CPA: Constrained Partitioning Algorithm for Initial Assignment of Protein ^{1}H Resonances from MQF-COSY

Jun Xu[†] and B. C. Sanctuary[*]

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

Received October 27, 1992

A new constrained partitioning algorithm (CPA) for initial assignment of protein ^{1}H resonances from MQF-COSY is proposed in this paper. First, the graph theory properties of spin system 2D patterns and a method of searching for these patterns from COSY spectra are discussed. These lead to an unconstrained partitioning method which uses a binary tree generation algorithm. In practice, the data set is assumed to be incomplete and to contain redundancies. Partitioning of this data requires techniques that deal with overlap and the data incompleteness. CPA algorithm is implemented in C and Pascal Languages on a SUN sparc station, and tested on the 2D NMR data set of melittin protein. At the end of the partitioning, the individual spin patterns are displayed and resonances assigned. The graphic program needs the support of the Open Windows system.

INTRODUCTION

A crucial step in the determination of structures by solution-state NMR spectroscopy is the sequence-specific assignment of the ^{1}H resonances. Furthermore, cross-peaks in NOESY spectra must be unambiguously assigned before they can be used to obtain distance constraints for secondary and tertiary structure elucidations.[1] Manual assignment is tedious and laborious, involving a considerable amount of data retrieval, comparison, and deduction. Therefore, automation of this process is needed.[2-6] To date, treatments have been aimed at implementing Wüthrich's sequential assignment strategy.[7] One of the key steps of this method is to detect and classify protein spin coupling systems. Our algorithms are philosophically based on Wüthrich's sequential assignment strategy, although more attention has been given to dealing with the heavy overlap problem.

In principle, protein spin coupling systems can be "read out" from correlation spectra (for example, DQF-COSY spectra) by recognition of spin topological patterns. If there are two cross-peaks (ω_i,ω_j) and $(\omega_{i'},\omega_k)$, and if abs$(\omega_i - \omega_{i'}) \leq T$ (T is a given tolerance), then spin i can be asserted to couple to spin j and spin k. Furthermore, if there is another cross-peak (ω_k,ω_l), more complex spin coupling topologies like j–i–k–l, etc., may exist. A major problem for protein studies is that a protein consists of many amino acids with some residues being of the same type. Therefore the overlap can be very heavy. Because of the overlap, the above assertions cannot be confirmed without further evidence from other cross-peaks or data sources. For example, it is not certain that cross-peaks (ω_i,ω_j) and $(\omega_{i'},\omega_k)$ belong to the same spin system $\{i,j,k\}$ rather than two separated spin systems $\{i,j\}$ and $\{i',k\}$ even though ω_i and $\omega_{i'}$ are within the tolerance. Tolerance evidence alone is not sufficient. Additional constraints such as a cross-peak (ω_j,ω_k), from, for example, TOCSY spectra or Relay COSY spectra are needed to resolve the overlap problem. On the basis of this idea, a constrained partitioning algorithm for automatically detecting and classifying the protein spin systems from MQF-COSY ($M \leq 4$), which uses as constraints data sets such as Relay COSY, MQ experiment, and/or TOCSY, is proposed.

UNCONSTRAINED PARTITIONING AND BINARY TREE GENERATION ALGORITHM

The algorithms are discussed by means of examples. Consider a proline residue which has a spin coupling topology illustrated in Figure 1. There are, in addition, 13 cross-peaks observed from DQF-COSY spectra (Table I). Only weak spin scalar couplings are retained.

The spin topology (Figure 1b) can be considered to be a mathematical graph as defined by graph theory.[8-10] Hence representation of the cross-peaks is given by an edge. For example, peak 1 coincides with edge α–β, peak 2 is edge α–β', etc. Conversely, by analyzing cross-peaks, the spin topology can be created.

A spin coupling system can be represented by a nondirectional strongly connected graph.[9] According to graph theory, a strong graph is traversable from any arbitrary starting node. By traversing a spin topological graph, all possible peaks belonging to the same spin system can be included. This is called a *path* which is a sequence of edges in a graph. In the case at hand, a *path* is a sequence of MQF-COSY cross-peaks in a graph. Walking from any node of a strong graph, along any *path* can include all edges of the graph. That is, every node in the graph can be visited.[10] A cross-peak data set can therefore be considered as a set of edges which can consist of different spin topological graphs. By implementing walks on a MQF-COSY cross-peak set, the edges belonging to the same spin topological graph can be partitioned; that is, this group of edges is assigned to a specific amino acid residue of a protein molecule.

Again, take Figure 1 as an example. If a walk from peak 1 is initiated, e.g., α–β, and the value of frequency β is retained for further walking, then α is used as the pointer to go ahead. Alternately, α can be retained, and β is used as the pointer, which although it generates a different *path*, gives the same result. This is because peak 2 has the same frequency as peak 1 (i.e., α); therefore, peak 2 is partitioned into the walking *path*, and so forth. When the walking finishes, there can be two kinds of paths: (1) 1 → 2 → 3 → 5 → 4 → 7 → 6 → 8 → 11 → 9 → 12 → 13 → 10 and (2) 1 → 2 → 5 → 6 → 11 → 4 → 9 → 10 → 12 → 8 → 7 → 13 → 3, which correspond

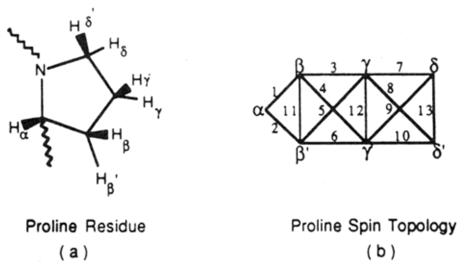
[*] To whom correspondence should be addressed.

[†] Present address: New Methods Research, Inc./Tripos Associates, Inc., 6035 Corporate Dr., East Syracuse, NY 13057-1016.

0095-2338/93/1633-0490$04.00/0

Table I. Possible Proline DQF-COSY Cross-Peaks According to Figure 1b

1	2	3	4	5	6	7	8	9	10	11	12	13
(α,β)	(α,β')	(β,γ)	(β,γ')	(β',γ)	(β',γ')	(γ,δ)	(γ,δ')	(γ',δ)	(γ',δ')	(β,β')	(γ,γ')	(δ,δ')



Proline Residue
(a)

Proline Spin Topology
(b)

Figure 1. Proline spin topology and DQF-COSY cross-peaks.

respectively to *Breadth-first* algorithm and *Depth-first* algorithm (see Figure 2).

On these binary trees, every walk is along an edge of the spin topological graph (physically, for example, a COSY cross-peak), which can have no more than two possible directions to go forward. If the original data set (called as a peak space) includes more than one spin system, the result of walking from different starting nodes can result in a binary tree forest. Unconstrained partitioning relies only on tolerance to extract spin graphs; i.e., two peaks having the same frequency within a given tolerance are partitioned to the same spin fragment. Reliable partitioning, however, requires further constraints to split up a peak space into many subspaces and at the same time avoids accidental overlap. These subspaces may correspond to complete spin systems or fragments of a complete spin system.

When the partitioning algorithm is successful, it generates binary trees. Every cross-peak (ω_1,ω_2) in the data set is considered as a node of a binary tree, such that, ω_1 and ω_2 are expected to produce the left subtree and the right subtree of their parent node. Breath-first strategy searches all possible left and right subtrees in the current level and then chooses one of the nodes to go ahead. Depth-first strategy searches all possible nodes in the left subtree and then back-tracks to the upgrade level to search the right subtree.

In Figure 2, although two different searching strategies produce two *paths*, both contain the same node set. That means the cross-peak set can be partitioned into individual spin systems by walking. In the ideal case of no overlap and data incompleteness the partitioning results are independent of the starting point and searching strategies.

In Figure 2, both path 1 and path 2 include all peaks of Table I, which can be easily transformed into a topological adjacency table or displayed on a screen to show the spin pattern, in this case, Figure 1b.

Unconstrained partitioning, however, has a number of difficulties which must be addressed for reliable results.

(1) TOLERANCE: suppose a peak set has been created from a MQF-COSY experiment; then the unconstrained partitioning algorithm will decide if two peaks (ω_i,ω_j) and (ω_k,ω_l) belong to the same spin system by tolerance alone. That is, the absolute value of the difference of two peaks is below a given tolerance condition T,

$$\mathrm{abs}(\omega_i - \omega_k) \leq T \quad \text{or} \quad \mathrm{abs}(\omega_j - \omega_l) \leq T \quad \text{or}$$
$$\mathrm{abs}(\omega_i - \omega_l) \leq T \quad \text{or} \quad \mathrm{abs}(\omega_j - \omega_k) \leq T$$

If the tolerance is too big, it may incorrectly include other spin topological fragments which do not actually belong to the current spin system. On the other hand, if the tolerance is too narrow, some peaks can be missed and important spin topological fragments may be lost. The latter case breaks a spin system into many fragments which are too small. For example, in Figure 2, the depth-first algorithm, if peak 4 is not partitioned to that fragment, then the spin system will become three fragments, namely, {1, 2, 3, 5, 6, 11} and {9, 10, 12, 8, 7, 13} and {4}.

