Refinement of the main chain directed assignment strategy for the analysis of ^{1}H NMR spectra of proteins

A. Joshua Wand* and Sarah J. Nelson‡

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

ABSTRACT The underlying basis of the main chain directed (MCD) resonance assignment strategy for the analysis of ^{1}H NMR spectra of proteins is reexamined. The criteria used in the construction of the patterns used in the MCD method have been extended to increase the robustness of the approach to the presence of variable protein secondary structure and significant spectral degeneracy. These criteria have led to the development of several dozen patterns exclusively involving the short distance relationships between main chain amide NH-C_α-H-C_βH (NAB) J-coupled subspin systems of the amino acid residues. The MCD patterns have been examined for fidelity and frequency of occurrence in a database composed of the high resolution crystal structures of 39 proteins. The analysis has identified several extremely robust patterns, suitable for initiating a hierarchical construction of units of secondary structure based upon a systematic analysis of two-dimensional nuclear Overhauser effect spectra. A formal procedure, suitable for the computer assisted application of the MCD strategy, is developed. This procedure, termed MCDPAT, has been applied to the analysis of the crystal structures of human ubiquitin, T_4 lysozyme, and ribonuclease A. It has been found that the MCDPAT procedure is conservative producing no significant errors and is globally successful in correctly identifying the appropriate units of secondary structure contained in these three proteins.

INTRODUCTION

The problem of the assignment of resonances to their parent nuclei remains a considerable challenge even in light of the continuing developments of multidimensional NMR techniques (1, 2). Historically, the ^{1}H nucleus has attracted the most attention owing to its high natural abundance, inherent sensitivity, and the high level of structural and dynamic information contained in the behavior of ^{1}H-^{1}H interactions. In the context of resonance assignment based purely on scalar ^{1}H-^{1}H interactions, the fundamental barrier presented by the polypeptide backbone is the inability to generate direct J-correlated information between neighboring residues in the primary sequence of a protein (3). In recognition of this, Wüthrich and co-workers have designed the sequential assignment strategy (3–6). In this approach, the distinguishing features of different side chain spin systems are used to identify the individual J-coupled spin systems of each residue as to exact type or, at least, place them in a restricted class of residues. The identification rests on the ability to discern predicted behavior of each spin system type or class under the action of different types of NMR experiments which invoke various properties of the through bond interaction. The classification of each residue, by analysis of J-correlated spectra, then provides a means to subsequently use the nuclear Overhauser effect to align individual residues in the primary sequence. This is successful owing to the fact that the conformational space accessible to amino acid residues linked by the trans peptide bond is sufficiently restricted to essentially guarantee that sequentially related residues will always have at least one interresidue distance, formed by combination of amide NH, alpha, and beta protons, within the range detectable by the nuclear Overhauser effect (4). This basic strategy of using the predetermined identity of each isolated J-coupled spin system to provide a means of error checking when determining the correct pathway of NOE connectivities along the polypeptide backbone has been quite successful in the assignment of a number of relatively small proteins (e.g., 7–11). It has become clear, however, that the cornerstone of the sequential assignment strategy, the definition and identification of amino acid J-coupled spin systems, is quite sensitive to the problems presented by proteins of significant size (12). The presence of variable structure, dynamics, and local environment renders the amino acid side chain spin system in extremely plastic entity (12). This plasticity is even more troublesome in the analysis of cross-peak fine structure as evidenced by simulations demonstrating the effect of variable cou-

Dr. Wand's present address is Department of Biochemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801.
Dr. Nelson's present address is Department of Radiology, University of California at San Francisco, San Francisco, CA 94143.

Abbreviations used in this paper: A, alpha hydrogen; B, beta hydrogen; MCD, main chain directed; NOE, nuclear Overhauser effect; NOESY, NOE-correlated spectroscopy.

0006-3495/91/05/1101/12 $2.00

pling constants, strong coupling, and the effect of spin–spin relaxation (13). These observations, in combination with the confusion introduced by the presence of extensive chemical shift degeneracy, makes the definition of J-coupled spin systems of individual amino acids a complicated exercise.

Recently, a complementary ^1H resonance assignment strategy has been proposed whose central goal is to reduce the *initial* reliance upon the analysis of J-correlated spectra (12, 14). This strategy, termed the main chain directed (MCD) method, relies heavily upon pattern recognition in NOE-correlated (i.e., distance correlated) spectra without the knowledge of the origin (residue type) of the protons participating in a given NOE. Such an approach must compensate for the absence of this knowledge. The main chain amide NH-C_αH-C_βH J-coupled subspin system (NAB set) has been determined to be the component of amino acid spin systems most easily identified in J-correlated spectra (12, 14). The NAB set is therefore chosen to provide the elements of each MCD pattern. In previous work a basic set of patterns were examined for their frequency and fidelity in a database of 21 high resolution crystal structures of proteins (12). However, in light of the structural variability presented by proteins and the level of spectral degeneracy present in spectra of proteins of significant size (14, 15) it becomes necessary to develop a more robust collection of patterns.

In this and the following paper we address three issues concerning the confident and efficient application of an MCD-based assignment strategy. Here, the library of patterns is enlarged and evaluated against an expanded database. This evaluation has been facilitated by formulating the procedure into a set of logical rules which provide the basis for a computer-assisted pattern search and lead naturally to assembling the patterns into units of secondary structure. The central issue addressed in this paper is whether the library of MCD patterns can be used to successfully analyze the *structures* presented by proteins. The second paper describes how the logical rules derived from the study of crystal structures have been expanded to treat the ambiguity which is introduced into the analysis by spectral degeneracy. Finally, the second paper also examines the application of the MCD procedure to experimental data and discusses how the difficulties associated with missing or imprecise information can be surmounted.

