Spec2D: A Structure Elucidation System Based on ^1H NMR and H−H COSY Spectra in Organic Chemistry

Hideyuki Masui* and Huixiao Hong[†]

Organic Synthesis Research Laboratory, Sumitomo Chemical Co., Ltd., Osaka 554-8558, Japan

Received July 10, 2005

A system for structure elucidation based on proton NMR spectra has been developed. The system, named Spec2D (system for *spec*tra from *2D*-NMR), incorporates ^1H NMR and H−H correlation spectroscopy (COSY) spectral information obtained from 2D-NMR experiments. 2D-NMR is important for the structure elucidation because it provides information about the relationships among differently situated protons in the structures of unknown compounds. The system uses the concepts of molecular graphs. The improved representation of substructures as well as several novel algorithms for structure generation have been devised to solve the combinatorial problem and to reduce the processing time. Spec2D consists of a knowledge base, an analysis module, and a candidate structure generator module. Spec2D proposes candidate structures from only ^1H NMR and H−H COSY spectral information of an unknown compound without any ^{13}C NMR spectral or structural information, such as molecular formulas. Spec2D has the capability to propose the "new" structure of an unknown compound, if the corresponding substructures are included in the knowledge base.

INTRODUCTION

^1H NMR is more sensitive than other measuring methods and is widely used for lots of molecules. ^1H NMR spectra have not only chemical shifts but also splitting signals or spin systems, which are dependent upon the instrument's magnetic field strength. There are some key issues in archiving ^1H NMR spectra and elucidating structures from the analysis of ^1H NMR spectra: the number of available ^1H NMR databases is smaller than that of ^{13}C NMR, and constructing ^1H NMR databases is costly. Consequently, few systems are able to propose plausible structures just by using ^1H NMR spectra.

Recently, combinatorial chemistry and high-throughput screening methods have been applied to discover compounds with desired properties. The compounds with expected properties should be analyzed to confirm their correct structures.

Spectroscopic methods are widely used to effectively elucidate the structure of organic compounds, including mass, infrared, and nuclear magnetic resonance spectroscopies. Three distinct approaches are used to automatically analyze spectra by using computers: database search, pattern recognition, and artificial intelligence (mainly expert systems). The expert system method elucidates the complete structure of the unknown compound by inferring from a knowledge base of spectra. The successful systems include DENDRAL,[1] CASE,[2] and CHEMICS.[3] However, those systems are not effective in real use.

Neudert and Penk[4] categorized the uses of computers in elucidating structures into three levels: constructing of and retrieving from spectral databases, predicting spectra, and proposing candidate structures by using spectral information of an unknown compound. This study focuses on proposing candidate structures.

This study concentrated on proton NMR spectra because protons have the highest sensitivity, less measuring time than other nuclei, and only a small quantity of a sample is required for the experiment. Although many ^1H NMR spectra are routinely measured, there is less information in the literature on ^1H NMR spectra than on ^{13}C NMR spectra as a result of the complicated patterns of ^1H NMR spectra and the difficulties in data handling. Nevertheless, ^1H NMR spectra have the potential to be effectively used in structure elucidation.

Currently, 2D-NMR spectra are routinely measured for structure elucidation using the pulsed field gradients method. 2D-NMR spectroscopy also has the advantage of acquiring information about the relationships among protons in different environments in the structure of an unknown compound. It is useful to incorporate into the structure elucidation system both the H−H correlation spectroscopy (COSY) spectral information, produced by the fastest experiments in 2D NMR spectroscopy, and the ^1H NMR information.

Commercial systems for the third-level structure elucidation, such as KnowItAll[5] and ACD StrucEluc,[6] are able to accept ^1H NMR spectra, but they mainly use ^{13}C NMR spectra for the structure elucidation. Those systems require a molecular formula or information about connections between atoms to propose candidate structures for unknown compounds. SpecSolv,[7] on the other hand, does not require any molecular formula, as it is based only on ^{13}C NMR spectral information.

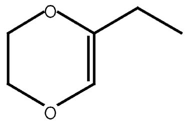
This study focused on developing a method for structure elucidation solely on the basis of NMR information of hydrogen atoms. A database containing ^1H NMR data with

* Corresponding author e-mail: hideyuki-masui@ya.sumitomo-chem.co.jp.

[†] Current address: Division of Bioinformatics, Z-Tech at National Center for Toxicological Research, U.S. Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079.

10.1021/ci0502810 CCC: $33.50
Published on Web 12/23/2005

Chart 1



1

the corresponding structural information was constructed and used by the structure elucidation system Spec2D. Spec2D has been developed to propose the plausible structures for unknown compounds only using ^{1}H NMR and H−H COSY spectral information without ^{13}C NMR spectral data and molecular formulas.

THEORETICAL BASIS

A chemical structure can be treated as a colored graph in which atoms are nodes and bonds between atoms are edges (eq 1).

$$\mathbf{G} = (\mathbf{D}, \mathbf{E}, \mathbf{X}_d, \mathbf{X}_e) \qquad (1)$$

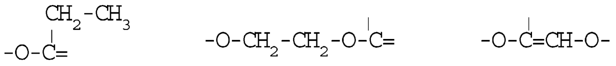
D is a set of nodes in the graph (atoms in a chemical structure), **E** is a set of edges (bonds in a chemical structure), and $\mathbf{X}_d$ and $\mathbf{X}_e$ are the coloring functions for **D** and **E**, respectively. If **D**, **E**, $\mathbf{X}_d$, and $\mathbf{X}_e$ are determined, **G** (chemical structure) is determined.

In H−H COSY NMR spectra, the relationship between individual hydrogen atoms can be revealed by the cross-peak nets, including the "buried" cross-peaks which cannot be clearly observed because of the resolution of the NMR instrument or similar chemical shifts of hydrogen atoms. The structure of a compound can be determined by analyzing its proton NMR. On the basis of its cross-peak nets, a compound can be represented as a colored graph in which substructures (not atoms) with complete cross-peak nets are nodes and substructures (not bonds) which have no NMR spectral information, and by which substructures derived from cross-peak nets are connected together, are edges. Theoretically, from eq 1, a colored graph (**G**) can be generated if all nodes (**D**), edges (**E**), and their coloring functions ($\mathbf{X}_d$ and $\mathbf{X}_e$) are known. Compared with conventional systems such as CHEMICS, in which nodes are heavy atoms and edges are chemical bonds between heavy atoms, Spec2D should be more efficient in generating candidate structures. In Spec2D, substructures are described logically (in concept) and physically (by the computer) to make structure elucidation possible. Spec2D processes include spectral analysis to obtain the substructure constraints and structure generation to build the complete candidate structures from the substructure constraints.

DESCRIPTION OF SUBSTRUCTURES

Logical Description of Substructures. A chemical structure can be described as a combination of atoms and bonds. To generate a chemical structure from a set of atoms and bonds, it is more efficient to integrate bond information into the description of atoms than individual descriptions of atoms and bonds. Structure **1** (Chart 1) is taken as an example to illustrate the difference between two descriptions for structure generation (hydrogen atoms are described implicitly). If individual descriptions of atoms and bonds are used, this compound contains six C atoms, two O atoms, seven single bonds, and one double bond. If bond information is integrated into the description of atoms, structure **1** can be described as three $-C-$, two $-O-$, one $=C-$, one $=C<$, and one $-C$ atom. Structure elucidation systems are expected to generate correct candidate structures from the description of substructures. It can be seen that fewer candidate structures are generated from the description of atoms, combined with bond information, than from individual descriptions of atoms and bonds.

In Spec2D, a chemical structure is described on the basis of fragments with and without proton NMR information, not on the basis of individual atoms and bonds. Fragments with and without proton NMR information are treated as atoms and bonds, respectively. In this way, structure **1** can be treated as an assembly of fragments that have proton NMR information (one CH_3-CH_2-, one $-CH_2-CH_2-$, and one $-CH=$) and a set of fragments that have no proton NMR information (one $-O-$ and one $-O-\overset{|}{C}=$). Fragments are connected together to make substructures. After integration of the fragments, the set of substructures derived from structure **1** contains



The structure generator of Spec2D combines substructures into all possible complete structures using an "overlapping" technique. For example, three substructures derived from structure **1** can be combined into a complete structure (Figure 1). During the combining process, larger substructures are generated by combining the bonds (substructures without a proton atom here) shared by two or more smaller substructures. Figure 1 shows that all three substructures can be used as a starting substructure leading to structure **1**. The number of possible complete structures generated from substructures should be much smaller than that via the methods, which use conventional atoms and bonds.

