Automation of Protein 2D Proton NMR Assignment by Means of Fuzzy Mathematics and Graph Theory

Jun Xu,† Suzana K. Straus, and B. C. Sanctuary*

Department of Chemistry, McGill University, 801 Sherbrooke Street West, Montreal, PQ, Canada H3A 2K6

Laird Trimble

Merck Frosst Canada Inc., P.O. Box 1005, Pointe Claire-Dorval, PQ, Canada H9R 4P8

Received January 27, 1993•

The novel methodology for protein 2D NMR assignment presented in this paper is based upon protein spin coupling graph theory analysis, fuzzy graph pattern recognition, and tree searching. The method required to formalize the whole assignment procedure into a logical system which can be properly processed by computer software is also discussed. Solutions for peak overlaps, spin coupling network overlaps, and details related to the automated assignment of BPTI are reported as well.

INTRODUCTION

To determine solution protein structures from multiple dimensional NMR spectra, the cross peaks in the spectra must be unambiguously assigned before they can be used for such purposes as distance constraint calculations. Two general strategies for the full elucidation of protein structures exist. The first, termed sequence assignment strategy, was proposed by Wüthrich and co-workers.[1] The second, the main-chain-directed assignment strategy (MCD), was reported by Englander and Wand[2] and involves determining the secondary structure initially by identifying characteristic patterns associated with helical and β-sheet structures by means of a combination of COSY and NOESY data. The former method involves the following three steps:

(i) Trace and identify spin coupling networks from COSY and TOCSY spectra or spectra obtained from other NMR experiments, e.g. MQF-COSY, HOHAHA, RCT-COSY, NOESY, ROESY, HNCA, HOHAHA-HMQC, HCACO, HCA(CO)N, HNCO, etc.[3]

(ii) Map the spin coupling networks to individual amino acid residues.

(iii) Make sequence-specific assignments using a combination of NOESY spectra and information obtained from the primary sequence.

In the ideal case, for a small protein, all three steps can easily be carried out. For larger proteins, however, step i becomes difficult due to line broadening and increase in the number of cross peaks which, in turn, respectively leads to serious overlap problems and to overwhelming amounts of complicated spin coupling networks that need to be deciphered. Step ii is also very complicated because of spin coupling topological overlap and significant variations in the chemical shifts in some residues. And step iii yields too many possibilities resulting from the multiple interpretations of NOESY cross peaks. Two solutions are possible for these problems. The first consists of improving the quality of the spectra by approaching the problem from an experimental point of view.[4] An extensive number of studies (too extensive to be covered here in great detail) have used a variety of methods for studying large molecular complexes by NMR: transferred NOE,[5] isotope-edited proton NMR,[6,7] 2D NOE difference spectroscopy using deuterated ligands,[8] perdeuterated receptors,[9] and heteronuclear 3D and 4D NMR experiments.[3,10-12] The other involves improving the quality of the assignment process by developing computer-aided automated techniques which can be used to identify spin coupling topologies.[13-17] Some recent progress has been collected in NATO ASI Series A, Vol. 225.[18] For instance, Pfändler and Bodenhausen described a method with which the topological fragments of spin coupling networks could be identified on the basis of an analysis of the fine structures of cross peaks by means of a pattern recognition method. With this method, which hinges on graph theory, overlapping multiplets can be separated by identifying "the D_2 symnmetry inherent to combined multiplets extracted from pairs of complementary two-dimensional spectra".[19] The fragments that are identified are subsequently connected together to obtain a global coupling network. The number of possible fragments for a single amino acid residue (e.g. Pro) is quite extensive, thus making this task of connecting them difficult. For large proteins, in particular, which consist of many different types of residues, this problem becomes especially serious.

Alternatively, other techniques relating to graph theory can be used. In this paper, we present three such algorithms: constrained partitioning algorithm (CPA), fuzzy pattern recognition algorithm (FPRA), and tree search algorithm (TSA). CPA can automatically extract and identify spin coupling networks from a combination of COSY spectra and TOCSY spectra (and/or other spectra) where the latter spectra are used as partitioning constraints. FPRA can map the spin coupling networks obtained from CPA to particular amino acid residues. Therefore, a graph residue-to-candidates relationship is produced where a candidate is a spin system to be assigned. The correct sequence specific assignment exists in this residue-to-candidates graph as a path which is found by TSA with the aid of NOESY cross peaks.

EXTRACTING SPIN COUPLING SYSTEMS FROM DQF-COSY AND TOCSY PROTON NMR SPECTRA

Basically, to extract spin coupling topological systems, CPA, described more fully in ref 20, scans a DQF-COSY cross-peak table from beginning to end. Some cross peaks form a starting point for a given spin system. The algorithm then attempts to find all other cross peaks which have frequencies

* To whom correspondence should be addressed.
† Present address: Tripos Analytical Group, 6035 Corporate Dr., East Syracuse, NY 13057-1016.
• Abstract published in *Advance ACS Abstracts*, August 15, 1993.

0095-2338/93/1633-0668$04.00/0

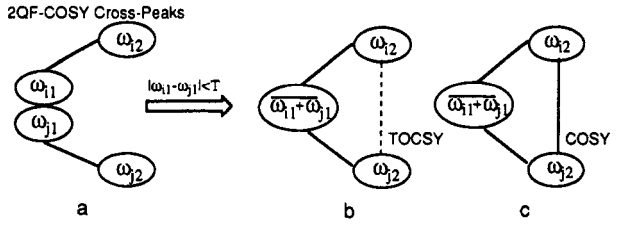


Figure 1. Spin coupling system created by CPA: (a) If ω_{i1} is considered to be a starting point of a spin system, CPA finds other cross peaks with frequencies in common with ω_{i1} (within a given tolerance); (b) the spin coupling system is formed if TOCSY evidence is found; (c) the spin coupling system is formed if COSY evidence is found.

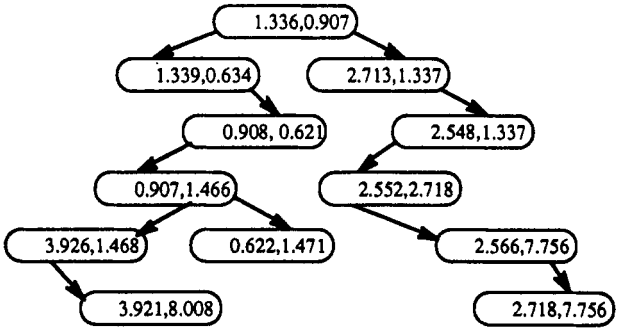


Figure 2. Binary tree representation of a subspace generated for the Lys21 residue in melittin (fuzzy graph).

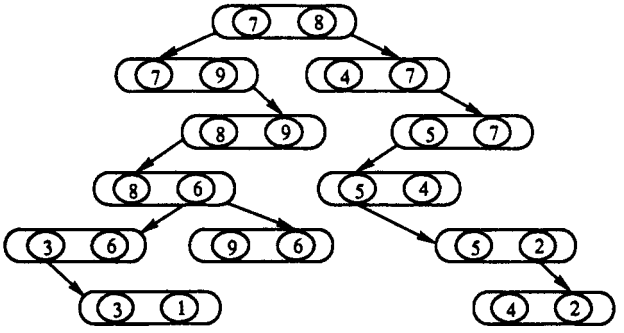


Figure 3. Binary tree representation of a subspace generated for the Lys21 in melittin (boolean graph).

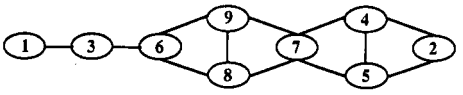


Figure 4. Spin topology extracted for Lys21 from raw data obtained from the DQF-COSY and TOCSY spectra of melittin.

in common with the starting point of a system within a given tolerance. It searches for TOCSY or COSY proof to decide if these new DQF-COSY cross peaks can be added to the spin coupling system (Figure 1).

Mathematically, a spin topological graph can be represented by a matrix, a connectivity table, a binary tree, or an edge set. A spin coupling system can be considered as a subspace of a DQF-COSY cross-peak set. In graph theory, this subspace is an edge set of a spin coupling graph. In the ideal case, a DQF-COSY peak subspace corresponds to a complete spin system which for a protein translates to an amino acid. If some peaks are missing, the subspace corresponds to only part of a complete spin system of a residue. The procedure involved in generating a spin topological graph is illustrated by means of the results obtained from actual data for the Lys21 residue in the protein melittin (Figures 2–4). From the partitioning result, which has been represented as a binary tree in Figure 2, a subspace consisting of twelve peaks is formed. Six redundant peaks are also found but are discarded since they do not provide any useful information.

Table I. Average Frequencies from the Binary Tree in Figure 2

no.	freq	no.	freq	no.	freq
1	8.01	4	2.72	7	1.34
2	7.76	5	2.56	8	0.91
3	3.92	6	1.46	9	0.63

Table II. Spin Topological Connectivity Table

no.	freq	connectivity	no.	freq	connectivity
1	8.01	3	6	1.46	3, 8, 9
2	7.76	4, 5	7	1.34	4, 5, 8, 9
3	3.92	1, 6	8	0.91	6, 7, 9
4	2.72	2, 5, 7	9	0.63	6, 7, 8
5	2.56	2, 4, 7			

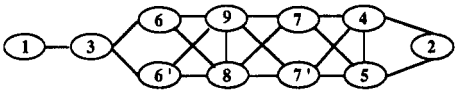


Figure 5. Complete spin topology of Lysine.

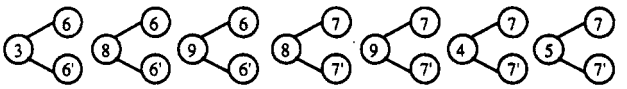


Figure 6. Equivalent spin coupling topological fragments which are not predicted by CPA due to the fact that the nodes 6 and 6′ and the nodes 7 and 7′ are degenerate.

