ISE-TR-81-25

LEARNING SYSTEM FOR AUTOMATIC STRUCTURAL ANALYSIS OF MASS SPECTRA

by

TAKASHI NAKAYAMA

and

YUZURU FUJIWARA *

April 1, 1981

INSTITUTE
OF
INFORMATION SCIENCES AND ELECTRONICS

UNIVERSITY OF TSUKUBA

LEARNING SYSTEM FOR AUTOMATIC STRUCTURAL ANALYSIS OF MASS SPECTRA

TAKASHI NAKAYAMA and YUZURU FUJIWARA*

Institute of Information Sciences and Electronics, University of Tsukuba, Sakura-mura, Niihari-gun, Ibaraki 305, Japan

A computer-assisted mass spectral interpretation system which has the capability of learning is described. The set of correspondences between substructure and spectral component (CSSC) is used for interpreting mass spectra. CSSC is generated, renewed and improved automatically in the system. This automatic generation is learning. Chemical structures are represented in terms of blocks.

I. Introduction

Many kinds of structural analysis methods of mass spectra have been presented and developed, such as the pattern recognition technique and structure generation.[1-6] In this paper, a chemical structure is regarded as a graph, and is described hierarchically by means of not only the structural unit "atom", but also the intermediate concept "block". This hierarchical representation simplifies and saves time in analysing process. A vertex of a graph is a "cutpoint" if its removal increases the number of connected components of the graph. A "block" of a graph is a maximal subgraph which has no cutpoints.[7] For example, the ring assembly corresponds to a block. A block is treated as a constituent unit of an ion giving some spectral pattern in case of structural analysis. Therefore, these is no need for substructure inference to use such an inefficient algorithm as one that enumerates all the combinations of atoms. This brings about great advantages in simplicity and time saving in the analysing process. The total structure of a chemical compound (connectivity among blocks) is represented by means of block-cutpoint tree (BCT)[8]. The structural analysis is to describe a sample spectrum in terms of couplings of substructure and spectral component. These couplings of substructure and spectral component are sets of pairs of substructure represented by blocks and corresponding spectral components. This set is not rigid (i.e. not a ready-made data set), but is automatically generated and renewed by the system. This property of self-organization of knowledge is the capability of learning. The

advantages of the structural analysis method described here are (1) that the unit of structural representation is a block, and (2) that the system has the capability of learning.

II. Analysing System

A block diagram of the system is shown in Figure 1. Structural data of chemical compounds, mass spectral data and the table of correspondence between substructure and spectral component (CSSC) constitute the data sets. When a sample spectrum is given, programs ANALYSIS/LEARNING perform analysis/learning referring to the data sets. The program LEARNING is activated by program ANALYSIS if necessary, but the two programs work independently. Therefore, the learning can be advanced even at the same time as ANALYSIS works.

1. Data Sets. A structural data set and spectral data set are shown in Figure 2. Structural data is stored in two files, FCF and VCF. The FCF record is a bit sequence of fixed length which gives block constitution of compounds. The length of the bit sequence is 540 bits and the i-th bit value (1/0) of the sequence specifies the presence/absence of the block whose identification number is i. Blocks which overflow the 540-bit table of FCF are described in VCF. Further details of block constitution are described in the VCF record, which consists of the number of kinds of blocks, the number of each block, the degree of each block in BCT, and so on. The blocks, which

are specified by bit sequence, are identified by block file BF. VCF and BF records are of variable length and are accessed by means of each directory file shown in Figure 2.

The spectral data set consists of IF and SF. IF is an information file which contains items such as compound name, molecular weight, and other conditions of the measurement. SF is a file of mass spectral data.

Since the identification numbers of compounds are common throughout these data sets, each file is considered as a set of attribute data of compounds.

CSSC consists of two kinds of records. One is records whose substructure item is known, and the other is records whose substructure item is unknown. The latter is detected as a pair of spectral component and unknown substructure in the structural analysing process, and is registered in CSSC. These records are also used for analysing sample spectra (i.e., the spectral component is treated definitely even though the corresponding substructure is unknown). It is possible that the unknown substructure is inferred by means of program LEARNING, if the spectral data set is renewed (e.g., new spectral data are added). The former is an ordinary CSSC record. This record format is shown in Figure 3. A pair of substructure and spectral component is represented as two fixed length items, CSST and CSSP. CSST is a 540-bit sequence and has the same meaning as the FCF record. The BCT code of CSST is linked to the CSSC record, which represents the connectivity among blocks. (Further strict representation of connectivity among blocks is given by specifying cutpoints in each block.)

CSSP is a 958-bit sequence whose i-th bit value (1/0) represents presence/absence of a peak at m/e = i. It does not indicate intensity. Spectral data in SF is regarded as a specific CSSP which has intensity information, and the corresponding structures are given by structural data set FCF/VCF. (Actually, spectral analysis is performed by matching sample spectrum with spectral data in SF, prior to description by CSSC.)

Thus, structural data of compounds is represented in terms of the block as an intermediate concept, and file organization is achieved by using the block as a processing unit.