METHODS

A total of 39 high resolution crystal structures of proteins were obtained from the Brookhaven protein data bank and were used without further refinement. The protein structures used along with their Brookhaven designation and stated resolution are summarized in Table 1. Protons were added using standard bond length and geometry with a revision of a computer program kindly provided by Dr. Peter Wright (Scripps Research Clinic, La Jolla, CA). For the analyses presented here each proton was assigned an arbitrary and nondegenerate chemical shift except for protons of each methyl group which were considered degenerate. Two primary data sets were constructed for each protein. One contained the chemical shift definitions for the amide NH-C_αH-C_βH of each residue. If a residue contained two beta protons only the beta proton closest to its own amide NH was included in this data set. It has been previously shown (12) that this proton will display the largest coupling constant to the alpha proton and is, therefore, the one most likely to appear in J-correlated spectra. The

TABLE 1 **Protein database for statistical analysis of MCD patterns***

Protein	Brookhaven code	Resolution Å	Number of residues
actinidin	2ACT	1.7	218
avian pancreatic polypeptide	1PPT	1.4	36
Bence-Jones immunoglobulin	1REI	2.0	107
concanavalin A	2CNA	2.0	237
cytochrome b_5	2B5C	2.0	85
cytochrome *c* (rice)	1CCR	1.5	113
cytochrome *c* (tuna)	4CYT	1.5	103
cytochrome c_2	3C2C	1.7	112
cytochrome c_3	2CDV	1.8	107
cytochrome c_{551}	351C	1.6	82
dihydrofolate reductase	3FDR	1.7	162
ferredoxin	1FDX	2.0	54
flavodoxin	1FX1	2.0	147
gamma II crystallin	1GCR	1.6	174
hemerythrin	1HMQ	2.0	113
hemoglobin (aquomet)	1ECA	1.4	136
hemoglobin (aquomet horse)	2MHB	2.0	287
hemoglobin (deoxy human)	4HHB	1.7	287
hemoglobin (sea lamprey)	1LHB	2.0	149
high potential protein	1HIP	2.0	85
lysozyme (human)	1LZ1	1.5	130
lysozyme (hen)	1LZT	1.9	129
lysozyme (T4)	2LZM	1.7	164
myoglobin (deoxy sperm whale)	1MBD	1.4	153
mellitin	1MLT	2.0	26
neurotoxin B	1NXB	1.4	62
ovomucoid third domain	2OVO	1.5	56
pancreatic trypsin inhibitor	5PTI	1.0	58
Parvalbumin B	1CPV	1.9	108
phospholipase A2	3BP2	2.1	122
plastocyanin	1PCY	1.6	99
prealbumin	2PAB	1.8	113
ribonuclease A	1RN3	1.5	124
rubredoxin	3RXN	1.5	52
scorpio neurotoxin 3	1SN3	1.8	65
staphylococcal nuclease	2SNS	1.5	141
superoxide dismutase	2SOD	2.0	151
ubiquitin	1UBO	1.8	76

*In the case of oligomeric proteins each subunit was considered separately.

second data set contained a list of all pairs of protons within the protein that were found to be within the given NOE detection limit (cutoff) distance of each other. Each pair of data sets for each protein listed in Table 1 were anayzed using computer programs described in the following paper (16).

RESULTS

The MCD patterns

There are five fundamental MCD patterns. All of these patterns are composed exclusively from scalar and NOE connectivities between amide NH, alpha, and beta protons of amino acid residues. The MCD patterns conceptually associated with beta sheet structure exclusively involve amide NH-alpha CH and amide NH-amide NH NOE interactions. Three patterns compose the fundamental set for the higher order pattern constructions described below for antiparallel strand orientations (Fig. 1). Two, the so-called inner (I) and outer (O) loops, may be combined to generate the hybrid (H) pattern (12). The final pattern, P, represents the fundamental MCD pattern associated with a parallel orientation between two polypeptide chains (Fig. 1). The fundamental helical MCD pattern, H3-1, is composed of three sequential NAB sets and involve amide NH-amide NH and amide NH-beta CH NOE interactions (Fig. 4).

As will be demonstrated explicitly in the following papers, it is advantageous to construct a library of patterns resulting from the fusion of the fundamental patterns. This enforces an additional level of self-correction upon the patterns to be ultimately used. The fusion is guided by following the principle that composite patterns must display at least two NAB sets and at least two NOE's that are common to both fundamental patterns. This is done to take advantage of the fact that although the MCD approach does not require a sequential orientation to the assignment it does require a comprehensive set of interlocking and internally consistent patterns to be globally successful and robust (14). We first consider 18 composite MCD patterns arising from various combinations of the fundamental beta sheet patterns I, O, H, and P defined in Fig. 1. These are summarized in Figs. 2 and 3. These patterns share the common feature that they arise from the fusion of fundamental patterns sharing elements (i.e., N, A, or B) and NOE's involving at least two distinct NAB sets. This property confers a high degree of robustness on nearly all of the patterns presented in Fig. 2. Similar patterns are constructed from fusions of three and four NAB sets involved in helical-like MCD patterns (Fig. 4) and progressively accommodates various combinations of alpha-amide and alpha-beta interactions.