Table III. Spin Coupling Topological Network and Evidence Provided by CPA for Ser-6 of NAc-t21a

premise		
TOCSY cross peak	DQF-COSY cross peak	conclusion
(8.966, 2.020)	(8.969, 4.431)(4.432, 2.022)	(8.969)–(4.432)–(2.022)
(4.438, 2.385)	(4.432, 2.022)(2.017, 2.386)	(4.432)–(2.022)–(2.386)
(2.130, 2.004)[a]	(2.116, 4.436)(4.432, 2.022)	(2.116)–(4.432)–(2.022)
(2.023, 2.385)[a]	(2.004, 2.130)(2.116, 2.385)	(2.004)–(2.120)–(2.385)
(2.027, 4.431)[a]	(2.004, 2.130)(2.116, 4.436)	(2.004)–(2.120)–(4.436)

[a] COSY cross peaks can also be used as constraints to automatically create spin coupling topological networks. This case is called self-partitioning.[20]

The twelve peaks have frequencies in common which are all within a tolerance of 0.02 ppm. In mathematical terms, this binary tree is a fuzzy graph with a fuzziness of 0.02 ppm. From Figure 2, average frequencies can be extracted and are listed in Table I.

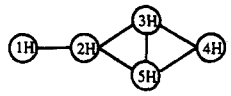
To each real number frequency value is associated an integer label. By replacing the peak frequencies by labels, the fuzzy graph is transformed into a boolean graph (Figure 3). Since the values in a boolean graph do not vary within a given tolerance, a boolean graph is not fuzzy. In turn, Figure 3 is transformed into a connectivity table as shown in Table II.

Finally, the connectivity table is transformed into a topological graph as indicated in Figure 4.

The theoretical topological graph for lysine is illustrated in Figure 5. A comparison of the graph obtained in Figure 4 with the complete one in Figure 5 indicates that nodes 6 and 7 should correspond to $-CH_2-$ groups in which the two protons are equivalent.

The reason for which a complete spin topology cannot be constructed directly by CPA is that certain spin topological fragments (Figure 6) are not predicted or checked against a multiple quantum experimental data set. The method used for predicting and detecting this kind of topology will be presented in a subsequent publication.

The output of CPA is a set of spin systems in the form of adjacency tables. The evidence for grouping a cross-peak subspace is also listed so that a user can verify that the results



Node	Shift	Adjacency Table
1H	8.969	2
2H	4.432	1, 3, 5
3H	2.022	2, 4, 5
4H	2.385	3, 5
5H	2.120	2, 3, 4

Figure 7. Spin coupling topology from Table III and its adjacency table.

obtained by CPA are reasonable. For example, to assign Ser-6 of NAc-t21a, CPA generates the spin coupling topological network and the evidence as in Table III.

CPA also converts the conclusion column of Table III into an integral spin coupling topological network in the form of an adjacency table (Figure 7).

FUZZY GRAPH THEORY AND SPIN COUPLING NETWORKS

A spin coupling network can be defined in terms of fuzzy graph (FG)[21,22] which is a mathematical representation that is defined by,

$$FG = \{V,\Delta_v,E,\Delta_E,\mu_v,\mu_e\} \quad (1)$$

where V is a cluster center which represents a group of chemical shifts ($V_\in$ {chemical shift set}), Δ_v is the set of distributions associated with every element in V (i.e. the element Δ_{vi} of Δ_v represents the range over which the ith chemical shift V_i in V can be obtained), E is also a cluster center which represents a group of J coupling constants and which has a distribution Δ_E associated with it ($E \in$ {V×V,J coupling constants set}), and finally, μ_v and μ_e are membership function sets for V and E, respectively. The latter parameters are essentially quality factors. For the components of V, for instance, the membership function set (μ_v), with a given ith membership function μ_{vi}, is used to determine whether or not an experimental frequency value belongs to a given spin system and to determine the quality of this assignment. Supposing the expected chemical shift values and J coupling constants obey a Gaussian distribution, the membership functions are

$$\mu_{Vi} = 1 - \exp\{-[(V_i - V_{xi})/\Delta_{vi}]^2\} \quad (2)$$

$$\mu_{ei} = 1 - \exp\{-[(E_i - E_{xi})/\Delta_{Ei}]^2\} \quad (3)$$

where V_i is the expected chemical shift value for spin i,[23] V_{xi} is the experimental chemical shift value to be assigned, and Δ_{vi} is the deviation of V_i.[23] The components in V and E are represented in fuzzy graphs by nodes and edges, respectively. The graphs are defined in this manner because chemical shifts and coupling networks are the basis for making resonance assignments.

Traditionally, only $H^N - H^\alpha$, $H^N - H^\beta$, and $H^\alpha - H^\beta$ J coupling constants are measured since they are required to determine the backbone conformation of a protein (Karplus equation). On the other hand, the J coupling constants of the side-chain protons in the amino acid residues are not frequently computed. In addition, the distribution of each of the J coupling constants is unknown. Given these facts, the fuzzy graph, as defined above, needs to be modified since all J coupling constant values and their distributions are required to build such a graph. Hence, the fuzzy graph is redefined in the following manner:

$$FG' = \{V,\Delta_v,E', \mu_v\} \quad (4)$$

where $E' \in \{V \times V\}$ and each edge in E' means spin i and spin j are coupled but the coupling constant (J_{ij}) is not considered. The other parameters are as defined previously. It is important to note that if J coupling constants were available, then the fuzzy graph would be more complete and thus more useful for the purpose of amino acid residue identification. An example of a fuzzy spin coupling graph is shown in Figure 8 for valine.

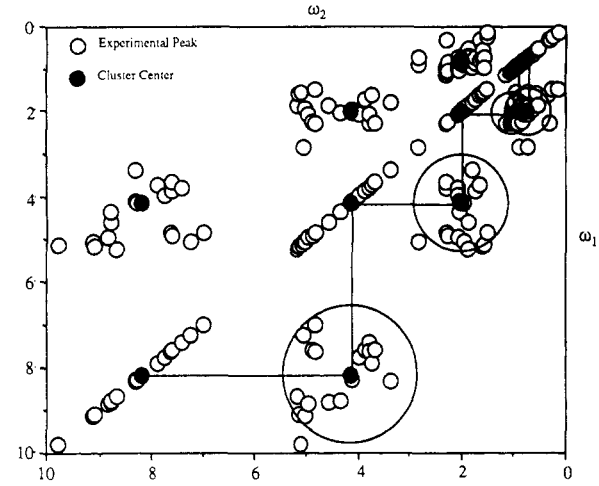


Figure 8. Fuzzy graph of the spin coupling network of valine.

Table IV. Comparison of the Expected Values[a] of the Proline Chemical Shift Distribution and the Experimental Values of four Prolines in BPTI

	chemical shift distribution for given proton						
	αH	βH	βH′	γH	γH′	δH	δH′
	Expected Values						
	4.48	1.88	2.18	1.92	2.02	3.62	3.77
	Experimental Values						
Pro2	4.317	2.011	**0.903**	1.592	1.885	3.600	3.730
Pro8	4.629	1.850	2.434		2.106	3.709	3.969
Pro9	3.714	**0.231**	**0.085**[b]	1.256	**0.149**	3.323	2.904
Pro13	4.550	2.100	2.180	1.991		3.630	3.990

[a] Expected values taken from Gross and Kalbitzer.[23] [b] Maximum difference = 2.095 ppm.

Experimental data gathered for twenty valines was used to generate Figure 8. Seventeen valine data points are from the protein glucose-specific enzyme IIA,[24] two are from the protein melittin,[25] and one is from a peptide of the protein 5-lipoxygenase activating protein (FLAP). The filled circles connected by lines are valine cluster centers. The areas surrounded by circles represent the distribution of the clusters. It can be seen that the deviation is different in different regions. It can also be noted that the two distribution regions for the two valine γ-methyl groups are heavily overlapped since the magnetic environments of the two kinds of protons are very similar.

The complexities of the above fuzzy graph are due to chemical shift variations. Many examples can be found in the literature in which chemical shift values are far from the expected values. An example of the extent of chemical shift variations is given in Table IV. The peak frequencies listed were obtained from raw data for BPTI.

These large variations make an assignment based solely on chemical shift data impossible and unreliable. It is important to note that heteronuclear experiments alleviate the problem here since for ^{13}C chemical shifts, for example, the situation for the type assignments better. In order to make a correct assignment, fuzzy graph pattern recognition must be used. In other words, patterns such as the one in Figure 8 must be found so that they can be compared with the standard coupling patterns of the twenty amino acids. If a reasonably good match is found between a standard coupling pattern and an experimental one, then the experimental pattern is assigned to an amino acid.

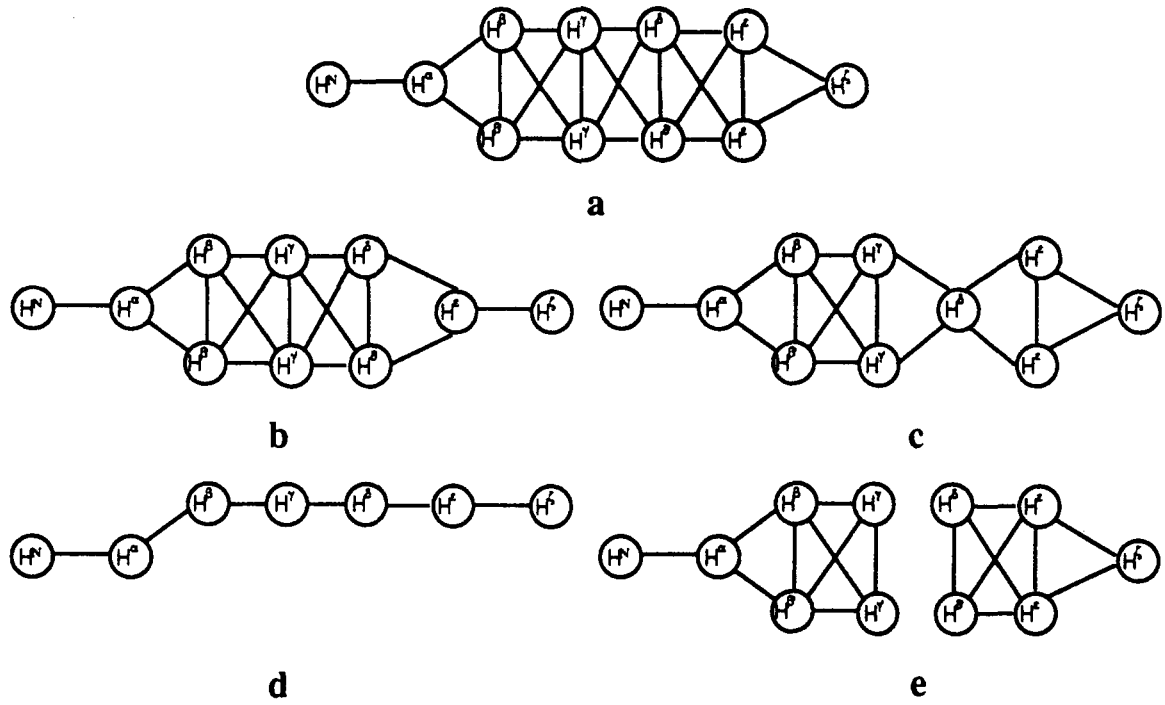


Figure 9. Examples of possible spin coupling topological patterns for a lysine residue: (a) the complete pattern; (b, c) patterns in which some of the protons are degenerate; (d, e) fragmented patterns.