The combinations shown in Figure 1 are based on the substructures without proton NMR information. This strategy is capable of generating "new" structures from the substructures derived from the "old" structures in the database, provided that all of the substructures in the "new" compound are contained in the "old" structures. For example, if structures **2** and **3** are included in the database and structure **4** is a "new" structure for which the proton NMR spectrum is supplied for structure elucidation, the substructures CH_3-CH_2-O- and $CH_3-CH_2-CH_2-O-$ can be obtained by analyzing the spectrum. Combining partial substructure $-O-$ shared by the substructures **2** and **3** should easily generate the new structure **4** (see Chart 2.). As another example, suppose compounds **5** and **6** are included in the database of Spec2D. Compound **7** is a "new" compound and is not included in the database. It can be seen that all the substructures of structure **7** can be found in compounds **5** and **6**. Therefore, it is expected that structure **7** could be generated from the substructures derived from compounds **5** and **6**, for which proton NMR spectral patterns are similar.

In some cases, more than one candidate structure may be generated. Therefore, all combining paths have to be explored to generate all possible candidate structures from a group of substructures.

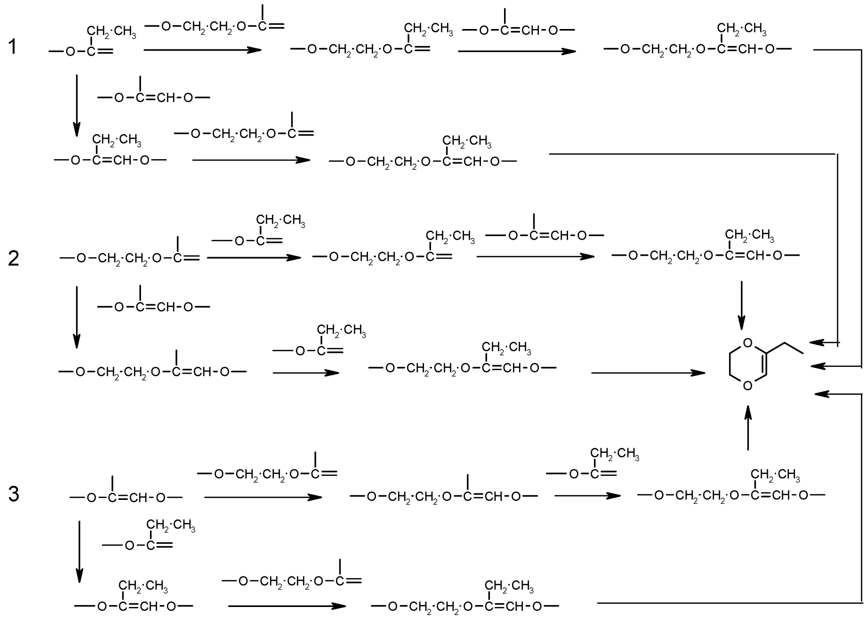


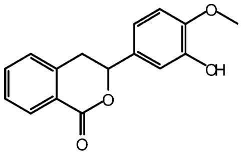
Figure 1. Generating complete structure from extracted substructures.

Chart 2

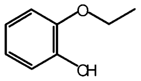
CH_3-CH_2-O-CH_3 **2**

CH_3-CH_2-CH_2-O-CH_3 **3**

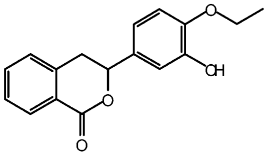
CH_3-CH_2-CH_2-O-CH_2-CH_3 **4**



5



6

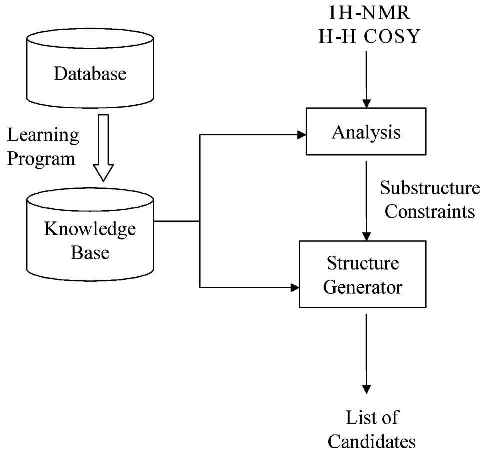


7

Physical Description of Substructures. Two types of fragments can be used to describe a structure: with and without proton NMR information. One fragment has hydrogen atoms attached to all heavy atoms; the other has no attached hydrogen atoms. Integrated substructures were used in Spec2D. Computer methods for developing physical descriptions of substructures are discussed in this section.

Integrated substructures were described by use of a modified connection table in which a chemical bond connected to other substructures might have a pseudoatom as one of its two connecting points. A pseudoatom is not a real atom but a "flexible" atom permitted to match with any type of heavy atom with hydrogen atoms attached to other substructures. In the structure generator, only pseudoatoms in one of two matched substructures need to be combined to heavy atoms with attached hydrogen atoms. Pseudoatoms in the other matched substructure can be matched to all heavy atoms, with or without hydrogen atoms. Partial combination is permitted to guarantee that the system can generate all possible "new" structures. This will be discussed further in the structure generator section.

Scheme 1. Overview of the System



In the connection table, atom order numbers of a substructure are unique to avoid identical substructures being treated differently. Substructures are canonically described in the modified connection tables. Canonical descriptions and the corresponding algorithm will be discussed in the knowledge base section.

OVERVIEW OF THE SPEC2D SYSTEM

Spec2D was designed as an expert system, which can learn from H−H COSY spectral analyses on the spectra in the database. Cross-peak net systems were efficiently used to reduce the number of possible substructure constraints and the time for candidate structure generation. Spec2D aims to emulate a human expert by providing the integrated description of substructures. An overview of the system is shown in Scheme 1.

In Spec2D, a knowledge base has been constructed, consisting of information about the correlations between chemical shifts of ^{1}H NMR spectra and the corresponding substructures. The knowledge base was constructed from the networks of spin−spin coupling information from the cross-peaks in H−H COSY spectral data derived from fully assigned ^{1}H NMR data in the databases. For the structure elucidation, Spec2D requires information on chemical shifts, the number of protons of each signal, and signal connections derived from the cross-peak information of H−H COSY spectra. No molecular formula is required. Requisite information includes only ^{1}H NMR and H−H COSY spectral information of the unknown compound. Spec2D extracts substructures consistent with the input spectral data from the knowledge base. A small number of remaining substructures should be consistent with cross-peak information. Furthermore, substructures are combined into larger substructures using constraints of the input spectral information. Substructures are then combined into complete structures in the final combination procedure. Spec2D calculates scores for each candidate structure by making comparisons with the input spectrum of the unknown compound. Finally, Spec2D proposes and ranks the plausible candidate structures. Details about the design and implementation of all modules will be discussed in the following sections.

DATABASE

A few electronic databases of ^{1}H NMR spectra are available. In this study, the published ^{1}H NMR spectral data in books and journals and our measured ones have been collected into a database. The database contains basic organic compounds, chemical reagents, dyestuffs, additives, intermediates of agricultural chemicals, and natural compounds except polymers. The largest number of heavy atoms of a compound in the database is 107, and its molecular formula is $C_{79}H_{140}N_4O_{22}S_2$ (MW: 1562.14Da). Measured frequencies of ^{1}H NMR spectral data in the database are 60−600 MHz. The database has no stereochemistry in the old data. However, the newly measured data have the stereochemical information. The database has about 17 000 ^{1}H NMR spectra. All ^{1}H NMR signals are fully assigned to the corresponding heavy atoms in the structures. The database provides training data for Spec2D.

The database in NMfile (*N*uclear *M*agnetic resonance spectra assignment information *file*) format[8] contains atom and bond information for the structures and the corresponding NMR spectral data. Representation of chemical structures in the database is based on Molfile. In Molfile, only heavy atoms are recorded explicitly to save space. Hydrogen atoms are implicitly represented. Any software using Molfile needs to derive the number of hydrogen atoms attached to a heavy atom by calculating the connectivity and valence of each heavy atom. Some atoms, such as N and S, have different valences in different structures, so care must be taken when calculating the number of hydrogen atoms attached to them. Proton NMR spectral analysis describes the relationship between spectral patterns and hydrogen atom patterns. Correct hydrogen atoms should be derived from Molfile or NMfile in order to guarantee the correct substructures are derived from the database.

By increasing the number and diversity of chemical species in the database, Spec2D's ability to propose plausible candidate structures will be improved because the knowledge base is dependent upon the database.

