Rigorous Deduction Theory for Assignment of Multidimensional NMR Spectra Using the Independent Spin Coupling Network Approach†

Jun Xu* and Philip N. Borer‡

TRIPOS Associates, Inc., 1699 South Hanley Road, Suite 303, St. Louis, Missouri 63144

Received May 18, 1993•

The concept of an ISNet (independent spin network), used to process various molecular connectivities which can be observed from multidimensional NMR spectra, is proposed. The conventional one-dimensional spectra-based computer-assisted structural elucidation (CASE) methods are based upon the correlation of subspectra-substructure which is often ambiguous. However, modern multidimensional NMR experiments offer us abundant molecular spin coupling connectivity information. This information can be directly or indirectly mapped to real chemical structure or substructure. This fact is known as "the correlation of spectra patterns and structures", where pattern implies connectivity. ISNets are the general graph theoretical representation for these patterns. By means of systematic graph theory and fuzzy mathematical analysis, a rigorous deduction theory for general purpose structural elucidation from multiple-dimensional NMR spectra is presented.

INTRODUCTION

Computer-assisted structure elucidation (CASE) can be divided into computer-assisted primary structure elucidation (CAPSE) and computer-assisted tertiary structure elucidation (CATSE). CAPSE is for studying the primary structures of unknown or partially known organic compounds, and CATSE is for the determination of the conformations of biomolecular structures of which the primary structures are normally known beforehand.

CASE began with the Stanford DENDRAL project in the late 1960s and has been studied for more than 20 years.[1] A number of CASE systems have been implemented. Unfortunately the original goals of the DENDRAL project were not completely achieved. The main reason is that the conventional structure elucidation theory was based upon the following principle:

> There is subspectra-substructure correlation in conventional 1D NMR, IR, and MS. By systematically analyzing subspectra, a group of substructures for an unknown structure can be obtained by means of computational methods and some chemical, mathematical, and other spectral constraints. A set of candidate structures is assembled, and the true structure is selected from these candidates.

In the ideal case, the above principle is true. However, in practice, spectral interpretation becomes difficult since in conventional 1D NMR signal splitting is not sufficiently resolved enough to determine the multiplicities; it is difficult to distinguish a doublet (CH) from a quartet (CH_3) signal when the signal/noise ratio is low; when many signals exist in a narrow chemical shift range, there is considerable signal overlap which obscures any identification of multiplets; some resonance values are far from their statistically expected values, which can lead to mistakes, and, as a result, the whole interpretation becomes false because of incorrect local subspectral interpretations.

The key point is that a chemical structure can be represented as structural components and their connectivities. The former correspond to subspectra which can be observed from conventional spectra. The latter cannot be always observed from conventional spectra. Due to the limitations of the instrumentation during the course of the DENDRAL project, the structural connectivities could not be deduced and hence the objectives of the project could not be met. For example, to elucidate the structure of the relatively simple compound warburganal (Figure 1A), the DENDRAL programs used 1D ^1H NMR, ^{13}C NMR and mass spectral data to deduce 12 substructure components. With use of an assembly algorithm, 42 structural candidates were produced! For more complicated molecules, the DENDRAL programs would face "combinatorial explosion" in the assembly procedure as well as problems in surveying and verifying the structural candidates.[1]

In the years since the DENDRAL project, multidimensional NMR experiments have been developed which provide many

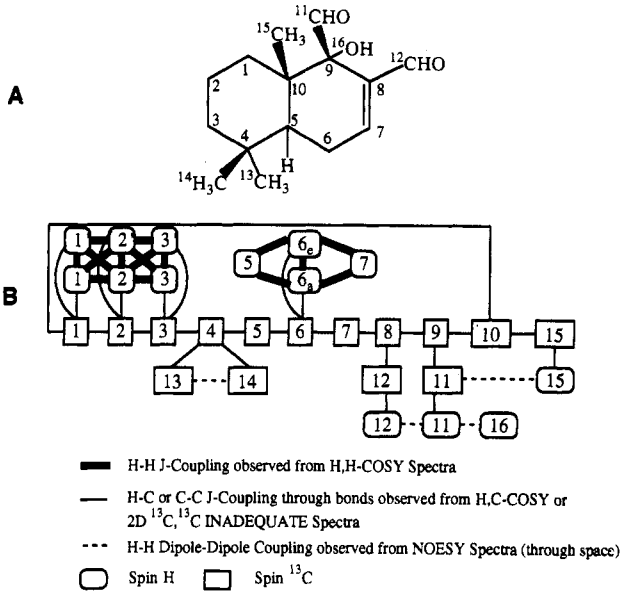


Figure 1. (A) Structure of warburganal. (B) Spin coupling network for warburganal in which each edge can be theoretically observed from multidimensional NMR experiments. In some cases, long-distance H–H–COSY correlations may be observed for specific configurations, for example, between H1 and H5 in the chair configuration of warburganal.

† Presented at The 5th International Conference on Mathematical and Computational Chemistry as a lecture.

‡ Current address: Department of Chemistry, Syracuse University, Syracuse, NY 13244-4100.

• Abstract published in *Advance ACS Abstracts*, February 15, 1994.

0095-2338/94/1634-0349$04.50/0

different kinds of structural connectivity information. We propose a structure elucidation principle that should realize the goals of the DENDRAL project. This principle is based on the correlation of spectral patterns and structural connectivity. Using this methodology, most of the experimental difficulties encountered by the DENDRAL group, such as overlap, ambiguities, and data incompleteness, can be overcome. On the basis of a similar principle, biochemists have determined a number of protein three-dimensional structures in solution.[2] This paper begins to systematically explore a structure elucidation theory to make CASE become rigorously logical, mathematical, and routine in both CAPSE and CATSE applications. From Figure 1, it is easy to see that a structure can be defined as a spin coupling network (or a set of spin coupling networks), which can be extracted from multidimensional NMR experiments, and that a direct relationship exists with the molecular structure.