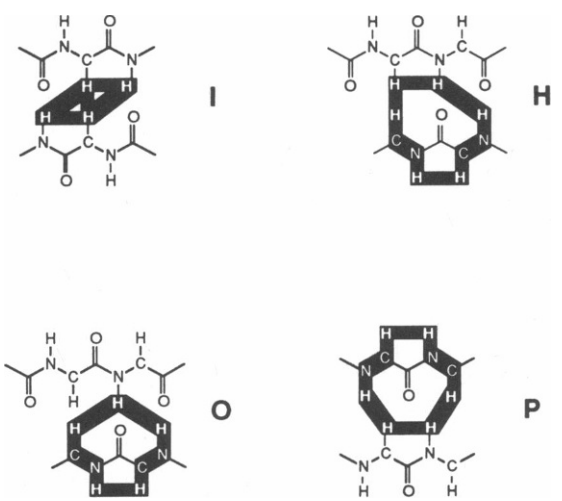

FIGURE 1 Definition of fundamental β-sheet like MCD patterns. Each pattern is highlighted on a polypeptide skeleton.

The empirical basis of the MCD patterns

Unlike the sequential assignment procedure, which requires no knowledge of the structures adopted by polypeptide chains beyond that dictated by the chemical constraints of bond lengths and angles, the MCD patterns, to be useful, must be empirically evaluated against the known structures of proteins. Accordingly, we have examined the behavior of the MCD patterns when applied to 39 high resolutions structures obtained by crystallographic methods. The intent of this is to determine if each pattern occurs with reasonable frequency and, if so, whether the discovered patterns actually represent the limited sequential and spatial inferences of the pattern.

Following definition of NB sets and nondegenerate frequencies for all hydrogens (see Methods), the structures were examined for *all* proton pairs which were within a given distance (ranging from 3 to 4.2 Å) of each other. Each of these pairs were assumed to show NOE connectivities. There is ample evidence that main chain protons 4.2 Å or less apart will consistently show a detectable NOE (6). The frequency and fidelity of each MCD pattern were then determined as a function of this cutoff distance. Lack of fidelity implied that the component NAB sets showed the required NOE's (short distances) but involved residues not meeting the implied sequence requirements. An example of such a false positive pattern is shown in Fig. 5. The performance of the primary and composite sheet-like MCD patterns is summarized in Tables 2 and 3. Similar statistics are presented for the helical-like MCD patterns in Table 4.

As with the helical MCD patterns, the antiparallel

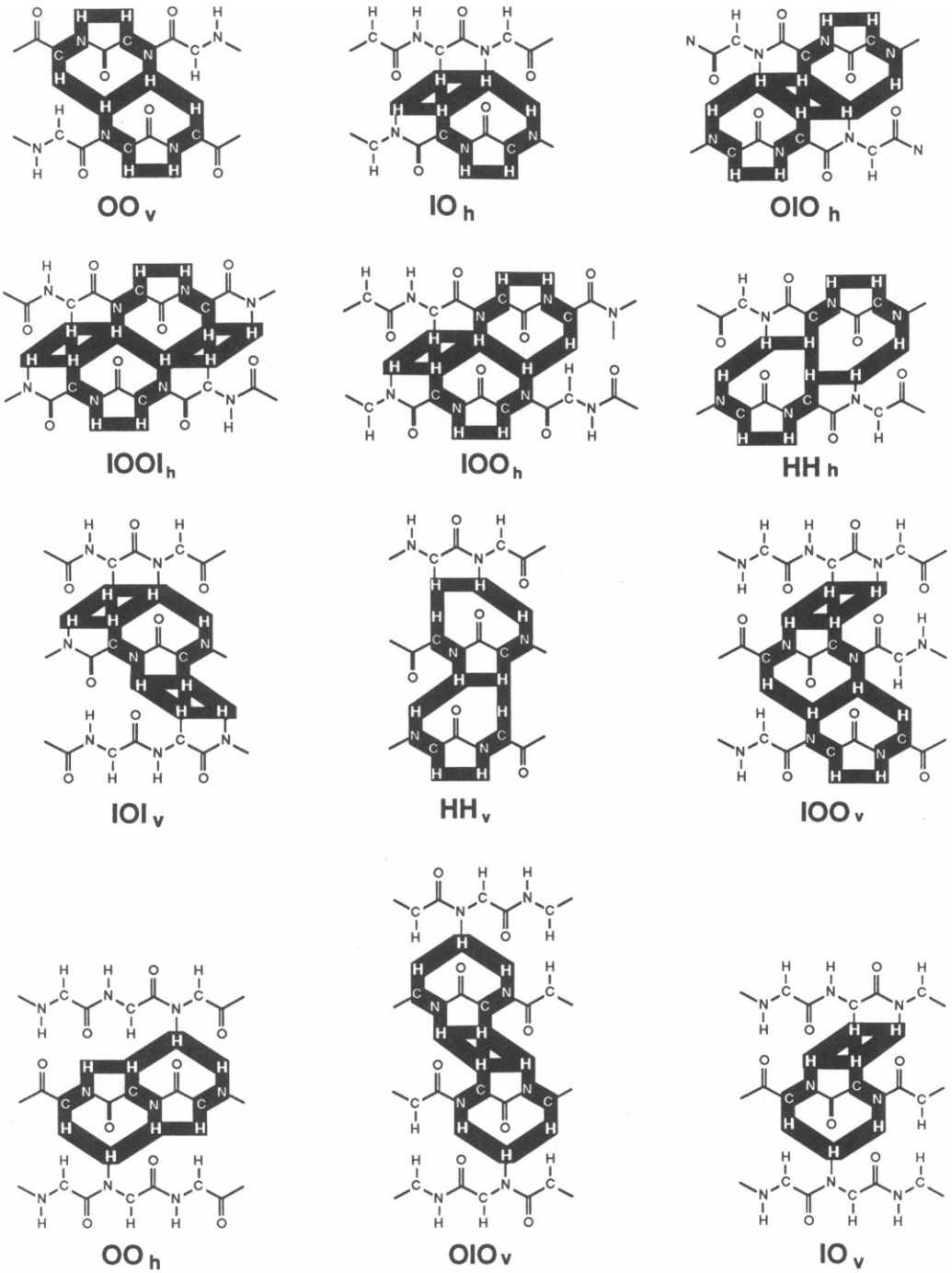

FIGURE 2 Definition of composite antiparallel β-sheet-like MCD patterns. Each pattern is highlighted in a polypeptide skeleton.