Moreover, as a result of these variations, the spin coupling networks can vary considerably. In some cases, one type of residue may even correspond to many possible spin coupling networks or spin fragments when, for instance, some COSY cross peaks are missing. For example, lysine should theoretically have a spin coupling network as illustrated in Figure 9a. However due to various magnetic equivalence cases and situations in which COSY peaks are missing, the possible spin coupling networks can be as those shown in Figure 9b–e. In these cases where fragmentation occurs, the fragments may overlap with spin coupling networks belonging to other amino acid residues if the topologies are identical.

Although a residue's spin coupling topological network can vary, all possible spin coupling topological patterns are the fuzzy subgraphs of the theoretical fuzzy graph for a given residue. From Figure 9, it can be seen that parts b–e are the subgraphs of graph a. A theoretical fuzzy spin coupling network is a fuzzy graph which consists of a spin coupling topological cluster center as well as a distribution. In addition, all possible magnetic nonequivalence cases are included. All together, this forms a spin coupling topological pattern data bank for a specific amino acid residue. To search this data bank, graph pattern recognition techniques are necessary. The theoretical spin coupling networks for twenty amino acid residues are listed in Appendix A. All fuzzy graphs for these residues are made in the form of extended connected tables (ECT),[26] which make up a fuzzy graph retrival space and a knowledge base. A cluster center in the knowledge base contains the name of an amino acid and reference information, a set of chemical shift distribution and their positions, a set of standard chemical shift deviation values, and a set of spin coupling topological adjacency tables. Table V and Figure 10 show an example of a cluster center.

FUZZY GRAPH PATTERN RECOGNITION ALGORITHM

Once the experimental spin coupling patterns have been found, they must be compared with the patterns in the spin coupling topological knowledge base mentioned above. In order to do this, a fuzzy graph pattern recognition algorithm is required. This algorithm involves two steps which are best illustrated by means of an example. Consider the case of leucine, illustrated in Figure 11.

The fragments C and D are mapped to the theoretical spin topology B by (i) topological pattern recognition to decide if C or D is the subgraph of graph B and (ii) calculation of the membership values, $\mu_{C\in B}$ and $\mu_{D\in B}$, to determine which

Table V. Example of the Spin Coupling Topological Knowledge Base for Isoleucine

	Ile/isoleucine			other ref inf			
				adjacency table			
no.	proton	shift	deviation				
1	NH	8.26	0.72	2			
2	αH	4.13	0.52	1	3		
3	βH	1.74	0.37	2	4	5	6
4	γ_MH	1.01	0.26	3	5	7	
5	γH	1.30	0.32	3	4	7	
6	γH′	0.78	0.24	3			
7	δH	0.69	0.25	4	5		

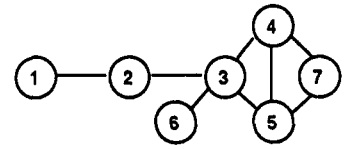


Figure 10. Graphical representation of the cluster center of isoleucine.

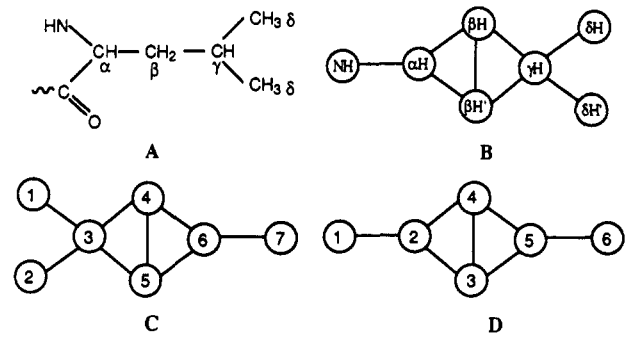


Figure 11. Leucine structure and *J* coupling networks: (a) structure; (b) theoretical *J* coupling network (cluster center); (c, d) possible experimental *J* coupling networks.

Table VI. Partial Order Route Generated for the Graph Illustrated in Figure 11C

partial order	degree	out degree	spin node/ POP[a]	partial order	degree	out degree	spin node/ POP[a]
1	1	1	7	5	7	†	POP
2	3	2	6	7	1	0	2
3	3	2	4	3	5	†	POP
4	3	1	5	5	4	0	3
5	4	2	3	2	4	†	POP
6	1	0	1	4	3	0	5

[a] "POP" is a stack instruction: stop the current walk direction, come back to the last cut-off point, and walk on another branch in the network. Numbers in this column are pointers to chemical shifts. When the POP instruction is given, degree and the following partial order value in the route represent the cut-off node.

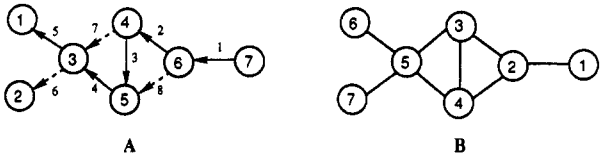


Figure 12. Walk on a graph and ordering: (a) result of walking on Figure 11C. The arrows represent the walking direction, the numbers over the arrows are the sequences of the walk, and the dotted arrows represent the walks driven by the "POP" instruction; (b) new order created from the walking sequence.

Table VII. Partial Ordering and Mappings Generated for the Graph Illustrated in Figure 11B

partial order of Figure 11C	1	2	3	4	5	6	7
spin node of Figure 11B	NH	αH	βH	βH′	γH	δH	δH′
spin node of Figure 11C	7	6	4	5	3	2	1
	7	6	5	4	3	2	1
	7	6	4	5	3	1	2
	7	6	5	4	3	1	2

subgraph of B is more closely related to the cluster center of leucine. To implement step i, a graph match algorithm is needed. Graph match algorithms have been studied for many years in chemical data base and expert system researches.[26,27] An efficient graph match algorithm named HBA (heuristic-back-tracking algorithm) was proposed in 1989.[28] According to this method, the J coupling topological cluster center B is a supergraph denoted by SG and an experimental J coupling network is a query graph denoted by QG. If QG $\in$ SG, i.e., if QG is the subgraph of SG, HBA can find at least one mapping between QG and SG. This means QG matches SG. HBA is based upon partial ordering relations.[22] These relations contain the topological characteristics of a graph. The basic strategy of HBA is that it gets a partial ordering set from QG. This set, denoted as a *ROUTE*, carries this ROUTE onto the graph SG. If it succeeds, then QG matches SG and a mapping is found. A ROUTE created by "walking" on Figure 11C is listed in Table VI. The starting point which can be chosen arbitrarily is, in this case, node 7. A partial order represents a walk sequence, degree is the adjacency of a node, and out degree is the number of edges from the given node to other nodes. A "walk ordering graph" is illustrated in Figure 12.

With the ROUTE in Table VI, HBA can walk in the same partial ordering sequence on Figure 11B. That is, HBA assigns the same partial set to the spin node set of Figure 11B (Table VII). Therefore, the partial set is the link between Figure 11B Figure 11C.

Theoretically, the twenty amino acids give twenty-six J coupling systems. The J coupling topological networks are listed in Appendix A. Tryptophan has three coupling topologies, asparagine has two, phenylalanines have two, and tyrosine has two. The twenty-six J coupling systems form a fuzzy topological space. The fuzzy pattern recognition algorithm finds a "best mapping" from a query topology to this space.

Due to possible degeneracies and missing COSY cross peaks, a query topology may be assigned to more than one amino acid, if only spin coupling topological patterns are considered. Take Figure 11D as an example. Topologically, it can be assigned to leucine, proline, lysine, and arginine (cf. Appendix A). Moreover, for each one, the assignment is not unique. For instance, when Figure 11D is assigned to leucine, four different mappings are found. The smaller a topological pattern is, the more residues it can be assigned to. From Appendix A, it can be seen that serine, phenylalanine, tyrosine, and histidine have the same spin coupling topological pattern. A number of theoretical spin coupling topologies have common spin coupling topological fragments (subgraph). An example is shown in Figure 13.

Thus in order to make correct assignments, it is not sufficient to simply consider spin coupling patterns. Recall that assignments based solely on chemical shift data are also unreliable. If however, a combination of chemical shift data and spin coupling patterns are used, as in the fuzzy graph pattern recognition algorithm, then more reliable assignments can be made.

To describe the fuzzy graph pattern recognition algorithm, let a query topological space QG represent experimental J coupling topologies and the fuzzy topological pattern space SG represent the J coupling fuzzy topological cluster center of the twenty amino acids. The following definitions can be made:

$$\mathrm{QG} = \{\mathrm{QG}(1), \mathrm{QG}(2), ..., \mathrm{QG}(k), ...\} \quad (5)$$

where QG(k) is a spin coupling graph that consists of a set of chemical shifts, V_{kl}, and the connectivities between them

$$\mathrm{QG}(k) = \{V_{k1}, V_{k2}, ..., V_{kl}, ...\} \quad (6)$$

SG is a set of cluster centers which contains the chemical shift and spin coupling topological properties of all twenty amino acids. It can also be changed for different kinds of compounds. SG is defined as

$$\mathrm{SG} = \{\mathrm{SG}(1), \mathrm{SG}(2), ..., \mathrm{SG}(i), ...\} \quad i \in \{\mathrm{Gly}, \mathrm{Ala}, ...\} \quad (7)$$

where SG(i) is a fuzzy graph,

$$\mathrm{SG}(i) = \{I_{i1}|\Delta_{i1}, I_{i2}|\Delta_{i2}, ..., I_{ij}|\Delta_{ij}, ...\} \quad j \in \{\mathrm{NH}, \alpha\mathrm{H}, \beta\mathrm{H}, ...\} \quad (8)$$

with I_{ij} as the expected value of a particular chemical shift and Δ_{ij} as the deviation of I_{ij}.