The basis for this structure elucidation methodology and some of its basic principles are outlined below:[3-6]

(1) Experimental data are often incomplete. This can be compensated by combining different kinds of experiments.

(2) Spectral peaks are often overlapped. They can be resolved by using higher dimensional experiments or heteronuclear experiments.

(3) A given substructure has a cluster center which includes not only a group of expected chemical shift values but also their connectivities. This cluster center is a fuzzy graph.

(4) By using a fuzzy graph pattern recognition algorithm, subspectral components and their connectivities can be mapped to a substructure even when some subspectral components have large deviations from their average values.

(5) By using different types of correlation spectroscopy (for example, TOCSY and/or NOESY), substructures can be integrated and a complete structure is deduced.

The key to a practical methodology is that the relations of subspectral pattern and substructure should be well understood and cast into a logical system. Moreover, this methodology should easily be extended to other complex compounds, such as biopolymers (polysaccharides, nucleic acids), drugs, etc.

Modern multidimensional NMR offers us so many useful and sophisticated experiments. However, some chemists are deterred from using them because of the difficulties involved in extracting a structure from a spectrum. Even an NMR expert cannot interpret all routinely used multidimensional spectra. Therefore, it is significant and important to develop a new generation of software for a computer-aided elucidation package to aid chemists in studying the structures of known, partially known, and unknown compounds.

CHROMATIC GRAPH REPRESENTATION OF AN ISNET

(1) ISNet. The concept of an ISNet (independent spin coupling topological network) is an extension of conventional spin coupling representations such as AX, AMX, AX_2, etc.[2] Such a conventional spin coupling system is a part of an ISNet or a fragment of an ISNet. Conventional spin coupling systems can be observed from 1D NMR spectra and correspond to a substructure. However, an ISNet can also be observed from multidimensional NMR spectra and often contains a group of conventional spin coupling systems and their connectivities. Normally there is a clear and unambiguous mapping from an ISNet to a real structure. Figure 2 shows the tryptophan structure and its three ISNets. The ideal J_{H-H}-coupling networks are also shown. The amide protons are labile and observed by NMR in aqueous solution under certain conditions.

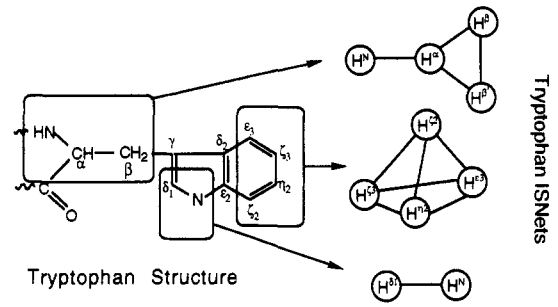


Figure 2. Tryptophan and its ISNets.

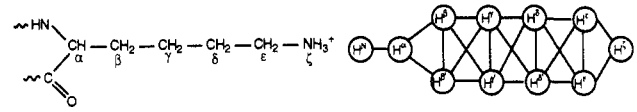


Figure 3. Lysine and its ideal proton homo-ISNet. Each edge may be observed from H,H-COSY spectra. An experimentally observed lysine homo-ISNet is a subgraph of this ideal ISNet.

Couplings H^N–H^α and $H^{\delta 1}$–H^N are not observed if D_2O is used as a solvent. Due to changes of the chemical structure in given environments, the tryptophan residue may give degenerate ISNets when it appears in a different protein or in a different position in the same protein. These degenerate ISNets are still subgraphs of tryptophan's ideal ISNets. As a comparison, tryptophan's J-coupling systems are conventionally represented as AMX + A(X)MP + A.[2] ISNet is not only an alternative graph theoretical representation of this conventional J-coupling systems' notation but is a more general methodology for representing structural connectivities, and maps multidimensional NMR spectra to real chemical structures.

(2) Chromatic Graph. A chromatic graph, CG, is defined as follows:[7]

$$CG = \{N, E, c_n, c_e\} \tag{1}$$

$$N = \{n_1, n_2, ..., n_k\} \tag{2}$$

$$E = \{e_1, e_2, ..., e_m\} \tag{3}$$

$$E \in N \times N \tag{4}$$

where N and E are a node set and an edge set, respectively, and c_n and c_e are a node domain and an edge domain, respectively. For example,

$$c_n = \{\text{spin-}^1\text{H}, \text{spin-}^{13}\text{C}, ...\} \tag{5}$$

$$c_e = \{J_{H-H}, J_{H-C}, D_{H-H}, D_{H-C}, ...\} \tag{6}$$

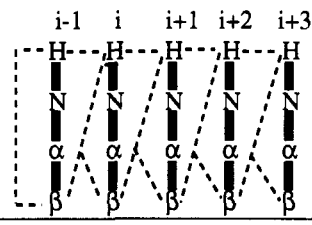
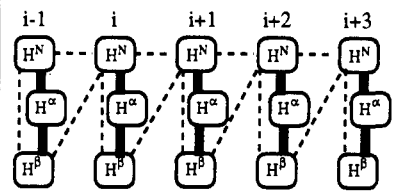
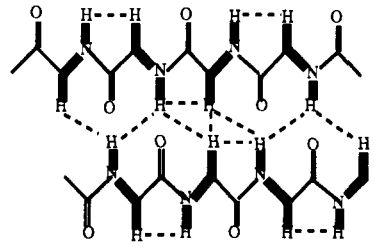
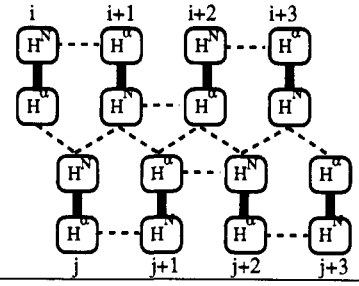
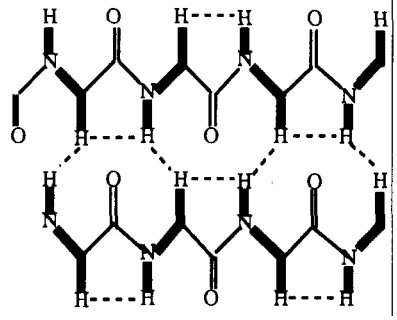
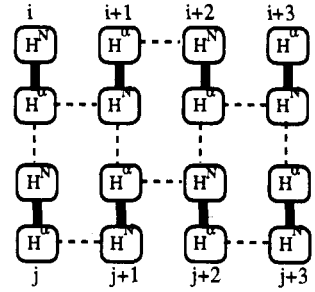
where J denotes a through-bond coupling and D denotes a through-space coupling (dipole–dipole coupling). If a graph G satisfies the above definitions, then G is an ISNet.