2. Learning. CSSC is a set of pairs of correspondence between substructure and spectral component. If these correspondences are precise and sufficient, any mass spectrum can be described by CSSC. Sample spectrum can be retrieved and identified in spectral data set, if spectral data of all compounds are prepared. However, the method described here identifies a compound constructively by CSSC. Therefore, it is possible to analyse as many compounds as the number of combinations of substructures in CSSC. It is necessary for practical analysis to prepare CSSC adequately, in both quality and quantity. Learning is the process to organize good CSSC. Namely, the generation and improvement of CSSC are performed by automatic judgement of the system. This generation/improvement process corresponds to the process of acquirement, refinement and accumulation of knowledge, i.e. learning. In other words, the framework of spectral analysis, which is the correspondence

between substructure and spectral component, is not given in the form of rigid input data, but is organized automatically from spectral data or CSSC records by the system.

Generation of initial CSSC The initial CSSC is generated from sample spectra by the program LEARNING shown in Figure 1. The basic idea of the generation method is, first, to make a set of compounds similar to a sample (there can be a variety of criteria for the similarity), then to extract common substructures and common spectral components from the set.

Using the similarity between the two spectra as a criterion, the outline of the generation procedure is as follows:
(1) Noise elimination of spectral data. The peak intensity is compared with the value of the function $f(x) = a + c/(x+b)$, and any peak smaller than that is eliminated. x represents m/e. Coefficients a, b and c are determined empirically.
(2) Computation of similarity. After the noise elimination of spectral data, the similarity between a sample spectrum (P_1) and a spectrum is SF(P_2) is computed. The similarity is defined by the expression below:

$$S(P_1, P_2) = \frac{(P_1, P_2)}{|P_1| \cdot |P_2|}$$

where P_1 and P_2 are 958-dimensional vectors (the positions where m/e = 1, 2, ..., 958 are regarded as the area where the spectra exist). The part of spectra which overflow the size of SF file of m/e = 958 are stored in additional SF. The spectrum P_2 in expression (1) is often filtered. The filtering vector $F = (f_1, \cdots, f_{958})$ is made from spectrum P_1 as follows.

f_i = 1 if there exists a peak of P_1 at m/e = i, f_i = 0 otherwise. Then $P_2 = (P_1^2, \cdots, P_{958}^2)$ is renewed by $P_i^2 = f_i \cdot P_i^2$.

(3) Extraction of common substructures and common spectral components. A set of similar compounds $C = \{C_1, \cdots, C_n\}$ is obtained through preprocessing (1) and (2), where C_i (i = 1, $\cdots$, n) is selected for a member of the set of the similarity between a sample spectrum and c_i's spectrum is greater than some standard value. The structural data of $C_1, \cdots, C_n$ are obtained from data set shown in Figure 2. The common substructure is extracted in the form of a common block set. The extraction procedure is performed rapidly using a file FCF whose record is a bit sequence. The spectral data of $C_1, \cdots, C_n$ are obtained from SF as shown in Figure 2, then the common spectral component is extracted. Though spectral data in SF contain peak intensity, the common spectral component consists of mass numbers (m/e values) where $C_1, \cdots, C_n$ give significant peaks (i.e. peaks which are not noise) in common.

(4) Check of extracted correspondence. Actually, there occur many cases in which the similar compound set is not generated, or common substructures/common spectral components is not extracted. The improvement procedure is applied to these cases through feedback technique (this procedure is described in detail in the next paragraph). When the correspondence is obtained, it is checked if it is proper as a CSSC record. The check points are (a) that the common substructure should be a connected subgraph for all members of the similar compound set, (b) that the common substructure should contain at least one terminal block, and (c) that the maximal mass numbers of the

common spectral components should not exceed the mass of the corresponding substructures. Conditions (a) and (b) stem from the supposition that the mass spectra reflect mainly the simple fragmentation. However, these conditions do not mean that the other types of fragmentation are eliminated.

The initial CSSC is generated in another way: first a common substructure is specified, then a set of compounds are chosen which contain the substructure in common, and the common spectral components from the set are extracted. This common spectral component should be checked by the condition (c) described above.

Improvement of CSSC The accuracy of the correspondence of CSSC record is checked when it is initially generated, but it is not always satisfactory. Even when that the substructure is accurately specified, it is still probable that the corresponding spectral component may contain noise peaks or may lack some peaks. The accuracy of the correspondence of CSSC depends on the size of the spectral data set as a whole. In general, the accuracy is expected to improve as the data set grows.

When the size of the data set is limited, it is still possible to improve the accuracy by reconstructing the compound set from which a CSSC record should be extracted. The compound set is constructed by collecting compounds in which a kind of similarity reaches a standard value. Three measures of similarity are prepared: (1) similarity between two spectra, (2) existence of common substructure, and (3) similarity of molecular weights. The compound set varies according as these three parameters vary, as shown in Figure 4. These are two

ways of constructing compound sets: reduction and expansion. The reconstructing method shown in Figure 4(a) is used for eliminating noise peaks of CSSP (i.e. the peaks which should not be given by the corresponding CSST), and the new compound set S_u is generated in the form of an expansion of the original set S_i which is generated on the basis of spectral similarity. The expanded set (S_u) consists of the members which contain the reference CSST as a substructure and are extracted from the whole compound set (FCF/VCF). In other words, the measure of the similarity is changed from the spectral similarity to the existence of a common substructure. It is certain that S_u includes S_i, so the common spectral component extracted from S_u does not contain more noise peaks than the original CSSP. That is to say, CSSP is improved.