FPRA (fuzzy pattern recognition algorithm), which uses HBA, can produce a fuzzy mapping set from QG to SG. The mapping set can be classified as a set of multiple mappings for one amino acid

$$\mathrm{M}_{k\to i} = \{M(1)_{k\to i}, M(2)_{k\to i}, ..., M(m)_{k\to i}, ...\} \quad (9)$$

This means that an experimental spin coupling topology SQ(k) is mapped to the ith amino acid by a set of mappings $\mathrm{M}_{k\to i}$. Each mapping $M(m)_{k\to i}$ belongs to the permutation set of QG(k) and SG(i)

$$M(m)_{k\to i} \in \{\mathrm{QG}(k) \times \mathrm{SG}(i)\} \quad (10)$$

A similarity for the mth mapping $M(m)_{k\to i}$, $S(m)_{k\to i}$, can be calculated by

$$S_{\mathrm{E}}(m)_{k\to i} = (\sum \mu(\mathrm{l,n})^2/\mathrm{n})^{1/2} \quad \text{[Euclidean similarity]} \quad (11)$$

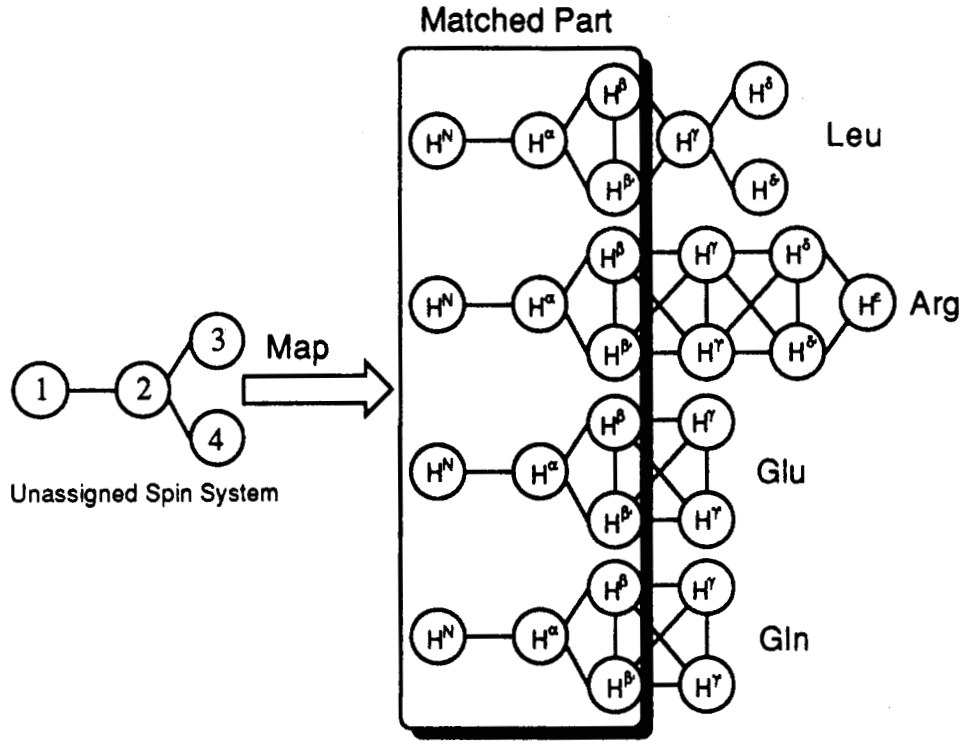
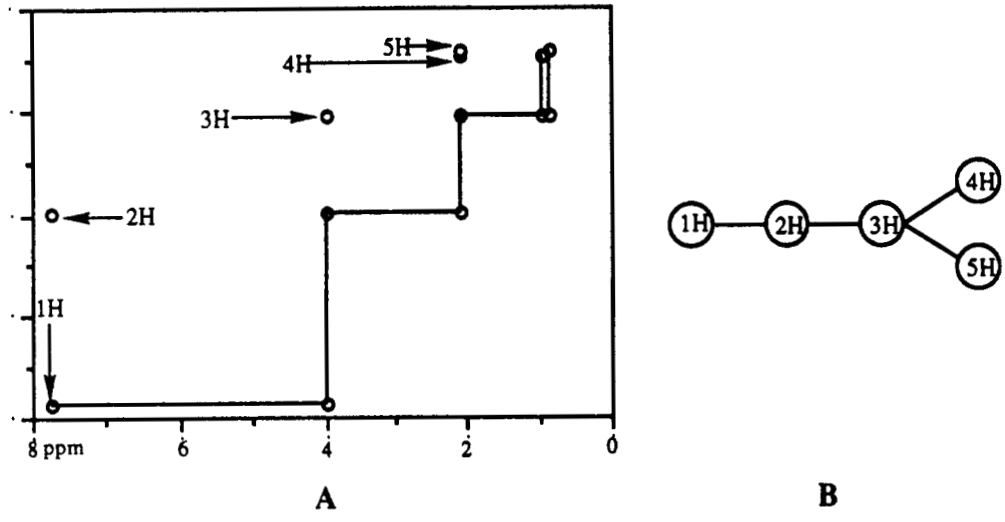


Figure 13. Small spin topological fragment assignable to a large number of residues. Fuzzy pattern recognition, tree searching, and NOESY data are used to correctly assign such a spin pattern to the appropriate amino acid.



Proton	1H	2H	3H	4H	5H
Chemical Shift (ppm)	7.754	3.984	2.074	0.956	0.826

Figure 14. Experimental J coupling topology from a peptide fragment (NAc-t21a) of 5-lipoxygenase activating protein in DPC micelles: (a) fuzzy pattern obtained from the DQF-COSY spectrum; (b) spin coupling topology found by CPA (QG); (c) peak frequencies (V_{kl}).

or

$$S_F(m)_{k\to i} = \min(\mu(l,n)) \quad \text{[fuzzy similarity]} \quad (12)$$

where $\mu(l,n)$ is the membership value of mapping the lth chemical shift of QG(k) onto the nth chemical shift of SG(i). The similarity of QG(k) and SG(i), $S(k,i)$, is given by

$$S(k,i) = \max(S_E(m)_{k\to i}) \quad (13)$$

or

$$S(k,i) = \max(S_F(m)_{k\to i}) \quad (14)$$

In other words, the mth mapping for which the query topological space and the fuzzy topological space match the most is the best mapping. For example, in the automated assignment procedure of the peptide NAc-t21a of 5-lipoxygenase activating protein in DPC micelles, an experimental J coupling topology such as the one illustrated in Figure 14 was found.

Figure 14B represents a spin coupling graph that consists of a set of chemical shifts and the connectivities between them. This QG can be mapped to valine, leucine, glutamic acid, or arginine because it is the subgraph of these amino acids' J coupling topological cluster graphs (cf. Appendix A). Hence, many mappings $\{M(m)_{k\to i}\}$ are possible. For example, two

Table VIII. Partial Mappings and the Computed Membership Values for a Pattern (Figure 14) Found in the DQF-COSY Spectrum of NAc-t21a (a Fragment of 5-Lipoxygenase Activating Protein) in DPC Micelles

SG(i)	1 H		2 H		3 H		4 H		5 H		$S_{QG\rightarrow AA}$[a]
Val	NH	0.77[b]	αH	0.95	βH	0.98	γH′	0.91	γH	0.96	0.77
	NH	0.77	αH	0.95	βH	0.98	γH	0.67	γH′	0.96	0.67
	no. of mappings$_{QG\rightarrow Val}$ = 2							S(QG, Val) = 0.77[c]			
Leu	NH	0.77	αH	0.86	βH′	0.50	γH	0.18	βH	0.11	0.11
	NH	0.77	αH	0.86	βH′	0.50	βH	0.22	γH	0.07	0.07
	NH	0.77	αH	0.86	βH	0.44	γH	0.18	βH′	0.02	0.02
	NH	0.77	αH	0.86	βH	0.44	βH′	0.05	γH	0.07	0.05
	NH	0.00	αH	0.00	γH	0.17	δH′	0.88	δH	0.94	0.00
	...	...	...	...	...	...	...	...	...	...	...
	no. of mappings$_{QG\rightarrow Leu}$ = 16							S(QG, Leu) = 0.11[c]			
Glu	NH	0.74	αH	0.70	βH	0.98	γH′	0.00	γH	0.00	0.00
	NH	0.74	αH	0.70	βH	0.98	γH	0.00	βH′	0.00	0.00
	αH	0.00	βH′	0.00	γH	0.62	γH′	0.00	βH	0.00	0.00
	...	...	...	...	...	...	...	...	...	...	...
	no. of mappings$_{QG\rightarrow Glu}$ = 24							S(QG, Glu) = 0.00[c]			
Arg	NH	0.87	αH	0.70	βH	0.71	γH′	0.21	γH	0.12	0.12
	NH	0.87	αH	0.70	βH′	0.71	γH′	0.21	βH	0.17	0.17
	NH	0.87	αH	0.70	βH′	0.71	γH	0.25	γH′	0.10	0.10
	αH	0.00	βH′	0.00	γH′	0.32	δH′	0.00	δH	0.00	0.00
	...	...	...	...	...	...	...	...	...	...	...
	no. of mappings$_{QG\rightarrow Arg}$ = 116							S(QG, Arg) = 0.17[c]			

[a] AA: Amino acid to be assigned. $S_{QG\rightarrow AA}$: similarity for each mapping of QG to AA (calculated using formula 12). [b] NH 0.77: Assigned proton and membership calculated using formula 2. [c] Calculated using formula 14.

mappings can arise for valine, sixteen mappings for leucine, and so on. Table VIII shows partial mappings and the membership values computed for each of them. The best similarity between QG and SG(i) is found to be the one when SG is the valine coupling topology. Hence QG can be assigned to Val.