DATABASE TO KNOWLEDGE BASE

A knowledge base contains a set of substructures and their corresponding spectral features. **S** is used to represent the set of substructures (eq 2).

$$\mathbf{S} = [S_i] \qquad i = 1, 2, ..., n \tag{2}$$

S_i is the *i*th substructure, and n is the number of knowledge rules or substructures in the knowledge base. Suppose the corresponding spectral feature in the 2D proton NMR of substructure S_i consists of a set of peaks P_{S_i} and a cross-peak net $\mathrm{CP}_{S_{i.}}$

$$P_{S_i} = [P_j] \qquad j = 1, 2, ..., m \tag{3}$$

m is the number of peaks of the spectral feature of

$$\mathrm{CP}_{S_i} = [P_k©P_l] \qquad k \neq l \tag{4}$$

$$P_k \in P_{S_i} \text{and } P_l \in P_{S_{i<}} \tag{5}$$

substructure S_i. $P_k©P_l$ indicates that there should be a cross-peak between the *k*th peak P_k and the *l*th peak P_l in the H−H COSY spectra. ∈ means "is included in".

Suppose 2D proton NMR spectra of a compound have a set of peaks **P** and a set of cross-peak nets CP. To generate possible candidate structures for the compound, possible substructures should be extracted from 2D proton NMR spectra. To determine if a substructure is possibly contained in a structure, the relationship in eq 6 is used.

$$P_{S_i} \subseteq \mathbf{P} \wedge \mathrm{CP}_{S_i} \subseteq \mathrm{CP} \rightarrow S_i \tag{6}$$

To derive substructure constraints for the structure generator from the input spectral data, spectral feature−substructure relationships need to be extracted. The set of substructures $\mathbf{S} = [S_i]$ and their corresponding spectral feature P_{S_i} and CP_{Si} need to be defined. There are two ways to define spectral feature−substructure relationships: by human experts or a database method. If enough data on structures and spectra exist in a database, the spectral feature−substructure relationships can be determined. In Spec2D, the spectral feature−substructure relationships are extracted from the database of proton NMR spectra. The algorithm and main procedures for converting the database to the knowledge base are discussed in the next section.

The knowledge base consists of three parts. The first part contains connection data and other structural information such as atom names and bond types, which describe the substructure. Only heavy atoms were recorded. The second part consists of expected ^{1}H NMR and H−H COSY spectral patterns for different substructures, generated by analyzing the whole database. The number of peaks, the value of chemical shifts, and the number of hydrogen atoms corresponding to each peak are main components of the spectral feature part of the knowledge rules. The knowledge base is constructed on the basis of cross-peak nets. One knowledge rule has one complete cross-peak net. The last part is the structure environment requirement of the substructures. This was designed for guiding structure generation (details in the structure generation section).

Two kinds of substructures are in the knowledge base of Spec2D: chain and aliphatic ring substructures (called chain substructures) and aromatic ring substructures. The perception of ring systems is based on the algorithm described in ref 9. Chain substructures have only one cross-peak net. Each

Promise: (Spectral Feature)

Peak No.	Chemical Shift	Dev.	Number of H atoms
1	7.460	0.100	1
2	7.570	0.110	1
3	7.310	0.090	1
4	8.150	0.180	1

Goal: (Substructure)
Connection Table of the following substructure

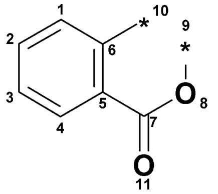


Figure 2. Spectral feature−substructure relationship rule. Dev.: deviation of chemical shifts in the knowledge base.

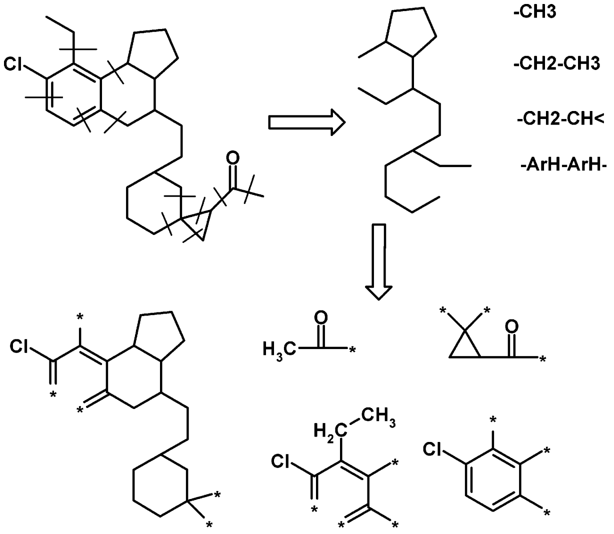


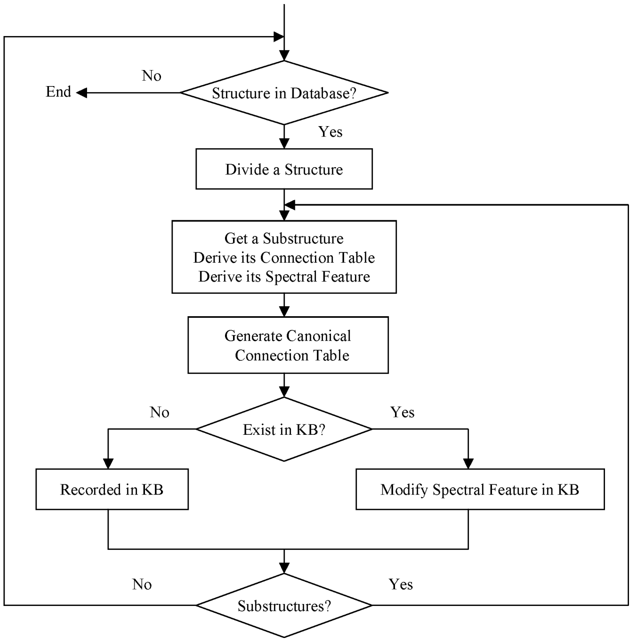
Figure 3. Breaking a structure into substructures. Dotted line: breaking point. *: pseudoatom. Ar: aromatic carbon.

kind of proton must be cross-peak linked to other kinds of protons in a chain substructure. Aromatic ring substructures may have more than one cross-peak net. Some protons may have no cross-peak to other protons in the same ring substructure. If some structures have no cross-peak, meaning there is only one peak in the substructure, the system is able to register it in the knowledge base as a substructure.

The program to build the knowledge base extracts all possible substructures and ^{1}H NMR spectral patterns from the database. No substructures can be duplicated in the knowledge base. ^{1}H NMR spectral information of each compound is integrated into the representation of its substructures in the knowledge base.

The conventional approach for spectral interpretation is based upon simplified models of the physical process underlying resonance and resulting spectral absorption. Those physical models permit specific spectral signals to be related to structural components of the molecule. Initial analysis of a spectral signal may identify a specific substructure in the unknown compound, and a more detailed analysis may determine aspects of the substructure environment. In Spec2D, the knowledge of 2D proton NMR spectral analysis was encoded in the form of spectral feature−substructure relationship rules, which comprise the knowledge base.

Scheme 2. Overview of Knowledge Base Creation Program[a]



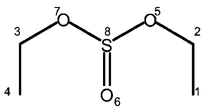
[a] DB: database. KB: knowledge base.

One of the rules is illustrated in Figure 2. In this “production” rule, the *Goal* (conclusion) is considered to be true only when the *Promise* is satisfied. All peaks in the *Promise* of a knowledge rule are in the same cross-peak net, and the detailed cross-peak net can be found in the connection table of the substructure. The peak number is labeled consistently with the atom order of heavy atoms to which corresponding hydrogen atoms are attached. To give an example of this rule, if four peaks exist around chemical shifts 7.46, 7.57, 7.31, and 8.15 ppm, with only one hydrogen atom corresponding to all peaks, and cross-peaks 7.46−7.57, 7.57−7.31, and 7.31−8.15 are found in 2D proton NMR spectra of an unknown compound, the substructure shown in Figure 2 is possible.

Describing substructures in Spec2D is based on cross-peak nets. Structural parts that have no hydrogen atoms are treated as connection parts, such as chemical bonds. The program converting the database to the knowledge base first extracts cross-peak nets in a structure, followed by connection tables of substructures and their spectral features. An overview of the program used to create the knowledge base is shown in Scheme 2.

