Employment of Fuzzy Information Derived from Spectroscopic Data toward Reducing the Redundancy in the Process of Structure Generation

Thierry Laidboeur, Isabelle Laude, and Daniel Cabrol-Bass*

LARTIC, University of Nice Sophia-Antipolis, 06108 Nice Cedex 2, France

Ivan P. Bangov†

Institute of Organic Chemistry, Building 9, Bulgarian Academy of Sciences, Sofia 1113, Bulgaria

Received June 29, 1993®

An approach toward the construction of *fuzzy sets* of vertices and bonds, hence of *fuzzy graphs*, based on ^{13}C NMR information is presented in this paper. A membership function m transforming the support vertex set V onto a fuzzy vertex set V^f is introduced. The formation of the fuzzy bond set T^f is carried out during the generation process. It was shown that the fuzzy information content of these sets can be exploited toward the discrimination of the unlikely extensions appearing at each step of the structure generation process. This leads to a *heuristic search procedure* which greatly reduces the redundancy and alleviates the combinatorial problem. The constructed fuzzy graphs at the end of this process fully satisfy the mathematical definition of these objects. However, it is argued that it is not possible for completely reliable conclusions to be derived from the additivity parameter based ^{13}C NMR chemical shift/substructure relationships, only. Fuzzy spectral information from additional sources should employed in order to assist in solving the structure elucidation problem.

INTRODUCTION

Several major projects in the field of computer-assisted structure elucidation (CASE) have been developed for more than a decade,[1-8] and some are still in continuous improvement[9-14] attesting to the great difficulty of the task. However, the main problem in the way of the development of fully automated structure elucidation systems emerges from the fact that the structural information derived from any available spectral data is more or less *fuzzy* by its nature. This means that the structural features derived from such information are incomplete, ambiguous, overlapping, and/or alternative. Even the most powerful 2D NMR INADEQUATE method does not always provide a definitive answer about the query structure. Several attempts to overcome these difficulties have been made. Thus, in GENOA[15] parts of the input substructures are considered alternative; other parts, overlapping. However, both the alternative, the overlapping, and the constant parts are handled as *crisp* substructures in reality. The real limits of their fuzziness are practically not known. Another approach to address this problem was employed in CHEMICS.[4,5,9] The first part of this system consists of an elegant procedure which automatically derives several sets of small fragments and chemical groups (called *segments*) taken from a small multispectral data base and consistent with the input molecular formula. Even the large fragments entered as an additional constraint are further decomposed to such segments. Each set may be regarded as a *hypothesis*, and it is further processed by the structure generator. So long as these fragments are quite simple, most of them having one or two non-hydrogen atoms, there is no need considering the overlapping and alternative substructures. However, this approach leads to severe combinatorial difficulties for molecules having more than 10 non-hydrogen atoms. A similar approach was devised by the group of Munk using atom centred fragments (ACFs).[16] In contrast to CHEMICS, the number and the size of those fragments is larger. Another approach, generation by reduction, has also been developed in this group[17] in order to encompass the structural uncertainty. An original approach using multiresonance data was devised by Dubois et al.[12] Here, the chemical structure of the unknown compound is constructed using formalized local and global knowledge described statistically by juxtaposing of the $\delta^{13}C \times \delta^{13}C$ correlation plane supporting the 3D occurrence distribution.

Recently, a new project for development of a structure elucidation system was initiated in our laboratory. It is based on the following considerations: First, it seems to us that the original goal of building a completely automatic system is too ambitious; consequently the system in development is better described as a *decision support system*. Thus the user is often prompted to take decisions and actions in critical parts of the elucidation process. Second, the performance of such a system can be greatly enhanced by using simultaneously several spectroscopic methods. On the one hand, all relevant spectroscopic information for an unknown might not be available; on the other hand, new spectroscopic methods such as 2D NMR have been rapidly developed. All this calls for the choice of a modular architecture capable of working even in the case that some particular spectroscopic information is not available and in which new modules designed to handle new spectroscopic methods could be easily incorporated. This architecture is presented in Figure 1. It allows the cooperation of several knowledge sources (named the "specialists", Figure 1c), each of them being "specialized" in one spectroscopic method or even in one particular aspect of a given spectroscopy. During the process of structure elucidation, the content of a *workplace* (Figure 1b) holds all the structural information available at a given time. The specialists are activated by an *interface module*, and they either provide information at the beginning of the structure elucidation or evaluate the degree of support of the new substructures created at each stage of the structure generation process. One main advantage of this architecture is that each knowledge source can be implemented

* To whom all correspondence should be addressed.

† Presently, a Senior Guest Scientist on a NATO grant at the University of Nice Sophia-Antipolis, France.

® Abstract published in *Advance ACS Abstracts*, January 15, 1994.