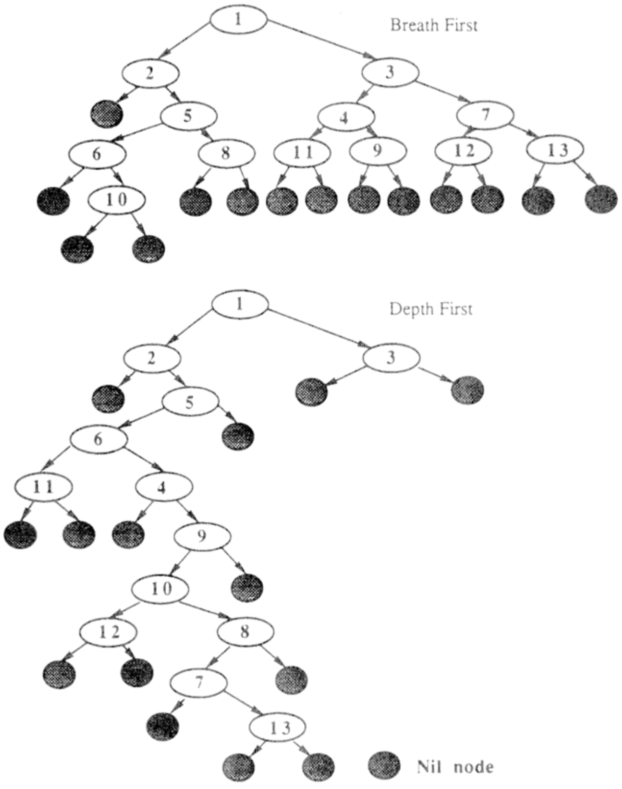


Figure 2. Walking on the cross-peak data set to find the *path*.

(2) HEAVY OVERLAP: a protein molecule consists of many amino acid residues; some residues are of the same type, and the proton chemical shifts of most residues appear in a narrow frequency band except for the ones on aromatic groups. It is quite common for a reasonable tolerance T, $\mathrm{abs}(\omega_i - \omega_{i'}) \leq T$, that peaks (ω_i,ω_j) and $(\omega_{i'},\omega_k)$ belong to different spin systems but are incorrectly included in the same path produced by the unconstrained partitioning algorithm. Therefore, using tolerance alone to decide peak partitioning is necessary, but not sufficient. The unconstrained partition algorithm is only valid when there is almost no overlap and the data set is complete. The unconstrained partitioning algorithm is not useful for usual protein NMR spectra assignments, thereby leading to the need for partitioning constraints other than tolerance.

CONSTRAINED PARTITIONING ALGORITHM

To improve the tolerance-only partitioning, an algorithm must be found which treats heavy overlap situations. Various partitioning constraints can be considered. The strategies to use these constraints, the experimental data incompleteness, and data redundancy are treated by graph theory.

Walking from a peak $(\omega_{i1},\omega_{i2})$ to another peak $(\omega_{j1},\omega_{j2})$ means in graph theory that two edges in the spin topological graph are merged; see Figure 3.

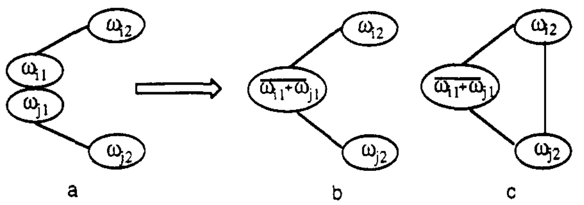


Figure 3. Walking and merging.

The unconstrained partition algorithm carries out this merge only when the condition P0 is satisfied:

(0) tolerance: $\text{abs}(\omega_{i1} - \omega_{j1}) \le T$ (P0)

Constrained partitioning algorithms need more evidence than P0 to justify this merge. Consider three data sources from which more evidence can be found.

(1) TOCSY/Relay COSY:

cross-peak: $(\omega_{i2},\omega_{j2})$ (P1)

(2) 2Q spectra: $(\omega_{i2}+\omega_{j2},\omega_{i1})$ IIM pathway[3] (P2)

or $(\omega_{i2}+\omega_{j2},\omega_{i2})$ DDM pathway[3] (P3)

or $(\omega_{i2}+\omega_{j2},\omega_{j2})$ DDM pathway (P4)

3Q spectra $(\omega_{i1}+\omega_{i2}+\omega_{j2},\omega_{i1})$ DDM pathway (P5)

(3) MQF-COSY cross-peak: $(\omega_{i2},\omega_{j2})$ (P6)

Theoretically, if the merging of Figure 3 is valid, i.e., from Figure 3a to Figure 3b, any one of the above-mentioned P1–P5 constraints should be satisfied. Constraint P6 may appear when three spins consist of a triangle coupling topology, i.e., from Figure 3a to Figure 3c. If P6 is true, then MQF-COSY cross-peak $(\omega_{i2},\omega_{j2})$ itself is also added into the current spin system. Therefore, (P6) should be used together with one more other constraints; otherwise, it may produce an incorrect result for the heavy overlap. Because this constraint evidence is from a MQF-COSY cross-peak data set and not other experiments, (P6) is also a self-partitioning constraint.

(P0)–(P6) are seven possible constraints of the partitioning algorithm. When logical operators "and" ("$\wedge$") "or" ("$\vee$") are applied to these constraints, many combinations of these constraints can be used, as the following examples show

$$\text{P0} \wedge \text{P1} \vee \text{P2} \vee \text{P3} \vee \text{P4} \vee \text{P5} \vee \text{P6} \quad \text{(C1)}$$

$$\text{P0} \wedge \text{P1} \vee \text{P2} \vee \text{P3} \vee \text{P4} \vee \text{P5} \wedge \text{P6} \quad \text{(C2)}$$

$$\text{P0} \wedge \text{P1} \vee \text{P2} \vee \text{P3} \vee \text{P4} \wedge \text{P5} \wedge \text{P6} \quad \text{(C3)}$$

$$\text{P0} \wedge \text{P1} \vee \text{P2} \vee \text{P3} \wedge \text{P4} \wedge \text{P5} \wedge \text{P6} \quad \text{(C4)}$$

$$\text{P0} \wedge \text{P1} \vee \text{P2} \wedge \text{P3} \wedge \text{P4} \wedge \text{P5} \wedge \text{P6} \quad \text{(C5)}$$

$$\text{P0} \wedge \text{P1} \wedge \text{P2} \wedge \text{P3} \wedge \text{P4} \wedge \text{P5} \wedge \text{P6} \quad \text{(C6)}$$

Moving from (C1) to (C6), the constraints become more strict, and the partitioning results are more reliable. On the other hand, when the value of the tolerance is changed, the strictness of the constraints will be changed. Clearly, when the tolerance is assigned as 0, (C6) will be the most strict constraint. In the initial phases of partitioning when there are more peaks with heavy overlap, a narrow tolerance and strict constraints are imposed. In the last phases of the partitioning fewer peaks are overlapped. Therefore, a wider and more spacious constraint is used. This partitioning strategy is called "constrained repartition".

In the ideal case, all necessary peaks are observed, and the data set is complete. Furthermore, no overlap occurs. Hence, only a boolean type data entry is needed to mark a peak, and partition it. Subsequent partitioning scans will not partition this peak further. When the data are complete, the partitioning result is not affected by the starting point of the algorithm. After partitioning, the full peak set for a spin system is always obtained. Such an ideal case does not, however, exist. Some peaks may be too weak to be observed, or a peak is found in more than one spin system because of the heavy overlap. For example, in the depth-first algorithm in Figure 2, if peak 9 belongs to the current spin system, but, because of heavy overlap, also belongs simultaneously to another spin system, and, more over, if this peak has previously been partitioned to that other spin system, then a walk through peak 9 to arrive at the fragment {10 ,12 ,8 ,7 ,13} will not occur. This is called experimental data incompleteness. Because of this incompleteness, the algorithm will give different partitioned results when a walk starts at different peaks.

