Journal of Biomolecular NMR, 5 (1995) 183–192
ESCOM

J-Bio NMR 233

Computer-assisted assignment of peptides with non-standard amino acids

Jun Xu[a,*], P.L. Weber[a] and P.N. Borer[b]

[a]*TRIPOS, Inc., 1699 South Hanley Road, Suite 303, St. Louis, MO 63144, U.S.A.*
[b]*Chemistry Department, Syracuse University, Syracuse, NY 13244-4100, U.S.A.*

Received 12 October 1993
Accepted 12 September 1994

Keywords: 2D NMR; Proteins; Automated assignment; Graph theory; Fuzzy mathematics

Summary

A comprehensive peptide assignment program and its application to a cyclic peptide, cyclosporin A, are presented in this paper. A group of graph theoretical algorithms using fuzzy logic are discussed with the aid of examples from cyclosporin A. The algorithms deal with heavily overlapped peaks, recover disjointed and distorted spin coupling networks, and include strategies for sequence-specific assignment. A procedure to extend the Protein Knowledge Base for automatically assigning non-standard amino acid residues is also presented. The program is capable of completely automated assignment for small peptides (~20 residues). For such molecules, it is insensitive to whether the peptide chain is cyclic or acyclic, and to whether amide protons are present or absent. For larger peptides/proteins, more user interaction is required and the sequence-specific assignment step usually must proceed through fragments smaller than the full length to avoid problems due to occurrence of a combinatorial explosion. The program can be applied as a rigorous tool to check manual assignments. The fuzzy graph theoretical concepts built in the program are illustrated with 2D proton spectra of a peptide, but may be extended to higher-dimensional spectra, other biopolymers, natural products and other organic structures.

Introduction

For most aspects of the process of determining high-resolution structures of small biopolymers from multidimensional (mD) NMR, efficient computer programs exist (Ernst, 1991). Programs for time-frequency transformation, signal enhancement, peak picking, peak list accounting, NOE and torsional constraint determination, distance geometry and restrained molecular dynamics calculation, and modelling have been well developed and are commercially available. The crucial step of computerized resonance assignment, however, is not well developed. Several computer-assisted proton assignment software packages have been reported for proteins, such as ANSIG (Kraulis, 1989), EASY (Eccles et al., 1991), CLAIRE (Kleywegt et al., 1991), etc. ANSIG is essentially an assignment support system, or 'electronic drawing board' (Kleywegt et al., 1991). The others are experience-based systems that include a number of programs which can be of assistance in the process of assigning 2D proton NMR spectra of proteins. These experience-based systems emulate manual assignment procedures, starting to identify spin coupling patterns from the amide-α proton coupling region, then ranking the spin patterns with experience-based scoring rules.

We recommend that rule-based assignment software should meet the following design criteria:

(1) The software should automate the creation of spin coupling patterns from spectra.

(2) It should provide automated and user-interactive identification of spin coupling patterns.

(3) It should automate the process of sequence-specific assignment.

(4) The software should be robust toward common overlaps, including peak overlaps and spin coupling pattern overlaps.

(5) The software should handle experimental data set incompleteness, redundant information, artifacts and impurity peaks.

*To whom correspondence should be addressed at: Sadtler Division, BIO-RAD, 3316 Spring Garden Street, Philadelphia, PA 19104-2596, U.S.A.

(6) The software should allow the user to control the assignment procedure at each step.

(7) The software should be general purpose, and should not reject any substructures or residues such as prolines or non-standard amino acids.

This paper will focus on ^{1}H assignments in peptides or proteins, for which the principles of manual assignment are well known (Wüthrich, 1986). Encoding these principles into a software system, however, is not easy. Fortunately, concepts in graph theory and fuzzy logic provide a useful framework. Each type of amino acid has a unique expected spin coupling pattern, which is called the 'Cluster Center'. The Cluster Center is 'fuzzy', where fuzzy implies that the chemical shift set of the Cluster Center may be incomplete and the values may have larger deviations than expected; also, the spin coupling connectivity may be incomplete due to missing peaks, spin degeneracy or other reasons. This fuzzy graph should be described in a rigorous mathematical framework. The program must map an experimentally observed spin coupling pattern onto the Cluster Center of a specific amino acid residue or a set of amino acids. If the mapping is ambiguous, the program should choose the best mapping. If one spin coupling pattern can be mapped onto more than one type of residue due to pattern incompleteness and spin coupling pattern overlaps, then the program should include all possibilities. This mapping problem is addressed in Fuzzy Graph Pattern Recognition, which is an NP-Complete problem in computer science (NP = nondeterministic procedure; Xu and Zhang, 1989).

Peak overlap, artifact peaks and missing peaks are difficult and realistic problems in NMR assignment. In the manual assignment process, skilled spectroscopists are

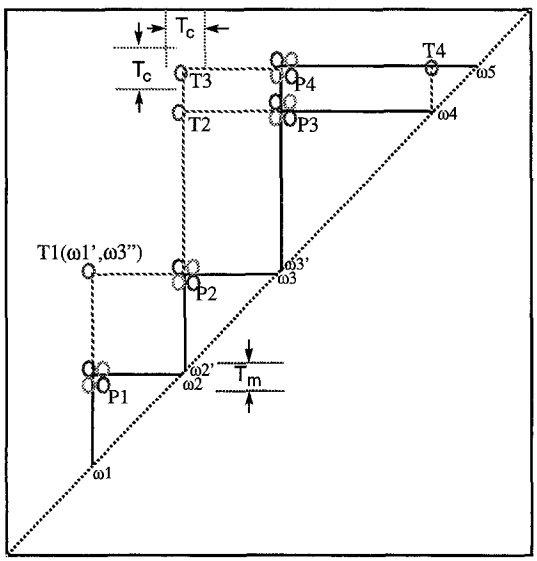

Fig. 1. Superimposition of a TOCSY and a COSY spectrum. T_m is the intra tolerance, which is used to determine if two COSY cross peaks can be merged; T_c is the inter tolerance, which is used to determine whether a TOCSY cross peak can be used to prove that two COSY cross peaks belong to the same spin system. Normally, $T_m < T_c$.