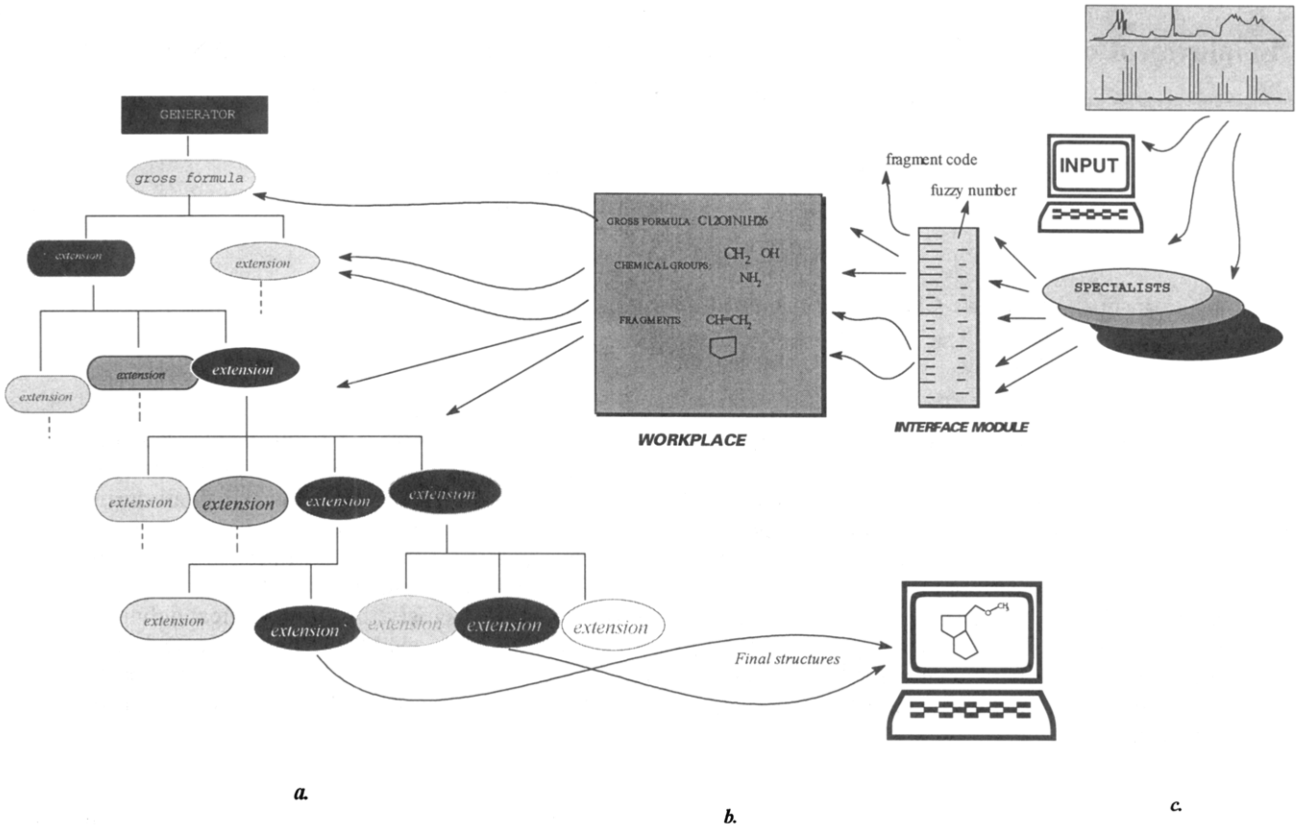


Figure 1. Modular architecture of the structure elucidation system: (a) Multilevel structure generation procedure, producing one or several final structures; (b) workplace and interface module; (c) "specialists"-program modules converting the spectral information into a structural result.

using different approaches (rule-based expert systems, pattern matching methods, neural networks classifiers, etc.). Thus, each specialist remains entirely independent of the others, both at the level of its specific knowledge representation and of its own reasoning mechanism.

A flexible structure generator was developed[18–20] which is based on a hierarchical multilevel procedure (Figure 1a). At each level a new atom, chemical group, or fragment is added to the free valences of the substructure(s) formed at the lower levels, thus producing several extensions. Each extension is characterized by the formation of a new chemical bond. All extensions generated at a given level are compared with the available spectral information (a heuristic search in the generation tree presented in Figure 1a) and some fuzziness values in the intervals [0...1] (or [0...100] in percentage) indicating either lower (lighter grey in Figure 1a) or higher (darker grey) support of each extension are generated. Only extensions having higher degrees of support are retained for the next level structure generation. Thus, whole branches of the generation tree are pruned and the generation process is highly alleviated. The use of 2D NMR spectral data within this generation approach has been reported in a recent paper.[21] The further development of this architecture and the implementation of new spectroscopic sources (specialists) will be discussed in a future article of ours. This paper is devoted to the employment and the practical implementation of *fuzzy logic* and *fuzzy graphs* to the problems of structure elucidation. It should be mentioned here that the notion of fuzzy graphs has become rather fashionable, but it is often used in literature intuitively, with no relation to the mathematical definition of these objects.

FUZZY MOLECULAR GRAPHS

In his basic paper Zadeh[22] introduced the concept of *fuzzy set* as a generalization of a classical set in order to deal with categories which cannot be precisely defined. In classical set theory an item is either a member of a set or it is not. For example, either a particular atom is a carbon or it is not. But there are cases where such a crisp membership cannot be so definitively determined. For example it is not so easy to decide if a particular element is a member of the set of "strongly electronegative elements". Fuzzy set theory deals with the representation of such sets whose boundaries are not precisely defined by means of a *membership function*. If $\boldsymbol{S}$ is a classical set defined on a universe $\boldsymbol{U}$, one can define the fuzzy set $\boldsymbol{S}^f$ by associating to each element $\mathbf{e} \in \boldsymbol{S}$ a membership function $m(\mathbf{e})$, taking real values in the interval [0,1]. The value of the membership function is often quite incorrectly interpreted as follow: $m(\mathbf{e})$ is equal to 1 when $\mathbf{e} \in \boldsymbol{S}^f$ and equal to 0 if $\mathbf{e}_i \in \boldsymbol{S}^f$, intermediate values reflecting the degree of membership of the element $\mathbf{e}$ to the set S^f. If the membership function is such that it maps $\boldsymbol{S}$ on the two values set {0,1}, then the set $\boldsymbol{S}^f$ is called crisp. The set $\boldsymbol{S}$ is called the *support set* of $\boldsymbol{S}^f$ in the universe $\boldsymbol{U}$. The subset of $\boldsymbol{S}$ which contains only those elements of $\boldsymbol{S}$ which have nonzero values of the membership function is called the projection of the fuzzy set $\boldsymbol{S}^f$. It should be noted that while the support set and the projection of a fuzzy set are both ordinary sets, a crisp set is indeed a particular case of a fuzzy set.

Since 1965 fuzzy set theory has been considerably extended and has served as basis for the development of fuzzy logic and the *theory of possibility*.[23] While in classical logic a proposition is either *true* or *false*, these Boolean values being frequently represented by the two integers 0 and 1, fuzzy logic makes use of a *thrust function* using real values in the interval [0,1] to represent the degree of support of the proposition. A value equal to 1 indicates that the proposition is considered as true for sure; conversely a value equal to 0 indicates that the proposition is not true. Thus, fuzzy logic and fuzzy set theory are tightly linked. Classical set operations (intersection, union, complement) and corresponding logical ones (AND, OR, NOT) have been adapted to fuzzy sets and logic and are applied to membership and thrust functions. A good tutorial introduction to fuzzy set theory and its applications to chemistry can be found in ref 24.

These concepts were further applied to graph theory leading to the generalized concept of fuzzy graphs.[25] For the sake of our further representation the application of this concept to molecular graphs will be outlined here.

A graph G is defined by two finite nonempty sets V and E as

$$G = (V,E)$$

Here V is the set of *vertices* (v_i) and E is the set of *edges* (e_{ij}) which is a subset of the Cartesian product ($V \times V$). In the case of directed bonds, named *arcs*, in the graph theory we have *directed graphs* (*or digraphs*), which can be defined as

$$D = (V,T)$$

Here T is the set of *arcs* (t_{ij}). The generator described below has been developed to handle directed rather than nondirected graphs.