The compound set is expanded/reduced by varying a standard value of the spectral similarity as shown in Figure 4(b). This construction technigue is used for the same case as (a). The measure of the similarity is the spectral similarity and is unchanged. While the members of the compound set are extracted from the compound data set (FCF/VCF), the set is not always the proper one for extracting CSSP, because the possibility that specific compounds are included in the set increases as the set size becomes larger. In this case, or when CSST/CSSP is not extracted at initial generation, the compound set is reduced. The set reduction is performed by selecting particular members of a compound set, specifying parameters appropriately. Figure 4(c) shows that S_i/S_u is reduced by using molecular weight as a new parameter. There are two kinds of

parameter (molecular weight) setting: one is to collect compounds whose molecular weight should become nearly equal (this implies that each member compound is required to be more similar), and the other is conversely to collect compounds whose molecular weights should scatter widely (this implies that the similarity measure other than the existence of common substructure should be excluded as much as possible). Figure 4(d) shows a kind of partitioning of S_i. When on common substructure can be found in S_i, the following partitioning procedure is applied: Suppose that $S_u^{(1)}, \cdots, S_u^{(l)}$ be all the substructures contained in the compounds of S_i, the subset of S_i which consists of the members containing $S_u^{(k)}$ is constructed for $k = 1, \cdots, l$. $(S_i = S_u^{(1)} \cup \cdots \cup S_u^{(l)})$. A common spectral component is extracted from these subsets.

The reconstruction procedure of these compound sets is applied dynamically in the generation/improvement procedure of CSSC, and it is intended to construct an optimal compound set.

Figure 5 shows the process of constructing compound sets and extracting a pair of common substructure and common spectral component when the spectral similarity is varied as a parameter (for a sample spectrum). Figure 5(a) shows that the size of compound set C_1 is 4, that CSST (common substructure) is benzene ring, and that CSSP (common spectral component) is given as a mass number set (50, 51, 74, 77, 78, 123), for reference similarity of 0.90. It is found that the mass number m/e = 123 is irrelevant to this CSSP (benzene ring) by the check of correspondence, so the noise elimination procedure is applied as

shown in Figure 5(b). This shows that CSSP is refined by varying reference similarity from 0.90 to 0.75. The expanded compound set for eliminating noise peaks is constructed also by means shown in Figure 4(a), and this set gives the same result as described above. A part of CSSC obtained is shown in Figure 6.

3. Analysis. If CSSC is provided with records that are adequate in both quality and quantity, it is possible to analyse any sample spectra. The analysis of mass spectra by CSSC means to describe given sample spectra in terms of CSSP's. Given that S is a sample spectrum; it is expressed as follows:

$$S = \sum' P_i + \sum' q_i + S' \qquad (2)$$

where P_i is a CSSP for which correspondence between CSST and CSSP is established, q_i is a CSSP for which the correspondence is not established, and S' is the spectral component which cannot be explained by the present CSSC. If S' = o, the description of a sample spectrum is considered complete. $\sum'$ is interpreted as follows:

$$\sum_{i=1}^{n} P_i = \sum_{i=1}^{n} P_i + \sum_{j}\sum_{k} P_j \oplus P_k + \ldots + P_1 \oplus P_2 \oplus \ldots \oplus P_n \qquad (3)$$

where $\sum_{i=1}^{n} P_i$ represents the vector summation of P_i (P_i is implemented as a 958-dimensional vector). If the peak intensity

is not taken into consideration, $\sum_{i=1}^{n} P_i$ represents the logical summation of the mass position of P_i. The operator $\oplus$ represents the composition of substructures, so $P_j \oplus P_k$ represents the spectral component corresponding to the substructure $t_j \oplus t_k$ composed of substructures t_j and t_k. Therefore, $\sum_j \sum_k P_j \oplus P_k$ represents the vector summation of spectral components corresponding to all the possible structures composed of two CSST's. Similarly, spectral components up to $P_1 \oplus P_2 \oplus \cdots \oplus P_n$ (this corresponds to the total structure) are computed, and the total vector summation of these components gives $\sum' P_i$.

when $S = \sum' P_i$ (i.e. $\sum' q_i = S' = 0$ in expression (2)), the chemical structure of the sample spectrum contains substructure $t_1, \cdots, t_n$ (t_i corresponds to P_i), composed substructure $t_j \oplus t_k, \cdots$, and $t_1 \oplus \cdots \oplus t_n$ as a total structure. The example of the analysis is shown in Figure 7. The input sample spectrum is shown in Figure 7(a). The analysis procedure is applied to the noise cut spectrum shown in Figure 7(b). Given that this spectrum is S, it is expressed as follows:

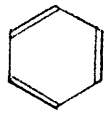
$$S = \sum_{i=1}^{3}{}' P_i \tag{4}$$

where P_1, P_2, P_3 are the CSSP's shown in Figure 6. Expression (4) is expanded according to expression (3):

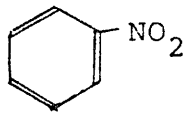
$$\begin{aligned} S &= \sum P_i + \sum \sum P_j \oplus P_k + P_1 \oplus P_2 \oplus P_3 \\ &= P_1 + P_2 + P_3 + P_1 \oplus P_2 + P_2 \oplus P_3 + P_3 \oplus P_1 + P_1 \oplus P_2 \oplus P_3 \end{aligned} \tag{5}$$

Substructures t_1, t_2, t_3 and composed substructures are found as follows:

