JOURNAL OF MAGNETIC RESONANCE, Series B 103, 53–58 (1994)

NOTES

Use of Fuzzy Mathematics for Complete Automated Assignment of Peptide ^{1}H 2D NMR Spectra

JUN XU,* SUZANA K. STRAUS,† B. C. SANCTUARY,*,‡ AND LAIRD TRIMBLE§

Department of Chemistry, McGill University, 801 Sherbrooke Street West, Montreal, Quebec, Canada H3A 2K6; †Laboratorium für Physikalische Chemie, ETH, 8092 Zurich, Switzerland, and §Merck Frosst Canada, Inc., P.O. Box 1005, Pointe Claire-Dorval, Quebec, Canada H9R 4P8

Received December 15, 1992; revised April 7, 1993

Modern NMR has become a powerful tool for the determination of protein and polypeptide structures. For small proteins (<100 residues), Wüthrich's strategy (*1*), which involves tracing and identifying spin coupling networks, mapping the spin coupling networks to individual amino acid residues, and detecting interpattern correlations, is sufficient to make complete sequence-specific assignments. If done manually, these are tedious and time consuming. Therefore, there is a great need for a completely automated assignment methodology. For larger proteins, multinuclear and higher-dimensional NMR experiments are used with concomitant assignment strategies. Given that the assignment is even more complicated for these large biopolymers, having such an automated methodology is even more important.

To develop an automated methodology, the following obstacles must be taken into consideration: (i) distinguishing random noise and artifacts from real signals, (ii) processing cross peaks that are heavily overlapped, (iii) processing spin systems that are heavily overlapped, and (iv) selecting the correct spin-coupling-network sequence for sequence-specific assignment using NOESY cross peaks. Moreover, NOESY cross peaks which arise from neighboring residues and more distant residues cannot be distinguished. This may lead to incorrect sequence assignments.

Experimental techniques and numerical data-processing methods may improve problem (i). Problem (ii) can be partially solved by peak-picking programs and pattern-recognition analysis (*2*). Problem (iii) can be solved by a fuzzy-graph pattern-recognition algorithm, which goes far beyond the limits of human ability to recognize these types of patterns. Problem (iv) requires an algorithm to search a large space of all possible sequence-specific assignments. This space can be likened to a forest of spin systems in which we seek the trees.

Most current computational methods consist of "bookkeeping." Bookkeeping, however, is not equivalent to making assignments and serves only as an aid to manual assignment strategies. Some recently reported methods (*3–6*), such as programs developed by Poulsen *et al.* (*5*), rely on the database function of a computer. Other approaches include semiautomated strategies, such as the one presented in Ref. (*4*), which leaves most of the assignment procedure up to the user. What we seek are "optimum-search-assignment" algorithms that work to the limit of the data, not the limit of human ability to analyze the data. The main goal of fully automating assignment strategies is therefore to find the best and most logically reliable assignments based upon a given data set (e.g., DQF-COSY, TOCSY/HOHAHA, and NOESY spectra). For example, to completely assign the peptide NAc-t21a (21 residues), our program checks more than 7000 possible assignment sequences in 20 minutes. Even allowing for human ability to recognize patterns, it is impossible for a person to compare even a fraction of these 7000×21 spin graphs in any practical time period. The crux of the "optimum search assignment" is thus pattern recognition. Some pattern-recognition algorithms for the automated evaluation of two-dimensional NMR spectra of peptides and proteins have been reported in the literature (*2, 7, 8*). The algorithms in Ref. (*7*), for instance, carry out the following steps: (i) searching for all the maxima and minima in a 2D spectrum (COSY or SECSY), (ii) searching for coupling patterns, (iii) eliminating weak patterns, (iv) replacing multiplet patterns by their average, (v) comparing with 1D data, and (vi) searching additional knowledge sources (assignments from chemical modifications or pH variations) to remove ambiguities.

In this paper, in which fuzzy-graph pattern recognition and graph-search methodology are used, a new procedure is formalized for the assignment of 2D ^{1}H NMR spectra. Generally, this logical system is divided into two parts. The first part creates and identifies spin-coupling topological systems. Then, with a fuzzy-graph pattern-recognition algorithm, each residue is assigned a set of possible candidates, each of which is a spin coupling network. Up to this point, only DQF-

‡ To whom correspondence should be addressed.