To overcome this difficulty, all partitions from every starting peak must be done. If there are N peaks, then there are N paths. To get the complete peak set for the spin systems, the intersection of these N paths is calculated. For example, on the basis of the peak set given in Figure 1, 13 paths can be obtained starting at any of the 13 nodes:

path 1	{1, 2, 3, 5, 6, 11, 4, 9, 10, 12, 8, 7, 13}
path 2	{2, 1, 6, 4, 5, 11, 3, 12, 10, 9, 8, 7, 13}
path 3	{3, 7, 8, 5, 12, 9, 13, 10, 6, 4, 11, 2, 1}
	⋮
path 13	{13, 7, 12, 3, 1, 4, 2, 9, 11, 5, 6, 8, 10}

and

$$\text{path 1} \cup \text{path 2} \cup \text{path 3} \cup \ldots \cup \text{path 13} = \{1,2,3,4,5,6,7,8,9,10,11,12,13\}$$

This example is the ideal case. In practice, the different paths may include different elements, but they should also have common elements if they belong to the same spin system, namely,

$$\prod_{ij}\{[(\text{path}_i \cap \text{path}_j) \neq \phi]\wedge\} \quad (1)$$

where ϕ is the empty set. The assertion below is always true unless there is overlap

$$[(\text{path}_i \cap \text{path}_j) \neq \phi] \Rightarrow \text{path}_i \cup \text{path}_j \quad (2)$$

$\text{Path}_i \cap \text{path}_j \neq \phi$, however, can result from an overlap peak being in two different spin systems. Considering, then, the relations of path_i and path_j, there are the following possible situations:

(case a) $$\text{path}_i \cap \text{path}_j = \phi \quad (3)$$

This means path_i and path_j are without overlap, but this does not mean they do not belong to the same spin system because of the possibility of incomplete experimental data. For example, in the depth-first path of Figure 2, suppose peak 4 is missed; this path becomes two fragments, {1, 2, 3, 5, 6, 11} and {9, 10, 12, 8, 7, 13}, their intersection is null, but they

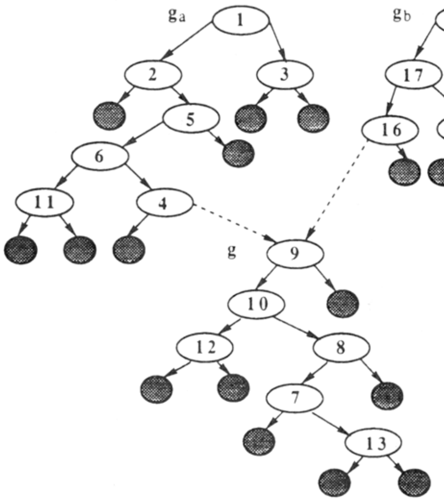


Figure 4. Graph theory representation of the peaks overlap.

belong to the same spin system. To justify merging the fragments, more evidence is needed.

(case b) $$\text{path}_i \cap \text{path}_j \equiv \text{path}_i \equiv \text{path}_j \quad (4)$$

implies path$_i$ and path$_j$ are isomorphic or redundant. One of them can be discarded,

$$\text{path}_i \supset \text{path}_j \vee \text{path}_j \supset \text{path}_i \quad (5)$$

In the case of path$_i$ ⊃ path$_j$, discard path$_j$; otherwise discard path$_i$.

(case c)

g, g_a, and g_b are graphs corresponding to the paths in Figure 4. We assume that g_a and g_b belong to different spin systems, but note

$$\text{path}_i \cap \text{path}_j \boxminus g \neq \phi,$$
$$g_a \supseteq \text{path}_i \wedge g_b \supseteq \text{path}_j \wedge g_a \neq g_b \quad (6)$$

This is the most common situation. To partition these results to the correct graphs g_a and g_b, more evidence is needed to decide how to deal with path$_i$ and path$_j$. This situation is explained by reference to the example in Figure 4.

g_a = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13} ⊃ P_i is carried over from Figure 2, the depth-first case. Suppose there is another spin system g_b = {9, 14, 15, 16, 17, 19, 20}, because peak 9 is an overlapped peak. So the partitioning algorithm obtains $g_{b'}$ = {9, 10, 12, 8, 7, 13, 14, 15, 16, 17, 19, 20} ⊃ P_j and $g = (g_a \cap g_{b'})$ = {7, 8, 9, 10, 12, 13} rather than g_b. Because of this overlap, the partition algorithm incorrectly takes subgraph g into g_b. This "avalanche effect" is introduced by an overlap within the tolerance and must be removed.

To describe how this is done, additional data from NMR experiments are considered. For example, TOCSY, Relay COSY MQ, etc., provide additional information which can be used to confirm or reject individual peaks in the overlap graph g. The set of peaks from the different types of experiments are called τ; that is,

$$\tau \in \{\text{TOCSY, MQF-COSY, Relay COSY, MQ, etc.}\}$$

Any node in g is called $g(i)$, and represents the ith element in g, i.e., a MQF-COSY cross-peak $(\omega_{ix},\omega_{iy})$, where x and y represent the two frequency domains. Other experimental evidence is needed to verify whether $g(i)$ belongs to g_a or g_b or both (in general, $g_{a/b}$ is called g_k).

The existence of correlations between g and g_k is deemed to be sufficient evidence for changing the earlier assignment based on weaker constraints. This is done by using a peak generator $\mathcal{R}(\tau,g_k,g(i),T_m)$ and peak-comparer $\Omega_k(\text{EP}_\tau,P_\tau,T_c)$, where EP$_\tau$ is the peak set from a τ type experiment, P_τ is the experimental peak set of all theoretically possible peaks from τ type experiment (P_τ is produced by $\mathcal{R}$ $(\tau,g_k,g(i),T_m)$, and T_c is the tolerance used by $\Omega_k(\text{EP}_\tau,P_\tau,T_c)$ for comparison.

$\mathcal{R}(\tau,g_k,g(i),T_m)$ is an operator which generates or predicts all theoretically possible peaks from various experiments τ correlated with peak $g(i)$ and peak subsets g_k and store the results in P_τ for use with the peak-comparer. That is, peak $g(i) = (\omega_{ix},\omega_{iy})$ is theoretically coupled by various experiments to a set of other peaks in g_k. For example, any node in g_k is called $g(j) = (\omega_{jx},\omega_{jy})$; operator $\mathcal{R}(\tau,g_k,g(i),T_m)$ will first

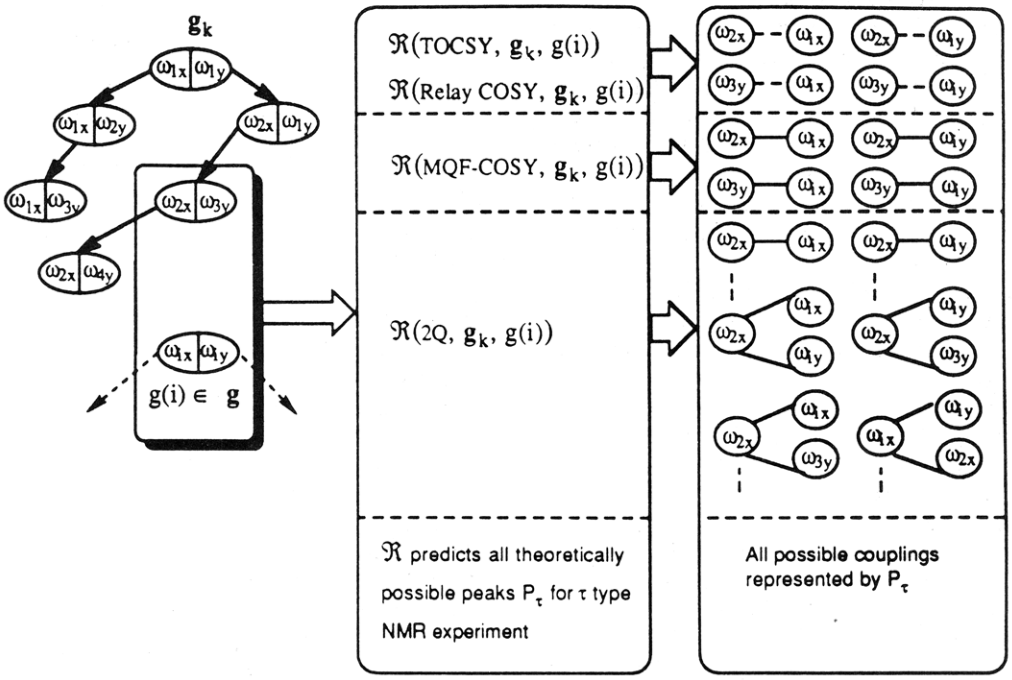

g_k
ω_{1x} ω_{1y}
ω_{1x} ω_{2y}
ω_{2x} ω_{1y}
ω_{1x} ω_{3y}
ω_{2x} ω_{3y}
ω_{2x} ω_{4y}
ω_{1x} ω_{1y}
g(i) ∈ g
ℜ(TOCSY, g_k, g(i))
ℜ(Relay COSY, g_k, g(i))
ℜ(MQF-COSY, g_k, g(i))
ℜ(2Q, g_k, g(i))
ℜ predicts all theoretically possible peaks P_τ for τ type NMR experiment
ω_{2x} ω_{1x} ω_{2x} ω_{1y}
ω_{3y} ω_{1x} ω_{3y} ω_{1y}
ω_{2x} ω_{1x} ω_{2x} ω_{1y}
ω_{3y} ω_{1x} ω_{3y} ω_{1y}
ω_{2x} ω_{1x} ω_{2x} ω_{1y}
ω_{1x} ω_{2x} ω_{1y} ω_{1y} ω_{2x} ω_{3y}
ω_{1x} ω_{2x} ω_{3y} ω_{1y} ω_{1x} ω_{2x}
All possible couplings represented by P_τ

Figure 5. Illustration of operator $\mathcal{R}$.