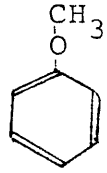
t_1 =



t_2 =



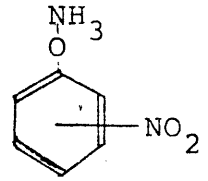
t_3 =



$t_1 \ominus t_2 = t_2$

$t_1 \ominus t_3 = t_3$

$t_2 \ominus t_3 = t_1 \ominus t_2 \ominus t_3 =$



Therefore, expression (5) gives the mass position derived from the last term $P_1 \ominus P_2 \ominus P_3$ in addition to the mass position of P_1, P_2 and P_3:

$$P_1 \ominus P_2 \ominus P_3 = \sum P_i + (153)$$

where the second term of the right side means that $P_1 \ominus P_2 \ominus P_3$ includes mass position m/e = 153. The synthesized spectrum is shown in Figure 7(c).

III. Conclusion

As the arguments so far indicates, if the mass spectral data and the structural data of compounds are adequate in both quantity and quality, it is possible to generate a CSSC which is able to analyse any sample spectra. That is to say, the more data increase in quantity, the more available substructure increase (quantity of knowledge), and the less the noise of spectral data is, the faster the speed of learning becomes (quality of knowledge). The experimental CSSC was generated using EPA/NIH Mass Spectral Database (1975 edition). The mass spectral data of this database contains systematic noises such as spectral pattern of solvents, air, etc., so they are not always appropriate to generate a CSSC. However, they can become available by eliminating such noises previously, or by selecting only such data that do not contain those noises from the beginning.

Chemical structures, therfore CSST (an entry of CSSC for substructures) are represented by means of BCT, so the processing efficiency has been improved largely for CSSC generation (extraction of substructures) and structure generation.

The description of a sample spectrum by CSSC is to infer the constituent substructure of the compound. Structure generation based on BCT representation of chemical structure is the subsquent step of structural analysis, and can be referred in our next paper.

(1) G. L. Ritter, S. R. Lowry, C. L. Wilkins, and T. L. Isenhower, "Simplex Pattern Recognition", Anal. Chem., 47, 1951 (1975).

(2) S. R. Lowry and T. L. Isenhower, "Comparison of Various K-Nearest Neighbour Voting Schemes with Self-Training Interpretive and Retrieval System for Identifying Substructures from Mass Spectral Data", Anal. Chem., 49, 1720 (1977).

(3) H. E. Dayringer, G. M. Pesyna, R. Venkataroghavan, and F. W. McLafferty, "Computer-Aided Interpretation of Mass Spectra", Org. Mass Spectrom., 11, 529 (1976).

(4) A. M. Duffield, A. V. Robertson, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, and J. Lederberg, "Application of Artificial Intelligence for Chemical Inference. II. Interpretation of Low-Resolution Mass Spectra of Ketons", J. Am. Chem. Soc., 91, 2977 (1969).

(5) B. G. Buchanan, D. H. Smith, W. C. White, R. J. Gritter, E. A. Feigenbaum, J. Lederberg and C. Djerassi, "Application of Artificial Intelligence for Chemical Inference. 22. Automatic Rule Formation in Mass Spectrometry by Means of the Meta-DENDRAL Program", J. Am. Chem. Soc., 98, 6168 (1976).

(6) S. Sasaki, H. Abe, Y. Hirota, Y. Ishida, Y. Kudo, S. Ochiai, and T. Yamasaki, "CHEMICS-F: A Computer Program System for Structure Elucidation of Organic Compounds", J. Chem. Inf. Comput. Sci., 18, 211 (1978).

(7) Frank Harary, "Graph Theory", Addison-Wesley, Reading,

Mass., 1969.

(8) T. Nakayama and Y. Fujiwara, "BCT Representation of Chemical Structures", J. Chem. Inf. Comput. Sci., 20, (in press)(1980).

FIGURE LEGENDS

Figure 1. Block diagram of the system.

Figure 2. Organization of structural data file and spectral data file. Each record is referred by identification number of a compound in case of fixed length records, and/or directory file in case of variable length records.