able to solve most of these problems by assessing line shapes, long-distance coupling connectivities, global spin coupling patterns and other spectral information. A mathematical procedure for this spin pattern identification must be clearly defined. A protein may contain a number of the same type of residues, and a spectroscopist assigns them to specific residues by checking NOESY peak connectivities (Wüthrich, 1986), or NOESY and COSY peak pattern connectivities (DiStefano and Wand, 1987; Englander and Wand, 1987; Wand et al., 1989). For larger proteins (number of residues > 40), a 'combinatorial explosion' is possible in the sequence-specific assignment step. This normally occurs when some sequential NOEs are missing because of the local conformation, are hidden below other peaks, or are broadened by internal motion in the molecule (Billeter et al., 1988). The normal strategy is to identify fragments consistent with a chain of NOE connectivities, then search the unused peaks for chains consistent with other fragments of the sequence.

In our previous papers (Xu et al., 1993a,b; Xu and Sanctuary, 1993; Xu and Borer, 1994; Xu et al., 1994), three primary algorithms were developed and described:

(1) a Constrained Partitioning Algorithm (CPA) was developed for automated and rigorous connection of COSY cross peaks to produce spin system patterns;

(2) a Fuzzy Graph Pattern Recognition Algorithm (FGPRA) for mapping these spin system patterns onto residues; and

(3) a Tree Search Algorithm (TSA) for sequence-specific assignment.

Since then, several improvements have been implemented. These include:

(1) a Constrained Partitioning of Spin Patterns Algorithm (CPSPA) for merging disconnected spin patterns;

(2) a Constrained Tree Search Algorithm (CTSA) for more efficient and reliable sequence-specific assignment;

(3) the extension of the residue knowledge base to include non-standard residues or substructures; and

(4) a number of methods to deal with artifact peaks, distorted spin patterns, and disconnected patterns.

All of these will be discussed in this paper. These new developments will be illustrated by assigning the ^{1}H resonances of cyclosporin A in benzene-d_6 using 2D homonuclear spectra.

Algorithms

CPSPA (Constrained Partitioning Spin Patterns Algorithm)

Our original CPA was a globally competitive algorithm. That is, if a COSY cross peak P can be partitioned to more than one spin system, P will be partitioned to the one which has the best MD (match degree) value. The MD is defined in Eq. 1 and illustrated in Fig. 1:

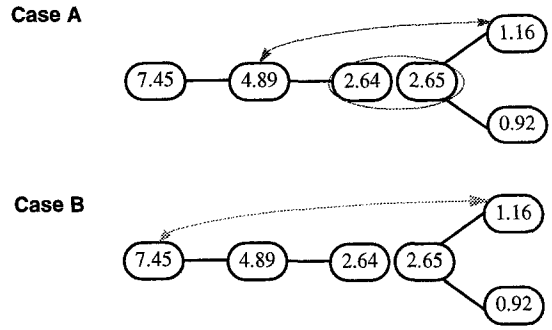

Fig. 2. For a valine spin system, CPSPA can merge Case A (four-bond TOCSY transfer), but not Case B (five-bond TOCSY transfer). Case B will be suggested by the program, and merged with the permission of the user (see the application part of this paper).

TABLE 1
STATISTICS OF SHORT ^{1}H-^{1}H DISTANCES IN PROTEIN CRYSTAL STRUCTURES

Type	Distance (Å)	NOE intensity	Probability j − i = 1 (%)
$d_{NN}(i,j)$	≤ 2.4	Strong	98
	≤ 3.0	Medium	88
	≤ 3.6	Weak	72
$d_{\alpha N}(i,j)$	≤ 2.4	Strong	94
	≤ 3.0	Medium	88
	≤ 3.6	Weak	76
$d_{\beta N}(i,j)$	≤ 2.4	Strong	79
	≤ 3.0	Medium	76
	≤ 3.6	Weak	66

$$MD = 1 - \sqrt{\left(\frac{\Delta_m}{T_m}\right)^2 + \left(\frac{\Delta_1}{T_c}\right)^2 + \left(\frac{\Delta_2}{T_c}\right)^2} \qquad (1)$$

where Δ_m is the deviation of the common frequency between two COSY cross peaks to be merged, and Δ_1 and Δ_2 are deviations of the TOCSY cross peak and the two COSY cross peaks. In Fig. 1, for example, TOCSY peak T1 can be used as the proof to merge COSY peaks P1 and P2, then $\Delta_m = |\omega 2 - \omega 2'|$, $\Delta_1 = |\omega 1 - \omega 1'|$ and $\Delta_2 = |\omega 3' - \omega 3''|$.

When CPA is applied to Fig. 1, COSY peaks P1 and P2 will be merged because of TOCSY peak T1; however, P3 may be merged to the P1-P2 spin system, or to P4, depending on the MD values. If MD(T2, P3, P2) > MD(T4, P3, P4), then eventually P3 and P4 will be partitioned to the same spin system. However, if MD(T2, P3, P2) < MD(T4, P3, P4), then the P3-P4 spin system will be produced first, and the P1-P2 and P3-P4 spin systems will remain disconnected. The TOCSY peaks T2 and T3 would indicate combining P1-P2 and P3-P4 together, but CPA cannot work on two spin coupling networks; it works only on COSY peaks.

CPSPA was developed to solve this problem. The algorithm is described as follows:

```
For (all spin systems from CPA)
{
  i = current spin system;
  for (all spin systems from the (i+1)th spin system to
  the last spin system)
  {
    j = current spin system;
    If (i and j have the same chemical shift within the
    given intra tolerance)
    {
      If (a four-bond TOCSY correlation between i
        and j is found within a given inter tolerance)
      merge spin system j to spin system i;
    };
  };
};
```

CPSPA uses the same intra and inter tolerances as used by CPA. To keep the algorithm rigorous, CPSPA merges only case A in Fig. 2, whereas CAPRI uses another algorithm to deal with case B in Fig. 2.