Dividing a structure into the corresponding substructures consists of two steps. The first step is to break bonds between heavy atoms with and without attached hydrogen atoms to obtain structural parts with complete cross-peak nets. The second step is to combine the structural parts without hydrogen atoms and those with complete cross-peak nets, to give complete substructures. Figure 3 provides an example of how a structure can be broken into substructures in an integrated way. The structure is first broken into five parts, which correspond to cross-peak nets. This is done by breaking bonds between heavy atoms with and without attached hydrogen atoms, as shown by the dotted lines in Figure 3. The second step integrates the parts with and

Chart 3



8

Table 1. Assignments of Compound **8**

atom number	chemical shift (ppm)
1	1.260
2	4.030
2	3.970
3	4.030
3	3.970
4	1.260

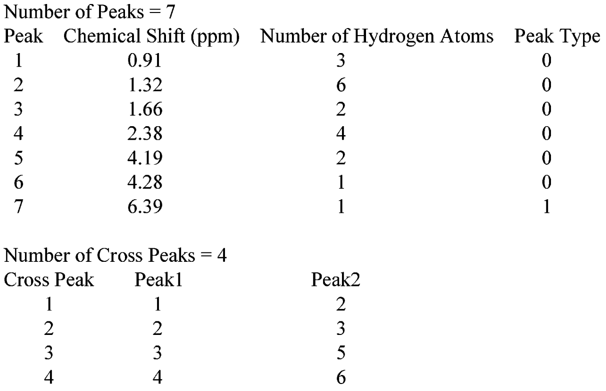
without attached hydrogen atoms, generating five integrated substructures in which * indicates a pseudoatom.

After the integrated substructures are extracted, their corresponding spectral features are derived from the spectral data. All hydrogen atoms are assigned to individual peaks on the basis of the spectral data recorded in the database.

Some ^{1}H NMR spectra with fine splitting peaks from couplings were recorded in the database. Therefore, the learning program has to treat this information carefully because the user would be required to input the fine splitting of ^{1}H NMR. For consistence and ease, the fine splitting when chemical shifts are similar should not be represented in the knowledge rules. For example, compound **8** (Chart 3) has two methylene groups, which possess fine splitting. In the database, peaks in ^{1}H NMR spectra of this compound were assigned to methyl and methylene groups (Table 1).

If the knowledge rules about methylene groups had two distinct peaks for each, the user would need to input two distinct peaks, but sometimes it is difficult to identify distinct peaks with a difference of only 0.06 ppm. The software accounts for this because if the difference in chemical shift between two peaks in a fine structure is too small, the average value of the chemical shift is used.

Before incorporating a substructure and its spectral feature into a knowledge rule, the program needs to determine if the substructure already exists in the knowledge base. It is very time-consuming to check if two substructures are isomorphic by conventional atom-to-atom matching methods when the order number of atoms is different. The canonical order number for atoms in a substructure has to be determined before making comparisons. The canonical order number of atoms in a substructure can be obtained using the algorithm described in ref 10. Usually, after the canonical order of atoms in a substructure is obtained, comparisons can be made with atoms of the same order number in different substructures. If all atoms with the same order number in two substructures are same, they are believed to be the same substructures. As the number of substructures in the knowledge base increases, the time to locate a substructure in the knowledge base increases. Because all substructures generated from structures in the database have to be checked to guarantee no duplicate substructures are recorded in the knowledge base, computing time is significant. To reduce the computing time, a characteristic number (CN) was used, which incorporates the information such as number of heavy atoms, hydrogen atoms, unsaturation, unconnected valences, and type of substructure. Only when two CNs of substructures are the same is the atom-to-atom comparison conducted.



```
Number of Peaks = 7
Peak   Chemical Shift (ppm)   Number of Hydrogen Atoms   Peak Type
 1          0.91                        3                    0
 2          1.32                        6                    0
 3          1.66                        2                    0
 4          2.38                        4                    0
 5          4.19                        2                    0
 6          4.28                        1                    0
 7          6.39                        1                    1

Number of Cross Peaks = 4
Cross Peak   Peak1          Peak2
    1          1              2
    2          2              3
    3          3              5
    4          4              6
```

Figure 4. Example of input data format. Chemical shift: mean value. Peak type: is (1) or is not (0) an exchangeable hydrogen atom.

SPECTRAL ANALYSIS

The first step in the structure elucidation is to analyze spectra of an unknown compound, to obtain the substructure constraints. Substructure constraints are derived from an analysis of ^{1}H NMR spectra, including cross-peak data, using the knowledge base.

Before the analysis, chemical shifts and the number of protons of each signal in the ^{1}H NMR spectrum, together with the H−H COSY information, need to be input into the system in a digital format. An example is shown in Figure 4. Peak type is defined as either an exchangeable hydrogen atom (1) or not (0). If it is not clear, peak type should be 0.

Users can input information about some elemental compositions. For example, if a user wants to specify the range of an atom number, constraints can be set to speed up the analysis. If no information about structure is specified, Spec2D derives the substructure constraints solely from the input ^{1}H NMR and H−H COSY spectra.

The goal of the analysis module is to find the subset of substructures in the knowledge base with spectral features consistent with the input peaks **P** (chemical shifts) and the cross-peak nets **CP** of the unknown compound. All substructures with spectral features satisfying eq 6 should be taken from the knowledge base as possible substructure constraints. The output of this module includes the possible substructures and the number of substructures calculated from the number of hydrogen atoms of the input peaks. Because a structure may have more than one copy of the same substructure, it is important for the structure generator to have the possible number of substructures. An overview of the analysis module is shown in Scheme 3.

A target-driving strategy was used in the analysis module. Spectral features corresponding to the substructures in the knowledge base are considered as the targets and matched to the input ^{1}H NMR spectrum and cross-peak net of the unknown compound. If the ^{1}H NMR spectral feature of a substructure in the knowledge base is consistent with the input data, the substructure may be contained in the unknown compound.

The program to check the existence of individual substructures in the structure of an unknown compound adopts the depth-first strategy. In this strategy, the prerequisites for

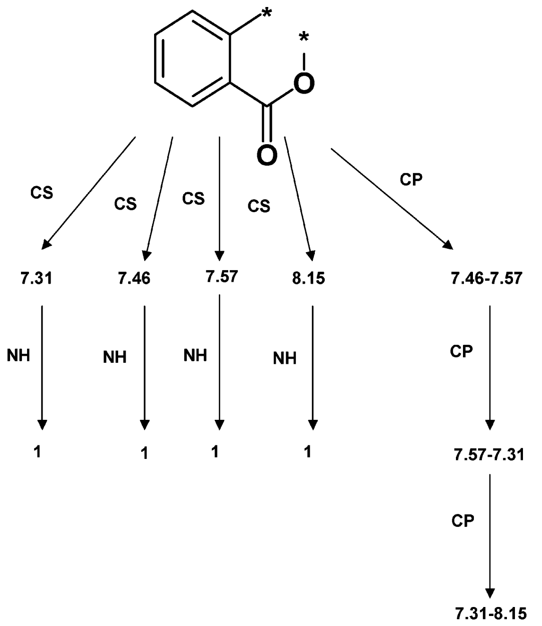
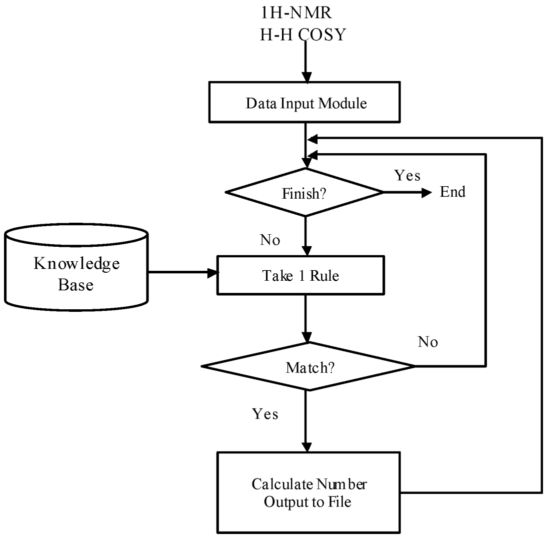


Figure 5. Inferring network of a substructure. CS: chemical shift. CP: cross peak. NH: number of hydrogen atoms.

Scheme 3. Overview of Analysis Module



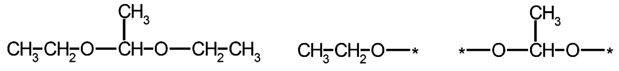
a target in the same branch are tried consecutively. If one of them is false, the target is discarded. For example, an inferring network is shown in Figure 5. The program first tries to find the peak with a chemical shift at around 7.31 ppm. If this chemical shift cannot be found in the input data, the substructure is discarded. If the peak with chemical shift at around 7.31 ppm is found, the number of hydrogen atoms corresponding to the peak is then checked. Sometimes the number of hydrogen atoms is inconsistent, requiring the substructure to be discarded. If the number of hydrogen atoms for a peak in the spectral feature of a knowledge rule is equal to (nonoverlapped peak) or less than (overlapped peak) the input number of hydrogen atoms corresponding to the matched peak, it is a possible substructure. After checking the peak at 7.31 ppm, peaks at 7.46, 7.57, and 8.15 ppm are checked in a similar fashion. If all chemical shifts and the number of hydrogen atoms are consistent with the input data, a search is conducted for the expected cross-peak linkage net. After all prerequisites for the existence of a substructure are verified, the substructure should be incorporated into the structure being generated.