TREE SEARCH ALGORITHM

In the previous section, the procedure used to assign a spin coupling pattern to an amino acid was outlined. The assignment is based on both chemical shift data and pattern recognition. The best assignment is considered to be the one that most closely matches the query topological graphs with the standard spin topological graphs of the twenty amino acids found in the knowledge base. Despite the fact that an assignment based on both elements given above is more reliable, the assignment is not necessarily error free. If the chemical shift variations are extensive and if proton degeneracies modify the coupling topologies significantly, then there exists a possibility that the pattern may be assigned to the wrong amino acid. More specifically, the spin coupling graph may be mapped to an amino acid which is the "best" assignment even though it is not the "correct" assignment. By best assignment, it is understood the SG(i) for which the similarity is a maximum (e.g. Val in Table VII). Thus in order to avoid making incorrect assignments, the fuzzy pattern recognition algorithm does not directly assign a pattern to an amino acid. Rather, it acts as a sort of a filter and produces a graph-to-residues relationship. That is, each spin coupling graph is assigned a set of candidate residues, one of which is the correct assignment. An example of a graph-to-residues relationship is given in Figure 15. This graph-to-residues relationship is then converted into a residue-to-graphs relationship, as illustrated in Figure 16 for NAc-t21a, which is used in conjunction with NOESY data to create a set of supergraphs, where a supergraph consists of an edge set and a node set. In this case, an edge is a NOESY connection identified by one or more NOESY cross peaks and a node is the number of a spin

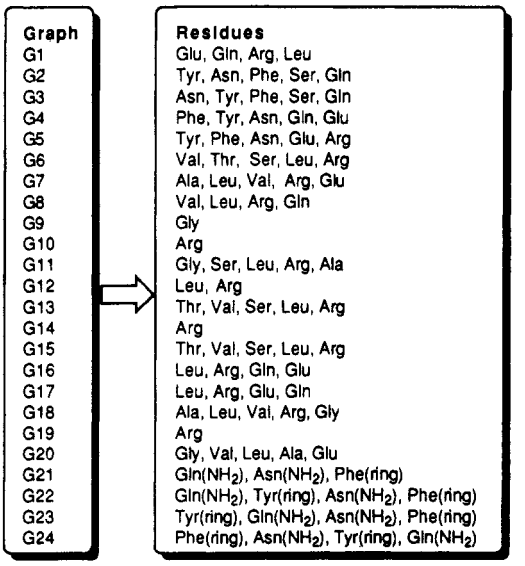


Figure 15. Graph-to-residues relationship for a set of spin coupling topologies obtained for the peptide NAc-t21a.

coupling topology or graph, produced by CPA and mapped to an amino acid by FPRA. The sequence of every residue in a peptide or a protein is determined by finding NOESY connectivities. In other words, the correct graph in the set of spin coupling candidates for a given residue has nodes which are connected to nodes in the correct graphs of the preceeding and subsequent amino acid residues in the sequence. So the correct spin coupling topologies can be determined by finding the connectivities between the topologies which link the amino acid residues together. These neighboring assignment candidates are connected pairwise. The supergraphs created are represented as trees. A group of these supergraphs is referred to as a forest. Figure 17 shows a typical example of a tree.

The tree search algorithm (TSA) is then used to find the specific sequence of spin coupling topologies which is the best match for the primary sequence of a given peptide or protein. In other words, the resulting supergraph network is mapped

Residues		Spin Coupling Graph Candidates
Thr1	→	13 15 6
Gln2	→	1 16 17 8 4 3 2
Asn3	→	2 4 5 3
Gly4	→	11 9 18 20
Arg5	→	10 14 19
Ser6	→	13 15 2 3 6 11
Phe7	→	4 2 5 3
Gln8	→	1 16 17 8 4 3 2
Arg9	→	10 14 19
Thr10	→	13 15 6
Gly11	→	11 9 18 20
Thr12	→	13 15 6
Leu13	→	12 17 16 1 8 13 15 6 11 18 7 20
Ala14	→	18 7 11 20
Phe15	→	4 2 5 3
Glu16	→	1 16 17 4 5 7 20
Arg17	→	10 14 19
Val18	→	8 6 5 13 18 7 20
Tyr19	→	4 2 5 3
Thr20	→	13 15 6
Ala21	→	18 7 11 20

Figure 16. Residue-to-graphs relationship for NAc-t21a. Note that only the first twenty graphs in Figure 15 are used. The other four are ignored so that the tree search algorithm (TSA, discussed in text) can be more efficient.

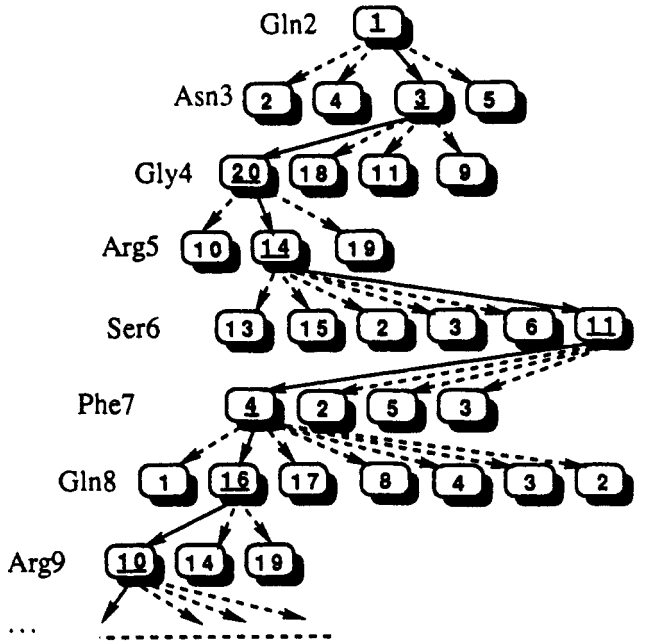


Figure 17. Possible assignment tree generated from the residue-to-graphs relationship (Figure 16). Note that the solid arrows represent the path that is the most likely sequence assignment, i.e. the best supergraph.

Residues	Q	N	G	R	S	F	Q	R	T	G	T	L	A	F	E	R	V	Y	T	A
Graphs	1	3	20	14	11	4	16	10	13	9	15	12	18	5	17	19	8	2	6	7
Number of NOESY Peaks	1	2	2	1	2	3	2	1	3	3	2	4	4	3	4	2	3	3	1	

Figure 18. Sequence assignment path and NOESY connectivities for NAc-t21a.

to the primary sequence supergraph, in a manner that is exactly analogous to the case of FPRA discussed above, where a candidate graph is mapped to an amino acid. TSA finds the best supergraph by (i) searching for the maximum NOESY correlations between neighboring residues, (ii) searching for the maximum supergraph similarity between a candidate supergraph and the primary sequence supergraph, and (iii) by picking the supergraph that has the maximum number of frequencies assigned.

In order to create these supergraphs, good NOESY data are required. As in the case of COSY and TOCSY spectra, NOESY spectra are more often than not far from ideal. In fact, it is very common that some neighboring residue protons do not give the expected cross peaks. Moreover, overlap problems arise frequently in NOESY spectra, making peak-picking difficult. Heteronuclear experiments are useful in this regard. Often, problems can also arise as far as the assignment is concerned since NOESY cross peaks arising from neighboring residue interactions cannot be distinguished from those arising from other effects (e.g. nonneighboring residues).

In terms of creating the supergraphs themselves, other problems arise because a systematic application of NOESY data to assign the spin patterns can lead to a "permutation explosion" problem, i.e. to search every possible path in the residue-to-graphs relation requires an enormous amount of time. In order to avoid this problem, the tree search algorithm searches for the path in the residue-to-graphs relationship with only the best NOESY connectivities. TSA tries to find the maximum number of connectivities so that despite the fact that one peak arising for neighboring interactions might be missing two residues can still be connected. As shown in Figure 18, for each sequence assignment pair, for example, G1 and G3, there is at least one NOESY cross peak found as the proof for identifying the space connection of this pair. If such evidence is found, then the connection between the two graphs is 1. Otherwise it is 0. Therefore, the total connection of the path in Figure 18 is 19. The total number of NOESY cross peaks for this path is 46. The total connection and the total number of NOESY cross peaks are always maximized by TSA. Before finding the path with the best NOESY connectivities, TSA must systematically search many trees which have different starting points. The amount of possible different paths is enormous. Strategies can however be used to sharply reduce the number of possibilities: (i) calculating the NOESY correlations in advance to avoid repeat computing; (ii) assigning the sequence fragment by fragment (similar to the manual assignment procedure). When the proteins are large, searching for the best path can lead to a serious combinatorial explosion problem. In this case, TSA allows the user to (i) do the sequence-specific assignment fragment by fragment, (ii) do the sequence-specific assignment from beginning to end, or (iii) search for NOESY connectivities with or without NOESY restrictions. A potentially more complete and useful method is also to use a parallel algorithm. This option is currently being explored. More on TSA will be reported in a subsequent paper.

PROGRAMS AND IMPLEMENTATION

CPA is written in SUN Pascal language and FPRA and TSA are in the SUN C language. All algorithms run on a SUN Sparc station 2. The basic steps involved in the automated assignment procedure which makes use of this software is summaried in Figure 19.