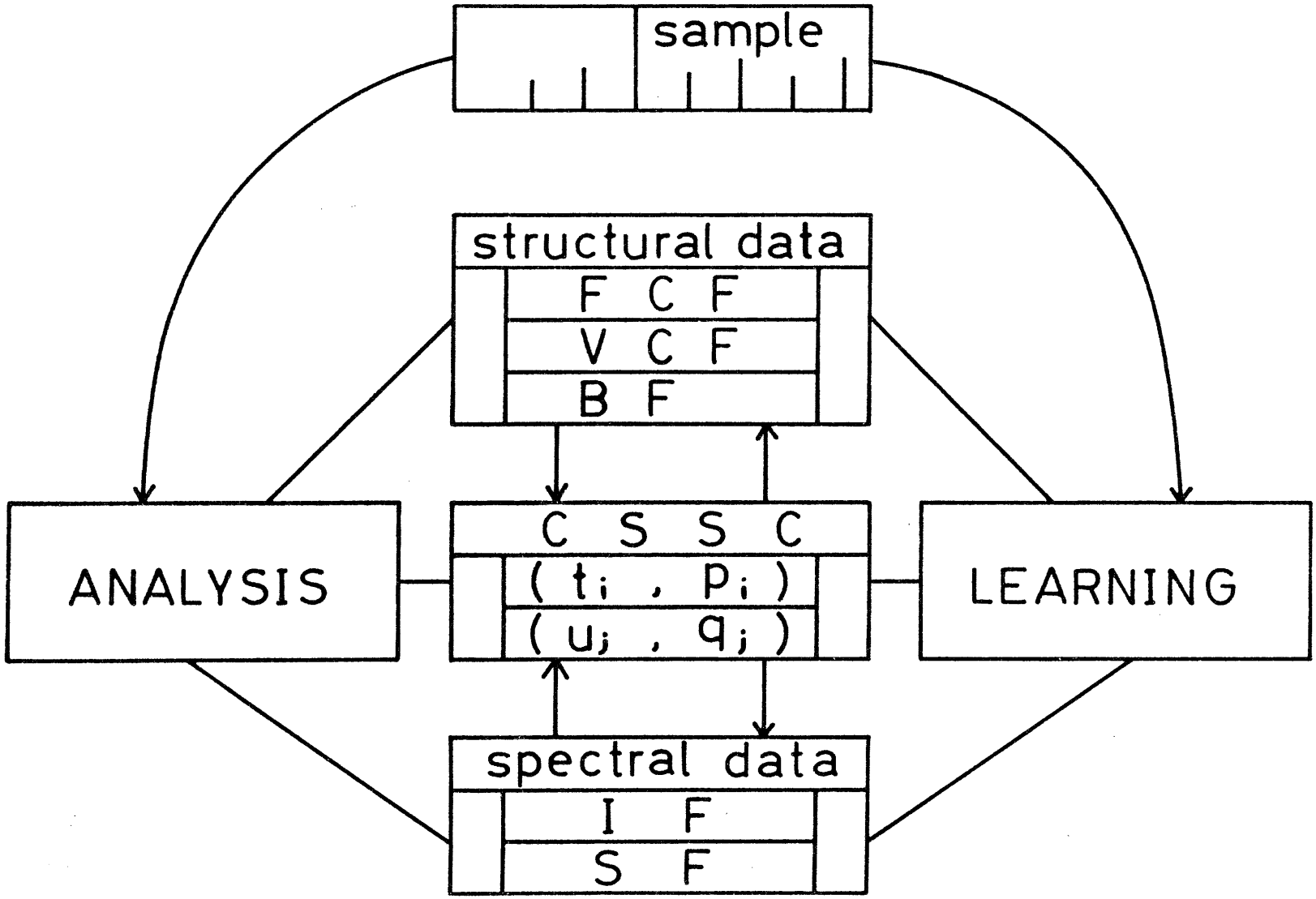
Figure 3. CSSC record format.

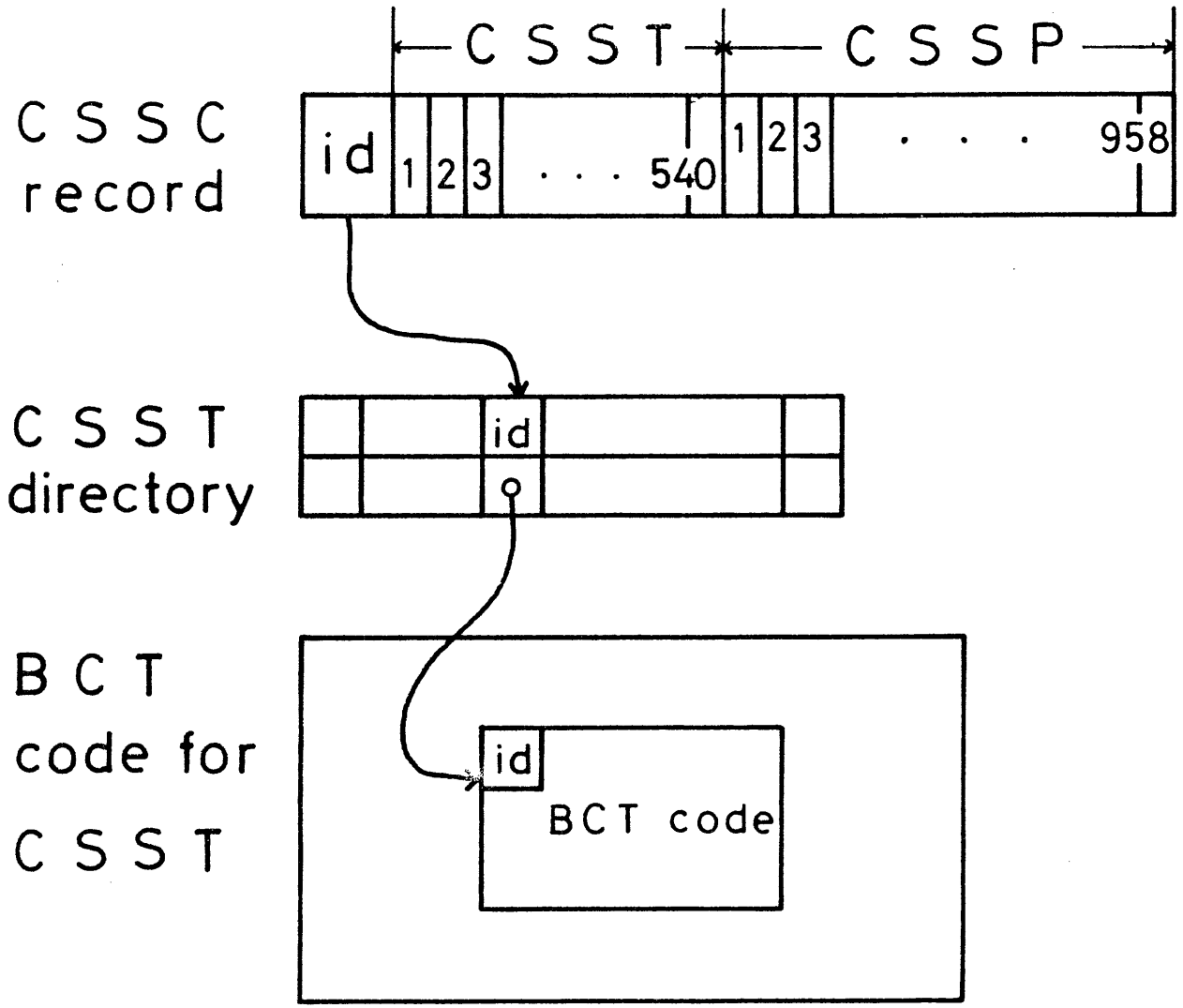
Figure 4. Reconstruction of compound sets according to parameters similarity (S_i), substructure (S_u) and molecular weight (M).