The algebraic object graph appeared extremely useful for the description of a series of chemical phenomena such as *chemical structure* (*molecular graphs*), *reaction paths* (*reaction graphs*), *chemical processes*,[26] etc. Hence, all these graphs have the general name of *chemical graphs*. One of the commonly used representations of a graph is the *adjacency matrix*. This matrix can be constructed in the following way: each vertex $v_i \in V$ is labeled by a number from the integer set $[1...N]$. Here N is the number of vertices, and it forms the rank of the matrix. Formally speaking there may exist $N!$ different such numberings, i.e., $N!$ different adjacency matrices representing the *abstract graph*.[27] This creates the so called *isomorphism problem*.[28] This problem within the framework of our generator has been discussed elsewhere.[29]

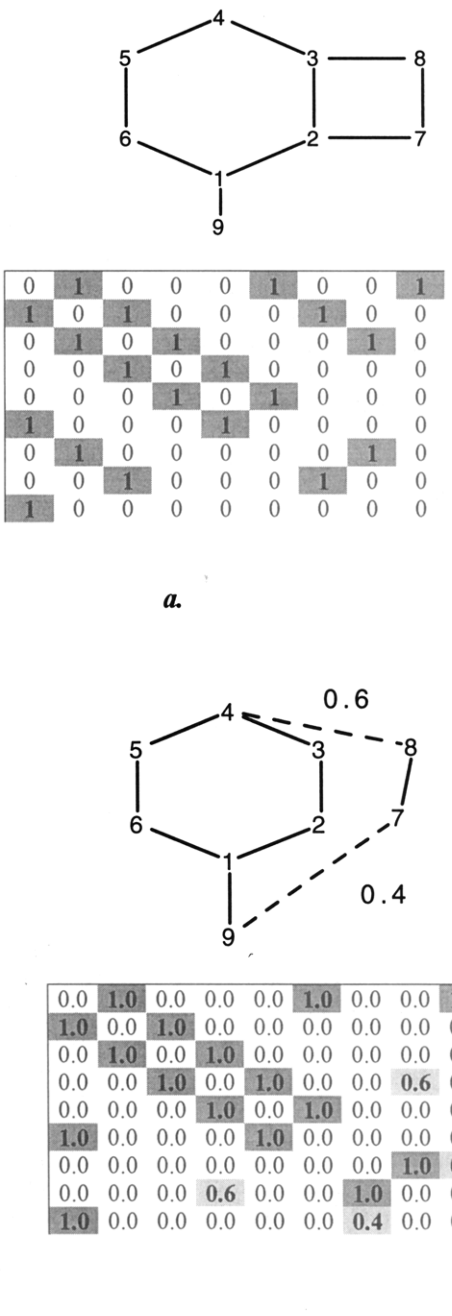
The adjacency matrix **A** has entries which may be viewed as mappings of a *characteristic function* of E denoted as $\mathbf{m}^E$. In the case of classical (crisp) graphs, this function may take either value from the set of integer numbers $\{0,1\}$ and it forms the usual adjacency matrix having entries

$$a_{ij} = \begin{cases} 1, & \text{if } (i,j) \text{ form an edge (arc)} \\ 0, & \text{otherwise} \end{cases}$$

This can be translated in chemical terms that if $a_{ij} = 1$, then a chemical bond between atoms i and j is present; otherwise such a bond is absent.

In the context of fuzzy set theory, both the set of vertices V and the set of edges E forming the graph G will be considered as fuzzy sets (respectively V^f and E^f) which constitute the *fuzzy topological graph* $G^f = (V^f,E^f)$. The consequences are that each vertex $v_i \in V$ on the one hand and each pair of the Cartesian product ($V \times V$) on the other hand will be associated with a membership function which maps these two sets on the range of real values [0,1]. Hence, the entries of the adjacency matrix corresponding to the fuzzy set E^f of edges will contain the values of the associated membership function. Here ($V \times V$) is the support set of E^f. In brief, attaching a membership function to the set of vertices and to the set of edges transforms a classical graph into a fuzzy graph.

As indicated above, a fuzzy set can be transformed into an ordinary set by *projection*; here the projection of E^f is E. Similarly the classical graph G is the *projection of the fuzzy graph* G^f, if and only if V is the projection of V^f and E is the projection of E^f.[30] Hence the isomorphism problem of the fuzzy graphs G^f can be reduced to the isomorphism problem of their *projection graphs* G. Two cases of common and fuzzy graphs are depicted in Figure 2 with the corresponding adjacency matrices.



0	1	0	0	0	1	0	0	1
1	0	1	0	0	0	1	0	0
0	1	0	1	0	0	0	1	0
0	0	1	0	1	0	0	0	0
0	0	0	1	0	1	0	0	0
1	0	0	0	1	0	0	0	0
0	1	0	0	0	0	0	1	0
0	0	1	0	0	0	1	0	0
1	0	0	0	0	0	0	0	0

a.

0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.6	0.0
0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.4
0.0	0.0	0.0	0.6	0.0	0.0	1.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0

b.

Figure 2. (a) Crisp and (b) fuzzy graphs given with their adjacency matrices.

Chemical molecular graphs are *irregular* (having different vertex degrees), *chromatic* (the vertices and edges may be differently labeled), having no *loop* graphs. These properties are related to chemical structures formed of atoms, chemical groups, or fragments with different valences, having a different chemical nature (carbon, nitrogen, oxygen atoms, methyl, hydroxyl groups, etc.) and with no bonding of an atom or chemical group with itself allowed. Consequently in addition to the fuzziness on topology, it is necessary to define the fuzziness on chromatism. The coloring a particular vertex v_i consists of labeling this vertex with the available information related to the atom, chemical nature, hybridization state, number of attached hydrogen atoms, etc. These labels are usually given of discrete variables corresponding to the different possible values of each label. Then the *fuzzy coloring* of a vertex will consist of attaching to each label the corresponding membership function value.

In some representations the multiple bonds are represented by *multi*graphs (having multiple edges). In our structure representation, the multiple bonds are represented as single bonds between atoms having different hybridization and valence states, i.e., by simple arcs between vertices of degrees different from the vertices forming single bonds. Hence, they have different labeling. Thus, a double bond between two carbon atoms is represented as a simple bond between two sp^2 olefinic atoms. However, this differentiation requires an additional labeling of the molecular graph vertices according to the different atom hybridization states and the definition of the corresponding membership functions. The main problem remains is how the membership functions reflecting the structural content of the spectral information can be constructed.

An attempt to implement the fuzzy logic reasoning within the structure elucidation scheme outlined above is reported in this paper. The considerations given below are based mainly on ^{13}C NMR spectral information, but intensive research in other fields of spectroscopy is currently ongoing. Thus, the results from IR and MASS spectrum trained neural networks obtained by us[31] also have fuzzy character, and they can be easily incorporated into this scheme. These latter developments will be reported in the near future.