CTSA (Constrained Tree Search Algorithm)

In our previous work (Xu et al., 1993b), TSA generated a so-called RTG (Residue To spin network Graph relations) supergraph. In RTG, each node is a spin system and edges are NOESY correlations among spin systems. Normally, RTG is a very complicated network. Because there is no way to physically distinguish NOE peaks from neighboring or non-neighboring residues, TSA searches for the sequence-specific assignment by finding the best NOESY connection path in RTG. This non-constrained search is not very robust when it searches through larger peptides.

In order to improve the automated sequence-specific assignment, CTSA also uses as constraints the probability that short ^{1}H-^{1}H backbone distances in proteins are between nearest-neighbor residue pairs (Wüthrich, 1986). These constraints are listed in Tables 1 and 2.

TABLE 2
COMBINED INTERRESIDUE PROBABILITY ESTIMATION BASED ON NOESY CROSS PEAKS

Type	Distance (Å)	NOE intensity	Type	Distance (Å)	NOE intensity	Probability j − i = 1 (%)
$d_{\alpha N}(i,j)$	≤ 3.6	≥ weak	$d_{NN}(i,j)$	≤ 3.0	≥ medium	99
$d_{\alpha N}(i,j)$	≤ 3.6	≥ weak	$d_{\beta N}(i,j)$	≤ 3.4	≥ weak	95
$d_{NN}(i,j)$	≤ 3.0	≥ medium	$d_{\beta N}(i,j)$	≤ 3.0	≥ weak	90

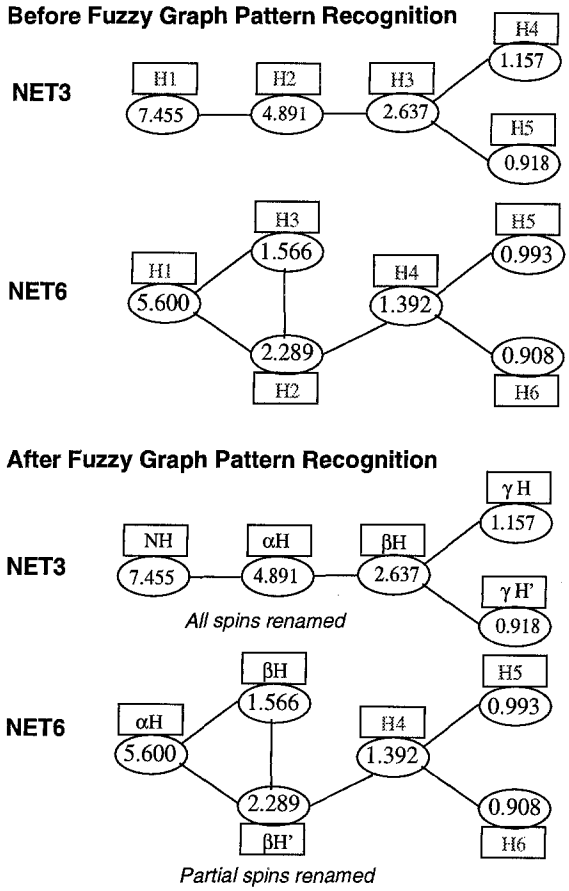

Fig. 3. Backbone protons are recognized and marked by FGPRA. NET3 is uniquely matched to a valine residue, therefore, all spins are renamed. NET6 can be matched with more than one residue and thus it can only be partially renamed, due to potential ambiguities.

In order to apply these constraints to the search procedure, CTSA must be able to distinguish the backbone and non-backbone spins. This is done by FGPRA (Fuzzy Graph Pattern Recognition Algorithm). FGPRA can map spin coupling systems to residue space, i.e., each spin coupling network will be assigned to a set of possible residues. If a spin coupling network is uniquely assigned to a residue, then all chemical shifts will be renamed as the corresponding atom names, such as amide proton, alpha proton, etc. However, if a spin coupling network is mapped to more than one residue, FGPRA cannot rename all chemical shifts because of possible ambiguities, but it can always rename the backbone atoms by marking these backbone protons. CTSA can then use the constraints shown in Tables 1 and 2. For non-peptide residues, the 'backbone' protons can be defined by the user to enable the program to deal with other types of molecules. This procedure is illustrated in Fig. 3.

Sequential NOEs occur between the amide, α, and/or β protons of one residue and the amide proton of the following residue in the chain, which can be represented by Neighboring NOE Matrices (NNMs). For example, if we have a peptide which has the following sequence: Gly-Arg-Gln-Ala-Gly, we can generate the theoretical NNMs as listed in Tables 3–5.

It should be noted that NNM(H^{α}-NH) and NNM(H^{β}-NH) are asymmetrical, and provide efficient constraints for the Tree Search. These matrices are generated from the NOESY peak table and the experimental, backbone-marked spin coupling networks. The matrices are stored to a file in a compressed format. With these constraints, CTSA is much more efficient and robust than the previous TSA.

The extension of the Residue Knowledge Base (RKB)

The Residue Knowledge Base (RKB) contains the spin coupling pattern cluster centers for all 20 standard amino acids. Each cluster center contains a set of expected chemical shifts, their standard deviations and theoretical spin coupling connectivities. Groß and Kalbitzer (1989) have reported a statistical analysis of proton chemical shifts in proteins; their work was used in our previous program (Xu et al., 1993b). Currently, the RKB data

TABLE 3
THEORETICAL NEIGHBORING NOE MATRICES FOR NH-NH CONNECTIVITIES IN THE SEQUENCE Gly-Arg-Gln-Ala-Gly

	Gly^1	Arg^2	Gln^3	Ala^4	Gly^5
Gly^1	0	1	0	0	0
Arg^2	1	0	1	0	0
Gln^3	0	1	0	1	0
Ala^4	0	0	1	0	1
Gly^5	0	0	0	1	0

The matrix elements represent the number of theoretical neighboring NOEs.