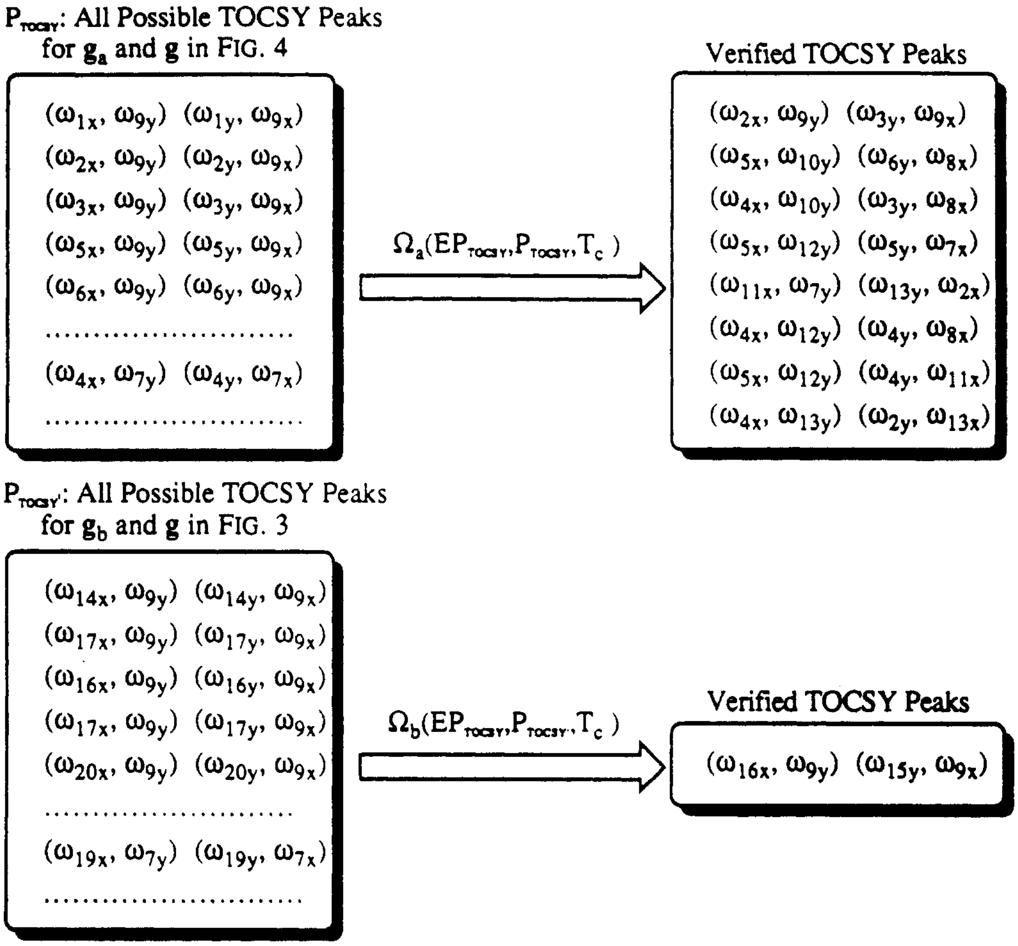

P$_{TOCSY}$: All Possible TOCSY Peaks for g$_a$ and g in FIG. 4
(ω$_{1x}$, ω$_{9y}$) (ω$_{1y}$, ω$_{9x}$)
(ω$_{2x}$, ω$_{9y}$) (ω$_{2y}$, ω$_{9x}$)
(ω$_{3x}$, ω$_{9y}$) (ω$_{3y}$, ω$_{9x}$)
(ω$_{5x}$, ω$_{9y}$) (ω$_{5y}$, ω$_{9x}$)
(ω$_{6x}$, ω$_{9y}$) (ω$_{6y}$, ω$_{9x}$)
(ω$_{4x}$, ω$_{7y}$) (ω$_{4y}$, ω$_{7x}$)
Ω$_a$(EP$_{TOCSY}$,P$_{TOCSY}$,T$_c$)
Verified TOCSY Peaks
(ω$_{2x}$, ω$_{9y}$) (ω$_{3y}$, ω$_{9x}$)
(ω$_{5x}$, ω$_{10y}$) (ω$_{6y}$, ω$_{8x}$)
(ω$_{4x}$, ω$_{10y}$) (ω$_{3y}$, ω$_{8x}$)
(ω$_{5x}$, ω$_{12y}$) (ω$_{5y}$, ω$_{7x}$)
(ω$_{11x}$, ω$_{7y}$) (ω$_{13y}$, ω$_{2x}$)
(ω$_{4x}$, ω$_{12y}$) (ω$_{4y}$, ω$_{8x}$)
(ω$_{5x}$, ω$_{12y}$) (ω$_{4y}$, ω$_{11x}$)
(ω$_{4x}$, ω$_{13y}$) (ω$_{2y}$, ω$_{13x}$)
P$_{TOCSY}$: All Possible TOCSY Peaks for g$_b$ and g in FIG. 3
(ω$_{14x}$, ω$_{9y}$) (ω$_{14y}$, ω$_{9x}$)
(ω$_{17x}$, ω$_{9y}$) (ω$_{17y}$, ω$_{9x}$)
(ω$_{16x}$, ω$_{9y}$) (ω$_{16y}$, ω$_{9x}$)
(ω$_{17x}$, ω$_{9y}$) (ω$_{17y}$, ω$_{9x}$)
(ω$_{20x}$, ω$_{9y}$) (ω$_{20y}$, ω$_{9x}$)
(ω$_{19x}$, ω$_{7y}$) (ω$_{19y}$, ω$_{7x}$)
Ω$_b$(EP$_{TOCSY}$,P$_{TOCSY}$,T$_c$)
Verified TOCSY Peaks
(ω$_{16x}$, ω$_{9y}$) (ω$_{15y}$, ω$_{9x}$)

Figure 6. Example to illustrate how Ω_k works and overlap verified or rejected.

produce the frequency combination

$$C = \{(\omega_{ix},\omega_{jx}),(\omega_{iy},\omega_{jx}),(\omega_{ix},\omega_{jy}),(\omega_{iy},\omega_{jy})\}$$

on the basis of C, theoretically possible peaks from τ type experiments can be predicted and stored in P_τ. Operator $\mathcal{R}$ is illustrated by Figure 5.

Figure 5 g_k can be g_a or g_b of Figure 4; the frequency pairs with the dotted line mean that it is certain that these two frequencies belong to the same spin system, but, it is uncertain if they couple with each other. In Relay COSY, e.g., (ω_{2x},ω_x) could imply spin ω_{2x} couples with spin ω_x, or a spin topology ω_{2x}–ω_{3y}–ω_x exists. For MQ experiment, there are more spin topologies which can be predicted, and not all are enumerated in Figure 5.

$\Omega(\mathrm{EP}_\tau,P_\tau,T_c)$ searches the various experimental data sets EP_τ within a given tolerance T_c for peaks that are predicted by $\mathcal{R}$. Successful searches enables $g(i)$ to be correctly partitioned. In addition, $\Omega(\mathrm{EP}_\tau,P_\tau,T_c)$ records the origin of experimental peaks found. $\Omega(\mathrm{EP}_\tau,P_\tau,T_c)$ will return value "true" if it finds that $g(i)$ can be partitioned to g_k; otherwise it returns "false".

Therefore, if $\Omega_a(\mathrm{EP}_\tau,P_\tau,T_c)$ is "true", then peak $g(i)$ is assigned to g_a; if $\Omega_b(\mathrm{EP}_\tau,P_\tau,T_c)$ is "true", then peak $g(i)$ is assigned to g_b; if both $\Omega_a(\mathrm{EP}_\tau,P_\tau,T_c)$ and $\Omega_b(\mathrm{EP}_\tau,P_\tau,T_c)$ are "true", then peak $g(i)$ is assigned to both g_a and g_b. This processing is illustrated by an example in Figure 6, referring to Figure 4. The conclusion drawn is that the fragment g, intersection of g_a and g_b, belongs to g_a rather than g_b. The node 9 of g, however, is partitioned to both g_a and g_b, because it is an overlap peak.

Prediction for TOCSY, Relay COSY, and MQF-COSY cross-peaks based upon the predicted spin topologies in Figure 5 is straight forward. The method to predict MQ experimental peaks is illustrated in Table II.

Table II. Prediction for MQ Experimental Peaks

		MQ peak	
predicted topology	quantum order[a]	MQ domain[b]	SQ domain
$i-j$	2	$2 \times C$–ω_i–ω_j	ω_i
	2	$2 \times C$–ω_i–ω_j	ω_j
$k-i-j$	2	$2 \times C$–ω_i–ω_j	ω_i
	2	$2 \times C$–ω_i–ω_k	ω_i
	2	$2 \times C$–ω_i–ω_j	ω_j
	2	$2 \times C$–ω_k–ω_j	ω_j
	2	$2 \times C$–ω_i–ω_k	ω_k
	2	$2 \times C$–ω_j–ω_k	ω_k
	2	$2 \times C$–ω_j–ω_k	ω_i
$k-i-j$	3	$3 \times C$–ω_i–ω_j–ω_k	ω_i

[a] Quantum order ≥ 4 cases are not listed here. [b] C: transmitter offset.