The methodology is based upon a "good enough" input data set, namely, DQF-COSY, TOCSY, and NOESY cross-peak tables. The software also provides a number of interactive functions to help a user refine the peak set. Moreover, it gives the user the number of maximum and minimum DQF-COSY cross peaks required for assigning backbone and side-chain protons given the primary sequence. This information can serve as a guideline in determining the quality of the experimental data sets. By means of fuzzy spin pattern recognition, the software can predict some missing cross peaks which a user can look for in the experimental spectra. If these cross peaks are found, then fragmented patterns created by CPA can be united to form more complete spin coupling

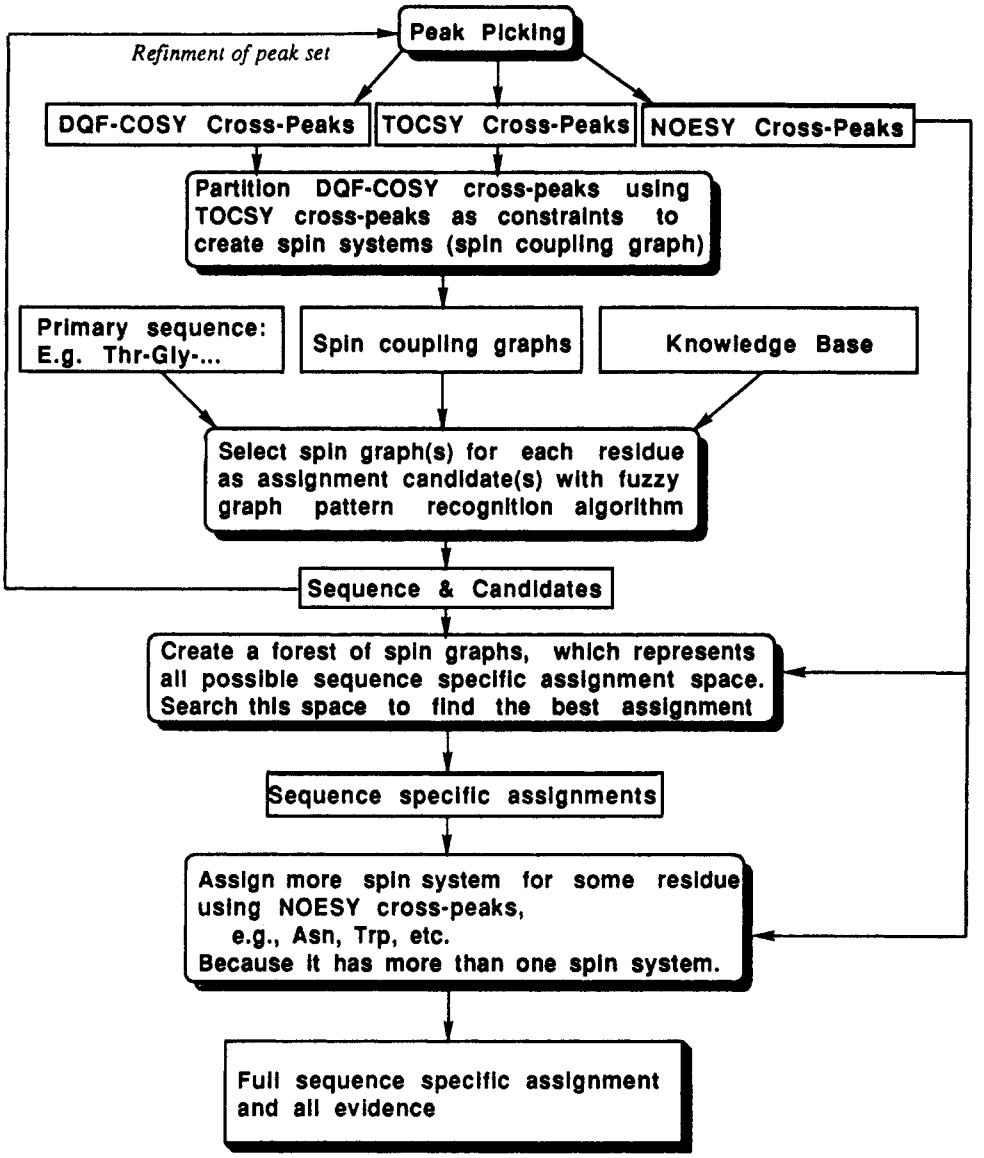


Figure 19. Flow chart of the completely automated sequence specific assignment of a protein from DQF-COSY, TOCSY, and NOESY cross-peak data sets.

patterns which can be easily assigned to an amino acid. The software provides a user–friend interface to assist chemists in controlling each step of the automated assignment procedure.

AUTOMATED ASSIGNMENT OF BPTI

To date, our programs have been tested on the raw data from NAc-t21a (cf. Appendix B) in three different environments, obtained from Peninsula Laboratories Inc. (Belmont, CA), and on basic pancreatic trypsin inhibitor (BPTI), obtained from Sigma Co. (St. Louis, MO). NMR samples of the synthetic peptide typically contained 4 mM protein in 90/10% H_2O/D_2O, and the pH was adjusted to 4.5 (uncorrected meter reading). In the case of BPTI, a 3.3 mM solution was used and the pH adjusted to 4.6. NMR spectra were recorded on a Bruker AMX 500 spectrometer at 295 K for the NAc-t21a and at 309 K for BPTI. Two-dimensional spectra were acquired in a phase sensitive mode using presaturation to remove the water signal and mixed states-TPPI in the t_1 dimension. In each case 600 t_1 increments were recorded with a size of 2K and 48 transients. The two dimensional data sets were then processed using Felix (Hare Research Inc., Bothell, WA). The data were Fourier transformed with mild resolution enhancement to 2K × 2K real matrices.

The manual assignment of BPTI (58 residues) has been done by Wüthrich's group.[29] Using the same experimental conditions, we carried out DFQ-COSY, TOCSY, and NOESY experiments on a Bruker NMR spectrometer. With the data being processed with FELIX, the following cross peak set was extracted:

2DF-COSY cross peaks: 338

TOCSY cross peaks: 855

All peak sets were saved as ASCII files. After these peak sets and the primary sequence were input into the system software, the following output was generated:

Arg x 6	Pro x 4	Asp x 2	Phe x 4
Cys x 6	Leu x 2	Glu x 2	Tyr x 4
Thr x 3	Gly x 6	Lys x 4	Ala x 6
Ile x 2	Asn x 3	Gln x 1	Val x 1
Ser x 1	Met x 1		

Knowledge Base: 22 cluster centers

Required number of 2QF-COSY cross-peaks: 439 ~ 190 (One side)

Actual number of 2QF-COSY cross-peaks: 227 (One side)

Required number of 2QF-COSY cross-peaks in NH-αH region: 60 ~ 54 (One side)

Required number of 2QF-COSY cross-peaks in αH-βH region: 92 ~ 52 (One side)

where "twenty-two cluster centers" means that the program chooses twenty-two types of amino acid spin coupling topological cluster centers from the knowledge base. This number is expected since there are eighteen kinds of residues in the primary sequence and PHe, Asn, Tyr, and Gln have two spin coupling cluster centers (aromatic rings and $-NH_2-$ groups).

After the raw data are processed by CPA with a tolerance of 0.02 ppm, a group of spin coupling systems are produced. A sample output is:

```
/*1st G/          Total Number of Peaks = 3
//Peak 1 (10.535, 4.285 ) -> 2 (4.286,10.528)
//Peak 192 (4.286, 2.724) -> 193 (2.724,4.285)
//Peak 190 (4.287, 3.454) -> 194 (3.463,4.285)
//TOCSY 356 (2.727, 10.532) => 1 (10.535, 4.285 ) +192 (4.286, 2.724)
//TOCSY 171 (2.727, 10.532) => 190 (4.287, 3.454) +192 (4.286, 2.724)
//Spin Coupling Topological Graph
1H,10.535,2
2H,4.285,1,3,4
3H,2.724,2
4H,3.454,2
/
/*2nd G/          Total Number of Peaks = 4
//Peak 3 (9.899,5.118) -> 4 (5.124,9.889)
//Peak 122 (5.125, 3.398)
//Peak 121 (5.125, 2.775) -> 123 (2.781, 5.125)
//Peak 254 (3.398,2.781) -> 255 (2.781, 3.393)
//TOCSY 358 (3.398,9.901) => 3 (9.899,5.118) + 122 (5.125, 3.398)
//TOCSY 248 (3.398, 2.781) => 122 (5.125, 3.398) + 121 (5.125, 2.775)
//TOCSY 701 (5.125, 3.398) => 254 (3.398,2.781) + 121 (5.125, 2.775)
1H,9.899,2
2H,5.118,1,3,4
3H,3.398,2,4
4H,2.775,2,3
```

.....

where the lines starting with "//" are the evidence for the partitioning and the lines without "//" produce an internal computer representation of a spin coupling topological graph, then used by FPRA to map to actual residues. FPRA produces a graph-to-residues relation which is converted into a residue-to-graphs relation as partially illustrated as follows: The best assignment path is found by TSA. Thus each graph (represented in bold as follows) is assigned to a particular amino acid residue in the sequence.

Arg1 → **54** 25 20 8 22 23
Pro2 → **55** 50 57 52
Asp3 → 33 34 51 48 6 4 1 35 2 9 39 **13** 5 44
Phe4 → **34** 6 51 33 2 48 4 1 9 44 13 35 39 5
Cys5 → **40** 14 18 49 46 53
Leu6 → **37** 47 16 19 41 9 33 34 13 51
Glu7 → 17 16 10 27 **38** 37 19 51 47 33 34
Pro8 → 55 **50** 57 52
Pro9 → 55 50 **57** 52
Tyr10 → 6 33 34 51 4 9 48 13 2 1 **35** 39 44 5
etc.

The sequence specific assignment along with the number of NOESY cross peaks found as evidence, given in square brackets, is given as follows.

Arg-[1]-Pro-[2]-Asp-[1]-Phe-[1]-Cys-[2]-Leu-[1]-Glu-[2]-Pro-[1]-Pro-[2]-Tyr-[2]-Thr-[1]-Gly-[1]-Pro-[1]-Cys-[2]-Lys-[1]-[2]-[3]-Ile-[1]-Ile-[2]-Arg-[2]-Tyr-[3]-Phe-[3]-Tyr-[1]-Asn-[1]-Ala-[1]-Lys-[2]-Ala-[2]-Gly-[2]-Leu-[2]-Cys-[4]-Gln-[2]-Thr-[4]-Phe-[2]-Val-[3]-Tyr-[2]-Gly-[3]-Gly-[3]-Cys-[1]-Arg-[2]-Ala-[2]-Lys-[4]-Arg-[4]-Asn-[4]-Asn-[3]-Phe-[2]-Lys-[3]-Ser-[3]-Ala-[3]-Glu-[2]-Asp-[5]-Cys-[5]-Met-[5]-Arg-[4]-Thr-[4]-Cys-[4]-Gly-[4]-Gly-[5]-Ala

Residues that have more than one spin system associated with them, such as Phe, which has two independent spin coupling networks, is assigned by using NOESY cross peaks.

