain directed assignment strategy for the analysis of ^{1}H NMR spectra of proteins. *Biophys. J.* 59:1101–1112.
16. Vijay-Kumar, S., C. E. Bugg, and W. J. Cook. 1987. Structure of ubiquitin refined at 1.8 Å resolution. *J. Mol. Biol.* 194:531–544.
17. Sternberg, M. J. E., and J. M. Thornton. 1977. On the conformation of proteins: an analysis of β-pleated sheets. *J. Mol. Biol.* 110:285–296.
18. Di Stefano, D. L., and A. J. Wand. 1987. Two-dimensional ^{1}H NMR studies of human ubiquitin. A main chain directed assignment and structure analysis. *Biochemistry.* 26:7272–7281.
19. Nelson, S. J., and T. R. Brown. 1987. A method for automatic quantification of one-dimensional spectra with low signal-to-noise ratio. *J. Magn. Reson.* 75:229–243.
20. Nelson, S. J., and T. R. Brown. 1989. The accuracy of quantification from 1D NMR spectra using the PIQABLE algorithm. *J. Magn. Reson.* 84:95–109.