(3) Types of ISNets. There are different classes of ISNets according to coupling types, structural features, and experimental types.

(a) Homo-ISNet. All spin nodes in a homo-ISNet come from data extracted from homonuclear NMR spectra. An example is illustrated in Figure 3.[6]

(b) Hetero-ISNet. In this type of ISNet, each edge will represent the coupling between heteronuclei (for example, ^{13}C–^1H coupling). Hetero-ISNets can be observed from ^1H-detected heteronuclear multiple-quantum coherence via direct coupling (HMQC)[8] and C,H-COSY spectra. An example is shown in Figure 4C.[9]

Table 1. MCD Patterns and J–D–ISNets

secondary structure	MCD pattern[a]	J–D–ISNet pattern[b]
Helix		
Antiparallel Sheet		
Parallel Sheet		

[a] Adapted from ref 10. [b] Heavy lines: through-bond J-coupled connectivities. Dotted lines: through-space couplings.

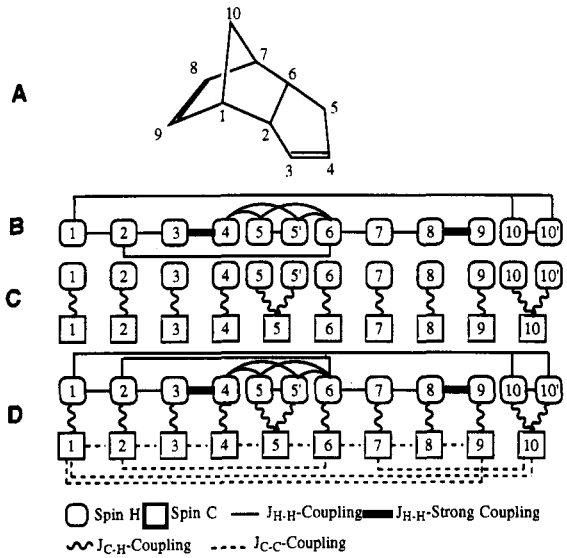


Figure 4. (A) Dicyclopentadiene structure. (B) Homo-ISNet. (C) Hetero-ISNets. (D) Combined J–ISNet.

(c) Combined J–ISNet. Because of structural features, a compound may have several smaller homo-ISNets and hetero-ISNets. This often makes it difficult to determine the structural backbone. If however, these homo-ISNets and hetero-ISNets are combined, we may have a strongly connected graph. This type of graph is a combined-ISNet. The mapping between a combined-ISNet and a real structure is usually direct (Figure 4D).[9]

(d) Combined J–D–ISNet. An example of this type of ISNet is in MCD (main chain directed) strategy developed by Stefano, Englander, and Wand[10,11] for the assignment of ^{1}H NMR spectra of proteins. This pattern contains both J_{H-H}-couplings and D_{H-H}-couplings. According to this methodology, the MCD patterns classify protein standard secondary structures as a helix, an antiparallel sheet, or a parallel sheet. These patterns compose three different combined J–D–ISNets and are shown in Table 1.

(e) Edge-Colored ISNet. An complete chromatic ISNet should include J-coupling constants. However, in many cases, J-coupling constants are very similar and hence indistinguishable and some J-coupling may be too weak to measure. The value of the J-coupling does not depend simply on the number of bonds between the nuclei but also on the details of the electronic distribution, especially the dihedral angle ϕ by which two neighboring J-coupling groups are twisted with respect to one another. Therefore, under unfavorable conditions the J-coupling between two spins may vanish. The Karplus equation[12] is often used in NMR CATSE applications. For example the equation to calculate J-coupling constants of the backbone of a protein is[13]

$$^3J(H^N,H^\alpha) = 6.7\cos^2(\phi - 60) - 1.3\cos(\phi - 60) + 1.5 \quad (7)$$

J-coupling constants play an important role in determining three-dimensional structures of proteins. It will be useful to study edge-colored (J-coupling constants included) ISNets. Figure 5 shows an edge-colored ISNet and a partial edge-colored ISNet.[14]

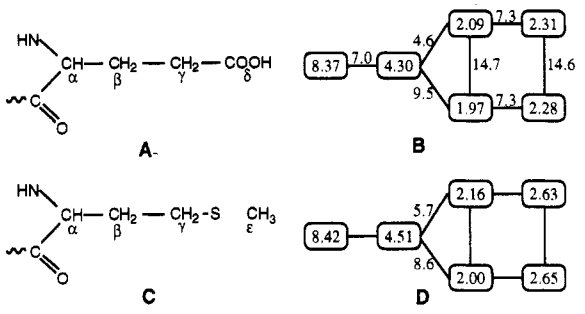


Figure 5. (A) Glu structure. (B) Edge-colored H–H ISNet for Glu. (C) Met Structure. (D) Partial edge-colored H–H ISNet for Met.