When an expected peak in the spectral feature of a knowledge rule is matched with the input data, it is not necessary for the expected and matched peaks to be exactly the same chemical shift. If the absolute value of the difference in the chemical shift between the expected peak and the input peak is less than or equal to the deviation value in the knowledge rule, the expected peak is successfully matched to the input peak. The chemical shift of a hydrogen atom in a proton NMR spectrum is determined by its global structural environment in a compound, whereas a substructure is in a local structural environment. Hydrogen atoms in the same local structural environment (substructure) with different global structural environments (structures) should have similar but not identical chemical shifts in the proton NMR spectra. Therefore, matching a knowledge rule with the input peaks should have a tolerance for chemical shifts to account for the differences in global structural environments. Spec2D uses a default tolerance between 0.1 and 0.2 ppm.

The number of hydrogen atoms of individual peaks in the input data is used in the analysis process. It does not have to be exactly matched with one kind of proton in a substructure. However, the number of hydrogen atoms with a certain chemical shift range in the input data has to be greater than or equal to the expected values in the matched knowledge rule. If it is larger than the expected number of hydrogen atoms in the matched knowledge rule, overlapping of some peaks may exist in the input spectra because of similar chemical shifts. For the input peak, which is matched to more than one expected type of hydrogen atom in the substructures, special attention has to be paid to the matching of the cross-peak linkage net between the expected and the input data. In this case, the number of cross-peaks in the input data might be less than the expected data because of overlapping of the peaks. Actually, the cross-peak between two peaks with similar chemical shifts is “buried” in the spectrum. Therefore, an overlapped peak (usually with many hydrogen atoms) can have cross-peaks within it and can be matched to peaks to which it has multiple cross-peak links.

Cross-peaks from 2D NMR spectra can be used to determine relative distances between coupled hydrogen atoms. This information is fully used in Spec2D. After the chemical shifts and the corresponding number of hydrogen atoms of a substructure have been matched with the input data, the expected cross-peak net of the substructure in the knowledge rule needs to be checked with the input cross-peak nets. All peaks on a cross-peak net need to be checked individually. To compare the cross-peak patterns of the expected peaks with the input data, two different cases may occur which require different strategies. In the first case, the expected peak in a knowledge rule is matched with an input peak with the same number of hydrogen atoms. The expected cross-peak pattern in the knowledge rule must be exactly the same as that in the input data. In the second case, the expected peak in a knowledge rule is matched with an input peak with more hydrogen atoms. Checking cross-peak patterns for the expected peak is more flexible than in the first case. Cross-peaks of the expected peak do not have to be found in the input data. An input peak which has been matched with the expected peak may have other cross-peaks

Chart 4



9 10 11

if the number of hydrogen atoms is more than that of the expected peak.

Information density in proton NMR spectra is high, usually within a range of 15 ppm of a chemical shift. Therefore, a proton NMR spectrum is very crowded with frequent peaks, leading to peaks that overlap to give a broad peak. This makes it difficult to explain proton NMR spectra and to elucidate chemical structures. Special attention needs to be paid to the overlapped peaks in a structure elucidation system based on proton NMR spectral analysis.

The strategy of Spec2D allows an overlapped peak to be input as one independent peak, and the correct substructures can be derived from the analysis of the overlapped peak. As the expected cross-peak pattern is checked, the system knows that some cross-peaks are buried in the overlapped input peak. Therefore, the expected cross-peak pattern can be matched with the input data, and the substructure is treated as a possible constraint for the structure generator.

A structure might contain multiple copies of the same substructure. The spectral feature of this substructure should overlap to increase the corresponding number of hydrogen atoms. Knowing the number of copies of a substructure is very important for the structure generator. The analysis module should extract out this information from the input data. For example, from the proton NMR spectrum of compound **9**, the substructures **10** and **11** can be derived (Chart 4). Accounting for the number of hydrogen atoms corresponding to the peaks, the analysis module should determine that there are two copies of the **10** substructure. The structure generator then tries to combine **11** and two copies of **10** to produce the correct structure as a candidate structure.

In checking the expected cross-peak pattern in the input data, special attention needs to be paid to the hydrogen atoms attached to heteroatoms such as O, N, and S, because the hydrogen atoms attached to heteroatoms sometimes have no cross-peaks with their vicinal hydrogen atoms. The analysis module should deal with those hydrogen atoms as special cases.

The output from the analysis module of Spec2D consists of substructures, matched peaks, and the corresponding number of substructures. The connection table of a substructure, the corresponding spectral features, the chemical shifts, and the number of hydrogen atoms are used in the structure generator. The structure generator operates on the connection tables. The input data, including the chemical shifts, the number of hydrogen atoms, and the cross-peaks, are also used to guide the structure generation.

STRUCTURE GENERATOR

The function of the structure generator of Spec2D is to generate all possible candidate structures that are consistent with the input spectral data of an unknown compound. Therefore, the substructure constraints resulting from the analysis module and from the input spectral data guide the generation of candidate structures.

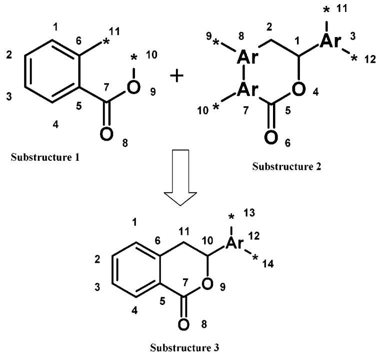


Figure 6. Example of combining. Ar: aromatic atom. *: Pseudoatom.

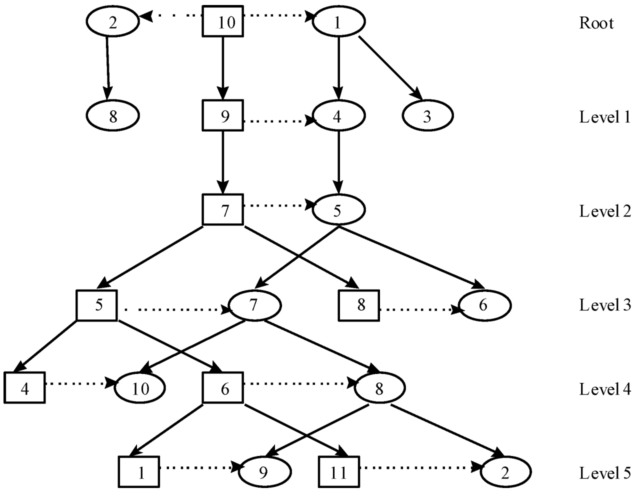


Figure 7. Level extended-match method. Rectangle: heavy atoms in one substructure. Ellipse: heavy atoms in the other substructure. Solid arrows: extension from an atom to an atom in the next level. Dotted arrows: match between two atoms.

The method for structure generation involves combining the substructures to generate larger substructures until the complete candidate structures are obtained. A substructure consists of two parts: one with and one without hydrogen atoms. Combining is only conducted on the part without hydrogen atoms. A special case involves pseudoatoms, which can be combined onto the part with hydrogen atoms. Combining for structure generation is based not only on the structural view of substructures but also on spectral data. Consistency of the input spectral data should be checked before combining is conducted. If two substructures conflict with the input spectral data, they do not need to be combined. This reduces computing time in comparison with the random combination.

Combining between two substructures uses the level extend-match method, starting from pseudoatoms in one substructure as root points, and the process is terminated when all atoms between pseudoatoms and the atoms with attached hydrogen atoms are successfully combined. By considering substructures 1 and 2 (Figure 6), the level extend-match method is explained in Figure 7.

In Figure 7, rectangles are heavy atoms in one substructure and ellipses are heavy atoms in the other substructure. Vertical arrows (solid lines) are extensions from atoms to atoms in the next level, and horizontal arrows (dotted lines) are matches between two atoms. In the matching process, different rules are used to determine if a match is successful.

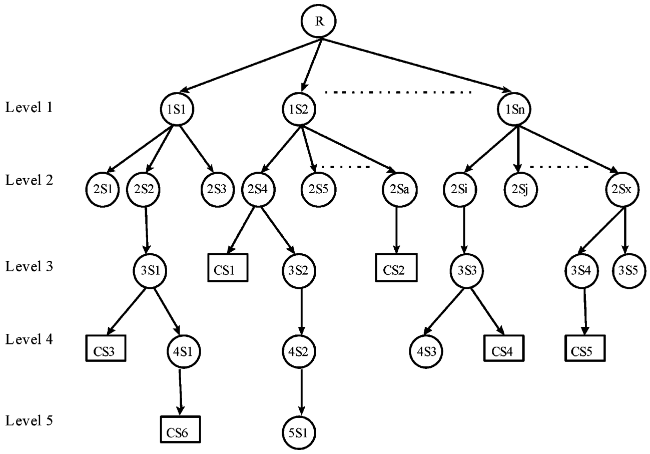


Figure 8. Structure generation tree. R: root substructure. S: combined substructure. CS: complete structure.