Following is the partial output for the full assignment of BPTI:

Arg1 :

H:	αH	βH	βH
E:	4.28	1.63	1.79
A:	4.358	1.799	1.879
M:	0.975	0.926	0.966

Similarity=0.956 RMS=0.112 Max_difference=0.169

Pro2 :

H:	αH	βH	βH	γH	γH	δH	δH
E:	4.48	1.88	2.18	1.92	2.02	3.62	3.77
A:	4.317	2.011	0.903	1.592	1.855	3.600	3.730
M:	0.871	0.932	0.006	0.806	0.935	0.997	0.991

Similarity=0.856 RMS=0.303 Max_difference=1.277

Asp3 :

H:	NH	αH	βH
E:	8.31	4.65	2.63
A:	8.673	4.240	2.759
M:	0.776	0.342	0.917

Similarity=0.721 RMS=0.301 Max_difference=0.410

.....

The entries which are underlined are the assigned chemical shift values.

CONCLUSION

When compared to the manual assignment procedure, the automated assignment methodology has many advantages. The first of these is that the automated method is very fast and effective. For NAc-t21a, for example, CPA and FPRA (fuzzy pattern recognition algorithm) only took a total of 2 min of CPU time (on a SUN Sparc station 2). Moreover, unlike for the manual assignment where the assignment procedure is restricted by the quality of the data and the experimenter's capabilities in extracting spin coupling patterns, the automated method is only restricted by the precision and number of peaks in COSY, TOCSY, and NOESY cross-peak sets. Although CPA does not require that all theoretical COSY cross peaks be found, it needs "enough" cross peaks. Take NAc-21a as an example. The number of theoretical COSY cross peaks for this peptide are in the range of 68–137. Experimentally, 74 cross-peaks were picked. Therefore, CPA and FPRA fully assigned nine of the twenty-one residues and partially assigned twelve residues. If assigned manually, twelve residues are also partially assigned and some of the assignments are based on TOCSY cross peaks only. However, in the first step of the automated assignment, our software suggested some possible assignments. Second, once more peaks were picked, more assignments were made. One of the reasons that missing peaks exist is that some COSY spectral regions were not considered as being important for the manual assignment. These regions are however useful for the automated method. Third, the assignment of the amino acids that have more than one spin coupling system (Phe, Tyr, Trp, and His) have more than one spin coupling system is easily carried out using the automated method once the main side chain is assigned. In addition, as the weight of a protein increases, the spin coupling networks become increasingly more complicated. This renders the manual assignment more difficult. The automated method, on the other hand, is powerful in this case. Finally, a practical automated system should tolerate deficiencies in the experimental data sets to some extent. This is one of the characteristics of our software.

Table IX. Theoretical *J* Coupling Networks for the Twenty Amino Acids

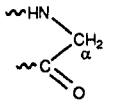
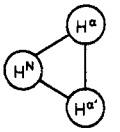
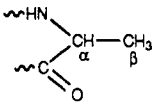
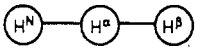
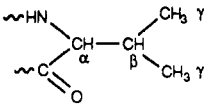
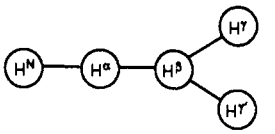
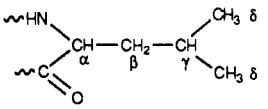
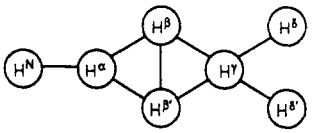
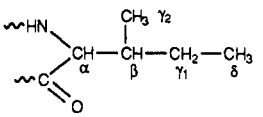
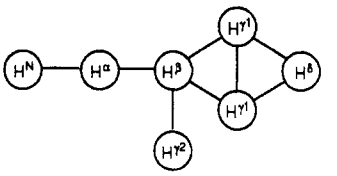
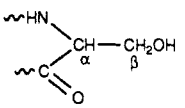
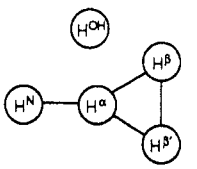
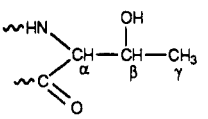
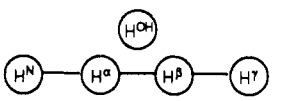
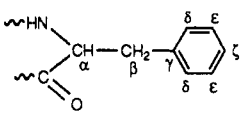
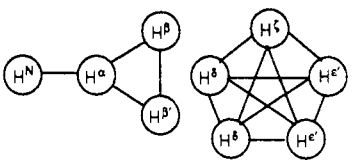
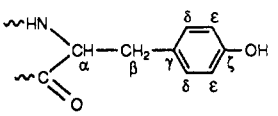
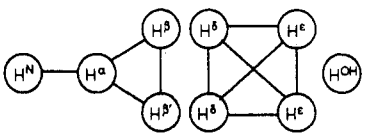
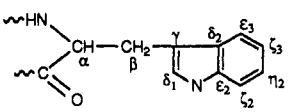
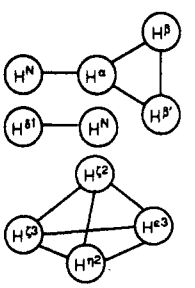
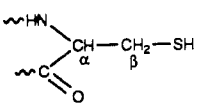
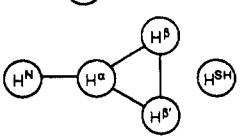
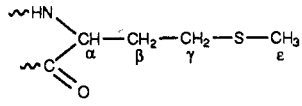
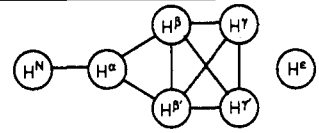
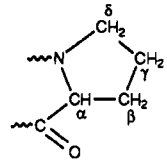
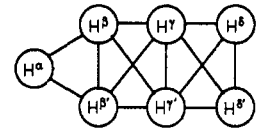
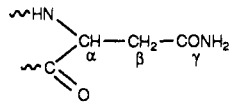
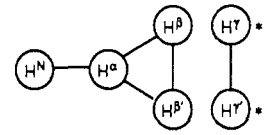
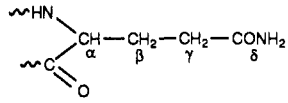
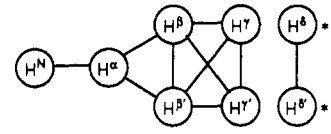
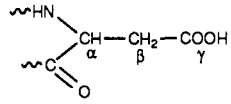
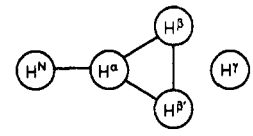
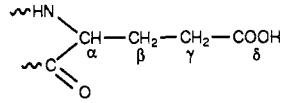
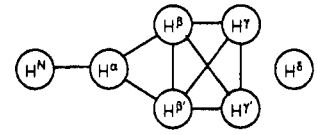
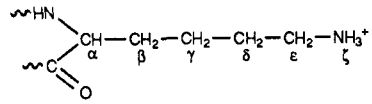
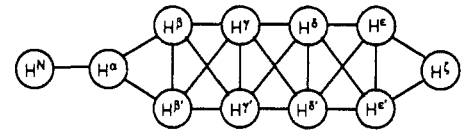
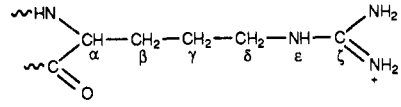
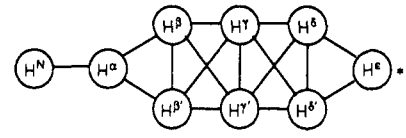
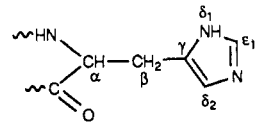
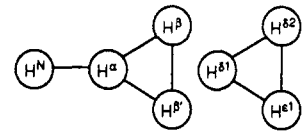
residue	chemical structure	spin coupling graph[a,b]
Gly (G)		
Ala (A)		
Val (V)		
Leu (L)		
Ile (I)		
Ser (S)		
Thr (T)		
Phe (F)		
Tyr (Y)		
Trp (W)		
Cys (S)		

Table IX (Continued)

residue	chemical structure	spin coupling graph[a,b]
Met (M)		
Pro (P)		
Asn (N)		
Gln (Q)		
Asp (D)		
Glu (E)		
Lys (K)		
Arg (R)		
His (H)		

[a] Notes: (1) The amide proton in the NH–αH coupling is labile but often observable in NMR spectra, especially when H_2O is used as the solvent. (2) In His, $H^{\epsilon 1}$ and $H^{\delta 2}$ often appear as two singlets, but the connectivity through the small four-bond coupling of approximately 1 Hz was observed through the small four-bond coupling of approximately 1 Hz was observed in spectra of several proteins. The singlets will appear in the COSY spectra as diagonal peaks, which due to heavy overlap, are very difficult to pick however. Moreover, $H^{\delta 1}$ is very labile. [b] Protons marked with an asterisk are side-chain labile protons that are not observable in some cases, particularly when a protein sample is dissolved in D_2O.

The automated assignment procedure relies heavily on good peak sets. In fact, the time-consuming aspect of our approach is making sure the peaks are properly picked. However, our program helps in the peak-picking because it is capable of predicting peaks that a user might have overlooked.

The assignment procedure outlined here could prove to be even more useful when used in conjunction with heteronuclear 3D experiments. Such experiments hold considerable promise for the study of larger proteins because one of the coherence transfer steps involves scalar couplings that are much larger than the proton line widths.[10] Moreover, heteronuclear spectra are typically more complete, and 13C chemical shifts of the side chains, for instance, are very good indicators of the side-chain type. Alternatively, data obtained from such 3D experiments as NOESY–COSY and NOESY–TOCSY[30] could also improve the assignment procedure. That is in a NOESY–COSY experiment, for example, both the 2D NOESY and 2D COSY data would be contained in the same data set, thus making the data easier to interpret.