^{13}C SPECTRAL INFORMATION AND FUZZY MOLECULAR GRAPHS

As stated above the structural information derived from any spectroscopic source is fuzzy. This fact should be reflected by the mathematical representation of the generated structures; i.e., the latter should be represented by fuzzy rather by crisp graphs. Accordingly, the structure elucidation may be viewed as a transformation of the fuzzy spectral input into crisp structural output (one or several chemical structures).

$\boldsymbol{K}_a$ denotes a set of atom types. We can define a function f which maps the set $[1...N]$ of integer numbers into the set $\boldsymbol{K}_a$; i.e., $f(i) = m$, where $m \in \boldsymbol{K}_a$. This function produces a *numbered* gross formula.[27] In chemical terms this means that the vertices are labeled by integer numbers and additionally labeled as carbon, oxygen, etc., atom identifiers. More complex is the problem in the cases of chemical groups and fragments. All they are called here *segments*.[5] Thus, formally we can consider two type of graphs; the first one is the *segment graph*, having the segments as vertices. The second one is the usual *atomic graph*, having the atoms as vertices. The segment graph can be used in the process of structure generation, while the atomic graph, in the process of structure identification. Those two types of molecular graphs are implemented in a common structure representation in our systems, which has been discussed in a series of recent papers.[18–21] It will be outlined below with respect to the generation of chemical bonds.

We consider the numbered molecular formula (gross formula) a crisp set of numbered atoms (vertices). It is expected that the information for its determination is not ambiguous—intersecting data from element analysis, MASS spectrometry, and ^{13}C NMR spectroscopy could provide a definitive answer. Each carbon atom is arbitrarily associated with a signal from the ^{13}C NMR spectrum of the unknown (query) structure. Consequently, each carbon atom adopts two attributes: chemical shift and multiplicity. Any signal overlapping must be user-recognized. The overlapping signals are considered different but have the same chemical shift values. Hence, the carbon atoms are transformed either into quaternary C atoms (multiplicity 1) or into CH (multiplicity 2), CH_2 (multiplicity 3), and CH_3 (multiplicity 4) groups according to the *multiplicity* attribute. The multiplicity information is also considered crisp as it can be unambiguously determined using different NMR techniques, such as *off-resonance* or *DEPT*. However, more ambiguous is the information content of the *chemical shift* attribute. For the further construction of the unknown structure we need to distinguish between the sp^3, sp^2 (olefinic, carbonylic and arylic), and sp signals. This discrimination produces additional chromatism among the molecular graph vertices. Thus, an olefinic sp^2 carbon atom of valence 3 (denoted as =C in our representation) is considered different than a sp^3 carbon atom of valence 4 (denoted as C) or than an arylic carbon atom of valence 3 (:C) or than a sp (#C) carbon atom of valence 2. So long as the chemical shift ranges of the different hybridizations are overlapping, their information content is fuzzy. Hence, the fuzzy logic formalism was used for their differentiation. The method used hereafter has been outlined in a previous paper.[21] It is presented here in more detail. The chemical shift ranges observed for carbon atoms of a given hybridization state (sp^3, sp^2, sp) do not define fixed intervals but show overlapping distributions. Even treating separately the cases of signals of different multiplicity M, corresponding to different numbers of attached hydrogen atoms, does not lead to a complete separation of the distribution curves. Thus instead of crisp chemical shift ranges, we define fuzzy ranges for each multiplicity $M = [1,2,3]$ and hybridization state leading to eight different fuzzy sets $\boldsymbol{S}_k$, namely, $M = 1$, sp^3; $M = 1$, sp^2,; $M = 1$, sp; $M = 2$, sp^3; $M = 2$, sp^2; $M = 2$, sp; $M = 3$, sp^3; $M = 3$, sp^2. The case $M = 4$, corresponding necessarily to sp^3 hybridization, raises no ambiguity. Each carbon atom number i assigned to a ^{13}C signal of known multiplicity M and chemical shift δ_i is considered as an element $e_i = \langle M,\delta_i \rangle$. The problem is then to determine to which of the above mentioned fuzzy sets this element belongs. This is achieved by defining a membership function $m_k(\delta_i)$ for each set $\boldsymbol{S}_k$. Each value of this function is an estimate of the degree of support of the hypothesis that an atom i is a member of the fuzzy set $\boldsymbol{S}_k$, indeed. The membership functions are built up by comparing the distribution curves of the observed chemical shifts. When these distribution curves do not overlap, the membership function takes the value 0 for regions outside of the observed ranges and 1 for regions inside. Within overlapping regions, an estimate of relative confidence is made by calculating the ratio of the number of the observed cases in each set in the considered region. For transitions between regions (outside, overlapping, inside) a sigmoid function is used to model the membership function.

$$m(e_i) = P + \frac{H}{1 + e^{-a(\delta_i - b)}}$$

For each fuzzy set $\boldsymbol{S}_k$ and each transition region, the parameters P, H, a, and b are determined empirically to fit the ratio in the chemical shifts distributions. If necessary the functions are normalized for each point so that the sum for all k obeys the relation $k \leq 1$.

Figures 3–5 show the variation of the membership functions with the chemical shift δ for multiplicity $M \in [1,2,3]$. This estimate incorporates the a-priori probability and is therefore dependent on the size and composition of the spectral collection used. In this work, the distribution curves have been built by analysis of an in-house collection of 1300 spectra corresponding to 8899 carbon atoms assigned to NMR signals. The work is being extended to a larger collection of spectra recently made available to us. However, an extensive study based on a large database containing 15 867 reference structures[11] leads

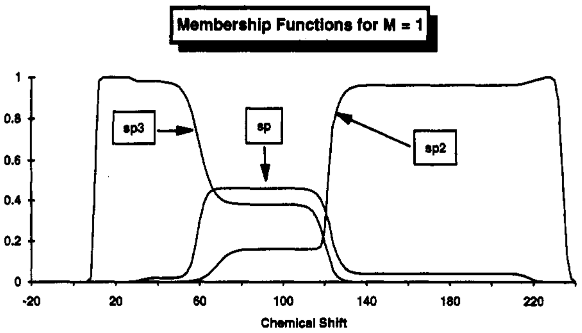


Figure 3. Membership functions of sp^3, sp^2, and sp carbon atoms for the multiplicity = 1 case.