sheet-like MCD patterns are sensitive to the distance cutoff employed (Table 2). Significant correlation between involved distances as a function of NOE distance detection limit is observed for high fidelity patterns. A simple illustration is provided by comparison of the behavior of the outer (O) and hybrid (H) patterns. The hybrid pattern results from the absence of the cross-strand αH ↔ NH interaction required to satisfy the companion inner or outer pattern (see Fig. 1). At cutoff distances >3.8 Å, the fidelity of the hybrid patterns

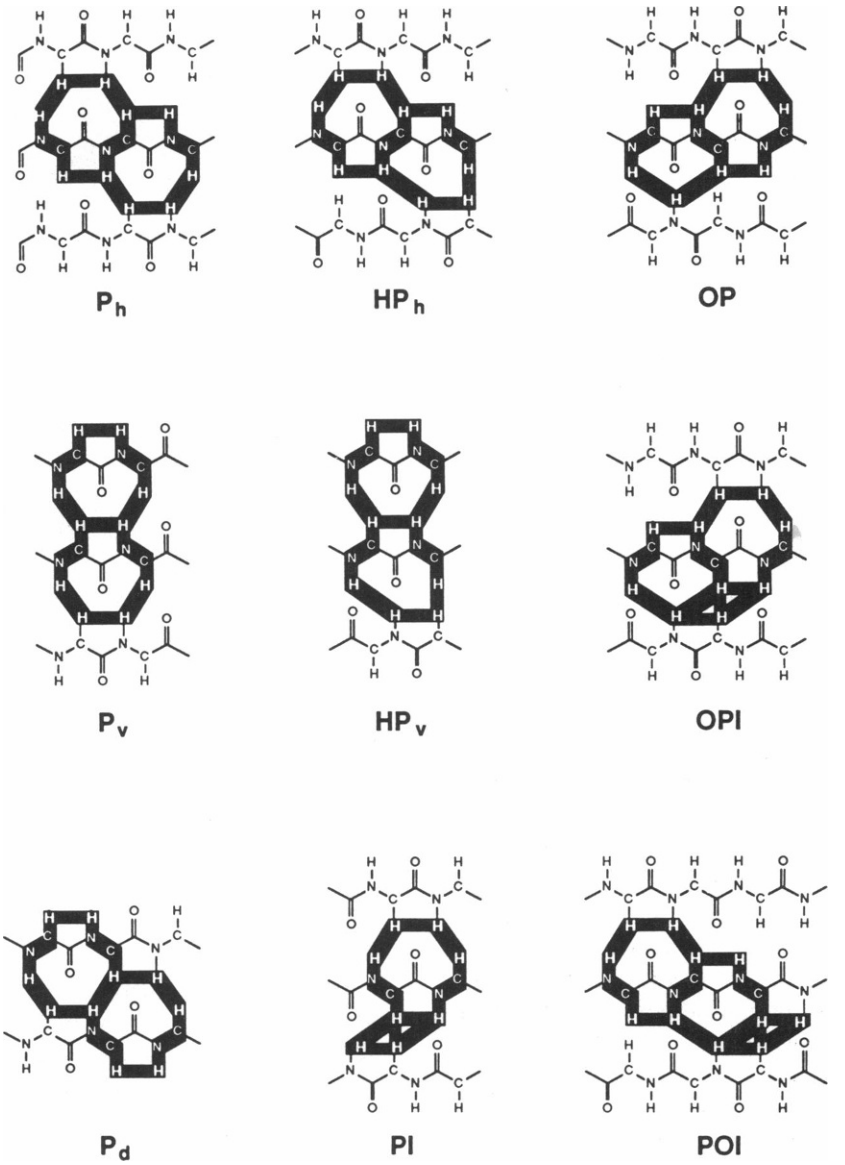

FIGURE 3 Definition of composite parallel sheet-like MCD patterns and composite patterns arising from fusion of I, O, and H fundamental MCD patterns with the fundamental MCD pattern P.

found drops significantly. In contrast, the fidelity of the outer pattern is maintained to 4.2 Å. While not as dramatic as the distance dependence of the helical patterns (see below), correlation of the involved distances is evident. Hierarchical constructions of composite patterns involving the inner and outer patterns display uniformly high fidelity while those involving the hybrid pattern show decreasing fidelity with increasing NOE detection limit. Most importantly, many of the new composite patterns presented here (e.g., OIO_h, OIO_v) display very high fidelity at all cutoff distances examined.

Following the conceptual link between classical protein secondary structure and the general principles guiding combinations of interactions comprising MCD patterns we have also constructed patterns involving parallel sheet orientations (Figs. 1 and 3). The fundamental parallel sheet MCD pattern (P) and its higher order composites show generally poor fidelity at NOE detection distances larger than 3.4 Å.