TABLE 4
THEORETICAL NEIGHBORING NOE MATRICES FOR H^{α}-NH CONNECTIVITIES IN THE SEQUENCE Gly-Arg-Gln-Ala-Gly

	Gly^1	Arg^2	Gln^3	Ala^4	Gly^5
Gly^1	0	2	0	0	0
Arg^2	0	0	1	0	0
Gln^3	0	0	0	1	0
Ala^4	0	0	0	0	1
Gly^5	0	0	0	0	0

The matrix elements represent the number of theoretical neighboring NOEs.

TABLE 5
THEORETICAL NEIGHBORING NOE MATRICES FOR H^{β}-NH CONNECTIVITIES IN THE SEQUENCE Gly-Arg-Gln-Ala-Gly

	Gly^1	Arg^2	Gln^3	Ala^4	Gly^5
Gly^1	0	0	0	0	0
Arg^2	0	0	2	0	0
Gln^3	0	0	0	2	0
Ala^4	0	0	0	0	1
Gly^5	0	0	0	0	0

The matrix elements represent the number of theoretical neighboring NOEs.

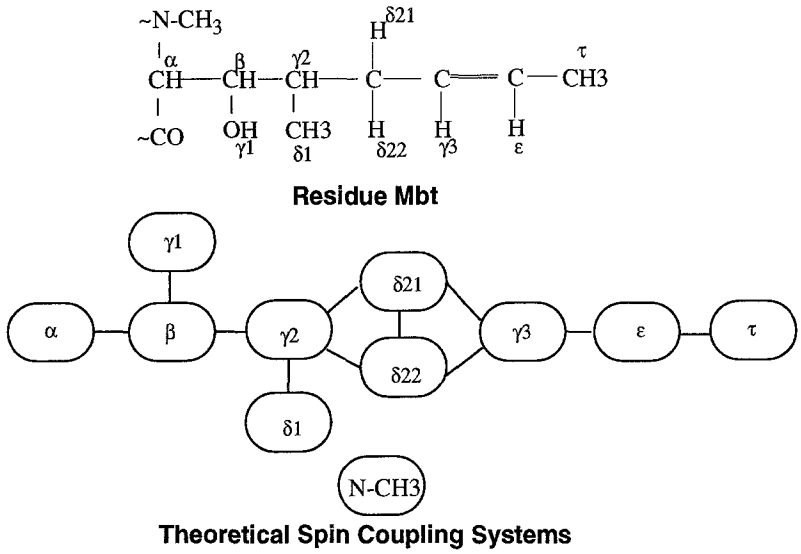

Fig. 4. Theoretical spin coupling network prediction.

come from B. Sykes and D. Wishart (Department of Biochemistry, University of Alberta, Canada).

In order to recognize non-standard amino acid residues or other substructures, RKB needs to be extended. The critical part of this extension is to predict the theoretical spin coupling networks based upon a given structure/substructure. These spin coupling networks are represented in adjacency tables (AT), which should be error-free. An example is given in Fig. 4.

Currently, our prediction algorithm for spin coupling networks considers two- or three-bond proton–proton couplings; it can be extended to consider couplings over more than three bonds if an unsaturated bond is present. When the theoretical spin coupling networks are predicted, the user may add expected chemical shifts and their standard deviations (normally 0.50 ppm is a good default value) to the table, and save this to the Residue Knowledge Base.

Application to cyclosporin A

The structure of cyclosporin A has been well studied, both in the crystal and in solution (Kessler et al., 1985, 1989; Loosli et al., 1985; Lautz et al., 1989,1990; Fesik et al., 1991; Weber et al., 1991). For our work, SIMPLE-COSY, TOCSY and NOESY spectra of cyclosporin A were measured in C_6D_6 at 303 K at 500.1 MHz on a GN-500 spectrometer (Pelczer, 1991). Spectral processing and peak picking were accomplished with TRIAD 6.1.

Nine of the 11 amino acid residues in cyclosporin A are non-standard residues, and seven of them have no amide protons. Based on substructures, we have predicted the theoretical spin coupling networks for these non-standard amino acid residues, namely, MeBmt, Abu, Sar, Me-Val and MeLeu. Residue 8 is a D-alanine, which should have the same theoretical spin coupling network as for the standard L-alanine residue. The expected chemical shifts for these new theoretical spin coupling networks were adapted from similar residues in the RKB. For example, the expected chemical shifts for methylvaline are adapted from those of valine in the RKB. The theoretical standard deviations were derived from the observed standard deviation of similar groups (for example, the amide proton standard deviation was the average of those of all 19 amino acids), which in our experience allows sufficiently large deviations of observed frequencies.

The assignment process refines the peak list. To test the capability of our program, we chose a very low threshold (just above the noise level), and picked all peaks in the COSY and TOCSY spectra. The first output of the program is the list of distinct theoretical spin coupling networks:

```
1 MBT = {MBT1}
1 ABU = {ABU2}
1 SAR = {SAR3}
4 NML = {NML4 NML6 NML9 NML10}
1 VAL = {VAL5}
1 ALA = {ALA7}
1 DA = {DA8}
1 NMV = {NMV11}
```

This is followed by some preliminary statistics:

```
Theoretical COSY Peaks = 57. Experimental COSY
Peaks = 70.
Theoretical Amide-Alpha COSY Peaks = 4.
Experimental Amide-Alpha COSY Peaks = 7.
Connecting COSY Peaks to create Spin Systems
.....
```

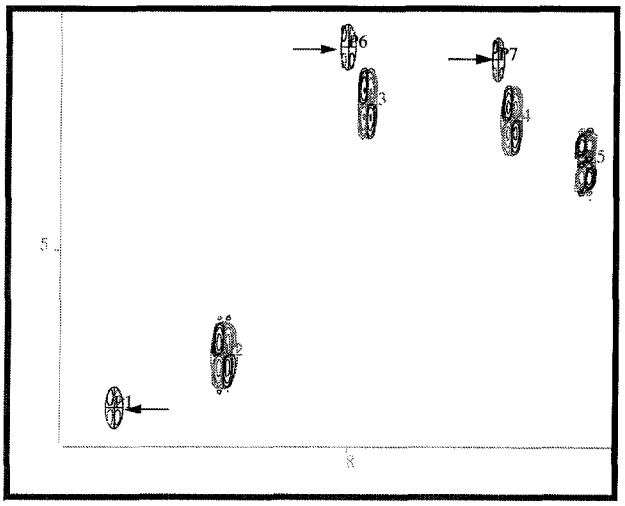

Fig. 5. Three artifact peaks (indicated by arrows) highlighted (in the red circles) in the backbone region of the COSY spectrum of cyclosporin A.