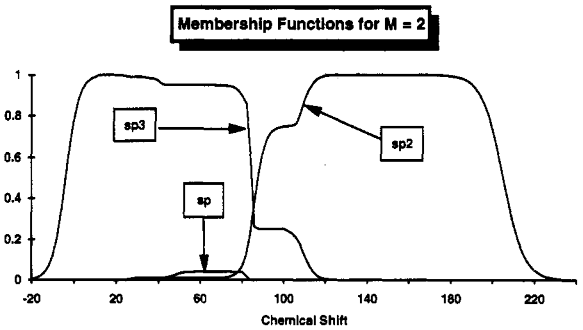


Figure 4. Membership functions of sp^3, sp^2, and sp carbon atoms for the multiplicity = 2 case.

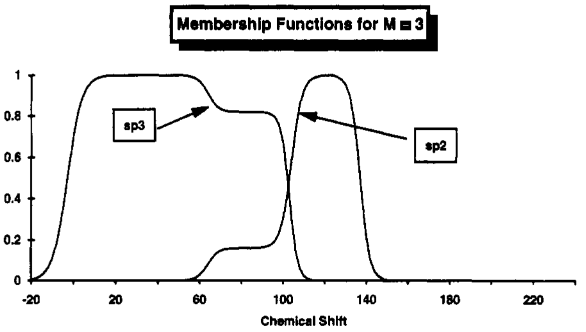


Figure 5. Membership functions of sp^3 and sp^2 carbon atoms for the multiplicity = 3 case.

the authors to the conclusion that a sample of 2000 structures shows a very similar distribution to that of the complete population.

After having defined the fuzzy set $\boldsymbol{V}^f$ of the vertices, our next step is to define the fuzzy set of edges (arcs) $\boldsymbol{E}^f$ ($\boldsymbol{T}^f$). This is directly related to the problem of structure elucidation because it is supposed that the most probable structure will have higher values of its A^f matrix entries. The formation of the different graphs is carried out in the course of the generation of different *support* sets of bonds. Accordingly, hereafter we shall define the set $\boldsymbol{E}^f$ ($\boldsymbol{T}^f$) within the terms of our structure generation method. A detailed description of the generation scheme has been given elsewhere.[18-21]

In contrast to other methods, this method is based on directed graphs $\boldsymbol{D}$. As mentioned above, atoms, chemical groups, and structural fragments are considered in a uniform way as segments. *Bonding sites* (BSs) instead of the segments themselves are considered to be the basic elements of this representation. The notion of a bonding site refers as to the classical notion of *free valence* in chemistry. A structure having directed bonds is presented at the bottom of Figure 6. One can see that each bond is formed by two formally different bonding sites depicted by crosses and arrows. In previous papers they were named saturating valences (SVs) and saturation sites (SSs), respectively. Each segment, except the first has 1 SV and (n – 1) SSs. Here n is the full valence of the given segment, i.e., the number of BSs. The first segment has all its n BSs as SSs. Each atom where a cycle closes provides one additional BS being transformed into SV. Accordingly, each segment except the first is characterized by its SV. The SVs form a separate set, and they are ranked according to some rules (see Figure 6) which are given in the previous papers on this method. Thus, a multilevel hierarchical scheme is formed, and the generation process is carried out as each SV on a given level forms bonds with all free SSs on this level, thus producing a set of extensions. The process of structure generation is depicted in Figure 6. For each *father extension* a plurality of *successor extensions* is produced (the SV–SS representation of the bonding sites practically describes the father/successor relationship in rooted trees). The structure is represented by a two-row matrix where the first row is the SSs given with their atom numbers and at the second row are the free BSs denoted by crosses, which will be saturated by SVs during the process of the structure construction. A bond is represented by juxtaposing a first row SS to a second row element saturated by a SV. As a result of this representation, only the SVs participate in the combinatorial process. On the one hand, this process is carried out between whole segments rather than between single atoms. On the other hand, the two-row matrix representation provides the atomic constitution of the molecular graph which is helpful for its further identification and manipulation.

The problem is how a real number t_{ij} is to be associated with each bond newly generated. Thus, a real-number vector having the dimension of the two-row matrix can be formed. Furthermore, while the two-row matrix represents the support of $\boldsymbol{T}^f$, this vector will represent the fuzzy part $\boldsymbol{T}^f$. The elements t_{ij} must reflect in a way the chemical shift/structure relationship. It should be stated here that the use of a structure generator within a structure elucidation system implies that there is a lack of information. Thus, if N structures are generated from any initial spectral data, then each generated bond (ij) may be associated with a number $t_{ij} = 1/N$. However, as all these numbers are equal, no discrimination of more or less *possible* structures can be carried out during the process of structure elucidation. Hence, this does not contribute very much to the elucidation process.

So as the chemical shifts depend both on the nearest and on the longer range environments, the following approach toward the formation of the numbers t_{ij} was adopted:

Each one of the successor extensions generated at a given level is a substructure (or several substructures) which differs from the father extension generated at the previous (lower) level by a new bond. The chemical shifts δ^{calc} of the carbon atoms may be evaluated by using any of the additivity parameter schemes (the parameters from ref 32 are used here). For each extension generated after a bond (ij) is formed, the Hamilton agreement (R) factor[33]

$$R_{ij} = [\sum_k (\delta_k^{\text{calc}} - \delta_k^{\text{expt}})^2]^{1/2} / [\sum_k (\delta_k^{\text{expt}})^2]^{1/2} \qquad (1)$$

is calculated where δ_k^{calc} and δ_k^{expt} are the calculated and experimental chemical shift values. While the latter, as discussed above, are attributes to the carbon atoms and they are constant during the whole generation process, the former change with each new extension. All the R_{ij} generated at a given level are summed (summation over S), and for each extension the following real number is generated:

$$t_{ij} = 1/R_{ij} / \sum_S 1/R_S \qquad (2)$$

Accordingly, each new bond (i,j) created at the current level

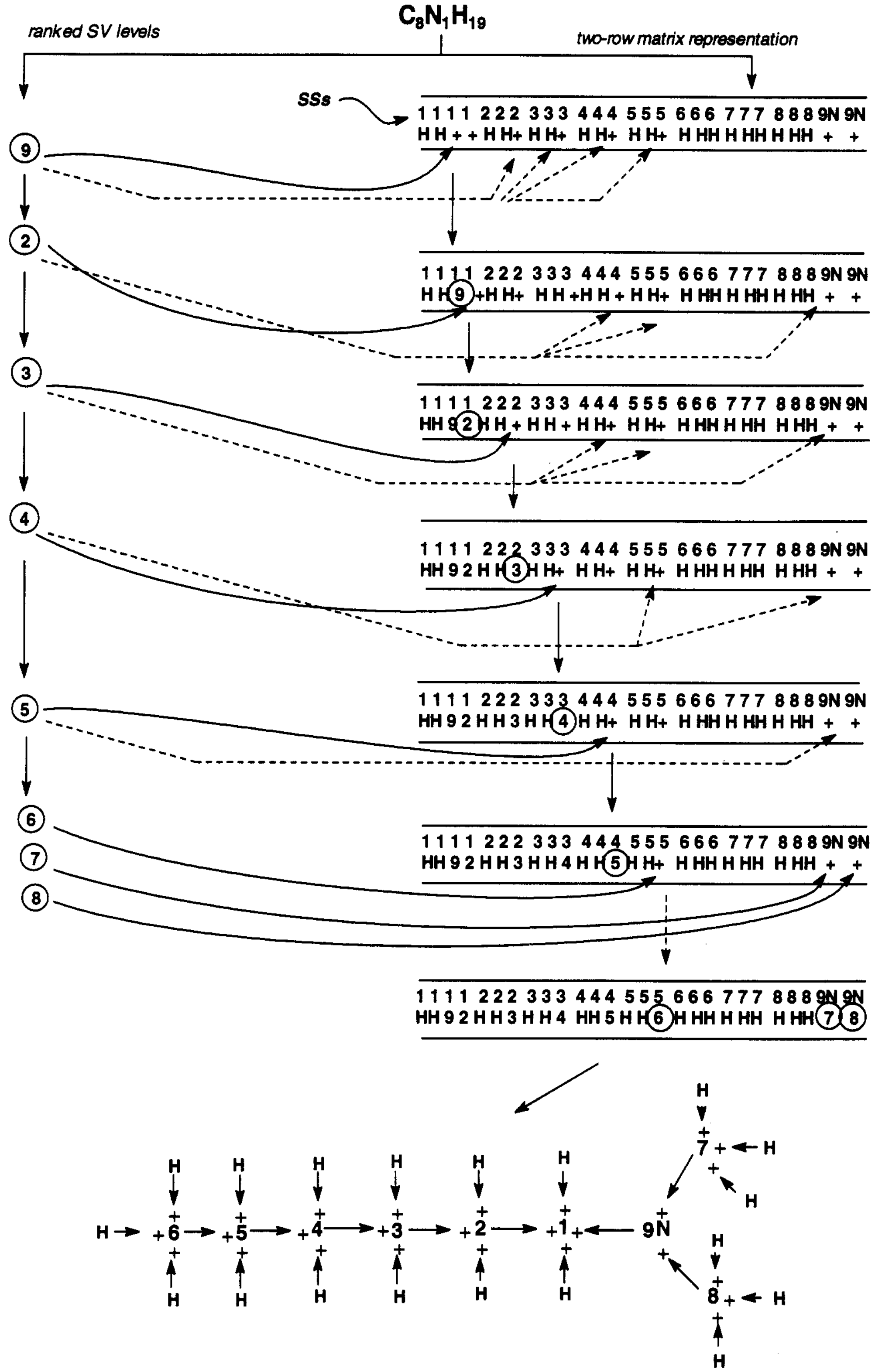

$C_8N_1H_{19}$
ranked SV levels
two-row matrix representation
SSs
1111 222 333 444 555 666 777 888 9N 9N
HH++HH+HH+ HH+ HH+ HHHHHHHHH + +
9
2
3
4
5
6
7
8

Figure 6. Two-row matrix representation of the process of multilevel structure generation. The solid curved arrows give the current, and the dashed arrows, the other possible successor extensions generated from each father extension.

is associated with such a fuzzy number t_{ij}. Precautions have been taken in the program so that the term $1/R$ does not produce an error in the cases of R_{ij}s approaching zero (practically never occurring).

The structure elucidation process is carried out by eliminating the extensions having lower t_{ij} and retaining for the next level structure generation those of higher t_{ij}. This process, depicted schematically in Figure 1a, is exemplified by a simple real example in Figure 7 (^{13}C NMR spectrum data taken from ref 34). Each bond newly generated at a given level (given in Figure 7 in dashed line boxes) is characterized by a fuzzy number. As discussed above, the extensions having the highest numbers (encircled in Figure 7) are the only ones retained for the next level structure formation, the others being discarded. Obviously, in the real cases more than one extension having close t_{ij} numbers may be retained. This leads to a *heuristic search procedure* which highly reduces the structural redundancy and thus alleviates the combinatorial problem. Consequently, whereas the *projection graph* is represented by the two-row matrix representation, the t_{ij}s form a vector T^f representing the fuzzy part of the graph. The fuzzy graph representation of the structure generated in Figure 7 is given

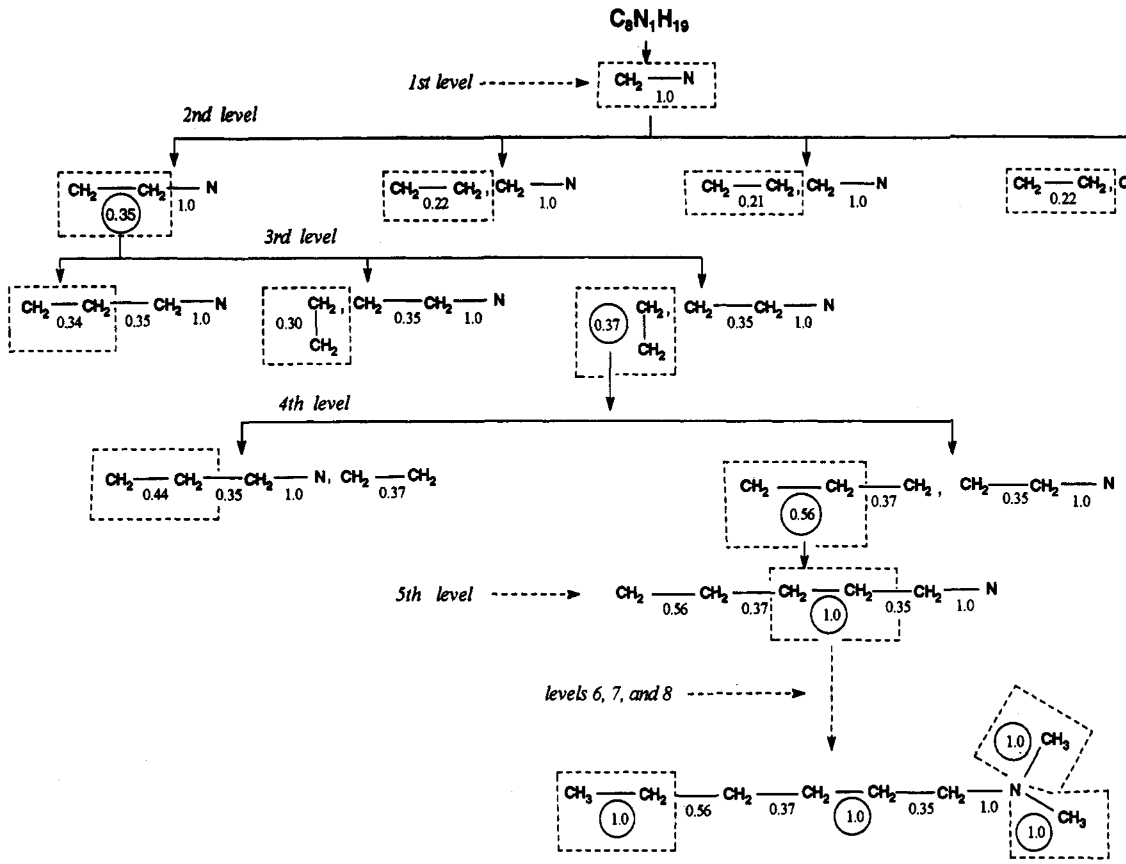
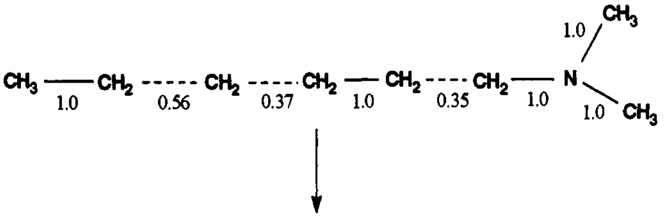