In MQF-COSY spectra, a peak observed in higher quantum order, e.g., 4QF-COSY, theoretically should also be observed in 2QF- and 3QF-COSY spectra. For a number of reasons, this does not always happen. To get the complete partitions, 2QF-, 3QF-, and 4QF-COSY spectral cross-peaks are combined and then partitioned. Since some cross-peaks may appear in 2QF-, 3QF-, and 4QF-COSY spectra more than once, the data are often duplicated. This, however, is not redundant information, because the same peak has different consequences when it appears in the different quantum order COSY spectra, providing important structural elucidation data. For example, if peak (ω_i,ω_j) is observed in 2QF-COSY spectra, it means that spin ω_i couples with ω_j, and both of them should concurrently couple with at least one or more other spin. This duplicity slows down the performance of the

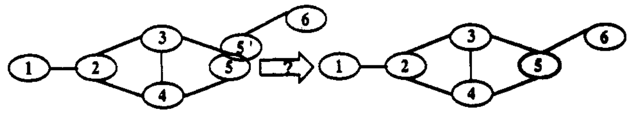


Figure 7. Incompleteness of constraint data.

algorithm. To speed up the algorithm, the duplicated peaks are averaged and treated as a single node. We call such a peak a super peak. Their average values are used as pointers for walking. When the walking procedures finish, the super peaks are recovered by "unaveraging" the super peaks. Hence, every partitioned peak still keeps its quantum order for the further interpretation.

As mentioned above, CPA partitions MQF-COSY spectral cross-peaks using TOCSY, Relay COSY, and MQ, as well as MQF-COSY, as constraints. Rarely, however, are these complete. Again, take the spin topology in Figure 1, which, theoretically, gives 21 TOCSY cross-peaks and 17 Relay COSY cross-peaks (peaks (α,β) and (β,α) are considered as one cross-peak). Thirteen of the TOCSY or Relay COSY cross-peaks give redundant information because they can also be observed from 2QF-COSY, which are listed in Table I. From Relay COSY, there are only four cross-peaks (α,γ), (α,γ'), (β,δ), and (β,δ'), which give information that is different from the other 13 cross-peaks. For TOCSY, there are 8 cross-peaks (α,γ), (α,γ'), (α,δ), (α,δ'), (β,δ), (β,δ'), (β',δ), and (β',δ'), which give new spin coupling information. If all possible cross-peaks are present, CPA will produce all correct partitions. Usually, these constraint data sets are incomplete. Consider an example shown in Figure 7.

In the left spin topology, frequencies 1–5 are assumed to be correctly partitioned into a spin system. Consider another MQF-COSY cross-peak (5′,6), where frequency 5′ and frequency 5 are within a given tolerance, but, at this stage, the assignment or the right hand side of Figure 7 cannot be made. The conditions of this assignment are as follows:

(1) unconstrained partition

$$\mathrm{abs}(\omega_5 - \omega_{5'}) \leq \text{tolerance} \qquad \text{(P0)}$$

(2) local spin coupling constrained partition using TOCSY/Relay-COSY

$$((\omega_3,\omega_6)_{\mathrm{TOCSY}} \wedge (\omega_4,\omega_6)_{\mathrm{TOCSY}})) \vee ((\omega_3,\omega_6)_{\mathrm{Relay\ COSY}} \wedge (\omega_4,\omega_6)_{\mathrm{Relay\ COSY}})) \qquad \text{(P7)}$$

(3) local spin coupling constrained partition using 2Q

$$((\omega_3+\omega_6,\omega_6)_{2Q} \wedge (\omega_4+\omega_6,\omega_6)_{2Q}) \vee ((\omega_3+\omega_6,\omega_3)_{2Q} \wedge (\omega_4+\omega_6,\omega_4)_{2Q}) \qquad \text{(P8)}$$

(4) local spin coupling constrained partition using 2Q

$$(\omega_3+\omega_6,\omega_5)_{2Q} \wedge (\omega_4+\omega_6,\omega_5)_{2Q} \qquad \text{(P9)}$$

(5) local spin coupling constrained partition using 3Q

$$(\omega_3+\omega_5+\omega_6,\omega_6)_{3Q} \wedge (\omega_3+\omega_5+\omega_6,\omega_6)_{3Q} \qquad \text{(P10)}$$

(6) long distance spin coupling constrained partition

$$(\omega_3,\omega_6)_{\mathrm{TOCSY}} \wedge (\omega_4,\omega_6)_{\mathrm{TOCSY}} \wedge (\omega_1,\omega_6)_{\mathrm{TOCSY}} \wedge (\omega_2,\omega_6)_{\mathrm{TOCSY}} \qquad \text{(P11)}$$

If all predicted cross-peaks are present, the constraint data set is complete, and the following constraint can be applied in CPA

$$\text{P0} \wedge \text{P7} \wedge \text{P8} \wedge \text{P9} \wedge \text{P10} \wedge \text{P11} \qquad \text{(C8)}$$

All overlaping cross peaks then disappear.

In practice, few spin systems can meet the needs of constraint C8 due to the incompleteness of constrained data. Take lysine 21 of the melittin protein as an example; its basic spin topology contains 9 frequencies and 12 edges (see Figure 10), and the maximum possible number of TOCSY peaks is 36, while only 25 of them are observed. Moreover, 11 of the cross-peaks give information similar to the 2QF-COSY cross-peaks. In other words, only 14 cross-peaks can be used as constraints for partitioning. The absent cross-peaks identifying the local spin coupling can be partially compensated by other kinds of spectral cross-peaks from MQ, MQF-COSY, Relay-COSY, etc. The long distance spin coupling evidence is presently only available from TOCSY.

Constraint C8 is too strict to be applied in practice. If the overlap is not heavy, only one constraint from (P7)–(P11) is needed, in addition to tolerance

$$\text{P0} \wedge (\text{P7} \vee \text{P8} \vee \text{P9} \vee \text{P10} \vee \text{P11}) \qquad \text{(C9)}$$

For a molecule like a protein, however, overlap is usually very heavy. To process this situation, the combinations of (P8)–(P12), "$\wedge$" and "$\vee$" are set to be varied according to the tolerance. The constraints are classified into local spin coupling constraints and long distance spin coupling constraints. It is not necessary that all the evidence should be connected with logical "and" within local spin coupling constraints or within long distance spin coupling constraints. In contrast, to assign overlap peaks, the local spin coupling constraints and long distance spin coupling constraints must be connected with logical "and". Therefore, following constraints are obtained.

local constraints

$$(\omega_3,\omega_6)_{\mathrm{MQF\text{-}COSY}} \vee (\omega_4,\omega_6)_{\mathrm{MQF\text{-}COSY}} \vee (\omega_1,\omega_6)_{\mathrm{MQF\text{-}COSY}} \vee (\omega_2,\omega_6)_{\mathrm{MQF\text{-}COSY}} \qquad \text{(P7')}$$

$$(\omega_3+\omega_6,\omega_6)_{2Q} \vee (\omega_4+\omega_6,\omega_6)_{2Q} \vee (\omega_3+\omega_6,\omega_3)_{2Q} \vee (\omega_4+\omega_6,\omega_4)_{2Q} \qquad \text{(P8')}$$

$$(\omega_3+\omega_6,\omega_5)_{2Q} \vee (\omega_4+\omega_6,\omega_5)_{2Q} \qquad \text{(P9')}$$

$$(\omega_3+\omega_5+\omega_6,\omega_6)_{3Q} \vee (\omega_3+\omega_5+\omega_6,\omega_6)_{3Q} \qquad \text{(P10')}$$

long distant constraint

$$(\omega_3,\omega_6)_{\mathrm{TOCSY}} \vee (\omega_4,\omega_6)_{\mathrm{TOCSY}} \vee (\omega_1,\omega_6)_{\mathrm{TOCSY}} \vee (\omega_2,\omega_6)_{\mathrm{TOCSY}} \qquad \text{(P11')}$$

The constraint for the partitioning will be

$$P0 \wedge (P7' \vee P8' \vee P9' \vee P10') \wedge P11' \qquad (C10)$$

The protein melittin NMR data set was partitioned using (C10) with tolerance 0.01–0.02 ppm, and good conformity resulted.

REPARTITIONING AND OPTIMIZATION

From the above discussion, it is seen that the constraints criteria are not easy to set up. Overlap peaks have different relations to the constraints. Take the simplest constraint, tolerance, as an example. If it is set to be 0.01, it may be too fine for some overlap peaks, while it can be too coarse for other heavily overlapped peaks. When a constraint is too strict, it will produce too many spin subspaces and single peak subspaces (unpartitioned peaks). If the constraints are not strict enough, it will be at the risk of producing wrong partitioning results, i.e., partitioning different spin systems into the same partitioning subspace. This case is misleading, giving incorrect assignments.