Subsequently, CPA and CPSPA output is obtained:

28 Spin Coupling Networks are created. 14 unconnected peaks.

These spin coupling networks are stored in a set of adjacency tables.

Then FGPRA recognizes these spin coupling networks, and maps them onto the residue pattern space. It also indicates strange patterns as follows:

Fuzzy Graph Pattern Recognition
.....
NET8 is a strange spin system!
.....
NET13 is a strange spin system!
.....

The output of our program indicates that we have picked too many peaks (57 theoretical, 70 picked), and we should delete at least 13 peaks, three of which are extra H^{α}-NH peaks (four theoretical, seven picked). Of course, we should examine our peaks in the spectrum, but which peaks should be deleted? Fortunately, the program suggested these peaks by reporting that 14 peaks are unconnected, and these are put at the end of the Spin System list, and easily highlighted. For example, in Fig. 5, P1, P6 and P7 have been highlighted as unconnected peaks. It is easy to see that they are all weak, and might come from artifacts or a minor conformer.

The program also reports two 'strange' spin systems (NET8 and NET13, see above). By displaying the 'strange' spin systems on the COSY spectrum, we easily see that spin systems NET8 and NET13 are 'strange', because COSY peaks due to a minor conformation accidentally share some frequencies with the larger peaks resulting from the main conformer. In the spectrum display, these peaks are very weak, and we decided to delete them. As an example, NET13 is displayed in Fig. 6.

After examining the COSY peaks with the assistance of the program, a total of 19 COSY peaks were deleted, and the program was run again. The resulting output was as follows:

.....
Theoretical COSY Peaks = 57. Experimental COSY Peaks = 51.
Theoretical Amide-Alpha COSY Peaks = 4.
Experimental Amide-Alpha COSY Peaks = 4.
Connecting COSY Peaks to create Spin Systems
.....
13 Spin Coupling Networks are created. 2 Unconnected peaks.
.....

This time, the result is much improved. As expected, the total number of experimental COSY peaks is a bit less than the theoretical prediction. The number of NH-H^{α} COSY peaks is as predicted. The program should generate 11 spin coupling networks. However, 13 spin coupling networks were reported, two of which come from unconnected COSY peaks. From the theoretical spin coupling topological prediction, residue Sar should have only one COSY cross peak. However, there are still more spin coupling systems present than are theoretically allowed. One alternative is to examine the peak list again to delete possible artifact peaks. Another possibility is that there are incomplete spin coupling systems.

At this point it is often more fruitful to use the program to suggest mergers of spin systems based on TOCSY evidence. The merge condition is:

If ((NET_i and NET_j have at least one common chemical shift within intra_tolerance)&&

TABLE 6
MERGE SUGGESTIONS FOR CYCLOSPORIN A

	NET1	NET2	Pattern OK	TOCSY score	Common shift	Common shift deviation
Suggestion1	NET4	NET11	Yes	0.73	2.189	0.00
Suggestion2	NET5	NET13	Yes	0.82	4.201	0.00
Suggestion3	NET11	NET12	Yes	0.87	2.190	0.01

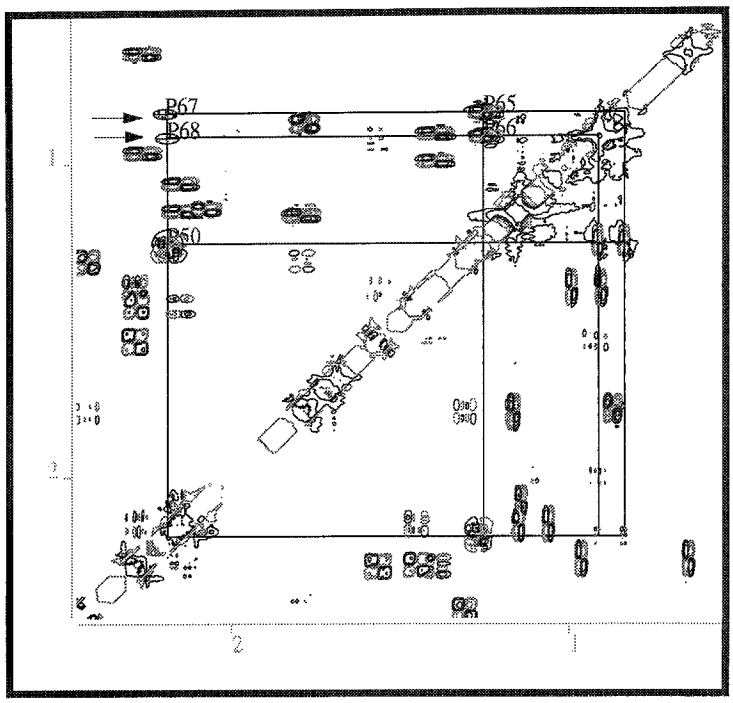

Fig. 6. The 'strange' spin coupling system NET13 is displayed on the COSY spectrum of cyclosporin A. P67 and P68 are weak, and come from a minor conformer or from four-bond couplings.