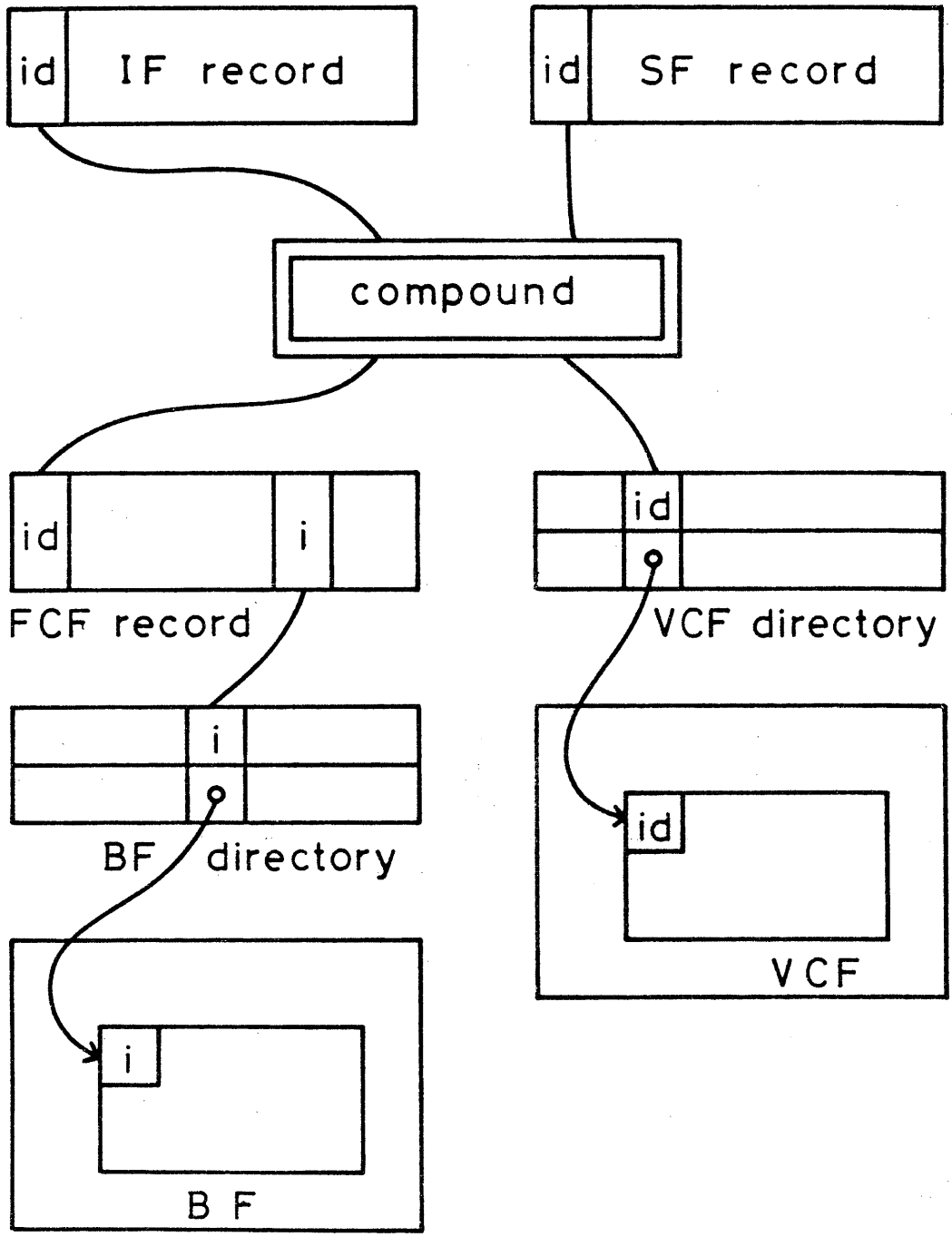
Figure 5. Improvement of a CSSC record by varying parameter similarity.