Pseudoatoms can only match to the heavy atoms with attached hydrogen atoms, which are also connected to the heavy atoms without attached hydrogen atoms. Heavy atoms without attached hydrogen atoms can only be matched to the same kinds of atoms in other substructures. The connectivity of an atom is also used to guide the match. All pseudoatoms can be selected as root nodes in the level extend-match method. For example, in substructure 1, pseudoatoms 10 and 11 can both be used as root nodes. When matching atoms, all possible matches should be tried. For example, in Figure 7, pseudoatom 10 in substructure 1 can match with atoms 1 and 2 in substructure 2. However, in the next level, atom 9 in substructure 1 is not able to match with atom 8 in substructure 2. Therefore, the path with pseudoatom 10 matching atom 2 is discarded.

The main algorithm for structure generation is the searching algorithm of the generation forest. This consists of individual generation trees in which candidate structures of an unknown compound are considered as leaves with a complete structure generated by combining different substructures. Partial structures, which cannot be grown further, are also considered as leaves. Structure generation involves searching the generation forest to get all the leaves. Searching a generation tree is the fundamental algorithm of the structure generator. Searching a generation forest involves repeating the process of searching a generation tree. Starting from the root, a generation tree is generated by growing each node and backtracking nonleaf nodes. A leaf node is a complete structure or a partial structure, which cannot combine with other substructures.

Searching a generation tree is conducted using the depth-first strategy and the backtracking technique, which guarantee that the whole tree is searched. For example, a generation tree is shown in Figure 8. Suppose one substructure is selected as the root. Theoretically, any substructure can be selected as the root; so, all substructures obtained from the analysis module are selected as roots. In reality, not all of these substructures need to be tried. Information from the input data and the knowledge base guide the searching process. Information includes the input chemical shift and the expected chemical shift in a substructure. Substructures with expected chemical shifts more similar to the input chemical shifts are used as roots of generation trees to reduce computing time. This information can also be used to guide the whole process of searching the structure generation forest.

Figure 8 is an imaginary structure generation tree. Circles represent substructures from the analysis module or larger substructures generated in the process. Rectangles represent complete candidate structures. Starting with the root node, possible combining paths with other substructures are searched, generating larger substructures 1S1, 1S2, ..., and 1S*n*. The search uses the depth-first strategy. Therefore, after 1S1 is generated, the system does not try to generate the partial structure 1S2 but searches the descendants of 1S1 first. Alternatively, the system generates larger partial structures 2S1, 2S2, and 2S3 by combining 1S1 with other substructures. The partial structure 2S1 should be tried first, to combine with other substructures. If it is impossible to combine with other substructures to generate either larger partial structures or a complete structure, the system then backtracks to its ancestor 1S1. The next descendant of 1S1, partial substructure 2S2, is then tried. 2S2 can combine with a substructure to generate 3S1, which is able to combine with a substructure to generate a complete structure CS3. The system cannot stop here. To search all possible candidates, it has to backtrack from CS3 to its ancestor, 3S1, which can combine with another substructure, obtaining 4S1. 4S1 can further combine with a substructure, generating CS6. The system backtracks to ancestors 4S1, 3S1, 2S2, and finally 1S1. After searching 2S3, it begins to backtrack to 1S1 again and finally to the root R. In a similar way, CS1, CS2, CS4, and CS5 are generated.

Each node in a structure generation tree has a father node (except the root), brother substructures, and son substructures. If a partial structure has no son substructure or if a complete structure has inconsistent spectra with the input spectral data, the node is a dead leaf, and the path from the root to this leaf is called a dead branch and is discarded. If a node does not have son substructures and is a complete structure having consistent spectra with the input spectral data, the node is a real leaf of the tree, and the path from the root to this leaf is a living branch. Real leaves are candidate structures. Before combining a node with its son substructures to generate larger partial substructures, the system needs to create a set of son substructures and a set of brother substructures, based on its father node and the input spectral data. Backtracking from a node involves combining its father node with one of its brother substructures.

As an example, suppose proton NMR spectra are input into the system as shown in Figure 4 with seven peaks and four cross-peak nets. From the analysis of the input NMR spectra by the analysis module, a set of substructure constraints is obtained. Several of them are shown in Figure 9, in which rectangles indicate the matched spectral patterns, the input peak numbers, and the corresponding number of hydrogen atoms.

In the structure generator, if **S7** (Figure 9) is selected as the root (focus), **S8** should be treated as the brother substructure to construct another generation tree because **S7** and **S8** share some peaks. These two substructures cannot be put into the same structure because the expected NMR spectral features are inconsistent with the input spectral data. Therefore, they have to be brothers, and two different generation trees or branches are constructed if they are not roots. **S1**, **S2**, **S3**, **S4**, **S5**, and **S6** should be treated as possible son substructures, which can be combined with **S7** and **S8**.

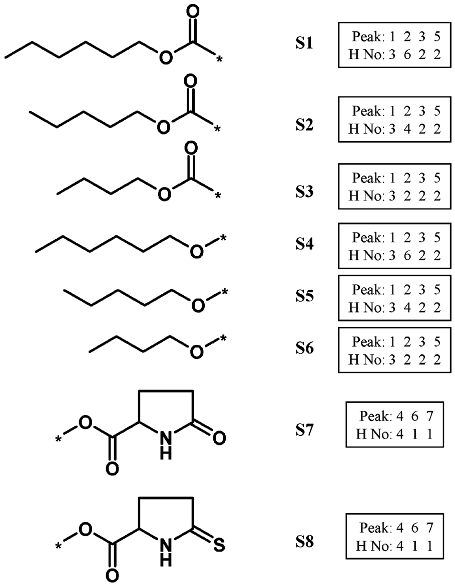


Figure 9. Substructures with matched spectral pattern from analysis module. Peak: matched input peak number. H No: corresponding numbers of hydrogen atoms.

Chart 5

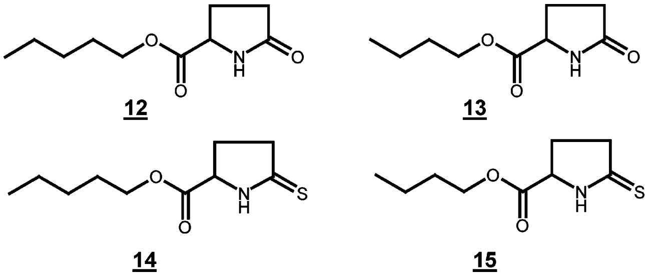
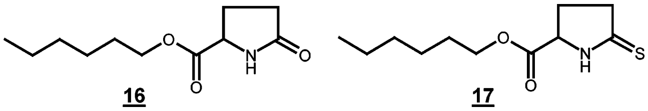


Chart 6



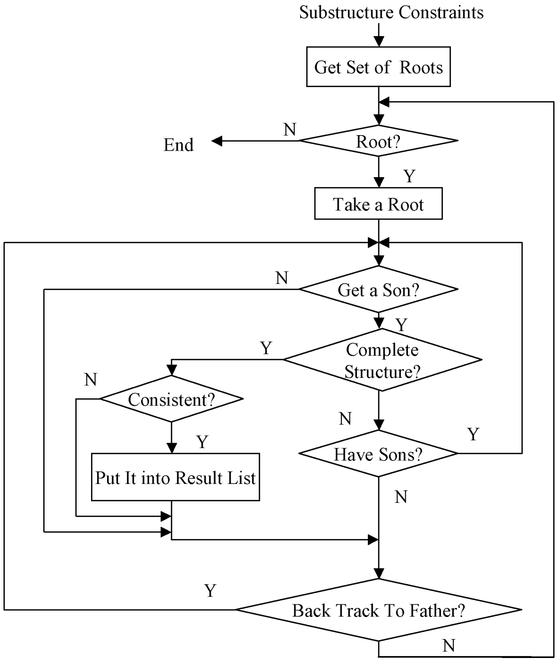
However, **S4**, **S5**, and **S6** cannot combine with **S7** and **S8** to generate larger substructures or complete structures. When **S2** and **S3** combine with **S7** and **S8**, respectively, complete structures **12**, **13**, **14**, and **15** are generated (see Chart 5). However, these structures possess expected spectral patterns which conflict with the input spectral data. Therefore, they are dead leaves.

When **S1** combines with **S7** and **S8**, the complete structures **16** and **17** are generated (Chart 6). Both structures are consistent with the input spectral data. Therefore, both are candidate structures.