(At least one long-distance TOCSY correlation peak is found within inter_tolerance))
{
NET_i and NET_j can be merged together;
}

The merge suggestions provided by the program for cyclosporin A are listed in Table 6, and each is examined in turn.

In comparing the three suggestions, there is no strong reason to prefer one over the others: (i) Column 4 ('Pattern OK') indicates that the putative merges are still sub-

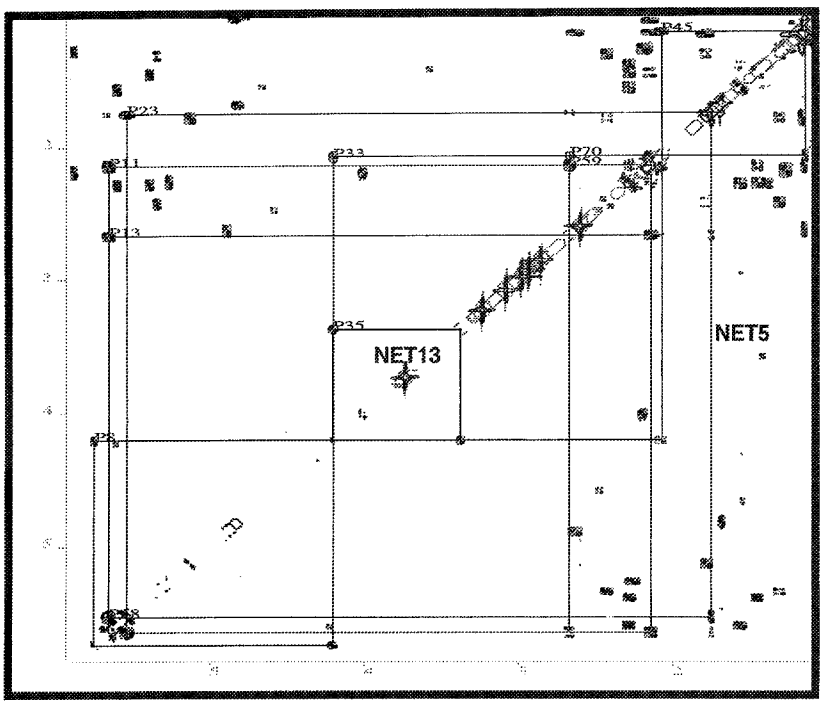

Fig. 7. According to the program's suggestion, spin coupling networks NET13 and NET5 are displayed, and verified to be the same spin coupling network. With the permission of the user, the program will automatically merge them.

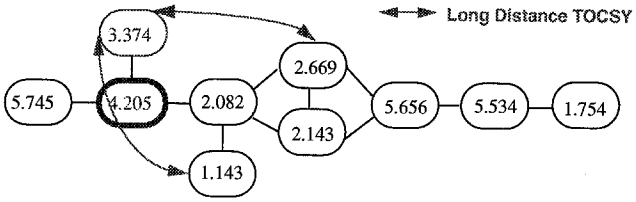

Fig. 8. The spin coupling network resulting from merging NET5 and NET13. Two long-distance TOCSY correlations (arrows) have been found. The reason NET13 was not previously connected to NET5 is that the expected TOCSY peak at (3.374, 5.745) occurs at (3.321, 5.741); the deviation between the expected peak and the observed peak is larger than 0.04 ppm.

graphs of the theoretical spin coupling patterns; (ii) the 'TOCSY Score' is a match degree similar to Eq. 1 and each of the three values are acceptable; (iii) the 'Common Shift Deviation' values are all small. However, only two of the suggestions, at most, can be accepted to leave 11 remaining spin systems.

User intervention now provides the best solution for choosing the most likely networks. As an example, Fig. 7 displays NET5 and NET13. Figure 8 shows how the suggested merger uniquely matches the Mbt spin system (Fig. 4), using TOCSY peaks for Hγ1-Hδ21 and Hγ1-Hδ1 (five-bond transfers). Similar examination of NET4 and NET11 suggests a close match with the *N*-methylleucine spin system. Upon acceptance of this suggestion the program automatically eliminates Suggestion3, which also uses NET11.

Long-range TOCSY transfers can follow many pathways, so at the present time we prefer user interaction over a fully automated approach to resolve such situations. It is interesting that the reason NET5 and NET13 were not joined previously is that the inter tolerance (<0.04 ppm) was too small for CPA and CPSPA to use the four-bond TOCSY peaks corresponding to Hγ1-Hγ2 and Hγ1-Hα, which are observed in the spectrum. It was only by good fortune that the peak picker located the Hγ1-Hδ21 and Hγ1-Hδ1 peaks within the inter tolerance for Hδ21 and Hδ1, thus allowing the program to suggest merging NET5 and NET13. Hγ1 is similar to a threonyl hydroxyl proton, so its chemical shift was substantially different between measurement of the COSY and TOCSY spectra. This sort of error will be common for situations where the chemical shifts of one or more protons are especially sensitive to environmental conditions, e.g., water content, temperature, pH, or isotope effects. This suggests that a useful strategy may also be to repeat CPSPA with looser tolerances after initial partitioning.

The tree search to provide sequential assignment is simpler if all of the spin system ambiguities have been resolved. However, CTSA (illustrated below) could be applied, even for the 13 NETs present before application of the mergers that were just discussed. The simple user interaction just described reduces the searching space and provides a more certain outcome for larger peptides where the possibility of a combinatorial explosion is a concern.

We now have 11 spin coupling networks. They are stored in the Spin System list and recognized by FGPRA. The results of this algorithm can be reviewed in two ways (i.e., GTR and RTG super graphs), see Table 7 and Fig. 9. The RTG (Residue To spin Graph) relation shown in Fig. 9 is the base data structure for the CTSA algorithm.