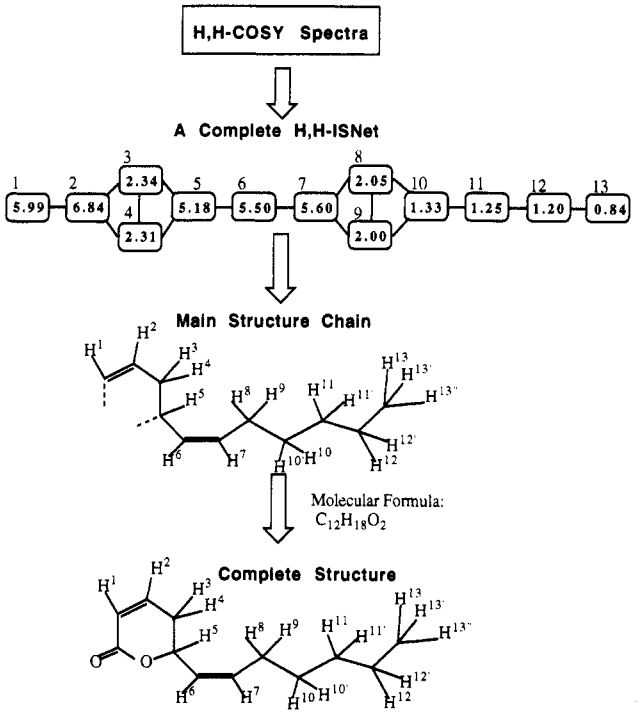


Figure 6. Direct mapping from a complete ISNet to the chemical structure.

(f) Complete ISNet. Given certain structural features, some compounds can have unique and complete homo-ISNets. If this type of ISNet is observed from 2D NMR spectra, its corresponding structural backbone can be easily determined. In Figure 6, a complete ISNet is illustrated. If data incompleteness is not taken into account, most protein amino acid residues have their own complete ISNets (Phe, Tyr, Trp, Met, Gln, and His have more than one ISNet; see Appendix A in ref 6). This fact allows us to make protein sequence-specific assignments. However, because of the separations of branched chains (for example, methyl groups), most of natural organic compounds will have more than one incomplete ISNets.

(g) Connectable or Unconnectable ISNet. In a computer-assisted structure elucidation procedure, determination of the molecular structural backbone is very important. For compounds with more than one ISNet, there are many ways of connecting the networks generated by using multidimensional NMR data from experiments such as NOESY, HMQC, H,C-COSY, 2D INADEQUATE (although higher concentration is needed), 2D heteronuclear NOE experiments,[15] and homo/hetero 2D/3D/4D NMR experiments[16–18]. As shown in Figure 4, H,C-COSY gives rise to many ISNets. When they are combined with COSY data, however, the combined ISNet is a connected graph, from which the structural backbone can be determined. Because of structural features, some broken

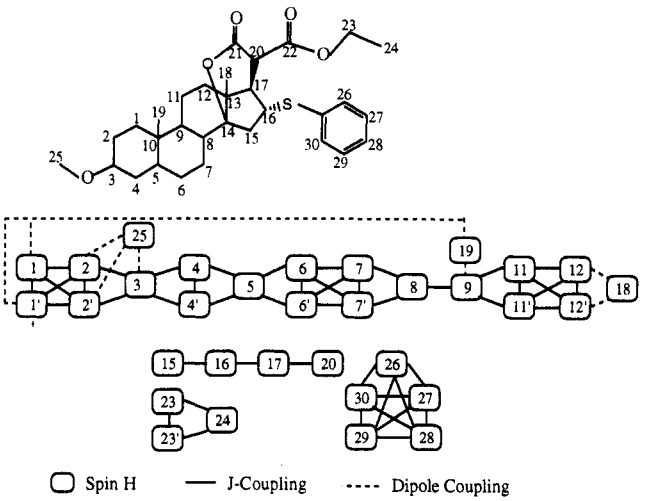


Figure 7. A steroid with seven H,H-ISNets. Four of them can be connected by dipole-coupling information. Others cannot be connected by dipole-coupling, but may be connected by heteronuclear NMR spectral information. The aromatic group of this structure is unconnectable. ISNets given in this figure are theoretical and include all possible experimental ISNets.

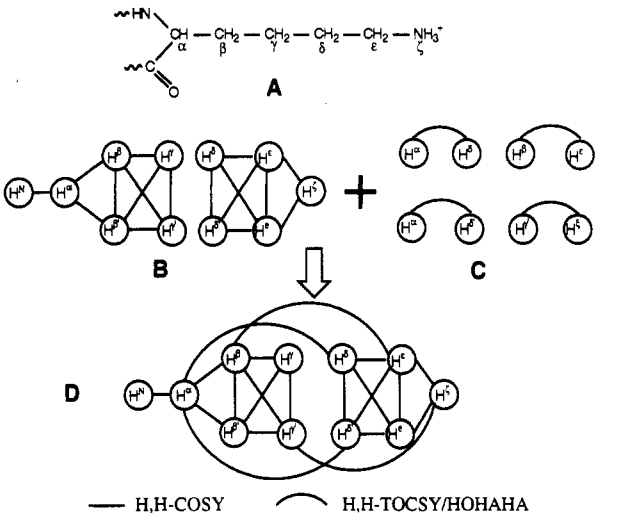


Figure 8. (A) Lysine structure. (B) Fragmented ISNets. (C) Long-distance coupling connectivities from TOCSY/HOHAHA. (D) Reconnected ISNet.

ISNets can be connected by another type of NMR experiment. An example is illustrated in Figure 7.

(4) Reconnection of Broken ISNets. In practice, peak sets are often incomplete because of phase cancellation, overlap, degeneracy, low intensity, etc. This case introduces difficulties into the structure elucidation procedure, since a group of unconnected ISNets may be observed for a given compound, even though its theoretical ISNet should be complete. Fortunately, there are a number of multidimensional NMR experiments that can be used to reconnect these broken ISNets by long-distance spin coupling information. Some examples are TOCSY (total correlation spectroscopy)/HOHAHA (homonuclear Hartmann–Hahn experiment), relay-COSY, HMBC (heteronuclear multiple bond correlation). These spectra have been used as constraints to partition 2QF-COSY spectral cross-peaks and used, for example, to form a set of ISNets from a heavily overlapped protein NMR spectrum.[5] For example, lysine and arginine residues have long spin coupling chains, which are often broken in the middle. By using long-distance coupling information from TOCSY, these long spin coupling chain can be connected (Figure 8).