The helical MCD patterns defined in Fig. 4 represent a hierarchy of interactions and build upon the basic closed loop of ($NH_i \leftrightarrow BH_i \leftrightarrow NH_{i+1} \leftrightarrow NH_i$) connectivities. The behavior of these patterns as a function of NOE detection distance (Table 4) reinforces two gen-

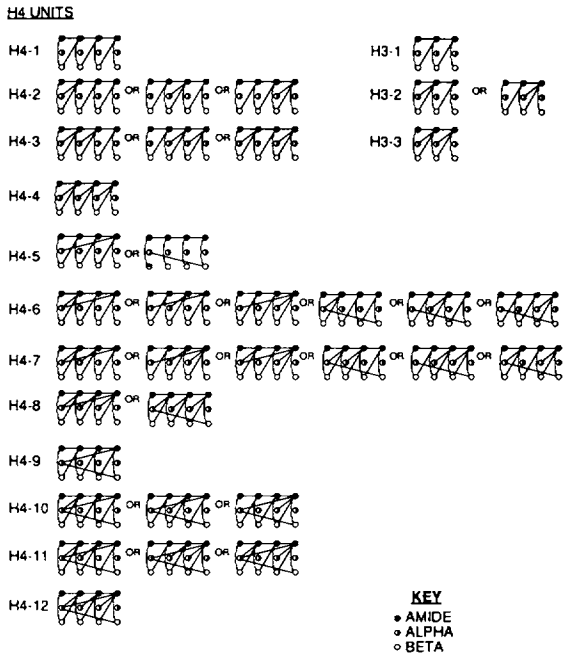

FIGURE 4 Definition of helical-like MCD patterns. Shown are the hierarchical H3 and H4 families of patterns.

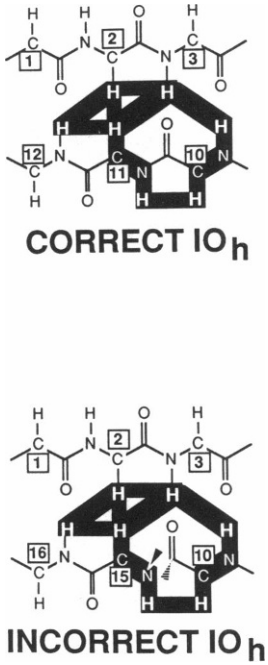

FIGURE 5 Example of a false positive MCD pattern arising from a fortuitous combination of residues satisfying the short distance relationships between main chain hydrogens while failing to satisfy the implied sequence relationships of the involved residues. The MCD patterns have been chosen to minimize the occurrence of these types of false positive patterns.

eral expectations for this hierarchical construction. First, the appearance of each pattern, as opposed to a more complex counterpart (e.g., H3-1 vs. H3-3), is sensitive to the distance cutoff employed. Second, the more complex the pattern the higher is its observed fidelity. The distance-dependent discrimination between helical patterns is due to the participation of NAB sets of both sequential nearest neighbors *and* nonnearest neighbors at distances above 3.6 Å. The appearance of the involved $\alpha H \leftrightarrow NH$ interactions is also highly correlated. For example, those true patterns found at distance cutoffs < 3.6 Å generally convert to the corresponding pattern of highest fidelity (e.g., H3-1 to H3-3; H4-9 to H4-12) at the larger distance cutoffs. Interestingly, the persistence of patterns of intermediate complexity (e.g., H3-2, H4-10) at large NOE detection distances generally correspond to false positive patterns (Table 4).

The MCD strategy

The foregoing results lead naturally to a hierarchical framework with which to construct a general resonance assignment strategy. It is important to emphasize that the NOE detection distances required to observe the necessary interactions comprising the MCD patterns are well within the range of the nuclear Overhauser effect. The use of increasing complexity and enforcing a minimum degree of interaction between fundamental MCD patterns in the composite patterns provides a suitably robust basis upon which to approach the systematic implementation of the MCD method. The expansion of the library of MCD patterns has led to the discovery of several patterns that are extremely robust to the structural variation presented by the protein structure database used. Although patterns showing both reasonably high fidelity and frequency were found previously (12), a quantitative examination of the effects of spectral degeneracy demands that the highest fidelity patterns be found and used (see the following paper). Accordingly, the H4-12, OIO_h, and OIO_v patterns are used to initiate the MCD assignment procedure. The high degree of correlation of the involved distances in true composite patterns causes the true patterns to occur with high frequency. The hierarchical consideration of classes of MCD patterns (i.e., helical, antiparallel, parallel) presented earlier is reinforced by the present results and is maintained. The set operations involving "mapping" and "reconciliation" are described in the following paper. Here we illustrate that this approach can be

TABLE 2 **Frequency and fidelity of antiparallel sheet-like MCD patterns applied to proteins***

MCD	Pattern	Cutoff distance (Å)						
		3.0	3.2	3.4	3.6	3.8	4.0	4.2
I	correct	1	6	25	64	120	199	249
	incorrect	0	1	5	7	20	41	91
H	correct	242	292	298	287	251	194	172
	incorrect	0	0	2	15	36	64	114
O	correct	8	37	118	276	523	963	1540
	incorrect	1	6	11	15	34	81	147
IO_h	correct	0	3	22	78	166	275	347
	incorrect	0	0	1	2	2	10	23
IO_v	correct	0	0	2	17	64	148	202
	incorrect	0	0	1	1	1	4	12
OO_h	correct	0	1	5	16	58	108	143
	incorrect	0	0	1	5	9	15	22
OO_v	correct	0	1	6	32	88	177	241
	incorrect	0	0	1	2	5	15	42
HH_v	correct	96	136	150	124	76	26	10
	incorrect	0	0	0	4	20	22	24
HH_h	correct	137	157	149	103	63	25	17
	incorrect	0	0	1	6	10	7	6
OIO_v	correct	0	0	0	1	9	25	43
	incorrect	0	0	0	0	0	0	0
OIO_h	correct	0	0	3	18	52	98	126
	incorrect	0	0	0	0	0	1	2
IOI_v	correct	0	0	2	11	31	86	126
	incorrect	0	0	0	0	0	3	11
IOO_v	correct	0	0	2	23	97	230	320
	incorrect	0	0	1	1	1	1	4
IOO_h	correct	0	0	3	12	72	145	201
	incorrect	0	0	0	1	4	13	20
$IOOI_h$	correct	0	0	1	4	23	51	77
	incorrect	0	0	0	0	0	0	0