In this example, the brothers are the roots, which are used as points for backtracking. Son substructures **S1**, **S2**, **S3**, **S4**, **S5**, and **S6** are brothers if partial substructures generated by combining them with their father substructures have descendants, because they share some input peaks. As a convention, brother substructures cannot appear in the same complete structure, and cannot be used as son substructures of each other. This constraint reduces the computing time.

The structure generator of Spec2D constructs the generation forest guided by the input spectral data using the algorithm described above. An overview of the process is shown in Scheme 4.

Scheme 4. Overview of Structure Generator



After generating candidates, the system predicts the chemical shifts for each candidate, compares them with the input spectrum of the unknown compound, and calculates scores describing similarities between them. A score is assigned to a complete structure as the verification result in Spec2D, on the basis of the consistence of the chemical shift values of the complete structure with the input NMR spectral data. We use the average value of chemical shift of the NMR peak and its deviation, which can be extracted from the knowledge base. The equation used to calculate the score in Spec2D is as in the following:

$$\text{Score} = 100\left\{1 - \sum\left[\left(\frac{A - S}{\text{MCSD}}\right)^2 \left(\frac{1 + \text{Dev}}{1 + \text{MD}}\right)\text{NH}\right]_i \frac{(10\text{Mx} - \text{NSS})}{\sum[\text{NH}]_i 10\text{Mx}}\right\} \quad (7)$$

where $i = 1, 2, ..., m$ is the number of peaks of a complete structure; NH is the number of hydrogen atoms corresponding to a peak; A is the average chemical shift value in the knowledge base; S is the chemical shift value of the input peak matched to the peak in the complete structure; Dev is half of the chemical shift range for a peak extracted from the knowledge base; MD is the maximum of half of the chemical shift range for a peak, temporarily postulated to be 0.4; Mx is the maximum of number of compounds from which a substructure is extracted in the knowledge base building up module, temporarily postulated to be 50; NSS is the number of the same substructure obtained in the knowledge base building up module; the maximum of NSS is set to 50; therefore, if it is more than 50 in the knowledge base, it will be set to 50. MCSD is the maximum of the chemical shift difference allowed between the average value in the knowledge base and the matched peak in the input spectral data, temporarily postulated to be 0.6. The most plausible structure is ranked at the highest or at a higher position in the list of candidates.

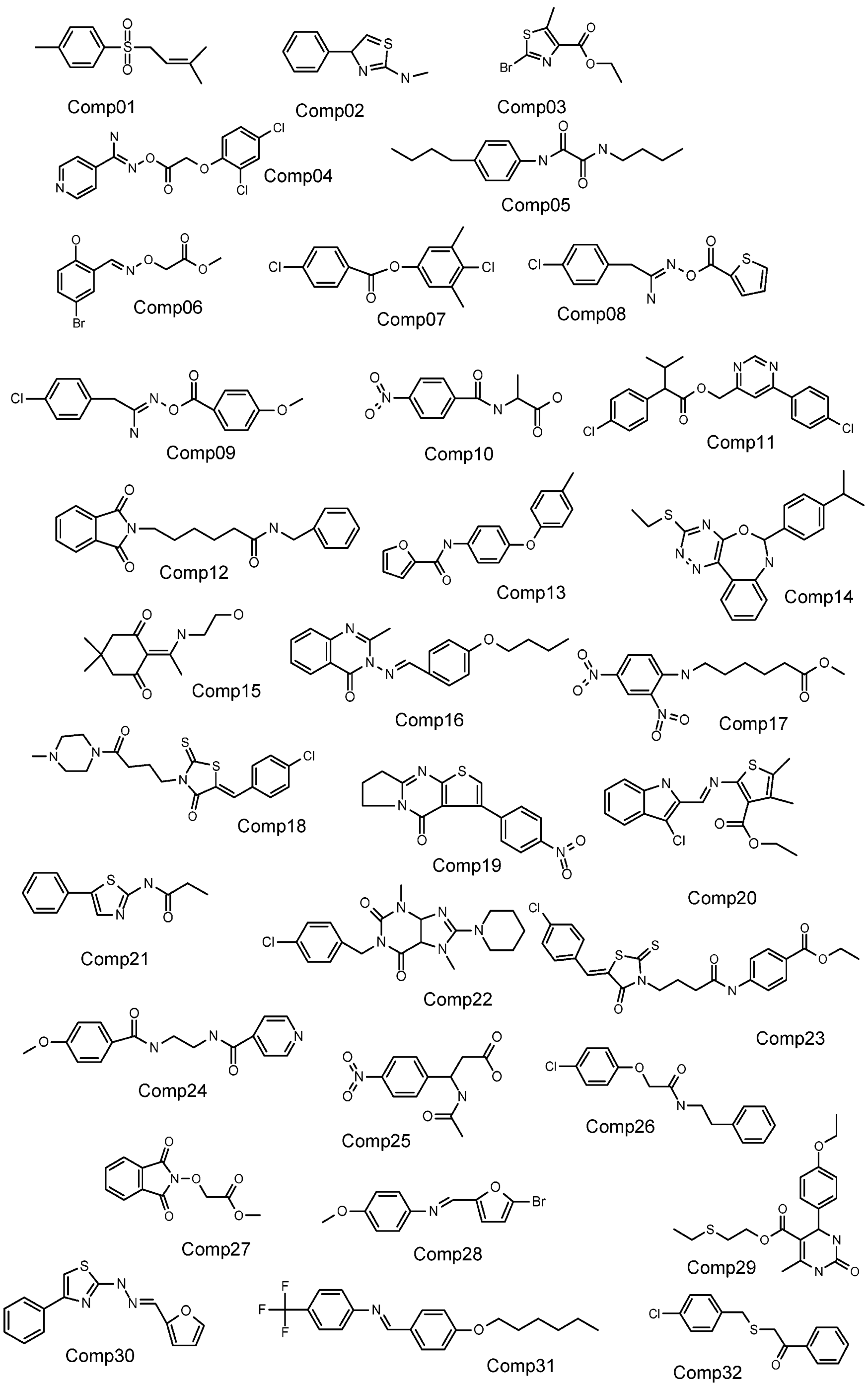


Figure 10. Structures for test examples.

IMPLEMENTATION

Spec2D has been developed on a UNIX-based operating system and installed on workstations such as O2 from Silicon Graphics. Spec2D programs have been developed using ANSI C. A Windows version of Spec2D has also been developed, running on Windows NT, 2000, and XP.

RESULTS AND DISCUSSION

Comp11 in Figure 10 was used as an unknown compound to demonstrate the structure elucidation process for Spec2D. The input data are shown in Figure 11, including nine peaks in the ^{1}H NMR spectrum and four cross-peaks in the H−H COSY spectrum. The analysis module extracted 340 substructures, consistent with the input spectral data (Table 2). One substructure was selected as the starting substructure ("focus"). Four candidate structures were proposed (Figure 12). According to the similarity scores, the correct structure is ranked above other candidate structures. The second ranked candidate is an isomer of the correct one. The third is a compound with one of the chlorines substituted by a nitrile group. As the system does not need a molecular formula, it is able to propose compounds with different molecular formulas. They are proposed, if consistent with the input spectral information. Although the fourth candidate has a peak with an 11.62 ppm chemical shift in the ^{1}H NMR spectrum corresponding to a carboxylic acid, it is still proposed as a candidate because of the system's flexibility

Number of Peaks = 9

Peak	Chemical Shift	Number of Hydrogen Atoms	Peak Type
1	0.700	3	0
2	1.040	3	0
3	2.310	1	0
4	3.540	1	0
5	5.230	2	0
6	7.420	4	0
7	7.600	3	0
8	7.970	2	0
9	9.150	1	0

Number of Cross Peaks = 4

Cross Peak	Peak1	Peak2
1	1	3
2	2	3
3	3	4
4	7	8

Figure 11. Input data of Comp11 to the Spec2D system.

Table 2. Results for Test Examples

sample	number of extracted substructures	number of focuses	number of candidates	processing time, s
Comp01	222	2	1	39
Comp02	57	20	1	10
Comp03	78	15	2	10
Comp04	150	8	2	11
Comp05	78	5	1	10
Comp06	198	14	2	16
Comp07	133	12	1	12
Comp08	112	12	1	11
Comp09	322	16	1	16
Comp10	36	6	1	9
Comp11	340	1	4	67
Comp12	446	2	1	366
Comp13	436	9	2	37
Comp14	280	1	2	11
Comp15	200	17	1	380
Comp16	387	4	1	16
Comp17	177	3	1	10
Comp18	650	1	4	38
Comp19	267	3	1	10
Comp20	495	24	1	34
Comp21	89	10	1	10
Comp22	521	5	9	1881
Comp23	477	2	14	122
Comp24	259	12	1	15
Comp25	120	2	1	9
Comp26	313	6	14	428
Comp27	182	9	8	15
Comp28	156	17	1	13
Comp29	453	13	2	173
Comp30	108	5	1	10
Comp31	85	5	1	11
Comp32	116	26	8	18
average	248.2	9.0	2.9	119.3

Focuses: starting substructures for generation.

in accounting for exchangeable protons. The processing time for Comp11 is 67 s in the UNIX environment.