1064-1866/94 $6.00

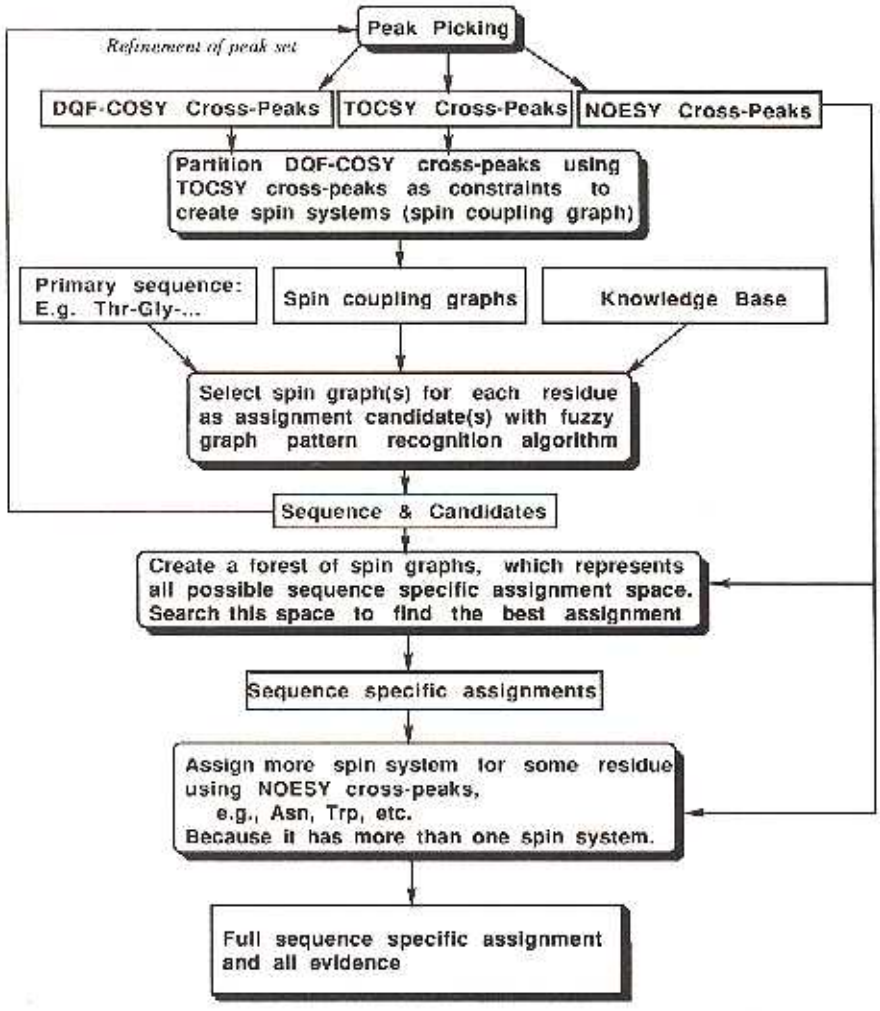


FIG. 1. Flow chart of the complete automated sequence-specific assignment program PSE.

COSY and TOCSY peak sets are used. The second part creates a forest of spin coupling networks such that each tree in the forest consists of sequential spin graphs. An algorithm then searches all of this forest and finds the optimum sequence-specific assignment based on a NOESY data set. The assignment procedure is summarized in Fig. 1.

Cross-peak lists of DQF-COSY, TOCSY/HOHAHA, and NOESY spectra are the input data sets. Although this input comes from a peak-picking program (for example, FELIX), our method can estimate the peak set required and prompt the user if not enough cross peaks are picked. For example, according to the sequence of peptide NAc-t21a, the number of DQF-COSY cross peaks should be in the range of 69 ~ 137 (note that only one of two symmetric peaks is considered in the estimation), where the upper limit is based upon the assumption that all the protons in the $-CH_2-$ groups in the protein are not magnetically equivalent, and the lower limit is based upon the assumption that all the protons in the $-CH_2-$ groups in the protein are degenerate. The number of experimental DQF-COSY cross peaks picked should be in this range.

Using TOCSY/HOHAHA cross peaks as constraints, the DQF-COSY cross peaks are partitioned to produce a subpeak set where each subset corresponds to a spin coupling network. This set is converted to the form of a spin coupling network. Each network is represented in an adjacency table (Fig. 2).