(5) Integration of a Molecule with Multiple ISNets. Because the ISNet of a molecule may be broken by branched chains (such as, quarternary carbons), and certain atoms are NMR "silent" (such as O, S atoms), an integration procedure is needed to connect separated ISNets by combining information from HMQC, H,C-COSY, 2D INADEQUATE, or NOESY/ROESY spectra. In most situations, the separations introduced by branched chains can be reconnected by using heteronuclear NMR spectra and NOESY/ROESY. The separations introduced by NMR silent atoms, however, can only be partially reconnected by NOESY/ROESY. Sometimes, isotope-labeling techniques (which are expensive) or information from IR and MS need to be applied.

ISNET-BASED STRUCTURAL ELUCIDATION STRATEGIES

Structure elucidation from multidimensional NMR spectra for all kinds of organic compounds should involve the following steps:

(1) Baseline Correction and Phasing. Software for these operations are normally provided by spectrometer manufacturers, although most NMR data processing software offers these functions. Baseline correction and phasing operations rely on human expertise.

(2) Spectral Signal Enhancement. The FFT can deform spectra due to initial delays, and truncated FIDs. These difficulties can be overcome by such methods as the maximum entropy method (MEM),[20] linear prediction,[21] and the maximum likelihood method (MLM).[21,22]

(3) Peak-Picking and Bookkeeping. A number of programs have been implemented for this purpose.[21,23-25] Although peak-picking and bookkeeping are not complicated, they are cumbersome and time-consuming and often introduce errors in ISNet analysis for larger biopolymers.

(4) ISNet Extraction and Identification from the NMR Peak Set. This is the key step in computer-assisted structure elucidation. For small molecules, the process may be trivial. For biomacromolecules, however, ISNet extraction and identification can be difficult because of spectral overlap. To solve this problem, a constrained partitioning algorithm (CPA) was proposed. This algorithm has been tested on a number of protein data sets and can be applied to other compounds without any restriction.[5]

(5) Integration of ISNets by Using Additional Information, Such as NOESY/ROESY, IR, MS, 1D NMR, and Chemical Properties.

(6) Conversion from Refined ISNet(s) to a Real Structure.

(7) Spectral Prediction and Simulation[24] Based upon a Hyperstructure. This step will further verify the structure just generated from the spectral data.

(8) Structure Conformation Refinement. There have been a number of methods for this purpose. Such as, the iterative relaxation matrix approach (IRMA),[31] molecular mechanics (MM), molecular graphics, and distance geometry (DG) calculations.

For CATSE, the main task is to make a correct assignment so the three-dimensional molecular structure can be determined from J-coupling constants, chemical exchange/relaxation rate, and the NOESY spectrum. Here the most difficult problems are cross-peak overlap, ISNet overlap, and sequence-specific assignment. The following strategies are, therefore, specifically for protein structure determinations.

Automated Extraction of ISNets from COSY and TOCSY/HOHAHA Spectra. The latter are used to distinguish heavily overlapped regions of COSY spectra. The CPA (constrained partitioning algorithm) is designed for this purpose and can briefly be described as follows:

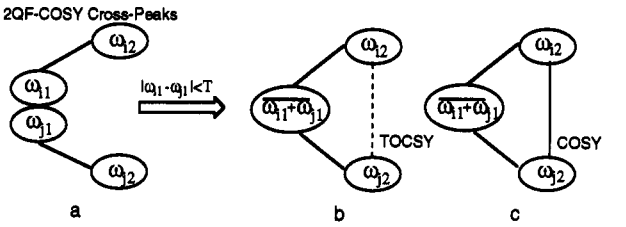


Figure 9. CPA creates a spin coupling system with rigorous deductions.

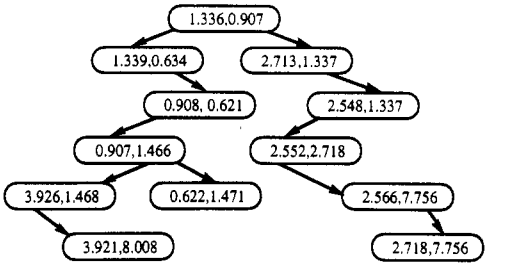


Figure 10. Binary tree representation of a sub-cross-peak set.

Table 2. Sub-Cross-Peak Set for the Protein Melittin

premise		
TOCSY cross-peak	DOF-COSY cross-peak	conclusion
(0.910, 0.632)	(1.336, 0.907)(1.339, 0.634)	(0.907)–(1.337)–(0.634)
(0.910, **2.712**)	(1.336, 0.907)(**2.713**, 1.337)	(**2.713**)–(1.337)–(0.907)
(1.338, 9.10)	(0.634, 1.339)(0.908, 0.621)	(1.339)–(0.634)–(0.908)
(**2.713**, 2.549)	(**2.713**, 1.337)(2.548, 1.337)	(**2.713**)–(1.337)–(2.548)
(0.621, 1.467)	(0.908, 0.621)(0.907, 1.466)	(0.621)–(0.908)–(1.466)
(1.337, **2.719**)	(1.337, 2.548)(2.552, **2.718**)	(1.337)–(2.552)–(**2.718**)
(0.910, 0.623)	(0.907, 1.466)(0.622, 1.471)	(0.907)–(1.471)–(0.622)
(3.923, 0.910)	(0.907, 1.466)(3.926, 1.468)	(3.926)–(1.468)–(0.907)
(8.010, 1.467)	(3.921, 8.008)(3.926, 1.468)	(8.008)–(3.921)–(1.468)
(**2.719**, 7.755)	(2.552, **2.718**)(2.566, 7.756)	(**2.718**)–(2.552)–(7.756)
(2.567, **2.718**)	(**2.718**, 7.756)(2.566, 7.756)	(2.566)–(7.756)–(**2.718**)

Table 3. Average Frequencies generated from the Binary Tree in Figure 10

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

To create an ISNet, at each step CPA searches for TOCSY proof or COSY proof and then adds a new COSY cross-peak to the current ISNet (Figure 9).