*Summary of the results of analysis of the protein structures listed in Table 1. The number of correct patterns found represents those combinations of short proton–proton distances whose associated amino acids reflect the sequential constraints of the applied pattern (see Fig. 2) and the given distance limit. For pattern involving hybrid loops (H) only those combinations of residues *not* satisfying the associated Inner loop (I) are included. This is in contrast to statistics presented earlier (12).

successfully used to define the collections of residues participating in units of secondary structure in proteins. Equivalently, we show that the procedure is globally robust in the analysis of ideal, nondegenerate NOE-correlated spectra. Three proteins have been analyzed: human ubiquitin, ribonuclease A, and T_4 lysozyme. Table 5 summarizes the number of through space interactions between NAB set elements that must be considered. The analysis was carried out with a data set constructed with a 4.2-Å NOE limit and followed the MCDPAT procedure defined in detail in the following paper (16). All but one of the 15 helical regions of the

TABLE 3 **Frequency and fidelity of MCD patterns involving parallel sheet-like structures***

MCD	Pattern	Cutoff distance (Å)						
		3.0	3.2	3.4	3.6	3.8	4.0	4.2
Pd	correct	20	23	49	56	65	69	62
	incorrect	0	0	0	415	837	1062	1338
P_h	correct	2	2	9	16	21	25	35
	incorrect	0	2	2	6	16	28	73
P_v	correct	2	4	10	17	23	27	35
	incorrect	0	1	1	6	14	30	60
OP	correct	0	0	0	7	19	89	138
	incorrect	3	3	4	8	20	67	67
OPI	correct	0	0	1	3	11	24	44
	incorrect	0	0	0	0	1	9	19
PI	correct	0	0	2	3	14	30	55
	incorrect	0	0	0	0	3	24	59
POI	correct	0	0	0	0	7	19	31
	incorrect	0	0	0	0	3	8	26
HP_h	correct	23	32	33	35	30	32	20
	incorrect	0	1	0	3	15	28	84
HP_v	correct	19	28	32	31	25	22	17
	incorrect	0	0	0	2	6	16	16

*Summary of the results of analysis of the protein structures listed in Table 1. The number of correct patterns found represents those combinations of short proton–proton distances whose associated amino acids reflect the sequential constraints of the applied pattern (see Fig. 3) and the given distance limit. For patterns involving hybrid loops (H) only those combinations of residues *not* satisfying the associated Inner loop (I) are included.

three proteins were recognized. No incorrect alignments were made. One helical region of T_4 lysozyme (residues 82 ↔ 90) was recognized by the MCD analysis as two helical stretches (residues 82 ↔ 85 and 87 ↔ 90) due to the presence of proline at position 86. The undefined helical region of T_4 lysozyme (residues 103 ↔ 108) is recognized as a nonstandard helical region in the crystal (17). All antiparallel sheets in the ribonuclease A and ubiquitin structures were found represented in the MCD analysis. Only one error was made, at the N-terminal end of one strand in the two-stranded sheet of ribonuclease A (see Table 6). The four-stranded sheet of T_4 lysozyme was not defined by the MCD analysis using the given criteria. However, only correct IOO_h and IOO_v structures were recognized and spanned the appropriate residues. In summary, the MCDPAT procedure performed remarkably well and correctly identified the units of secondary structure in these three proteins. No significant false positive helical and antiparallel sheet patterns survived the steps of reconciliation.

DISCUSSION

The library of MCD patterns has been significantly expanded over that presented previously (12). The robustness of the patterns have been evaluated against a database composed a 39 high resolution protein crystal structures. The database spans a wide range of protein size and distribution of type and amount of classical secondary structure. The statistics presented above allowed several patterns that are extremely robust to the structural variation presented by proteins to be identified. The high frequency of appearance of the most complex MCD patterns arises due to the high degree of correlation of involved distances comprising each true pattern. Fortuitous conformations which satisfy the distance relationships of less complex MCD patterns are generally unable to satisfy those of the more complex patterns. Thus, in conjunction with the high degree of correlation of involved distances in "correct" conforma-

TABLE 4 **Fidelity and frequency of appearance of helical-like MCD patterns***

MCD	Pattern	Cutoff distance (Å) 3.0	3.2	3.4	3.6	3.8	4.0	4.2
H3-1	correct	565	738	758	29	0	0	0
	incorrect	3	5	12	5	12	39	77
H3-2	correct	0	1	79	177	7	1	1
	incorrect	0	0	2	46	90	156	250
H3-3	correct	0	0	7	737	1109	1305	1413
	incorrect	0	0	0	0	2	7	19
H4-1	correct	109	115	87	0	0	0	0
	incorrect	0	0	1	0	0	3	13
H4-2	correct	0	0	14	7	0	0	0
	incorrect	0	0	0	0	1	11	28
H4-3	correct	0	0	5	13	0	1	1
	incorrect	0	0	0	13	40	77	127
H4-4	correct	2	0	0	65	79	71	72
	incorrect	0	0	0	0	2	2	4
H4-5	correct	239	330	209	0	0	0	0
	incorrect	0	0	0	0	0	0	0
H4-6	correct	0	0	21	7	0	0	0
	incorrect	0	0	0	0	0	1	2
H4-7	correct	0	0	3	28	0	0	0
	incorrect	0	0	0	1	5	10	15
H4-8	correct	0	0	0	92	89	96	96
	incorrect	0	0	0	0	0	0	0
H4-9	correct	7	65	225	2	0	0	0
	incorrect	0	0	1	0	0	0	0
H4-10	correct	0	0	25	48	0	0	0
	incorrect	0	0	0	0	0	1	2
H4-11	correct	0	0	2	103	5	0	0
	incorrect	0	0	0	1	1	3	3
H4-12	correct	0	0	0	305	658	859	952
	incorrect	0	0	0	0	0	0	0