For some spin coupling networks with long side chains, the chain cannot be fully observed due to poor coherence

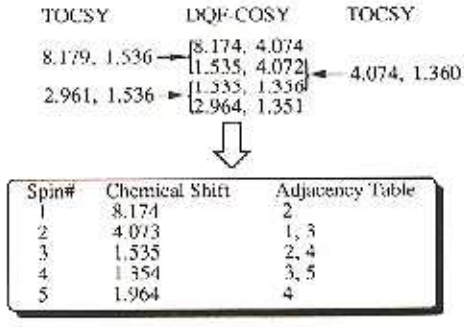
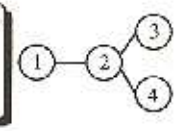


FIG. 2. Resulting subpeak set from the partitioning step and its associated spin-coupling topological graph.

Spin#	Chemical Shift	Adjacency Table
1	8.093	2
2	4.071	1, 3, 4
3	1.814	2
4	1.702	2

A



B

FIG. 3. A spin-coupling topology found from a DQF-COSY spectrum: A is the topological adjacency table and B is the corresponding graph.

transfer (e.g., Arg and Lys). This creates fragmented spin-coupling systems rather than complete ones. A program called Protein Structure Elucidation (PSE) contains a function that can reconnect these fragments by checking TOCSY/HOHAHA correlations. In heavy-overlap situations, a DQF-COSY cross peak could be partitioned, for example, into two different spin-coupling systems. These cases can be identified by using more TOCSY evidence and root mean square values to assign the peak to the best possible spin-coupling system.

Using a knowledge base of chemical-shift distributions in ^{1}H NMR spectra of the 20 amino acids (*9*), fuzzy mathematics (*10*) and pattern-recognition theory (*11*) can be used to map onto specific amino acid residues all the spin-coupling topological graphs which are found. Due to overlap and data incompleteness, these mappings must be further screened to be consistent with chemical-shift data and NOESY correlations. For instance, the spin-topological graph shown in Fig. 3 can be assigned to more than one amino acid residue.

Topologically, Fig. 3 can be mapped to Val, Ile, Leu, Asp, Glu, Gln, Lys, Arg, Asn, Met, Cys, Phe, Tyr, and His because it is a subgraph of the full theoretical spin-coupling graphs of all these amino acids. However, Ile, Asp, Lys, Met, Cys, and His are not in the primary sequence of the peptide considered in this example. So only Val, Leu, Glu, Gln, Arg, Asn, Tyr, and Phe need be considered as candidate assignments for Fig. 3. According to the fuzzy-graph similarity calculation, the similarity value for Val, Asn, Tyr, and Phe is 0, so these candidates are eliminated. The assignment candidates of Fig. 3 are reduced and become {Leu, Arg, Glu, Gln} (where the curly brackets are used to indicate a set). The fuzzy-graph pattern-recognition results are listed in Table 1. The membership of the kth spin of spin graph G_i belonging to the lth proton of amino acid residue R_j is represented by $\mu_{G_i-R_j}(k, l)$ (Table 1), assuming that the chemical cluster of the amino acid protons obeys a Gaussian distribution. The similarity is a general estimation of how well G_i belongs to R_j, which is calculated by using fuzzy mathematics (*10*) in the following manner:

$$\text{Similarity} = \min[\mu_{G_i R_j}(k, l)], \quad i, j, k, l = 1, 2, 3, \ldots.$$

Figure 4 shows how a spin coupling network such as the previously unassigned spin system in Fig. 3 is mapped to possible amino acid residues using chemical-shift data and the calculated spin-coupling topology.

Theoretically, Fig. 3 should be assigned to the amino acid which has the maximum similarity, that is, Leu. However, at this stage, Arg, Glu, and Gln cannot be eliminated since for large variations in chemical shift, the true candidate may not be Leu but some other amino acid. Hence, all one can say at this point is that the most possible assignment of Fig. 3 is Leu. Even if this were a complete spin system and not a fragment, it still can only be considered as a candidate because of extensive chemical-shift and spin-coupling-topology overlap. The final assignment can be made only once the NOESY correlations, which give the sequence-specific assignment, are included. In fact, Fig. 3 is finally assigned to Glu and not Leu. Therefore, this step is important as it selects and ranks the assignment candidates for each residue and eliminates incompatible sequence-specific assignments.