ACKNOWLEDGMENT

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

Table X. Comparison of Automated and Manual Assignments for the Peptide NAc-t21a in DPC Micelles[a]

				Thr 1					
T:	NH	αH	βH		γH				
E:	8.03	4.53	4.17		1.15				
μ:		0.379	0.995		0.576				
C:		3.931	4.201		1.318				
M:		3.941	4.205		1.330				
				Gln 2					
T:	NH	αH	βH	βH′	γH	γH′	δH	δH′	
E:	8.28	4.43	1.92	2.10	2.35	2.29	6.85	7.61	
μ:	0.582	1.000	0.934	0.993	0.984		0.855	0.983	
C:	8.969	4.434	2.020	2.123	2.386		6.965	7.663	
M:	8.968	4.437	2.030	2.133	2.389				
				Asn 3					
T:	NH	αH	βH	βH′	γH	γH′			
E:	8.29	4.73	2.69	2.95	7.18	7.78			
μ:	0.820	0.996	0.885	0.885	0.999	0.975			
C:	8.684	4.725	2.807	2.807	7.156	7.708			
M:	8.686	4.733	2.836	2.808	6.964	7.669			
				Gly 4					
T:	NH	αH	αH′						
E:	8.31	4.17	3.74						
μ:	0.582	1.000							
C:	8.355	4.134							
M:	8.494	3.974							
				Arg 5					
T:	NH	αH	βH	βH′	γH	γH′	δH	δH′	εH
E:	8.20	4.28	1.79	1.63	1.56	1.52	3.14	3.11	7.21
μ:	0.989	0.980	0.995		0.990		0.993		0.494
C:	8.321	4.351	1.756		1.607		3.170		7.400
M:	8.321	4.357	1.748	1.795	1.621				7.402
				Ser 6					
T:	NH	αH	βH	βH′					
E:	8.48	4.50	3.72	3.89					
μ:	0.998	1.000	0.971	0.996					
M:	8.446	4.495	3.835	3.928					
C:	8.444	4.496	3.827	3.928					
				Phe 7					
T:	NH	αH	βH	βH′	2/6H	3/5H	4H		
E:	8.49	4.69	3.16	2.85	7.12	7.17	7.08		
μ:	0.969	0.967	0.993						
C:	8.692	4.566	3.128						
M:	8.697	4.575	3.139		7.285				
				Gln 8					
T:	NH	αH	βH	βH′	γH	γH′	δH	δH′	
E:	8.28	4.43	1.92	2.10	2.35	2.29	6.85	7.61	
μ:		1.000	0.953	0.993	0.985		0.988	0.997	
C:		4.436	2.004	2.123	2.385		6.910	7.589	
M:	8.528	4.168	2.083	2.046	2.327		6.908	7.593	
				Arg 9					
T:	NH	αH	βH	βH′	γH	γH′	δH	δH′	εH
E:	8.20	4.28	1.63	1.79	1.52	1.56	3.11	3.14	7.21
μ:		0.992		1.000					
C:		4.237		1.790					
M:	8.025	4.240		1.940		1.695		3.227	7.514
				Thr 10					
T:	NH	αH	βH		γH				
E:	8.30	4.53	4.17		1.15				
μ:		0.623	0.951		0.826				
C:		4.118	4.268		1.249				
M:	8.256	4.162	4.233		1.244				
				Gly 11					
T:	NH	αH	αH′						
E:	8.31	3.74	4.17						
μ:	0.873	0.910							
C:	8.633	3.88							
M:	8.638	3.887							
				Thr 12					
T:	NH	AH	BH	GH					
E:	8.30	4.53	4.17	1.15					
μ:	0.979	0.635	0.951	0.826					
C:	8.144	4.120	4.268	1.249					
M:	8.142	4.128	4.272	1.263					

Table X (Continued)

					Leu 13				
T:	NH	αH	βH	βH′	γH	δH	δH′		
E:	8.19	4.25	1.60	1.71	1.51	0.68	0.83		
μ:	0.999	0.999		0.743	0.951	0.919			
C:	8.222	4.233		1.741	0.909	0.933			
M:	8.223	4.243	1.565	1.744	1.746	0.914	0.954		
					Ala 14				
T:	NH	αH	βH						
E:	8.15	4.24	1.32						
μ:	0.999	0.984	0.947						
C:	8.120	4.172	1.412						
M:	8.124	4.179	1.424						
					Phe 15				
T:	NH	αH	βH	βH′	3/5H	4H			
E:	8.49	4.69	2.85	3.16	7.17	7.08			
μ:		0.782	0.559	0.929	0.982	1.000			
C:		4.353	3.152	3.268	7.227	7.076			
M:	8.233	4.356	3.170	3.271					
					Glu 16				
T:	NH	αH	βH	βH′	γH	γH′			
E:	8.22	4.34	2.04	1.97	2.27	2.34			
M:		0.890	0.878		0.643	0.538			
C:		4.137	2.132		2.458	2.574			
M:	8.359	4.140	2.462	2.566	2.136				
					Arg 17				
T:	NH	αH	βH	βH′	γH	γH′	δH	δH′	εH
E:	8.20	4.28	1.79	1.63	1.56	1.52	3.11	3.14	7.21
μ:	0.968	0.999			0.941			0.871	0.447
C:	8.412	4.296			1.679			3.240	7.413
M:	8.416	4.303	1.972	1.874	1.621			3.244	7.412
					Val 18				
T:	NH	αH	βH		γH	γH′			
E:	8.20	4.16	2.02		0.88	0.76			
μ:	0.765	0.950	0.977		0.915	0.956			
C:	7.754	3.984	2.074		0.956	0.826			
M:	7.752	3.987	2.087		0.966	0.834			
					Tyr 19				
T:	NH	αH	βH	βH′		2/6H	3/5H		
E:	8.57	4.64	2.81	3.04		7.00	6.70		
μ:	0.831	0.968	0.816	0.860		0.995	0.906		
C:	8.028	4.515	2.689	2.886		7.019	6.789		
M:	8.028	4.522	2.692	2.889		7.024	6.788		
					Thr 20				
T:	NH	αH	βH		γH				
E:	8.03	4.53	4.17		1.15				
μ:			0.981		0.930				
C:			4.231		1.211				
M:	7.731	4.298	4.236		1.219				
					Ala 21				
T:	NH	αH	βH						
E:	8.15	4.24	1.32						
μ:	0.999	0.984	0.947						
C:	7.849	4.120	1.389						
M:	7.846	4.125	1.400						

[a] Remarks: T, type of proton; E, expected chemical shift value; μ, membership; C, computer's assignment; M, manual assignment.

APPENDIX A

The theoretical *J* coupling networks for the twenty amino acids are given in Table IX.

APPENDIX B

A comparison of automated and manual assignments for NAc-t21a is given in Table X.

REFERENCES AND NOTES

(1) Wüthrich, K. *NMR of Protein and Nucleic Acids*; Wiley: New York, 1986.
(2) Englander, S. W.; Wand, A. J. *Biochemistry* **1987**, *26*, 5953–5968.
(3) James, T. L.; Basus, V. J. *Annu. Rev. Phys. Chem.* **1991**, *42*, 501–542.
(4) Fesik, S. W. *J. Med. Chem.* **1991**, *34*, 2937–2945.
(5) Clore, G. W.; Gronenborn, A. M.; Carlson, G.; Meyer, E. F. *J. Mol. Biol.* **1986**, *190*, 259–267.
(6) Bax, A.; Weiss, M. A. *J. Magn. Reson.* **1987**, *71*, 571–575.
(7) Fesik, S. W.; Gampe, R. T.; Rockway, T. W. *J. Magn. Reson.* **1987**, *74*, 366–371.
(8) Fesik, S. W.; Zuiderweg, E. R. P. *J. Am. Chem. Soc.* **1989**, *111*, 5013–5015.
(9) Seeholzer, S. H.; Cohn, M.; Putkey, J. A.; Means, A. R. *Proc. Natl. Acad. Sci. USA* **1986**, *83*, 3634–3638.
(10) Fesik, S. W.; Zuiderweg, E. R. P. *J. Magn. Reson.* **1988**, *78*, 588–593.
(11) Griesinger, C.; Sørensen, O. W.; Ernst, R. R. *J. Magn. Reson.* **1989**, *84*, 14–63.
(12) Clore, G. M.; Gronenborn, A. M. *Prog. NMR Spectrosc.* **1991**, *23*, 43–92.
(13) Cieslar, C.; Clore, G. M.; Gronenborn, A. M. *J. Magn. Reson.* **1988**, *80*, 119–127.
(14) Weber, P. L.; Malikayil, J. A.; Müller, L. *J. Magn. Reson.* **1989**, *82*, 419–426.
(15) Eads, C. D.; Kuntz, I. D. *J. Magn. Reson.* **1989**, *82*, 467–482.

(16) Kleywegt, G. J.; Boelens, R.; Cox, M.; Llinas, M.; Kaptein, R. *J. Biomol. NMR* **1991**, *1*, 23–47.
(17) Kleywegt, G. J.; Lamerichs, R. M. J. N.; Boelens, R.; Kaptein, R. *J. Magn. Reson.* **1989**, *85*, 186–197.
(18) Hoch, J. C., Poulsen, F. M., Redfield, C., Eds. *Computational Aspects of the Study of Biological Micromolecules by Nuclear Magnetic Resonance Spectroscopy*; Plenum Press: New York, 1991.
(19) Pfändler, P.; Bodenhausen, G. *J. Magn. Reson.* **1988**, *79*, 99–123.
(20) Xu, J.; Sanctuary, B. C. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 475–489.
(21) Kaufmann, A. *Introduction to the Theory of Fuzzy Subsets*; Academic Press: New York, 1975; Vol. 1.
(22) Balaban, A. T., Ed. *Chemical Applications of Graph Theory*; Academic Press: New York, 1976; pp 333–365.
(23) Gross, K.-H.; Kalbitzer, H. R. *J. Magn. Reson.* **1989**, *76*, 87–99.
(24) Fairbrother, W. J.; Palmer, A. G., III; Rance, M.; et al. *Biochemistry* **1992**, *31*, 4413–4425.
(25) Inagaki, F.; Shimada, I.; Kawaguchi, K.; et al. *Biochemistry* **1989**, *28*, 5985–5991.
(26) Ash, J. E.,; et al. *Communication, Storage and Retrival of Chemical Information*; John Wiley & Sons: New York, 1985; pp 129–131.
(27) Tarjan, R. E. *Graphic Algorithm in Chemical Computation*; American Chemical Society: Washington, D.C., 1977. Vol. 46, pp 1–20.
(28) Xu, J.; Zhang, M. *Tetrahedron Comput. Methodol.* **1989**, *2*, 75–83.
(29) Wagner, G.; Braun, W.; Havel, T. F.; et al. *J. Mol. Biol.* **1987**, *196*, 611–639.
(30) Vuister, G. W.; Boelens, R.; Padilla, A.; Kleywegt, G. J.; Kaptein, R. *Biochemistry* **1990**, *29*, 1829–1839.