The natural and safe partitioning strategy is to apply CPA repeatedly with different tolerances. At the start, there is heavy overlap. A coarse tolerance cannot distinguish overlap peaks and produces incorrect partitioning results. There are two kinds of tolerances that can be chosen, merging tolerance T_m and comparison tolerance T_c. The merging tolerance is used to merge a new edge, i.e., a MQF-COSY cross-peak, with a spin coupling topology. For example, in Figure 7, T_m is used to determine if edge 5′–6, is partitioned, that is, $\text{abs}(\omega_5 - \omega_{5'}) \le T_m$. Comparison tolerance T_c is used to compare theoretically predicted peaks against experimental peaks. Theoretically, these two tolerances are the same and match the resolution of a NMR spectrometer. In practice, both of the tolerances have a varied range. When both are narrow, correct partitioning results are produced, but many unpartitioned peaks will exist. After the initial partitioning is complete, however, there will be fewer overlap peaks. Hence, CPA can be applied on the unpartitioned data set with larger tolerances, and less strict constraints. This is called repartitioning. With repartitioning, all possible spin topologies should be found within the limitations of the data, and overlap is reduced. One disadvantage of the repartitioning is that a spin topology may still be partitioned into several spin topological fragments; however, these spin topological fragments, if they belong in the same spin system, can be reconnected by further processing with some connection rules.[13]

Some peaks, no matter how the constraints are varied, cannot be partitioned. This is due to data limitations, such as missing cross-peaks. If the constraints are set up to be too coarse, they are partitioned; however, many unreliable partitionings are produced. To deal with this, a global optimum partitioning strategy is introduced. In order to realize this strategy, a parameter A is defined to be the weight of the partitioning of various spins according to the following criterion:

$$A = 1 - \frac{D_m + D_{c1} + D_{c2}}{T_m + T_{c1} + T_{c2}} \qquad (7)$$

For example, suppose there are two MQF-COSY cross-peaks $P_i(\omega_{i1},\omega_{i2})$, $P_j(\omega_{j1},\omega_{j2})$, and a TOCSY cross-peak $P_k(\omega_{k1},\omega_{k2})$, then they are satisfied with following assertion:

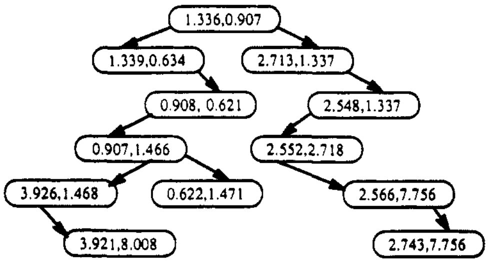


Figure 8. Binary tree of a subspace.

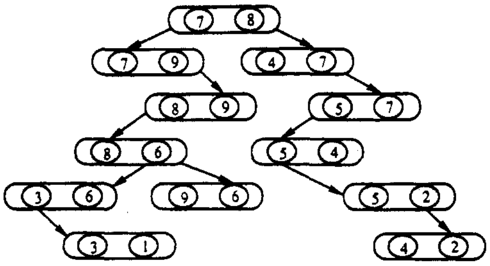


Figure 9. Binary tree following from Figure 8.

$$(\text{abs}(\omega_{i1} - \omega_{j1}) \le T_m) \wedge (\text{abs}(\omega_{i2} - \omega_{k1}) \le T_{c1}) \wedge (\text{abs}(\omega_{j2} - \omega_{k2}) \le T_{c2}) \qquad (8)$$

In this case, $D_m = \text{abs}(\omega_{i1} - \omega_{j1})$, $D_{c1} = \text{abs}(\omega_{i2} - \omega_{k1})$, and $D_{c2} = \text{abs}(\omega_{j2} - \omega_{k2})$. T_m is the tolerance for merging two MQF-COSY cross-peaks; T_{c1} and T_{c2} are for the comparisons in F_1 and F_2 frequency domains, respectively.

If condition 8 is satisfied, then, clearly, $D_m + D_{c1} + D_{c2}$ is always less than $T_m + T_{c1} + T_{c2}$. Therefore, formula 7 has three cases:

(1) $A = 1$ means evidence P_k indicates the partitioning of P_i and P_j should be correct.

(2) $A = 0$ means P_k is not evidence for partitioning P_i and P_j.

(3) $0 < A < 1$, gives a weight to P_k as evidence for partitioning P_i and P_j.

During CPA's partitioning, there can be several pieces of evidence supporting partitioning P_i and P_j. If TOCSY, MQF-COSY, Relay COSY, 2Q, and 3Q are all available as partitioning constraints, the number of theoretical peaks giving supporting evidence can be up to 11. For example, if P_i and P_j are satisfied with (8), then all the possible evidence peaks are as follows:

TOCSY cross-peak	$(\omega_{i2},\omega_{j2})$
Relay COSY cross-peak	$(\omega'_{i2},\omega'_{j2})$
MQF-COSY cross-peak	$(\omega''_{i2},\omega''_{j2})$
	duplicates in different quantum order are considered as one peak
2Q	$(\omega_{j2}+\omega_{i2},\omega_{j2})$, $(\omega_{j2}+\omega_{i2},\omega_{j1})$ $(\omega_{j2}+\omega_{i2},\omega_i)$, $(\omega_{j2}+\omega_i,\omega_{i2})$, $(\omega_i+\omega_{i2},\omega_{j2})$
3Q	$(\omega_i+\omega_{j2}+\omega_{i2},\omega_i)$, $(\omega_{i2}+\omega_{j2}+\omega_i,\omega_{i2})$ $(\omega_{j2}+\omega_i+\omega_{i2},\omega_{j2})$

where $\omega_i = (\omega_{i1}+\omega_{j1})/2$. Therefore, more than one assignment parameter A can exist with an overall weighting of greater than 1; this implies more than one evidence peak is found within given tolerances.

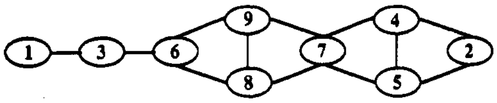


Figure 10. Protein residue spin topology.

Table III. Average Frequencies from the Binary Tree in Figure 8

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 IV. 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			

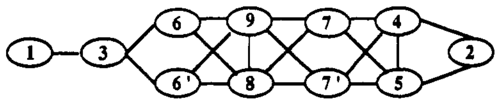


Figure 11. Complete spin topology of lysine 21 of melittin.

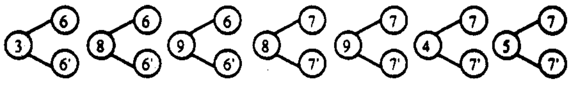


Figure 12. Equivalent spins of coupling topologies.

Under the global optimum partitioning strategy, before putting P_i and P_j together, CPA considers all cross-peaks which are satisfied with assertion 8 and estimates the corresponding assignment parameters. A peak pair which has a maximum value of A is partitioned, while other peak pairs are rejected, even though they may have big assignment parameters. With this strategy, most of the heavy overlap is resolved with the only negative aspect being the increase of CPU time from 10 to 28 min for melittin on the SUN sparc station IPX.

RESULTS AND CONCLUSION

The algorithms and programs have been tested on the two dimensional NMR data set of the protein melittin. Melittin is a major component of the venom of the honey bee, *Apis mellifera*, which is comprised of 26 amino acid residues:[11-13]

^{1}Gly-Ile-Gly-Ala-^{5}Val-Leu-Lys-Val-Leu-^{10}Thr-Thr-
Gly-Leu-Pro-^{15}Ala-Leu-Ile-Ser-Trp-^{20}Ile-Lys-Arg-
Lys-Arg-^{25}Gln-Gln

Resonance assignments of the 500-MHz 5 mM melittin bound to dodecylphosphocholine micelles at 40 °C (pH = 3, accuracy = ±0.01 ppm) have been done manually by Gray.[13] The structure of melittin was also studied with two dimensional NMR and distance geometry calculations by Inagaki.[12] To extract the melittin spin topologies, spectral peaks from 2-, 3-, and 4QF-COSY (in both water and D_2O) are combined in an experimental data set to be partitioned by the algorithm. TOCSY (including in water and in D_2O), 2- and 3Q experiments (in D_2O only), and MQF-COSY spectral peaks serve as the constraints. The tolerance for manual assignments is 0.03 ppm (15 Hz). The tolerance in CPA is 0.03 ppm. The tolerance used in the partitioning and comparison is within the range of 0.01–0.03 ppm. In applying these tolerances, a number of duplicate peaks can be found. These duplicate peaks give redundant information. With a tolerance set to 0.03 ppm, 62 duplicate peaks are found from a total of 189 MQF-COSY cross-peaks (excluding the peaks produced from impurities).