Viewed differently, a specific residue can have assigned to it more than one spin-coupling topological graph. Consider

TABLE 1
Fuzzy-Graph Pattern Recognition for Fig. 3[a]

From Fig. 3 Proton	8.093 H^N	4.071 H^α	1.814 H^α	1.702 $H^{\alpha'}$	Similarity
Leu	8.19 (0.60)[b]	4.25 (0.49)	1.71 (0.31)	1.60 (0.37)	
$\mu_{\text{Fig3-Leu}}$	0.987	0.935	0.945	0.943	0.935
Arg	8.20 (0.83)	4.28 (0.35)	1.63 (0.43)	1.79 (0.34)	
$\mu_{\text{Fig3-Arg}}$	0.992	0.837	0.913	0.967	0.837
Glu	8.22 (0.60)	4.34 (0.42)	1.97 (0.20)	2.04 (0.18)	
$\mu_{\text{Fig3-Glu}}$	0.987	0.935	0.945	0.607	0.407
Gln	8.28 (0.61)	4.43 (0.451)	1.92 (0.27)	2.10 (0.30)	
$\mu_{\text{Fig3-Gln}}$	0.954	0.727	0.722	0.360	0.360

[a] All chemical shifts are in ppm.
[b] Expected chemical shift and standard deviation.

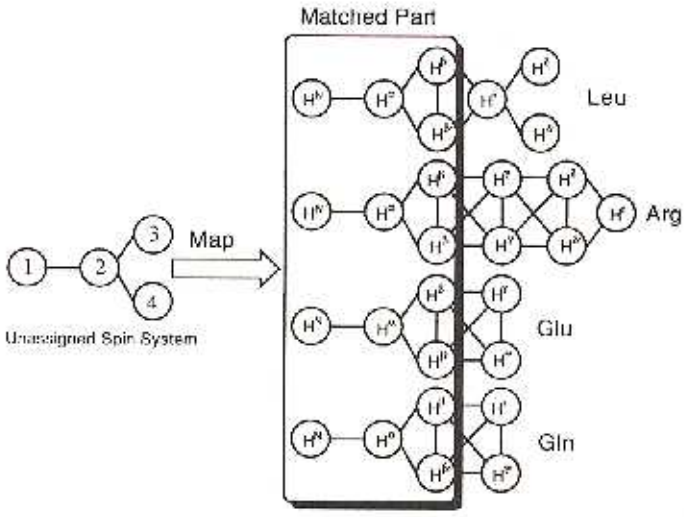


FIG. 4. The subgraph-match algorithm maps an unassigned spin system to possible amino acid residues. The matched part includes chemical-shift data. Topologically, the unassigned spin system can be mapped to many other fragments. By using the fuzzy-graph pattern recognition algorithm, only the coupling topologies with similarity > 0 remain.

Residues		Spin Coupling Graph Candidates
Thr1	→	13 14 6 11
Gln2	→	28 33 1 15 27
Asn3	→	2 4 3 5
Gly4	→	27 33 11 28 9
Arg5	→	28 33 10 15 7 16 17 27
Ser6	→	1 1
Phe7	→	2 5 4 11 3
Gln8	→	28 33 1 15 27
Arg9	→	28 33 10 15 7 16 17 27
Thr10	→	13 14 6 11
Gly11	→	27 33 11 28 9
Thr12	→	13 14 6 11
Leu13	→	28 33 15 10 12 27 7 16
Ala14	→	28 33 16 7 27
Phe15	→	2 5 4 11 3
Glu16	→	28 33 1 15 27
Arg17	→	28 33 10 15 7 16 17 27
Val18	→	33 28 27 8
Tyr19	→	2 4 5 3
Thr20	→	13 14 6 11
Ala21	→	28 33 16 7 27

FIG. 6. An amino acid sequence and their possible assignment candidates before using NOESY.

the primary sequence of a protein or a polypeptide as an ordered set **R** and the spin-coupling topological graphs as a graph set **G**. The relation between **R** and **G** consists of a space of all possible sequence-specific assignments. Each amino acid in the ordered set **R** can have one or more graphs in **G** associated with it. Ideally, each element in the set **R** should be related to a single element in **G**. Unfortunately, this is often not the case since, as mentioned previously, fragmentation of the spin topological graphs gives rise to a number of possible assignment candidates. In the peptide NAc-t21a, for example, Arg has eight possible spin-coupling-graph candidates since there are three Arg residues in the primary sequence and since some of the spin-coupling topological graphs overlap.