Using the above procedure, a COSY cross-peak set is partitioned into many sub-cross-peak sets (Table 2). For ease of illustration the boldface entries are for cross-peaks having one frequency at ≈2.72 ppm.

The above deductions can be represented as a dynamic binary tree (Figure 10), which can be converted into a table, such as the one in Table 3.

Based upon this table, the binary tree of Figure 10 is then converted into a Boolean graph (Figure 11).

This binary tree is equivalent to the ISNet in Figure 12, and the latter can be represented as an adjacency table (Table 4).

Given the influence of chemical environments, the same type of residue may have different sets of chemical shift values. Their connectivities may also vary when this type residue appears in a different protein or the same protein but in different positions. However, any given type of residue has

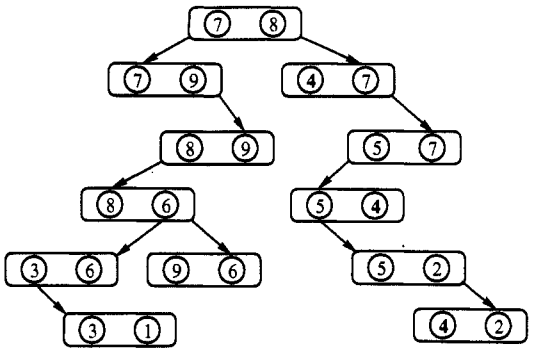


Figure 11. Binary tree representation of the frequency labels.

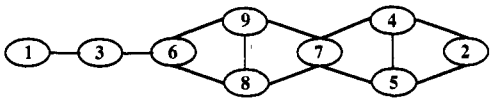


Figure 12. ISNet from Figure 11.

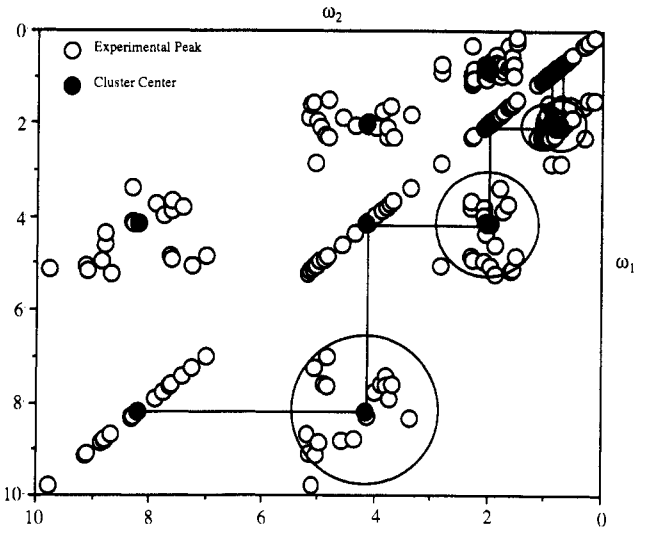


Figure 13. Fuzzy graph of the ISNet of valine residues with circles indicating deviation, $d \approx 1$ standard deviation.

Table 4. Adjacency Table of an ISNet

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

its statistical distribution of chemical shift value sets and their connectivities. This distribution can be represented as a cluster center,[27] which is a fuzzy graph (Figure 13).

Experimental data gathered for twenty valines was used to generate Figure 13. Two from the peptide melittin,[28] seventeen valine data points are from the protein glucose-specific enzyme IIA,[29] one from a peptide of the protein 5-lipoxygenase activating protein (FLAP). The filled-in 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 is 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.

Table 5. Comparison of the Expected Values[a] of the Proline Chemical Shift Distribution and the Experimental Values of 4 Prolines in BPTI

proton	αH	βH	βH′	γH	γH′	δH	δH′
expected	4.48	1.88	2.18	1.92	2.02	3.62	3.77
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 ref 30. [b] Maximum difference = 2.095 ppm.

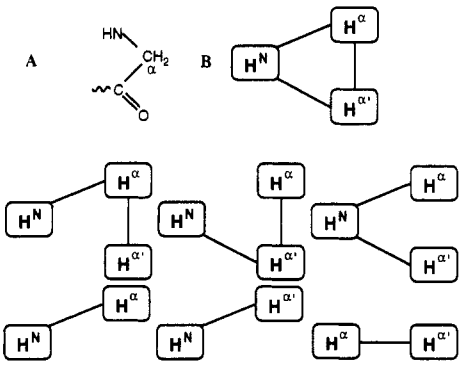


Figure 14. (A) Structure of Gly. (B) Ideal ISNet of Gly. The other graphs are possible experimental ISNet patterns.

From Figure 13, we can also see that some cross-peaks have large deviations from their cluster centers. This often happens in biomolecules and can be clearly seen from Table 5. These large deviations can lead to incorrect ISNet identifications and assignments. However, this difficulty can be overcome because fuzzy graph pattern recognition considers chemical shift values and their connectivities simultaneously.

The ideal ISNets for the twenty amino acids can be generated from theoretical prediction and statistical data.[6] Each ideal ISNet is a knowledge base because it includes all possible experimental ISNet patterns. The experimental patterns are subgraphs of the ideal ISNet for a given type of amino acid. The ISNet of Gly residue is the simplest case. Its ideal ISNet and all possible experimental ISNet patterns are illustrated in Figure 14.

Fuzzy graph pattern recognition algorithm (FPRA) maps experimental ISNets to fuzzy graph cluster center spaces. Each residue of the given protein has a set of experimental ISNets which are assignment candidates. FPRA includes graph matching, membership evaluation for each experimental chemical shift value as it is related to the expected chemical shift values, and graph fuzzy similarity calculation. It also chooses ISNet candidates for each residue (see Figure 15).

Finally, the correct assignment will be made by using a group of NOESY correlations to integrate these ISNets into a connected graph. This job is done by tree search algorithm (TSA). An example search tree based upon Figure 15 is shown in Figure 16. For a large protein, this tree becomes very large. However, when constrained search strategies are applied, the tree can be "well pruned".