When CTSA is applied to Fig. 9, it creates the asymmetric NNM matrices based on the input NOESY peak list, then searches for the best path in the figure with the sequence constraints represented in NNM matrices and the sequence information. Because most of the backbone amide protons have been substituted by methyl groups in cyclosporin A, only five neighboring backbone NOEs are found. These are MBT1.αH-ABU2.NH, MBT1.βH-ABU2.NH, NML4.αH-VAL5.NH, NML6.αH-ALA7.NH and ALA7.NH-D-ALA8.NH, shown in Fig. 9 with thick arrows. The thin arrows represent the assignments without NOESY evidence. These assignments are made by using unique matches (for example, NET12 is uniquely matched to SAR3) and spin coupling pattern recognition (for example, NET7 is the only candidate for NMV11, since NET12 has been assigned to SAR3). The tty output of our program for sequential assignment is as follows:

```
Present Assignment Candidates:
MBT1:   NET5*
ABU2:   NET10 NET1 NET7 NET3* NET12 NET2
SAR3:   NET12*
NML4:   NET8*  NET4*  NET6*  NET9  NET7
        NET12
VAL5:   NET12 NET7 NET10 NET1 NET3*
NML6:   NET8*  NET4*  NET6*  NET9  NET7
        NET12
ALA7:   NET1* NET2*
DA8:    NET1* NET2*
NML9:   NET8*  NET4*  NET6*  NET9  NET7
        NET12
```

TABLE 7
SUPER GRAPH GTR (SPIN GRAPH TO RESIDUE RELATION)

Graph name	Candidate1	Candidate2	Candidate3	Candidate4
NET1	Abu	Val	Ala	D-Ala
NET2	Abu	Ala	D-Ala	
NET3	Abu	Val	Mbt	Nml
NET4	Nml			
NET5	Mbt			
NET6	Nml			
NET7	Nmv	Val	Abu	Nml
NET8	Nml			
NET9	Nml			
NET10	Abu	Val		
NET12	Sar	Val	Abu	Nmv

Residue	Candidate1	Candidate2	Candidate3	Candidate4	Candidate5	Candidate6
MBT1	(NET5)					
ABU2	(NET10)	NET1	NET7	NET3	NET12	NET2
SAR3	(NET12)					
NML4	(NET8)	NET4	NET6	NET9	NET7	
VAL5	NET12	NET10	NET1	(NET3)		
NML6	NET8	NET4	(NET6)	NET9	NET7	
ALA7	(NET1)	NET2				
D-ALA8	NET1	(NET2)				
NML9	(NET4)					
NML10	(NET9)					
NMV11	(NET7)					

Fig. 9. Super graph RTG (residue to spin graph relation).

NML10: NET8* NET4* NET6* NET9 NET7 NET12
NMV11: NET7* NET12

Sequential_assignment from MBT1 to NMV11:
Cannot find connected sequentials from MBT1 to NMV11.

Possible sequentials are:
Assign NET5 to MBT1 - P1 P9
Assign NET10 to ABU2 - - -
Assign NET12 to SAR3 - - -
Assign NET8 to NML4 - P5 -
Assign NET3 to VAL5 - - -
Assign NET6 to NML6 - P3 -
Assign NET1 to ALA7 P132 - -
Assign NET2 to DA8 - - -
Assign NET4 to NML9 ???
Assign NET9 to NML10 ???
Assign NET7 to NMV11

The first block is the same as in Fig. 9; a candidate with '*' means that if this candidate is assigned to the corresponding residue, every spin in the residue will be assigned with a chemical shift value. The second block tells us that the program has not found connected sequentials, but it suggests the best possible sequential assignment. The last columns list the sequential NH-NH, H^{α}-NH, and H^{β}-NH NOESY peaks. '–' indicates that the corresponding sequential NOESY peak is not found. '???' means that no sequential evidence is present to distinguish NML9 and NML10. This is expected, as there are no HN atoms in residues 9–11, therefore no sequential NOE is possible. To assign NML9 and NML10, N-CH_3 groups at NML9 and NML10 have to be assigned from other spectra, such as $^{1}H,^{13}C$-COSY and NOESY spectra (Kessler et al., 1985).

Once the sequence-specific assignment is done, the program automatically gives each observed chemical shift with the corresponding atom name, and the chemical shift assignments are stored into a Master Assignment list. Based on this list, the other NOESY peaks are assigned and verified.

In this paper, all of the cyclosporin A protons have been automatically and correctly assigned (Kessler et al., 1985). It should be noted that the present version of the program makes no attempt at stereospecific assignment. The Residue Knowledge Base, CPA, CPSPA, FGRRA, CTSA, and a number of graph theory algorithms to reconnect spin coupling networks, indicate artifact peaks, etc., are merged into a commercial product named CAPRI, available from TRIPOS, Inc., St. Louis, MO. CAPRI is fully integrated with the TRIPOS NMR package, TRIAD, and the modeling package SYBYL.

Conclusions

Both manual and computer-assisted assignment of multidimensional NMR spectra must resolve a number of 'fuzzy' issues: the information is usually incomplete, overlapping, distorted and partially redundant. This makes it difficult to compose computer-assisted algorithms built around 'yes/no' choices. This paper describes a general approach using pattern recognition, built on a framework of graph theory and fuzzy mathematics.

This approach allows the construction of rigorous and easily extendible algorithms for assigning spectra, and combining information from related spectra. Spin coupling networks are graphs that are simple to construct from a molecular structure – these are called theoretical spin coupling networks or spin coupling network Cluster Centers. The spin coupling patterns distinguished in spectra must be fuzzy subgraphs of those predicted from the structure; thus, the pattern matching is quite direct.

Calculations using fuzzy mathematics are distinct from a rule-based approach that uses tolerances or thresholds. First, our program allows spin coupling patterns and subpatterns to be recognized, even if the chemical shifts are far from the expected values. Second, it is possible to map a spectrally observed spin coupling pattern onto more than one spin coupling network Cluster Center. In this case, we use fuzzy mathematics to rank the possibilities and reduce the searching space.

Assignment of multidimensional NMR spectra can be divided into a set of interlocking phases:

(1) peak picking;
(2) identification of spin coupling patterns;
(3) sequence-specific assignment; and
(4) structure determination.