The sequence-specific assignment process searches for the best NOESY correlation pathway in the forest of spin-topological graphs. For the peptide NAc-t21a, there are 33 spin-coupling topological graphs which can be considered as assignment candidates for the 21 residues (Fig. 6).

Specific spin-coupling topological graphs are selected by searching for the maximum number of NOESY correlations among them. For the example in Fig. 5, the sequential assignment is

R	1	2	3	4	5	6	7	8	9	10	11									
G	—	1	3	27	28	11	4	33	10	13	9									
				12	13	14	15	16	17	18	19	20	21							
				14	12	16	5	15	17	8	2	6	7							

where the numbers in **R** refer to amino acid residues in the sequence and where the corresponding numbers listed in **G** are the spin-coupling graphs. The order in which these graphs appear is determined by finding NOESY connectivities in the experimental spectrum. This involves searching the large space of possible assignments. In terms of graph theory formalism, this space consists of a forest of spin-topological graphs. To see how the space is searched, start at any residue i in Fig. 6 and choose the jth spin-topological candidate $G(m)$. Then "walk" from $G(m)$ to the next residue $i + 1$ and get the kth candidate $G(n)$, and so forth. As a result, a path, within which is a possible assignment, is created. All paths make up the possible assignment trees. Starting at different residues yields different trees. Taken together, these paths give the space of possible assignments. Figure 7 gives an example of a partial representation of one possible assignment tree given the amino acid sequence and the assignment candidates of Fig. 6.

The solid arrows in Fig. 7 are used when NOESY correlations have been found so that two graphs, such as $G(33)$ and $G(10)$, can be connected, whereas the dotted arrows, joining, for example, $G(33)$ to $G(15)$, are used to indicate how two spin-coupling topological graphs can be connected

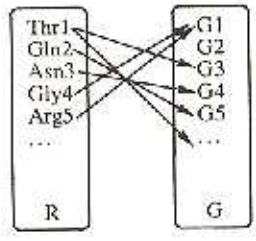


FIG. 5. Multiple mappings among residues and spin-coupling topological graphs before using NOESY.

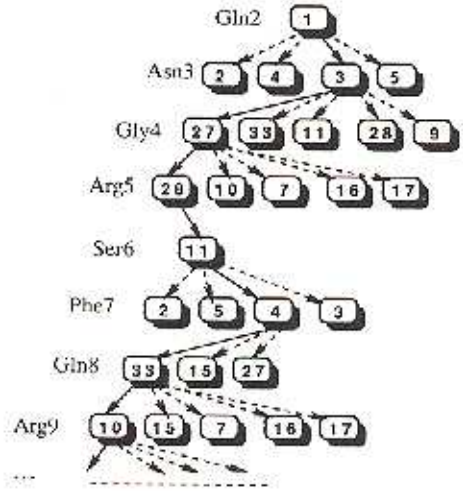


FIG. 7. An example of a possible assignment tree.

when no NOESY correlation evidence is found. These latter possibilities are not discarded at this point since NOESY cross peaks are often missing. Dimissing these possibilities may give rise to serious errors in the final assignment. The criterion used to connect the spin-coupling graphs $G(i)$ and $G(j)$ which are candidates for the residues $R(k)$ and $R(k+1)$, respectively, and which can be defined as a set of proton resonances

$$G(i) = \{H^N(i), H^\alpha(i), H^\beta(i), \cdots\}$$
$$G(j) = \{H^N(j), H^\alpha(j), H^\beta(j), \cdots\}$$

is the NOESY correlation evidence E_{ij}, defined as

$$E_{ij} = \{P_1[H^N(i), H^N(j)], P_2[H^N(i), H^\alpha(j)], P_3[H^N(j), H^\alpha(i)], P_4[H^\alpha(i), H^\alpha(j)], P_5[H^\alpha(i), H^\beta(j)], P_6[H^\beta(i), H^\alpha(j)], P_7[H^\beta(i), H^\beta(j)], P_8[H^N(i), H^\beta(j)], P_9[H^\alpha(i), H^N(j)]\},$$

where P_l is a predicted NOESY cross-peak and where in some cases there may be more than one H^α. It is important to note at this point that the correlation evidence E_{ij} must be defined differently for Pro and Gly, given the nature of these amino acids.