Figure 7. Generation of fuzzy graphs and application of the fuzzy logic to the problem of computer-assisted structure elucidation. The newly generated bond of each extension is given in a dashed-line box, and the t_{ij} value of the selected extension is encircled.



Two-row matrix	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	N	N
	1	1	1	1	2	2	2	3	3	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	9
	N	C	H	H	C	H	H	C	H	H	C	H	H	C	H	H	H	H	H	H	H	H	H	H	H	C	C
	9	2			3			4			5			6												7	8
T^f *vector*	1.0	0.35	1	1	1.0	1	1	0.37	1	1	0.56	1	1	1.0	1	1	1	1	1	1	1	1	1	1	1	1.0	1.0

Figure 8. Fuzzy T^f vector and the projection two-row matrix representation of the fuzzy structure generated in Figure 7. The C–H bonds are considered certain, having real numbers of 1.0 (given as 1 because of lack of space).

in Figure 8. The fuzzy graph adjacency matrix could be easily reconstructed from this representation.

Although this example leads to a correct result, use of only the additivity-rule-based fuzzy elements is not reliable. The problem is that the additivity parameter schemes are derived from complete structures. Consequently, the results strongly depend on the chemical environment of the resonating nuclei. For instance, whereas an increment value of 49 ppm is provided for the -O- substituent in alkanes, most carbon atoms adjacent to an oxygen resonate in the range of 55–70 ppm, the difference being due to the other substituents. However, these substituents might be generated at the higher levels of the generation process. Thus, the predicting reliability of the additivity parameter scheme is lower at the initial levels and it increases with the augmentation of the generated structure. Hence, we were forced to impose additional constraints here. So long as the heteroatoms exert large perturbations, it was assumed that only linkages of heteroatoms with carbon atoms having chemical shifts higher that 39 ppm are allowed.

The process of formation of the vector T^f depicted in Figure 7 starts with the generation of only one extension at the first level because of the restriction, mentioned above. Hence, from eq 2 follows that the only t_{ij} element is equal to 1.0. Four extensions are generated at the second level (the newly formed bonds are given in dashed line boxes). Three of them (the second, third and fourth) are chemically but not spectrally equivalent; i.e., eqs 1 and 2 produce different t_{ij} values (the t_{ij} values of the second and fourth extensions differ in the third place after the decimal point). Three nearly equivalent extensions (having close t_{ij} values) are generated at the third level, and the third extension was selected as having a slightly higher t_{ij} value. The fourth level provides two extensions where the second one is easily selected as having a markedly higher t_{ij} value. Only one extension is generated at each one of the next three levels due to the constraint discussed above. Hence, the corresponding t_{ij}'s equal 1.0. As a consequence, only one structure is singled out at the highest level which in this case coincides with the correct one.

It must be admitted here that even such crisp constraints cannot ensure reliable results in many cases. It is apparent that the values t_{ij} cannot be the only criteria and fuzzy values from other spectral techniques and physical methods must be implemented into this generation scheme. The development of such values is currently in progress.

CONCLUSIONS

It is shown in this paper that the concept of fuzzy graphs can be usefully exploited in the process of computer-aided structure elucidation. A practical implementation of the mathematical fuzzy graph theory into the molecular graph representation has been carried out. The uncertainty of the structural information employed during the elucidation process is critically considered. It is shown that whereas the molecular formula and the ^{13}C–^{1}H multiplicity may be considered, in most of the cases, as certain, the hybridization information for the carbon atoms derived from the ^{13}C chemical shifts has fuzzy character. A membership function has been introduced determining the different hybridization states of the carbon atoms; hence a fuzzy set of the vertices V^f has been defined. Additionally, an approach toward construction of the fuzzy set T^f of directed bonds during the generation of chemical structures has been devised. It is based on the use of the ^{13}C additivity incremental schemes. The practical application of this approach to the discrimination of the generated extensions leading to substantial reduction of the structural redundancy is exemplified. Accordingly, one or several fuzzy graphs are generated at the end of the generation process. It is argued that the fuzzy graph defined on the ^{13}C NMR chemical shift information is useful tool for the structure information. However, additional fuzzy information derived from other spectroscopic sources must be employed in order to achieve a more reliable elucidation.

REFERENCES AND NOTES

(1) Ledeberg, J. Topology of Molecules. *The Mathematical Science*; The MIT Press: Cambridge, MA, 1969; p 37.

(2) Nelson, D. B.; Munk, M. E.; Gash, K. B.; Herald, D. L. An Application of Computer Techniques to Structure Elucidation. *J. Org. Chem.* **1969**, *34*, 3800–3805.

(3) Carhart, R. E.; Smith, D. H.; Brown, H.; Djerassi, C. Application of Artificial Intellegence for Chemical Inference. An Approach to Computer-Assisted Elucidation of Molecular Structure. *J. Am. Chem. Soc.* **1975**, *97*, 5755–5762.

(4) Yamasaki, T.; Abe, H.; Kudo, Y.; Sasaki, S. CHEMICS: A Computer Program System for Structure Elucidation of Organic Compounds. In *Computer-Assisted Structure Elucidation*; Smith, H. D., Ed.; ACS Symposium Series 54; American Chemical Society: Washington, D.C., 1977; p 108.