However, the phases are not independent. It is important to perform the peak picking stage with care, to avoid the potential for confusion or 'combinatorial explosion' when critical cross peaks are missed or extraneous peaks are included. At the second stage, our program assists the user

in distinguishing impurity or artifact peaks from real peaks and predicts missing and overlapping peaks. Additional information regarding assignment and peak-list clean up can also appear at the third and fourth stages. Peak-list generation, assignment, structure determination and refinement should generally be regarded as an iterative process.

This paper illustrates the application of our program to the ^{1}H NMR assignment of cyclosporin A, a cyclic peptide where 9 of the 11 residues are non-standard amino acids. The knowledge base of spin coupling Cluster Centers is easily modified to recognize non-standard spectral patterns. 2D COSY, TOCSY, and NOESY are the input spectra. We have shown how CAPRI is used for:

(1) peak-list refinement;

(2) joining spin coupling networks that were fragmented due to initially missing and overlapping peaks, or distorted by partially overlapping spin coupling networks; and

(3) the improved sequence-specific assignment algorithm which was not described previously.

The programs are more developed than those we have reported on before (Xu et al., 1993b). With the CPSPA algorithm, we can extract more complete spin coupling networks than before. For example, if we use CPA only, the program will give 19 spin coupling networks for cyclosporin A, and with CPA and CPSPA combined, this figure is reduced to 13. On the other hand, with CTSA, the sequence-specific assignment becomes more robust. The simple TSA is sensitive to side-chain NOESY peaks. However, CTSA is only sensitive to the quality of NH-NH, H^{α}-NH, and H^{β}-NH NOEs (to deal with proline, NH-H^{δ}, H^{α}-H^{δ}, and H^{β}-H^{δ} NOEs are taken into account).

This methodology is not limited to homonuclear NMR spectra. For example, the TOCSY spectrum can be replaced by H,C-COSY, and no program code modification is needed. It can also work for some aspects of the assignment of natural products and other organic molecules, because these structures can be logically divided into generalized residues.

Acknowledgements

We gratefully acknowledge the contribution of Prof. I. Pelczer (Syracuse University) in measuring and processing the cyclosporin A spectra. The spectra displayed in this paper were processed with Triad 6.1 (TRIPOS, Inc.). P.B. received partial support from NIH Grant GM32691. Assistance from Mr. R. Inch in preparing the manuscript is also greatly appreciated.

References

Billeter, M., Basus, V.J. and Kuntz, I.D. (1988) *J. Magn. Reson.*, **76**, 400–415.

DiStefano, D.L. and Wand, A.J. (1987) *Biochemistry*, **26**, 7272–7281.

Eccles, C., Güntert, P., Billeter, M. and Wüthrich, K. (1991) *J. Biomol. NMR*, **1**, 111–130.

Englander, S.W. and Wand, A.J. (1987) *Biochemistry*, **26**, 5953–5958.

Ernst, R.R. (1991) In *Computational Aspects of the Study of Biological Macromolecules by Nuclear Magnetic Resonance Spectroscopy* (Eds, Hoch, J.C., Poulsen, F.M. and Redfield, C.) Plenum Press, New York, NY, pp. 1–25.

Fesik, S.W., Gampe, R.T., Eaton, H.L., Gemmeker, G., Olejniczak, E.T., Neri, P. and Holzman, T.F. (1991) *Biochemistry*, **30**, 6574–6583.

Groß, K.-H. and Kalbitzer, H.R. (1989) *J. Magn. Reson.*, **76**, 87–99.

Kessler, H., Loosli, H.R. and Oschkinat, H. (1985) *Helv. Chim. Acta*, **68**, 661–681.

Kessler, H., Bats, J.W., Wagner, K. and Will, M. (1989) *Biopolymers*, **28**, 385–395.

Kleywegt, G.J., Boelens, R., Cox, M., Llinas, M. and Kaptein, R. (1991) *J. Biomol. NMR*, **1**, 23–47.

Kraulis, P.J. (1989) *J. Magn. Reson.*, **88**, 601–608.

Lautz, J., Kessler, H., Blaney, J.M., Scheek, R.M. and Van Gunsteren, W.F. (1989) *Int. J. Pept. Protein Res.*, **33**, 281–288.

Lautz, J., Kessler, H., Van Gunsteren, W.F., Weber, H.P. and Wenger, R.M. (1990) *Biopolymers*, **29**, 1669–1687.

Loosli, H.R., Kessler, H., Oschkinat, H., Weber, H.P., Petcher, T.J. and Widmer, A. (1985) *Helv. Chim. Acta*, **68**, 682–704.

Pelczer, I. (1991) *J. Am. Chem. Soc.*, **113**, 3211–3212.

Wand, A.J., DiStefano, D.L., Feng, Y., Roder, H. and Englander, S.W. (1989) *Biochemistry*, **28**, 186–194.

Weber, C., Wider, G., Von Freyberg, B., Traber, R., Braun, W., Widmer, H. and Wüthrich, K. (1991) *Biochemistry*, **30**, 6563–6574.

Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York, NY.

Xu, J. and Zhang, M. (1989) *Tetrahedron Comput. Methodol.*, **2**, 75–83.

Xu, J., Gray, B.N. and Sanctuary, B.C. (1993a) *J. Chem. Inf. Comput. Sci.*, **33**, 475–489.

Xu, J. and Sanctuary, B.C. (1993) *J. Chem. Inf. Comput. Sci.*, **33**, 490–500.

Xu, J., Straus, S.K. and Sanctuary, B.C. (1993b) *J. Chem. Inf. Comput. Sci.*, **33**, 668–682.

Xu, J. and Borer, P.N. (1994) *J. Chem. Inf. Comput. Sci.*, **34**, 349–356.

Xu, J., Straus, S.K. and Sanctuary, B.C. (1994) *J. Magn. Reson. Ser. B*, **103**, 53–58.