*Summary of the results of analysis of the protein structures listed in Table 1. The number of correct patterns found represents those combinations of short proton–proton distances where associated amino acids reflect the sequential constraints of the applied patterns (Fig. 4). For any given group of NAB sets tested only the highest order pattern found was counted.

tions, the MCD patterns identified as seeds (e.g., H4-12) form a sufficient starting point for a general assignment strategy. The simple nature of the patterns has allowed the construction of a computer-assisted general assignment procedure (MCDPAT). This is an important result owing to the sheer number of short distance relationships expected in proteins of even moderate size (see Table 5).

A computer assisted implementation of the MCD PAT procedure has been applied to the comprehensive analysis of three proteins: human ubiquitin, T_4 lysozyme, and ribonuclease A. In all three cases, the MCDPAT

TABLE 5 **Summary of the main chain distance relationships in ubiquitin, T_4 lysozyme, and ribonuclease A**

Interaction‡	Protein	Number of expected NOEs* NOE detection limit (Å) 3.0	3.2	3.4	3.6	3.8	4.0	4.2
N × N	Ubiquitin	37	37	39	41	50	54	66
	T_4 Lysozyme	122	126	135	137	142	154	196
	Ribonuclease A	61	65	67	68	80	93	106
N × A	Ubiquitin	127	140	156	183	194	210	221
	T_4 Lysozyme	218	233	267	384	450	479	538
	Ribonuclease A	204	215	233	283	309	335	354
N × B	Ubiquitin	157	174	196	243	283	303	335
	T_4 Lysozyme	419	446	481	602	672	733	771
	Ribonuclease A	297	336	373	451	520	561	611
A × A	Ubiquitin	31	36	48	55	68	80	92
	T_4 Lysozyme	32	39	46	52	55	63	74
	Ribonuclease A	32	36	45	55	68	78	92
A × B	Ubiquitin	100	192	210	220	237	257	285
	T_4 Lysozyme	201	360	374	398	422	457	496
	Ribonuclease A	145	349	371	409	447	492	528
B × B	Ubiquitin	62	64	68	73	76	89	98
	T_4 Lysozyme	134	143	146	156	168	182	199
	Ribonuclease A	114	117	122	134	149	167	184

*Based on the crystal structures 1UBQ (ubiquitin), 2LXM (T_4 lysozyme), and 1RN3 (Ribonuclease A) with hydrogens added with standard geometry.
‡All combinations of amide (N), alpha (A), and beta (B) hydrogens within the given NOE-detection limit were assumed to show an NOE.

procedure unequivocally and uniquely identified all but one of the helical regions of the three proteins. As expected, there was no distinction between α and 3_{10} helices. The lone "helical" region not recognized (in T_4 lysozyme: see Table 6) is, in fact, a disordered nonstandard helical region. Most important, the MCDPAT procedure is highly conservative: no incorrect alignments were identified. A similar level of robustness of the procedure was found for the analysis of antiparallel sheet-like conformations. The success of the search for parallel and fused parallel–antiparallel MCD patterns is, in light of the foregoing statistics, quite remarkable. This is due to the elimination of many potential false positive P patterns arising from residues (NAB sets) participating in helical MCD patterns. This reinforces the hierarchical nature of the MCDPAT procedure.

The general strategy embodied in the MCDPAT procedure provides a convenient, simple, and direct approach to the comprehensive resonance assignment of the ^{1}H spectra of a protein. Though taking a completely different approach it is not expected to eliminate the need for side chain spin system identification in J-correlated spectra. Indeed, a central benefit of an MCD-based approach is the provision of a significant constraint during a *subsequent* analysis of J-correlated spectra. This constraint, presented as the identification of residue type associated with a given NAB set, provides an obvious advantage when used in the identification of a (partially) *defined* but *unknown* side chain spin system. The alignment of a given MCD-defined unit of secondary structure within the primary sequence of the protein does require the identification of a minimal number of residues (e.g., two per helix, two per strand in β-sheets). Fortunately, the dispersal of simple amino acids (e.g., Gly, Ala, Thr) throughout a protein makes the placement of each MCD defined unit of secondary structure a relative easy task. The intervening NAB sets are then identified by association within the MCD unit. Confidence in the approach is provided by the empirically observed statistics presented here and may be further supplemented by additional side chain spin system identification. Indeed, in conjunction with the approaches introduced by Chazin et al. (18), a main chain directed approach would appear to be extremely powerful. These comments are equally applicable to three-dimensional techniques employing ^{1}H-^{1}H interactions alone or in combination with ^{1}H-X interactions.