Spec2D has been tested with more than 30 compounds randomly selected from ^{1}H-NMR Organic Compounds Volume 1, first edition database.[11] The structures of the test compounds are shown in Figure 10. Table 2 shows the results of Spec2D for the test compounds. The average number of extracted substructures is 248, the number of focuses is about 9, and the number of candidates is about 3. Since all test compounds are included in the database from which the knowledge base was constructed, correct compounds were proposed as candidates, which proves that all components of Spec2D are functional. The average processing time was about 2 min.

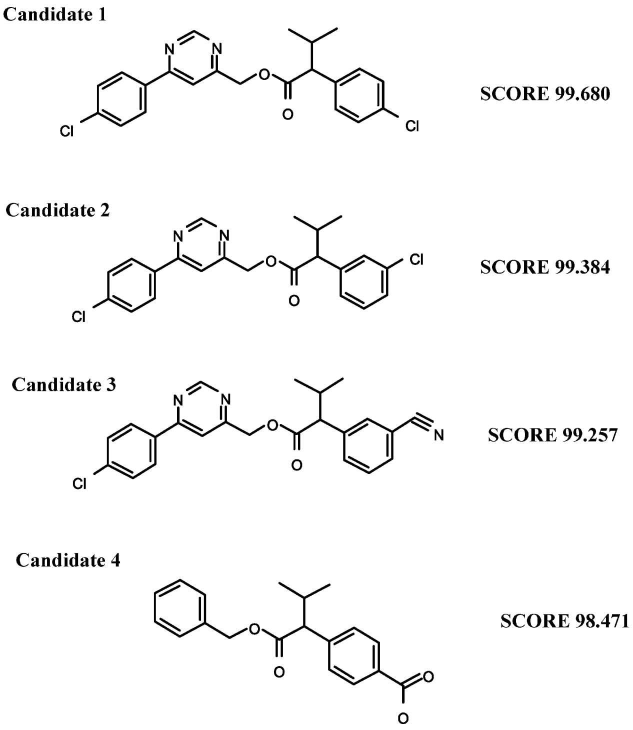


Figure 12. Candidates proposed by the Spec2D system for Comp11.

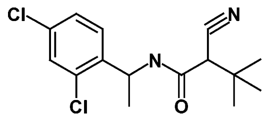


Figure 13. Structure of Comp33 {2-cyano-N-[1-(2,4-dichlorophenyl)ethyl]-3,3-dimethylbutanamide}.

Table 3. Results of 19 Test Compounds Using Leave-One-Out Method

	normal test			leave-one-out method		
sample	number of sub-structures	number of candi-dates	including correct one	number of sub-structures	number of candi-dates	including correct one
Comp33	49	2	yes	49	2	yes
Comp34	146	6	yes	145	2	no
Comp35	266	3	yes	266	3	yes
Comp36	53	2	yes	53	2	yes
Comp37	256	5	yes	256	5	yes
Comp38	208	2	yes	207	1	no
Comp39	280	4	yes	280	4	yes
Comp40	141	3	yes	141	3	yes
Comp41	238	4	yes	238	4	yes
Comp42	355	2	yes	353	0	no
Comp43	260	2	yes	257	0	no
Comp44	1026	25	yes	1026	25	yes
Comp45	14	2	yes	14	2	yes
Comp46	204	4	yes	203	3	yes
Comp47	208	2	yes	208	2	yes
Comp48	186	1	yes	186	1	yes
Comp49	437	2	yes	436	0	no
Comp50	837	2	yes	836	0	no
Comp51	275	3	yes	274	2	no

To test "new" compounds not present in the database, 19 compounds were randomly selected from our own data. All compounds were checked by the normal test mentioned above. The system proposed 1−25 candidate structures including the correct one for all the test compounds (Table 3). For testing "new" compounds, the leave-one-out method was applied. One such compound, Comp33 {2-cyano-N-[1-(2,4-dichlorophenyl)ethyl]-3,3-dimethylbutanamide}, is shown in Figure 13. Before construction of the knowledge base, Comp33 was deleted from the database. The knowledge base was then constructed and used to analyze Comp33 (leave-one-out method). The number of extracted substructures from the analysis for the input data of Comp33 was 49. The system proposed two candidate structures including the correct one.

A total of 18 other compounds were tested using a similar method. The results are shown in Table 3. Spec2D proposed the correct structures for about 63% (12/19) of the test compounds. Even for a "new" compound, if its corresponding substructures are in the knowledge base, the system proposes the correct structure. Otherwise, the system suggests similar compounds with useful substructure information. By increasing the amount and diversity of data in the database, the proposal capability of the system will be improved.

CONCLUSION

Spec2D was developed as a new and powerful 2D-NMR structure elucidation system. It proposes candidate structures for an unknown compound on the basis of H−H COSY and ^{1}H NMR spectra. A novel substructure representation method was used to construct the knowledge base from the database containing ^{1}H NMR data. To elucidate structures, the system analyzes NMR spectral information and extracts plausible substructures from the knowledge base. A structure generator then combines substructures into candidate structures. Spec2D has enhanced the efficiency of determining chemical structures for research and development in organic chemistry.

The features of the system are as follows:

1. The creation of novel substructures, which are appropriate for the corresponding ^{1}H NMR spectra.

2. The construction of a 2D-NMR knowledge base derived from 1D ^{1}H NMR data.

3. The practical application of 2D-NMR data, especially H−H COSY spectra in the analysis module.

4. The reduction of computing time using constraints on input spectral data for the generation process.

5. The proposal of plausible candidate structures by the generation module.

The scope of Spec2D is strongly dependent on substructures (not structures) contained in the database of spectra, though not totally. Human expert knowledge (not from inferring the database) can be integrated into the knowledge base to expand the scope of Spec2D. Compounds should be successfully predicted if all their substructures could be found in the knowledge base. Therefore, if any substructure contained in an "unknown" compound cannot be found, the correct structure would not be in the list of candidates.

ACKNOWLEDGMENT

H.M. is indebted to Professor Kimito Funatsu (Tokyo University) for valuable discussions and suggestions.

REFERENCES AND NOTES

(1) Duffield, A. M.; Robertson, A. V.; Djerassi, C.; Buchanan, B. G.; Sutherland, G. L.; Feigenbaum, E. A.; Lederberg, J. Application of artificial intelligence for chemical inference. II. Interpretation of low-resolution mass spectra of ketones. *J. Am. Chem. Soc.* **1969**, *91* (1), 2977−2981.

(2) Christie, B. D.; Munk, M. E. Structure elucidation by reduction−a new strategy for computer-assisted structure elucidation. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 87−93.

(3) Funatsu, K.; Miyabayashi, N.; Sasaki, S. Further development of structure generation in the automate structure elucidation system, CHEMICS. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 18−28.

(4) Neudert, R.; Penk, M. Enhanced structure elucidation. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 244−248.

(5) *KnowItAll*; BioRad Laboratories, Inc.: Hercules, CA. http://www.bio-rad.com (accessed Dec 2005).

(6) Elyashberg, M. E.; Blinov, K. A.; Williams, A. J.; Molodtsov, S. G.; Martin, G. E.; Martirosian, E. R. Structure Elucidator: A versatile expert system for molecular structure elucidation from 1D and 2D NMR data and molecular fragments. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 771−792.

(7) Will, M.; Fachinger, W.; Richert, J. R. Fully automated structure elucidation−a spectroscopist's dream comes true. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 221−227.

(8) Masui, H. A representation method for substructures associated with ^{1}H NMR spectra − HYPER Code. *Nippon Kagaku Kaishi* **1999**, 819−826.

(9) Hong, H.; Xin, X. ESSESA, an expert system for structure elucidation from spectral analysis Part II. Novel algorithm of perception of the linear independent smallest set of smallest rings. *Anal. Chim. Acta* **1992**, *262*, 179−191.

(10) Hong, H.; Xin, X. ESSESA, an expert system for structure elucidation from spectra. 4. Canonical representation of structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 730−734.

(11) *H NMR Organic Compounds Database*, 1st ed; Wiley-VCH, Chemical Concepts: Weinheim, Germany; Volume 1.

CI0502810