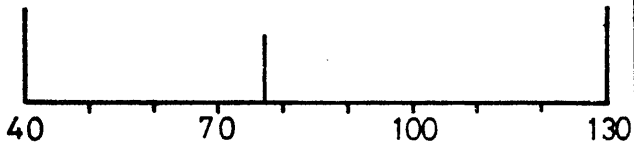
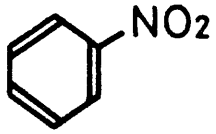
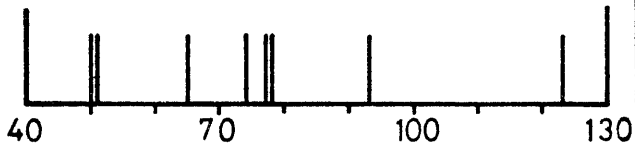
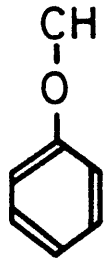
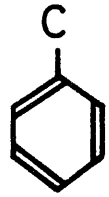
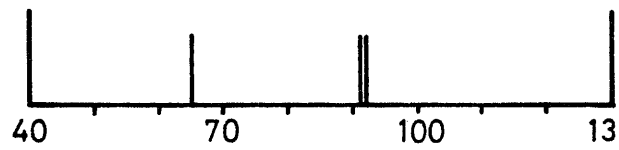
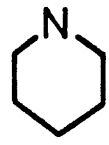
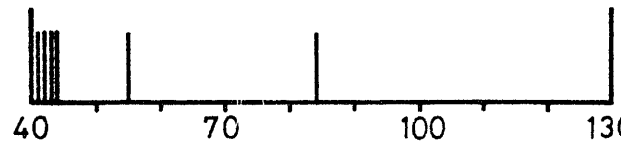
Figure 6. Example of CSSC.

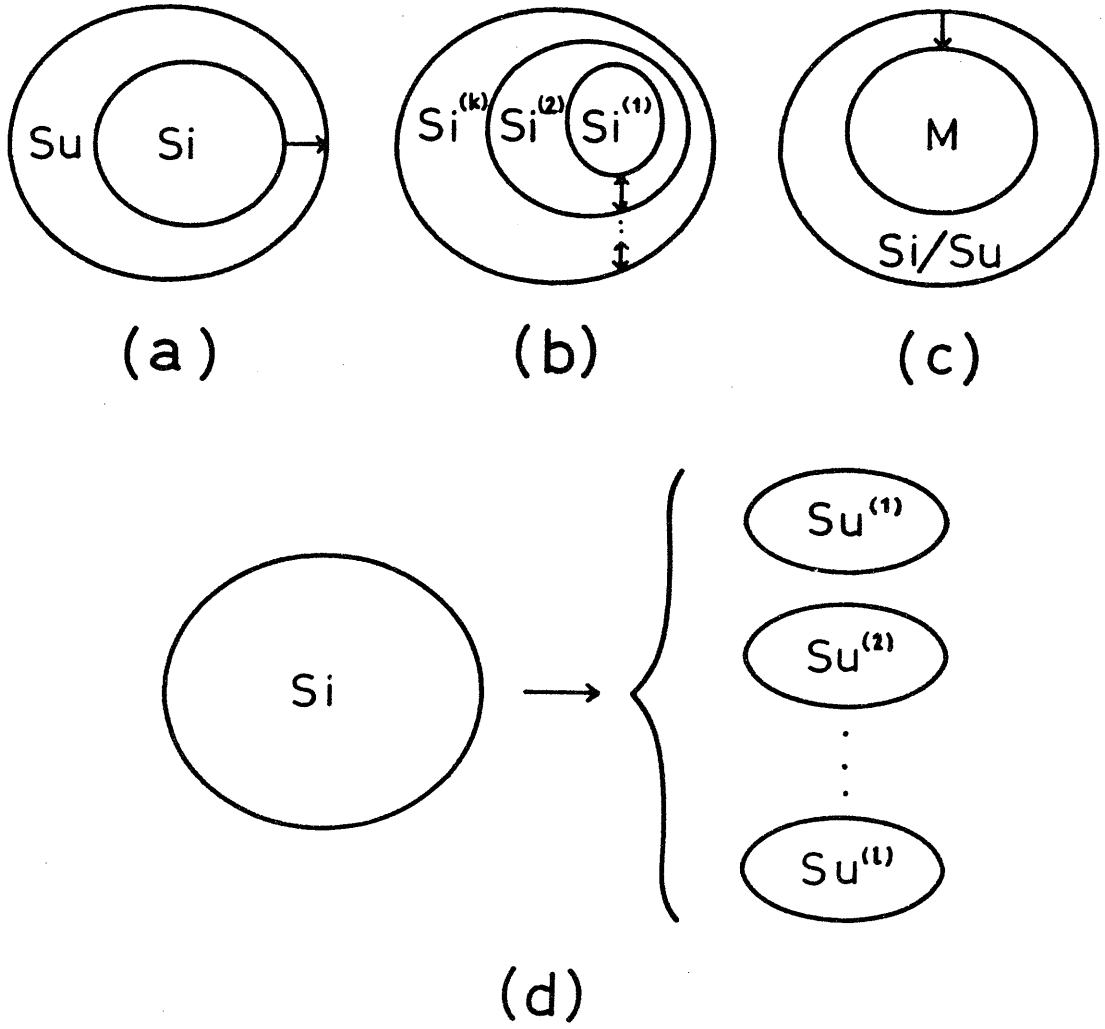
Figure 7. A sample spectrum and a synthesized spectrum by analysis.