In addition to the obvious application of the MCDPAT procedure during a general ^{1}H resonance assignment of a protein the approach has a significant albeit less globally informative application. Given that the vast majority of NAB sets can be efficiently and confidently

TABLE 6 **Comparison of the MCD and literature analysis of the crystal structures of ubiquitin, T_4 lysozyme, and ribonuclease A**

Protein	Structure type	Residues defined	
		MCD analysis*	Literature analysis‡
Ubiquitin	Helical§	23↔32	23↔34
		56↔59	56↔59
	Antiparallel‖	2↔7;17↔12	1↔7;17↔10
		69↔72;45↔41	68↔72;44↔40
		45↔43;49↔50	45↔43;48↔50
	Parallel¶	3(65),4↔7;65(3);66↔69	2↔7;64↔69
Ribonuclease A	Helical	4↔11	3↔13
		24↔32	24↔34
		51↔57	50↔60
	Antiparallel	45↔48;85↔81	42↔48;86↔80
		80↔85;104↔99	79↔85;104↔98
		72↔75;110↔107	71↔75;110↔106
		106↔109;124↔120	105↔109;124↔118
		70,72↔74;65↔62	n.d.
	Parallel	47↔49;13↔15	none
		35↔39;34,37↔38	
T_4 Lysozyme	Helical	3↔11	3↔11
		39↔50	39↔50
		60↔76	60↔80
		82↔85;87↔90	82↔90
		93↔105	93↔106
		n.d.	108↔113**
		115↔123	115↔123
		126↔135	126↔134
		137↔141	137↔141
		144↔154	143↔155
	Antiparallel	n.d.‡‡	56↔58;20↔14
			14↔20;27↔24
			24↔27;34↔31
	Parallel	§§	none

*The MCD analysis was carried out on data sets constructed using a 4.2-Å NOE detection limit. See Table 5.
‡Based on the remarks deposited with the crystal coordinates. In the case of ribonuclease A, a more recent analysis (19) was used and sheet interruptions due to bulges considered.
§The MCD analysis for helical regions used H4-12 patterns as seeds and a minimum H4-5 pattern as criterion for extension.
‖The MCD analysis for antiparallel sheet structures used OIO_h and OIO_v patterns as seeds and OO_h, IO_h, IOI_v, and OO_v patterns for extension.
¶The MCD analysis for parallel sheet structures used P_d patterns for both seeds and extension.
**This region is considered by both the MCD and literature analysis, not to be in a standard alpha helical conformation.
‡‡Correct IOO_h and IOO_v structures were found. No incorrect composites were found.
§§Two incorrect 3 residue/strand and two incorrect 4 residue/strand parallel sheets involving 14 residues were found.
n.d. not defined.

identified the MCDPAT procedure provides a means to rapidly and comprehensively assess the secondary structure content of a protein. This would provide a middle ground between structural analyses by, for example, circular dichroism and, for example, by comprehensive structure determination studies by NMR or crystallography.

Supported by National Institutes of Health (NIH) research grants DK-39806 and GM-35940 (Dr. Wand), National Science Foundation research grant DIR-04066 (Dr. Nelson), by NIH grants CA-06927 and RR-05539, by the Pew Memorial Trust and by an appropriation from the Commonwealth of Pennsylvania awarded to the Institute for Cancer Research.

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

REFERENCES

1. Ernst, R. R., G. Bodenhausen, and A. Wokaun. 1987. Principles of Nuclear Magnetic Resonance in One and Two Dimensional. Oxford University Press, Oxford.
2. Kessler, H., M. Gehrke, and C. Griesinger. 1988. Two dimensional NMR spectroscopy: background and overview of the experiments. *Angnew. Chem. Int. Ed. Engl.* 27:450–536.

3. Wüthrich K., G. Wider, G. Wagner, and W. Braun. 1982. Sequential resonance assignment as a basis for the determination for spatial proteins structures by high resolution proton nuclear magnetic resonance. *J. Mol. Biol.* 155:311–319.
4. 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.
5. Wüthrich, K. 1983. Sequential individual resonance assignments in the ^{1}H-NMR spectra of polypeptides and proteins. *Biopolymers.* 22:131–138.
6. Wüthrich, K. 1986. NMR of Proteins and Nucleic Acids. Wiley, New York.
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. Wemmer, D. E., and N. R. Kallenbach. 1983. Structure of apamin in solution: a two-dimensional nuclear magnetic resonance study. *Biochemistry.* 22:1901–1906.
9. Neuhaus, D., G. Wagner, M. Vasak, J. H. R. Kagi, and K. Wüthrich. 1985. Systematic application of high-resolution, phase-sensitive two-dimensional ^{1}H-NMR techniques for the identification of the amino-acid-proton spin systems in proteins. *Eur. J. Biochem.* 151:257–273.
10. Klevit, R. E., G. P. Drobny, and E. B. Waygood. 1986. Two-dimensional ^{1}H NMR studies of histidine-containing protein from *Escherichia coli.* 1. Sequential resonance assignments. *Biochemistry.* 25:7760–7769.
11. Redfield, C., and G. M. Dobson. 1988. Sequential ^{1}H NMR assignments and secondary structure of hen egg white lysozyme in solution. *Biochemistry.* 27:122–136.
12. 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.
13. Widmer, H., and K. Wüthrich. 1987. Simulated two-dimensional NMR cross-peak fine structures for ^{1}H spin systems in polypeptides and polydeoxynucleotides. *J. Magn. Res.* 74:316–336.
14. 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.
15. 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.
16. Nelson, S. J., D. M. Schneider, and A. J. Wand. 1990. Implementation of the main chain directed assignment strategy: a computer assisted approach. *Biophys. J.* 59:1113–1122.
17. Weaver, L. H., and B. M. Matthews. 1987. Structure of bacteriophage T4 lysozyme refined at 1.7 Å resolution. *J. Mol. Biol.* 193:189–199.
18. Chazin, W. J., M. Rance, and P. E. Wright. 1988. Complete assignment of the ^{1}H nuclear magnetic resonance spectrum of French bean plastocyanin. *J. Mol. Biol.* 202:603–622.
19. Wlodawer, A., and L. Sijòlin. 1983. Structure of ribonuclease A: results of joint neutron and x-ray refinement at 2.0-Å resolution. *Biochemistry.* 22:2720–2728.