Generally, the strategies for protein assignment can also be extended to other biopolymers, and the inclusion of information from 3D, 4D, ... spectra is straight forward.

For CAPSE, in contrast to biopolymer structure elucidation, the following characteristics exist:

(1) Normally synthetic intermediates and natural products are relatively small molecules; hence peak overlap, ISNet

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 15. Peptide NAc-t21a residues and their ISNet candidates which are denoted by numbers on the right. Each number represents an ISNet graph and is ranked by similarity.

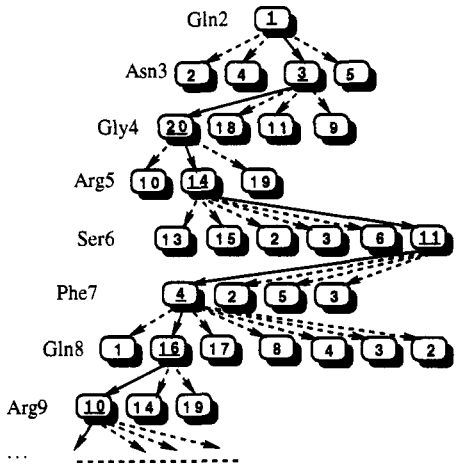


Figure 16. Possible assignment tree generated from Figure 15. Note that the solid arrows represent the path that is the most likely sequence assignment.

overlap, and resonance deviation are not as serious as in the case of biopolymers.

(2) Routine organic molecules involve uncountable structural components, and their ISNets are complicated. Therefore, to create a general functional group cluster center base, we want to focus on relatively small structural fragments.

CPSE will emphasize developing the following algorithms:

Connecting ISNets Algorithm: it combines all necessary spectral information and produces a connected ISNet.

A New Structure Generating Algorithm: This will get structures from ISNets. This algorithm is different from DENDRAL's GENOA because it is not based upon "blind combinations of substructural components". It is based upon the rigorous logical deduction of a real structure from ISNets.

Algorithms for Predicting Multidimensional NMR Spectra from a Hyperstructure: These algorithms will be used to verify the structure generated from a computer-aided elucidation procedure.

In the aspect of knowledge base (KB) development, the general relation of chemical shift and chemical environment should be better understood because it will be a powerful constraint to solve "combinatorial explosion". A general functional group knowledge base (GFKB) which includes expected chemical shift values and the corresponding ideal ISNet cluster centers should soon be created. A high-performance fuzzy pattern search algorithm should also be implemented.

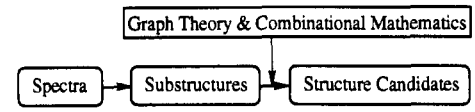


Figure 17. Conventional structure elucidation model.

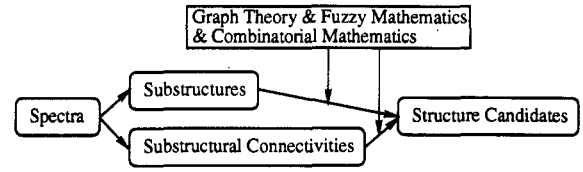


Figure 18. ISNet structure elucidation model.

CONCLUSIONS

ISNet analysis is used to explore a direct and rigorous mechanism to convert modern spectral data to a real three-dimensional structure. This methodology is a logical system which is based upon graph theory, fuzzy mathematics, and pattern recognition. Therefore, it is a powerful elucidation principle for computer-aided structure elucidation studies. By contrast, the conventional structure elucidation model (Figure 17) uses graph theory and combinatorial mathematics as tools and chemical component information as constraints. Therefore, serious "combinatorial explosion" problems are hard to avoid.

However, the ISNet structure elucidation model (Figure 18) includes the information of chemical components and structural connectivities as constraints, and hence the "combinatorial space" is sharply reduced. It also has rigorous rules to convert substructures and their connectivities into a real structure. In many cases, a unique structure candidate can be directly generated.

The ISNet structure elucidation model does not exclude 1D NMR, IR, and MS spectra. The information from these spectra are still very useful. The software based upon this method will provide different level interfaces so chemists from different fields can access and benefit from modern spectral technology.

The concepts related to ISNets, such as hetero-ISNets, partial-colored graphs, etc., are new concepts in graph theory. Moreover, the fuzzy ISNet pattern recognition has not been studied in fuzzy mathematics before.

REFERENCES AND NOTES

(1) Gray, N. A. B. *Computer-Assisted "Structure Elucidation"*; Wiley: New York, 1986.
(2) Wüthrich, K. *NMR of Proteins and Nucleic Acids*; Wiley: New York, 1986.
(3) Xu, J.; Straus, S. K.; Sanctuary, B. C. Use of Fuzzy Mathematics for Complete Automated Assignment of Protein ^{1}H 2DNMR spectra. *J. Magn. Res.*, in press.
(4) Xu, J.; Gray, B. N.; Sanctuary, B. C. Automated Extraction of Spin Coupling Topologies from 2D-NMR Correlation Spectra for Protein ^{1}H Resonance Assignment. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 475–489.
(5) Xu, J.; Sanctuary, B. C. CPA: Constrained Partitioning Algorithm for Initial Assignmentt of Protein ^{1}H Resonances from MQF-COSY. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 490–500.
(6) Xu, J.; Straus, S. K.; Sanctuary, B. C.; Trimble, L. Automation of Protein 2D Proton NMR Assignment by Means of Fuzzy Mathematics and Graph Theory. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 668–682.
(7) Chromatic graph concept and applications in chemistry are discussed in this monograph, but these discussions are for chemical structure processing. Balaban, A. T. *Chemical Applications of Graph Theory*; Academic Press: New York, 1976.
(8) Summers, M.; Luigi, G. M.; Bax, A. Complete ^{1}H and ^{13}C Assignments of Coenzyme B_{12} through the Use of New Two-Dimensional NMR Experiments. *J. Am. Chem. Soc.* **1986**, *108*, 4285–4294.
(9) Duddeck, H.; Dietrich, W. *Structure Elucidation by Modern NMR A Workbook*, Springer-Verlag: New York, 1989. The structure and corresponding 2D NMR spectra used here are taken from this book.