Two such peaks $P_A(\omega_{a1},\omega_{a2})$ and $P_B(\omega_{b1},\omega_{b2})$ should agree with the constraint as follows:

$$\text{duplicate}(P_A,P_B) \Rightarrow (\text{abs}(\omega_{a1}-\omega_{b1}) \le T) \wedge (\text{abs}(\omega_{a2}-\omega_{b2}) \le T) \vee (\text{abs}(\omega_{a1}-\omega_{b2}) \le T) \wedge (\text{abs}(\omega_{a2}-\omega_{b1}) \le T) \quad (9)$$

where T is a given tolerance. A 2QF-COSY cross-peak can have up to four theoretical duplicate peaks. That is, three of them can come from 2-, 3-, and 4-QF-COSY, and one can come from TOCSY peaks identified with MQ experimental peaks. When they are found, it is reasonable that the average frequency be used in the partitioning. Therefore, duplicate peaks and the average frequencies are saved in the following data structure:

```
Redundant struct
{
    redundant_peaks: array of peak_pointer;
    average_frequency_$\omega_1$: real;
    average_frequency_$\omega_2$: real;
};
```

where peak_pointer is a pointer to an element in the MQF-COSY peak list; average_frequency_ω_1, average_frequency_ω_2 will take part in the partitioning.

For the considerations of partitioning and comparing with the manual assignment result, a MQF-COSY peak is defined as follows:

```
MQF_COSY_Peak struct
{
    frequency_$\omega_1$: real;
    frequency_$\omega_2$: real;
    quntum_order: integer;
    partitioned: boolean;//"True" when the peak is
      partitioned, else "False"
    manual_assignment: integer;
};
```

Peaks should also have intensity attributes, but these data are unavailable and were not used in this work.

A MQF-COSY experimental peak set is considered as a spin space which contains all the spin subsystems. The goal of the partitioning algorithm is to divide it into subspaces, so that each subspace uniquely belongs to a real spin system. A spin system may however be split into several subspaces due to missing critical spin coupling linkages (linkage peaks). A subspace may contain an element which belongs to more than one spin systems, particularly when very heavy overlap exists. Therefore, a subspace should not be considered as a spin system at this phase.

The partitioning results produced by CPA are listed in Table V. From the total of 26 residues, CPA assigns 14 residues, which are identical to manual initial assignments; 8 residues are very similar to the manual initial assignments and, when compared to the actual spin topology, are better than the manual assignments. Four residues are very similar to manual initial assignments and, when compared the actual spin topology, worse than the manual ones. Generally, the assignment of CPA is much the same as the manual assignment.

The properties of CPA are discussed as follows:

(1) CPA can extract complicated spin coupling topologies directly from the partitioning results.

Mathematically, a spin topological graph can be represented by a matrix, a connectivity table, a binary tree, and an edge set, etc. As mentioned above, a subspace is an edge set. In the ideal case, a MQF-COSY peak subspace corresponds to a spin system, that is, for proteins, an amino acid. If some

Table V. Comparison of Spin Topologies

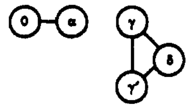
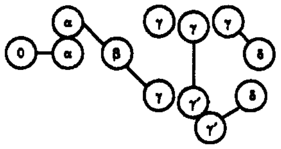
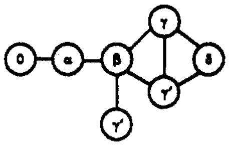
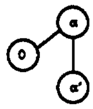
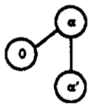
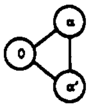
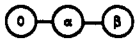
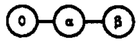
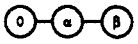
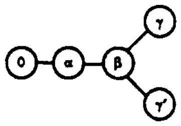
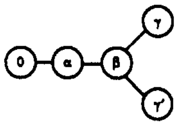
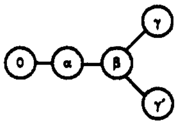
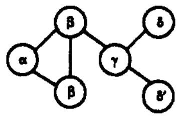
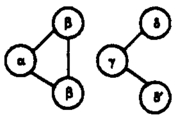
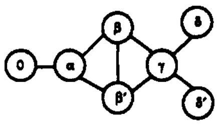
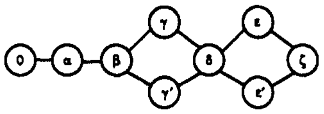
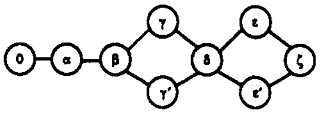
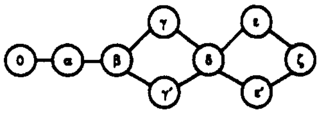
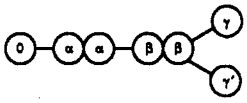
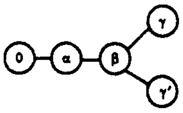
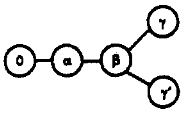
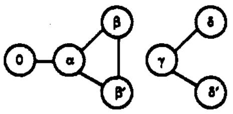
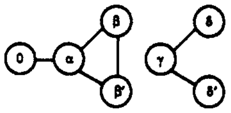
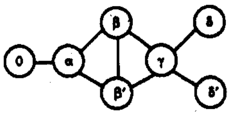
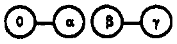
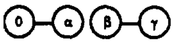
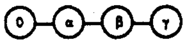
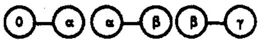
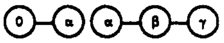
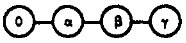
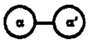
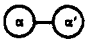
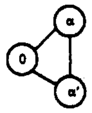
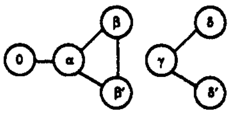
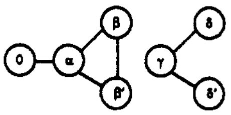
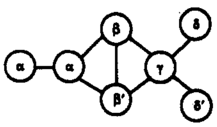
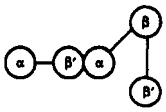
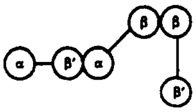
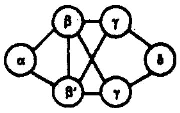
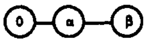
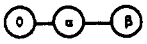
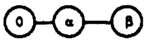
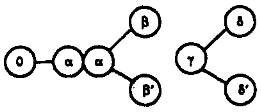
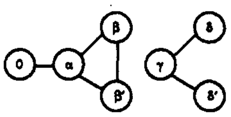
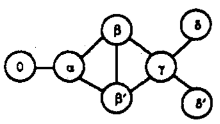
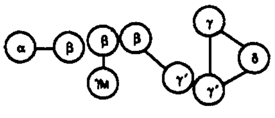
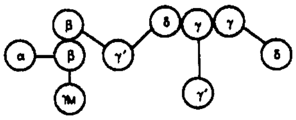
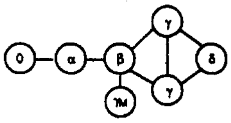
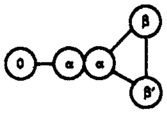
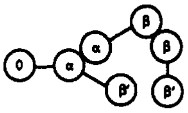
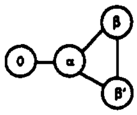
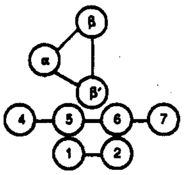
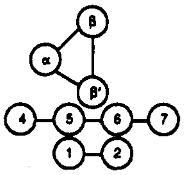
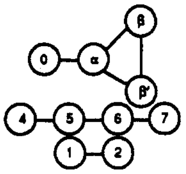
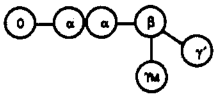
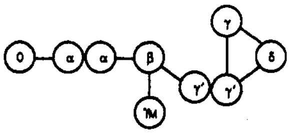
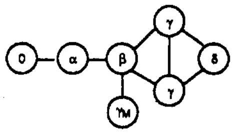
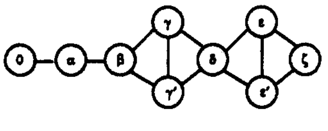
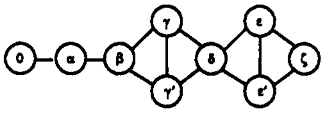
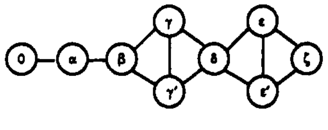
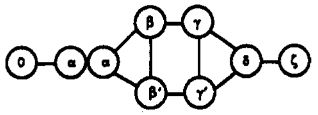
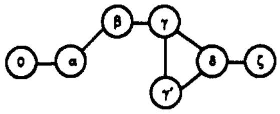
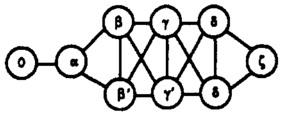
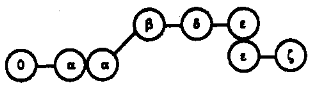
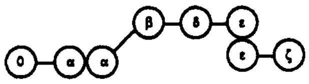
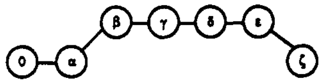
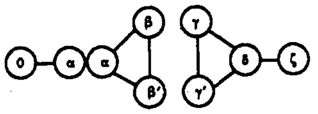
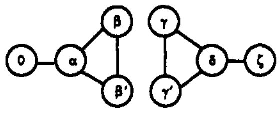
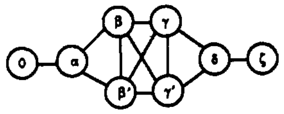
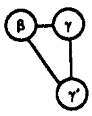
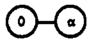
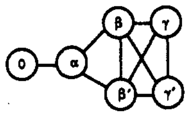
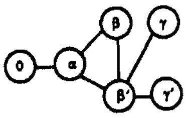
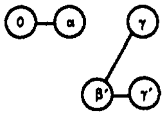
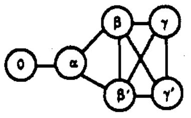
residue	CPA	manual[c]	actual[b]
Gly 1[a]			
Ile 2			
Gly 3			
Ala 4			
Val 5			
Leu 6			
Lys 7			
Val 8			
Leu 9			
Thr 10			
Thr 11			
Gly 12			
Leu 13			
pro 14			
Ala 15			
Leu 16			
Ile 17			