(5) Sasaki, S.-I.; Kudo, Y. Structure Elucidation System Using Structural Information from Multisources: CHEMICS. *J. Chem. Inf. Comput Sci.* **1985**, *25*, 252–257.

(6) Dubois, J.-E.; Carabedian, M.; Ancian, B. Elucidation Structural Automatique par RMN de Carbon 13: Methode DARC-EPIOS. Recherche d'une Relation Discriminante Structure-Déplacement Chimique (Automatic Structural Elucidation by ^{13}C NMR: DARC-EPIOS Method. Search for a Discriminating Structure Chemical Shift Relationship). *C. R. Hebd. Seances Acad. Sci., Ser. C* **1980**, *290*, 369–372.

(7) Dubois, J. E.; Carabedian, M.; Dagane, I. Computer-Aided Elucidation of Structures by Carbon-13 Nuclear Magnetic Resonance. The DARC-EPIOS Method: Characterisation of Ordered Substructures by Correlating the Chemical Shifts of Their Bonded Carbons. *Anal. Chim. Acta* **1984**, *158*, 217–233.

(8) Bremser, W.; Klier, M.; Meyer, E. Mutual Assignment of Subspectra and Substructures—A Way to Structure Elucidation by C-13 NMR Spectroscopy. *Org. Magn. Reson.* **1975**, *7*, 97–105.

(9) Funatsu, K.; Miyabayashi, N.; Sasaki, S.-I. Further Development of Structure Generation in the Automated Structure Elucidation System CHEMICS. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 18–28.

(10) Christie, B. D.; Munk, M. E. The Role of Two-Dimensional Nuclear Magnetic Resonance Spectroscopy in Computer-Enhanced Structure Elucidation. *J. Am. Chem. Soc.* **1991**, *113*, 3750–3757.

(11) Carabedian, M.; Dubois, J.-E. Single-Resonance Subspectra/Substructure Investigations of the ^{13}C DARC Databank. Representation of Local and Global Topological Knowledge. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 557–564.

(12) Dubois, J.-E.; Carabedian, M. Combined Model of Multi-Resonance Subspectra/Substructure and DARC Topological Structure Representation. Local and Global Knowledge in the ^{13}C NMR DARC Database. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 564–574.

(13) Munk, M. E. The Role of NMR Spectra in Computer Enhanced Structure Elucidation. In *Computer-Enhanced Analytical Spectroscopy*; Jurs, P. C., Ed.; Plenum Press: New York, 1992; Vol. 3, pp 127–148.

(14) Panaye, A.; Doucet, J.-P.; Fan, B. T. Topological Approach to 13C NMR Spectral Simulation: Application to Fuzzy Substructures. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 258–265.

(15) Carhart, R. A.; Smith, D. H.; Gray, N. A. B.; Nourse, J. G.; Djerassi, C. GENOA: A Computer Program for Structure Elucidation Utilizing Overlapping and Alternative Substructures. *J. Org. Chem.* **1981**, *46*, 1708–1718.

(16) Lipkus, A. H.; Munk, M. E. Automated Classification of Candidate Structures for Computer-Assisted Structure Elucidation. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 9–18.

(17) Christie, B. D.; Munk, M. E. Structure Generation by Reduction: A New Strategy for Computer-Assisted Structure Elucidation. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 87–93.

(18) Bangov, I. P. Computer-Assisted Structure Generation from a Gross Formula. 3. Alleviation of the Combinatorial Problem. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 277–289.

(19) Bangov, I. P. Toward the Solution of the Isomorphism Problem in Generation of Chemical Graphs: Generation of Benzenoid Hydrocarbons. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 167–173.

(20) Bangov, I. P. Computer-Assisted Structure Generation from a Gross Formula. 4. Fighting Against Graph Isomorphism Disease. *Commun. Math. Chem. (MATCH)* **1992**, *27*, 3–30.

(21) Bangov, I. P.; Simova, S.; Cabrol-Bass, D.; Laude, I. Computer-Assisted Structure Generation from a Gross Formula. 6. Reducing the Structural Redundancy by Employment of 2D NMR Spectral Information. *J. Chem. Inf. Comput. Sci.*, in press.

(22) Zadeh, L. A. Fuzzy Sets. *Inf. Control* **1965**, *8*, 338–355.

(23) Zadeh, L. A. Fuzzy Sets as a Basis for a Theory of Possibility. *Fuzzy Sets Syst.* **1978**, *1*, 3–28.

(24) Otto, M. Fuzzy Theory Explained. *Chemom. Intell. Lab. Syst.* **1988**, *4*, 101–120.

(25) Rosenfeld, A. Fuzzy Graphs. In *Fuzzy Sets and Their Applications to Cognitive and Decision Processes*; Zadeh, L. A., Fu, K.-S., Tanaka, K., Shimura, M., Eds.; Academic Press: New York, 1974; pp 77–95.

(26) Cabrol, D.; Caire, J. P.; Ozil, P. L'utilisation du langage PROLOG pour la décomposition des grands procédés chimiques (Using Prolog Language for Decomposing Large Chemical Process Flowsheets). *Comput. Chem.* **1988**, *12* (2), 165–170.

(27) Klin, M.; Zefirov, N. S. Group Theoretical Approach to the Investigation of Reaction Graphs for Highly Degenerate Rearrangements of Chemical Compounds. II. Fundamental Concepts. *Commun. Math. Chem. (MATCH)* **1991**, *26*, 171–190.

(28) Read, R. C.; Corneil, D. G.; Derek, G. The Graph Isomorphism Diseases. *J. Graph Theory* **1977**, *1*, 339–363.

(29) Bangov, I. P. Structure Generation from a Gross Formula. 7. Graph Isomorphism: A Consequence of the Vertex Equivalence. Paper presented at the 5th International Conference on Mathematical and Computational Chemistry. *J. Chem. Inf. Comput. Sci.*, in press.

(30) McAllister, M. L. N. Fuzzy Intersection Graphs. *Comput. Math. Appl.* **1988**, *15* (10), 871–886.

(31) Ricard, D.; Cachet, C.; Cabrol-Bass, D.; Forrest, T. P. Neural Network Approach to Structural Feature Recognition from Infrared Spectra. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 202–210.

(32) Pretsch, E.; Clerc, T.; Seibl, J.; Simon, W. *Tables of Spectral Data for Structure Determination of Organic Compounds*; Springer: Berlin, 1989.

(33) Hamilton, W. C. Significance Test on the Crystallographic R Factor. *Acta Crystallogr.* **1965**, *18*, 502–510.

(34) Kalinowski, H.-O.; Berger, S.; Braun, S. *Carbon-13 NMR Spectroscopy*; Wiley: Chichester—New York, 1988; p 223.