(10) Stefano, D. L. D.; Wand, A. J. Two-Dimensional ^{1}H NMR Study of Human Ubiquitin: A Main Chain Directed Assignment and Structure Analysis. *Biochemistry* **1987**, *26*, 7272–7281.
(11) Englander, S. W.; Wand, A. J. Main-Chain-Directed Strategy for the Assignment of ^{1}H NMR Spectra of Proteins. *Biochemistry*, **1987**, *26*, 5953–5958.
(12) Karplus, M. Contact Electron-Spin Couplintg of Nuclear Magnetic Moments. *J. Phys. Chem.* **1959**, *30*, 11–15.
(13) Ludvigsen, S.; Andersen, K. V.; Poulsen, F. M. Accurate Measurements of Coupling Constants from Two-Dimensional Nuclear Magnetic Resonance Spectra of Proteins and Determination of f-Angles. *J. Mol. Biol.* **1991**, *217*, 731–736.
(14) Date is taken from: Bundi, A.; Wüthrich, K. ^{1}H-NMR Parameters of the Common Amino Acid Residues Measured in Aqueous Solutions of the Linear Tetrapeptides H-Gly-Gly-*X*-L-Ala-OH. *Biopolymers* **1979**, *18*, 285–297.
(15) Kövér, K. E.; Batta, G. The Role of Mixing Time in 2D Hetero-nuclear NOE Experiments. *J. of Magn. Reson.* **1986**, *69*, 344–349.
(16) Some special experiments for protein structure elucidation, such as HNCA, HOHAHA-HMQC, HCACO, HCACO HCA(CO)N and HNCO, have been reviewed by: James, T. L.; Basus, V. J. Generation of High-Resolution Protein Structures in Solution from Multidimensional NMR. *Annu. Rev. Phys. Chem.* **1991**, *42*, 501–542.
(17) Griesinger, C.; Sfrensen, O. W.; Ernst, R. R. Three-Dimensional Fourier Spectroscopy. Application to High-Resolution NMR. *J. Magn. Reson.* **1989**, *84*, 14–63.
(18) Clore, G. M.; Gronenborn, A. M. Applications of Three- and Four-Dimensional Heteronuclear NMR Spectroscopy to Protein Structure Determination. *Prog. NMR Spectrosc.* **1991**, *23*, 43–92.
(19) Mirau, P. A. A Strategy for NMR Structure Determination. *J. Magn. Reson.* **1992**, *96*, 480–490.
(20) Laue, E. D.; Skilling, J.; Staunton, J.; Sibisi, S.; et al. Maximum Entropy Method in Nuclear Magnetic Resonance Spectroscopy. *J. Magn. Reson.* **1985**, *62*, 425–437.
(21) Hoch, J.; Poulsen, F. M.; Redfield, C. *Computational Aspects of the Study of Biological Macromolecules by Nuclear Magnetic Resonance Spectroscopy*; NATO Series A: Life Sciences, Vol. 225; Plenum Press: New York, 1991.
(22) Hoffman, R. E.; Kumar, A.; Bishop, K. D.; Borer, P. N.; Levy, G. C. Application of the Maximum Likelihood Method to a Large 2D NMR Spectrum Using a Parallel Computer. *J. Magn. Reson.* **1989**, *83*, 586–594.
(23) Kleywegt, G. J.; Boelens, R.; Kaptein, R. Aversatile Approach Toward the Partially Automatic Recognition of Cross Peaks in 2D ^{1}H NMR Spectra. *J. Magn. Reson.* **1990**, *88*, 601–608.
(24) Kraulis, P. ANSIG: A Program for the Assignment of Protein 1H 2D NMR Spectra by Interactive Computer Graphics. *J. Magn. Reson.* **1989**, *84*, 627–633.
(25) Zolnai, Z.; Westler, W. M.; Ulrich, E. L.; Markley, J. L. Drafting Table and Light-Box Software forr Multidimensional NMR Spectral Analysis (PIXI). Ther Personal Computer Workstation. *J. Magn. Reson.* **1990**, *88*, 511–522.
(26) Majumdar, A.; Hosur, R. V. Simulation of 2D NMR Spectra for Determination of Solution Conformations of Nucleic Acids. *Prog. NMR Spectrosc.* **1992**, *24*, 109–158.
(27) Kaufmann, A. *Introduction to the Theory of Fuzzy Subsets*; Academic Press: New York, 1975; Vol. 1.
(28) Inagaki, F.; Shimada, I.; Kawaguchi, K.; et al. Structure of Melittin Bound to Perdeuterated Dodecylphosphocholine Micelles As Studied by Two-Dimensional NMR and Distance Geometry Calculations. *Biochemistry* **1989**, *28*, 5985–5991.
(29) Fairbrother, W.; Palmer, A. G., III; Rance, M.; et al. Assignment of the Alephatic ^{1}H and ^{13}C Resonances of the Bacillus subtilis Glucose Permease IIA Domain Using Double-Sand Triple-Resonance Hetero-nuclease Three-Dimensional NMR Spectroscopy. *Biochemistry* **1992**, *31*, 4413–4425.
(30) Gross, K. H.; Kalbitzer, R. Distribution of Chemical Shifts in ^{1}H Nuclear Magnetic Resonance Spectra of Proteins. *J. Magn. Reson.* **1988**, *76*, 87–99.
(31) Boelens, R.; Koning, T. M. G.; van der Marel, G. A.; van Boom, J. H.; Kaptein, R. Interative Procedure for Structure Determination from Proton-Proton NOEs Using a Full Relaxation Matrix Approach. Application to a DNA Octamer. *J. Magn. Reson.* **1989**, *82*, 290–308.