The set E_{ij} includes all possible predicted NOESY cross peaks. If the predicted peaks are found in an experimental NOESY spectrum, a connection between graphs $G(i)$ and $G(j)$ is identified locally, within a given tolerance. As mentioned above, not all predicted NOESY correlation evidence is observable in practice. Some observed NOESY peaks may even identify two spin systems which are spatially close but which in fact are not directly connected. Therefore, care must be taken when considering connected and unconnected graphs (local assignments) in order to avoid making mistakes.

The final step involved in making the sequence-specific assignment, which is to determine the best path out of the entire space of possible assignments, requires that local assignments be consistent with all other assignments. Hence local and global criteria are needed. These are: (i) the number of NOESY cross peaks which identify a pair of spin systems as neighbors and the value of the corresponding match degree (II) among predicted peaks and experimental peaks should be a maximum (local optimum condition), and (ii) the total number of NOESY cross peaks which identifies all pairs of spin systems as neighbors and the value of the total corresponding match degree among predicted peaks and experimental peaks should be a maximum (global optimum condition). Two strategies which incorporate these criteria can be used to create and search the paths. (i) Connected-path strategy: Every branch in a tree should have more than one NOESY peak as supporting evidence. Therefore, all dotted arrows in Fig. 7 will be eliminated. With this strategy, the possible assignment space is sharply reduced. However, it cannot be error proof because the correct assignment might have been missed due to NOESY data incompleteness, giving accidental disconnections. (ii) All-path strategy: All possible assignment paths are created and checked. The main advantage of this strategy is that all possible assignments are included, so the result is reliable. However, this requires considerable CPU time as the search space increases.

The "all-path" strategy becomes especially difficult for larger proteins ($M_r \approx 10$ kDa) since the number of possible paths increases dramatically with size. To circumvent this problem, a segment-by-segment assignment technique is applied. Because the protons in terminal residues are more flexible, it is usually better to begin to make sequence-specific assignments in the middle of the protein primary sequence. In this way the search space is reduced by assigning fragments of the protein. Hence the all-path strategy is the strategy of choice for determining the best path and hence the best sequence-specific assignment.

To facilitate the sequence-specific assignment procedure, NOESY correlations of the candidates which are expected to be neighbors can also be determined in advance. These correlations can then be used to pick out the graphs which correspond to these neighboring residues. This speeds up the procedure since fewer assignment paths need to be verified. This method is of course limited to the cases where the sequence is already known and to the cases where the proton resonances are close to the expected values and where no NOESY cross peaks are missing.

The Protein Structure Elucidation (PSE) Program has been written in Sun C language on a SparcStation to implement these ideas. The source codes for partitioning DQF-COSY spectra and creating spin system graphs are written in SUN Pascal language. The total program, described else-

where (*12, 13*), contains about 25,000 lines of code. A raw data set of COSY, TOCSY, and NOESY peak sets for NAc-t21a was used to test PSE. It took 2.5 minutes to create all of the spin-coupling topological graphs, 3.3 minutes to select candidates for the residues, and 5.5 minutes to do the sequence-specific assignments. PSE tested approximately 173.5 million fuzzy-graph patterns and compared NOESY cross peaks 16,500 times. Such a task is beyond unassisted human capabilities. In the end, the best sequence-specific assignment was found to be

Residues	Q–N–G–R–S–F–Q–R
Number of NOESY peaks	1 2 2 1 2 3 2
	–T–G–T–L–A–F–E–R–V–Y–T–A
	1 3 3 2 4 4 3 4 2 3 3 1

These results agreed completely with the manual assignments that were also made.

PSE also includes a number of functions which can reconnect long spin systems that were incomplete after the partitioning step in which TOCSY peaks are used as constraints. It also includes functions which can assign additional spin-coupling systems to residues which have more than one spin coupling network (e.g., Gln).