Table V (Continued)

residue	CPA	manual[c]	actual[b]
Ser 18			
Trp 19			
Ile 20			
Lys 21			
Arg 22			
Lys 23			
Arg 24			
Gln 25			
Gln 26			

[a] Both CPA and manual method do not find a peak for assigning Gly 1. [b] Equivalent spin topologies are not considered at this stage. [c] Adopted from MQF-COSY cross-peak manual assignment.[13]

peaks are missing, a subspace may correspond to only a part of a complete spin system of a residue. This is shown by the example in Figures 8–10, using actual data for the residue lysine of melittin. From the partitioning result, there is an actual subspace consisting of 12 peaks (there are also 6 redundant peaks, which are not listed in Figure 8). A binary tree in Figure 8 shows the partitioning result.

These 12 peaks, are all within the tolerance of 0.03 ppm. This is a fuzzy graph with fuzziness of 0.03 ppm. From Figure 8, 9 average frequencies can be reported in Table III.

The real number frequency values are replaced by integer labels in Table III. The fuzzy graph of Figure 8 is transformed into a boolean graph, Figure 9. A fuzzy graph has a tolerance; that is, a boolean graph is not fuzzy.

In turn, Figure 9 is transformed into a connectivity table as shown in Table IV.

It is straight forward to transfer this connectivity of Table IV to a topological graph, as shown in Figure 10.

The spin topology of Figure 10 corresponds to lysine residue (it actually corresponds to lysine 21 of melittin). Therefore, 6 and 7 should correspond to $-CH_2-$ groups; giving the actual spin topology in Figure 11.

The reason a complete spin topology of Figure 11 cannot be constructed directly from CPA is that the spin topological fragments in Figure 12 are not predicted by $\mathcal{R}$ nor checked against MQ experimental data set by Ω_k. These spin topologies include equivalent spins. A method for predicting and detecting this kind of topology is to be reported in later work.

(2) CPA records the experimental origin for assignment of individual partitionings. This is of aid in identifying and following the computer automated assignments.

For protein structural elucidation, it is useful that the deduction traces are known for every step of the spin topology generation. Manual assignments make this combersome and complicated. The CPA can list these deduction steps at every partitioning stage for later reference.

Take the melittin arginine 22 residue as an example, and take two peaks (4.076,1.837) and (1.983,1.836) which are considered to be assigned to the same spin coupling system, since abs(1.836 – 1.837) = 0.001 is within the tolerance. CPA

Table VI. Symbols and Descriptions

symbol	description
abs( )	absolute value of
{ }	denotes a set
ˆ	logical "and"
ˇ	logical "or"
ω	chemical shift
$\cup$	operator union
$\cap$	operator intersection
ϕ	denotes an empty set
$\in$	is an element of
$g \supset e$	*e* belongs to graph *g*
$\Longrightarrow$	imply
T_m	tolerance for merging two peaks
T_c	tolerance for comparison of predicted and experimental peaks
$\mathcal{R}\ (\tau,g_k,g(i), T_m)$	predict τ type theoretical peaks based upon peak g_k and $g(i)$, i.e., arbitrarily choose a peak (ω_x,ω_y) from g_k, then combine them into (ω_{ix},ω_y), (ω_{ix},ω_x), (ω_{iy},ω_x), and (ω_{iy},ω_y); from then on, predict τ type theoretical peaks
$\Omega_k(EP_\tau,P_\tau,T_c)$	compare peaks predicted by $\mathcal{R}$ against P_τ with tolerance *T*; if matched experimental peaks are found, then return "true"; else return "false"

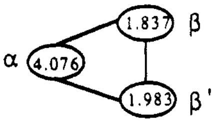


Figure 13. Melittin arginine 22 residue spin topological fragment.

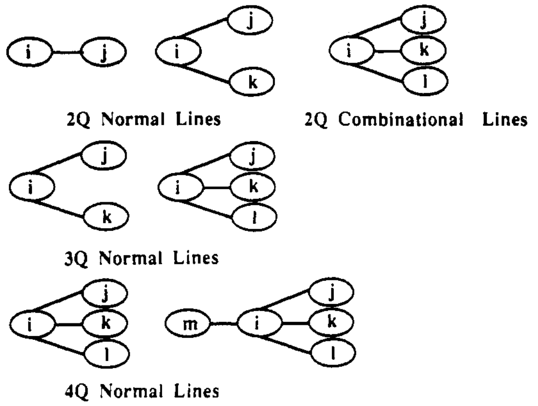


Figure 14. Some possible spin coupling topological fragments which can be identified from MQ spectra.

will search for more evidence to merge these two peaks. After partitioning, it lists the following data:

TOCSY	(1.972,4.092)
MQF-COSY	(1.993,4.018)
2Q	(–0.796,4.084) $\Leftarrow$ ([2×2.621–4.076–1.983],4.076)
3Q	(–0.021,4.076) $\Leftarrow$ ([3×2.621–4.076–1.983–1.837],4.076)

The conclusion is that the spin topological fragment in Figure 13 exists.

(3) CPA can determine equivalent spins, combinatorial lines, and more complex spin topologies.

Some possible spin coupling topological fragments which can be identified from MQ experimental spectra are shown in Figure 14.

Of these, the present version of CPA only uses 2Q normal lines and 3Q DDM[11] lines. That is, the prediction and comparison are only based on coupling topologies, *i–j*, and *j–i–k*. On the other hand, *j*, *k*, and *l* can be equivalent spins. Identifying equivalent spins, combinational lines, and more complex spin topologies involves other prediction methods[14] which are not considered here. These features can be included in CPA and will improve the algorithm.

ACKNOWLEDGMENT

The authors thank Dr. Larry R. Brown and Dr. Barry N. Gray for many helpful discussions. We also gratefully acknowledge Dr. Barry N. Gray for providing his melittin manual assignments used here for comparison with computed CPA results.

REFERENCES AND NOTES

(1) James, Thomas L.; Basus, Vladimir J. *Annu. Rev. Phys. Chem.* **1991**, *42*, 501.
(2) Cieslar, C.; Clore, G. M.; Gronenborn, A. M. *J. Magn. Reson.* **1988**, *80*, 119.
(3) Kleywegt, G. J.; Boelens, R.; Kaptein, R. *J. Magn. Reson.* **1990**, *88*, 601.
(4) Weber, P. L.; Malikayil, J. A.; Muller, L. *J. Magn. Reson.* **1989**, *82*, 419.
(5) Eads, C. D.; Kuntz, I. D. *J. Magn. Reson.* **1989**, *82*, 467.
(6) Kleywegt, G. J.; Boelens, R.; Cox, M.; Llinas, M.; Kaptein, R. *J. Biomol. NMR* **1991**, *1*, 23.
(7) Wüthrich, K. *NMR of Protein and Nucleic Acids*; Wiley: New York, 1986.
(8) Balaban, A. T. *Chemical Applications of Graph Theory*; Academic Press: New York, 1976.
(9) Harary, F. *Graph Theory*; Addison-Wesley: Reading MA, 1972.
(10) Even, S. *Graph Algorithms*; Computer Science Press: 1979.
(11) Gray, B. N.; Brown, L. R. *J. Magn. Reson.* **1991**, *95*, 320.
(12) Inagaki, F.; Shimada, I.; Kawaguchi, K.; et al. *Biochemistry* **1989**, *28*, 5985.
(13) Gray, B. N. Doctoral Thesis, The Australian National University, 1991.
(14) Xu, J.; Gray, B. N.; Brown, L. R. *Trahedron Comput. Methodol.*, submitted for publication.