In situations where serious overlap of peaks occurs and where, as a result, some spin coupling networks cannot be assigned, PSE prompts the user to attempt to assign them. For example, in processing the raw data set of NAc-t21a, PSE identified the spin system shown in Fig. 8A, a spin system that could not be directly assigned to an amino acid using PSE's knowledge base. The user is informed of the situation. If it happens that H^{β} and H^{γ} accidently have the same chemical-shift value (1.445 ppm), a spin system such as in the one in Fig. 8A would be created. As a result of this overlap, the DQF-COSY cross peak (1.352, 1.445) may represent two kinds of couplings, namely H^{β}–H^{α} and $H^{\beta'}$–H^{γ}. After carefully considering this, the user can rearrange the graph in Fig. 8A to look like the one in Fig. 8B. PSE then resumes its functions and the fuzzy-graph pattern-recognition algorithm maps the modified graph to the Leu residue (Fig. 8C) uniquely. Rather than having the user make decisions regarding such spin systems, an artificial intelligence technique involving setting up a knowledge base which would incorporate past assignments or information gathered from HMQC or 3D NMR spectra could also be used. These possibilities are currently being explored.

We have completely assigned the protein BPTI (56 residues, 4 of which are prolines) in a few hours. The assignments agree completely with those found in the literature (*14*). We have also done an assignment on an unknown sample containing 76 residues, 11 of which are glutamic acids. Details regarding the results for these proteins and a more detailed discussion of PSE can be found in Ref. (*13*).

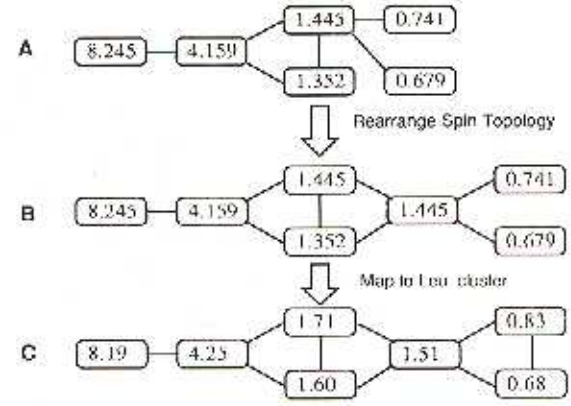


FIG. 8. Spin coupling network resulting from serious overlap and the strategy to solve it.

ACKNOWLEDGMENTS

The McGill group expresses their sincere thanks Dr. M. Bernstein of Merck Frosst Canada, Inc., for his help and encouragement during this project. This work is supported by a grant from the Natural Science and Engineering Research Council of Canada (NSERC). S. K. Straus thanks NSERC for an Undergraduate Summer Research Fellowship.

REFERENCES

1. K. Wüthrich, "NMR of Protein and Nucleic Acids," Wiley, New York, 1986.
2. P. Pfändler and G. Bodenhausen, *J. Magn. Reson.* **79,** 99 (1988).
3. P. J. Kraulis, *J. Magn. Reson.* **84,** 627 (1989).
4. M. Billeter, V. J. Basus, and I. D. Kuntz, *J. Magn. Reson.* **76,** 400 (1988).
5. J. C. Hoch, F. M. Poulsen, and C. Redfield (Eds.), "Computational Aspects of the Study of Biological Macromolecules by Nuclear Magnetic Resonance Spectroscopy," Plenum Press, New York, 1991.
6. F. J. M. van de Ven, *J. Magn. Reson.* **86,** 633 (1990).
7. K. P. Neidig, H. Bodenmueller, and H. R. Kalbitzer, *Biochem. Biophys. Res. Commun.* **125**(3), 1143 (1984).
8. P. Pfändler, G. Bodenhausen, B. U. Meier, and R. R. Ernst, *Anal. Chem.* **57,** 2510 (1985).
9. K. H. Groß and R. Kalbitzer, *J. Magn. Reson.* **76,** 87 (1988).
10. A. Kaufmann, "An Introduction to the Theory of Fuzzy Subsets," Vol. 1, Academic Press, New York, 1975.
11. J. Xu and M. Zhang, *Tetrahedr. Comput. Methodol.* **2**(2), 75 (1989).
12. J. Xu and B. C. Sanctuary, *J. Chem. Information Comput. Sci.* **33,** 490 (1993).
13. J. Xu, S. K. Straus, B. C. Sanctuary, and L. Trimble, *J. Chem. Information Comput. Sci.*, submitted.
14. G. Wagner, W. Braun, T. F. Havel, T. Schaumann, N. Gō, and K. Wüthrich, *J. Mol. Biol.* **196,** 611 (1987).