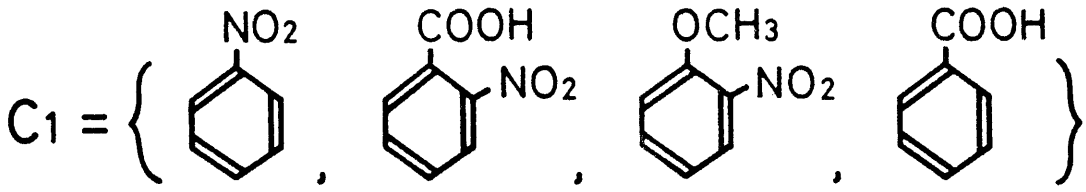

sample
structural data
F C F
V C F
B F
ANALYSIS
C S S C
(t_i , p_i)
(u_j , q_j)
LEARNING
spectral data
I F
S F


C S S T
C S S P
C S S C
record
id
1 2 3 . . . 540
1 2 3 . . . 958
C S S T
directory
id
B C T
code for
C S S T
id
BCT code

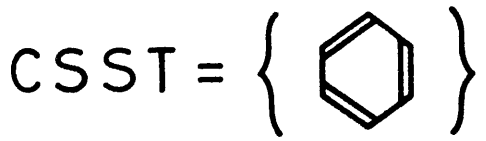

id
IF record
id
SF record
compound
id
i
FCF record
id
VCF directory
i
BF directory
id
VCF
i
B F

C S S T	C S S P (m/e)
t_1 = 	P_1 = (77) 
t_2 = 	P_2 = (50,51,65,74,77,78,93,123) 
t_3 = 	P_3 = (76,77,78,106) 
t_4 = 	P_4 = (65,91,92) 
t_5 = 	P_5 = (41,42,43,44,55,84) 

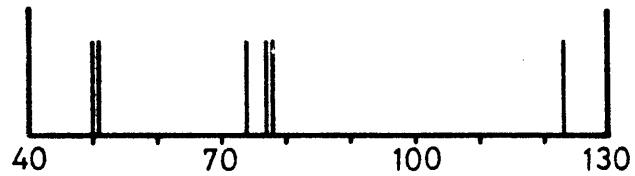

Su
Si
(a)
Si(k)
Si(2)
Si(1)
(b)
M
Si/Su
(c)
Si
Su(1)
Su(2)
Su(l)
(d)



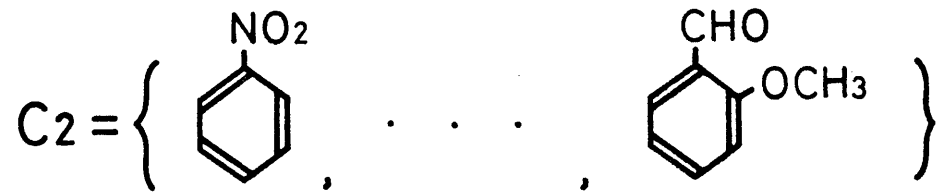
$|C_1| = 4$



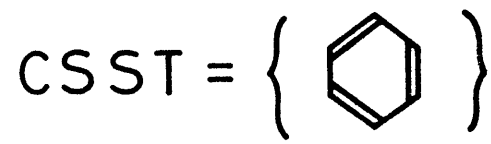
CSSP = (50, 51, 74, 77, 78, 123)



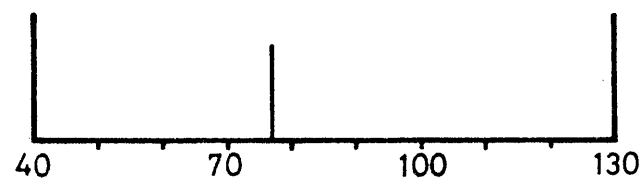
(a) similarity threshold = 0.90



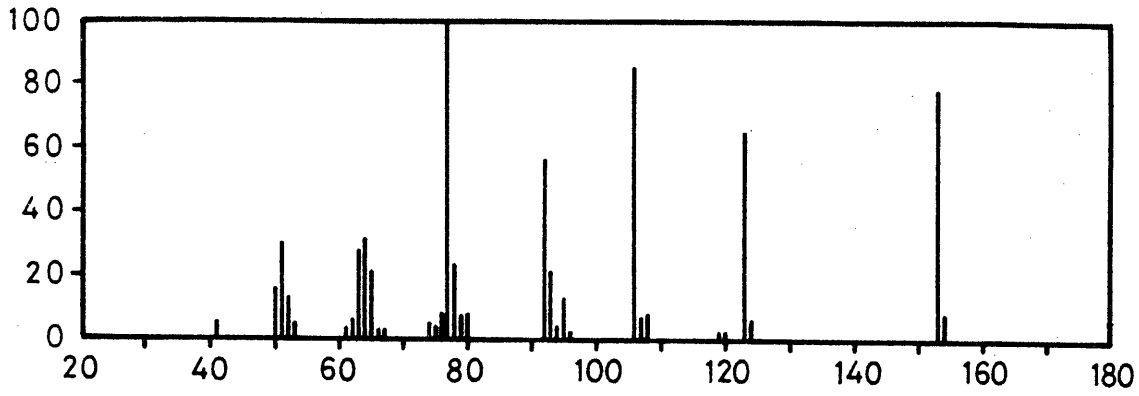
$|C_2| = 19$, $C_1 \subset C_2$



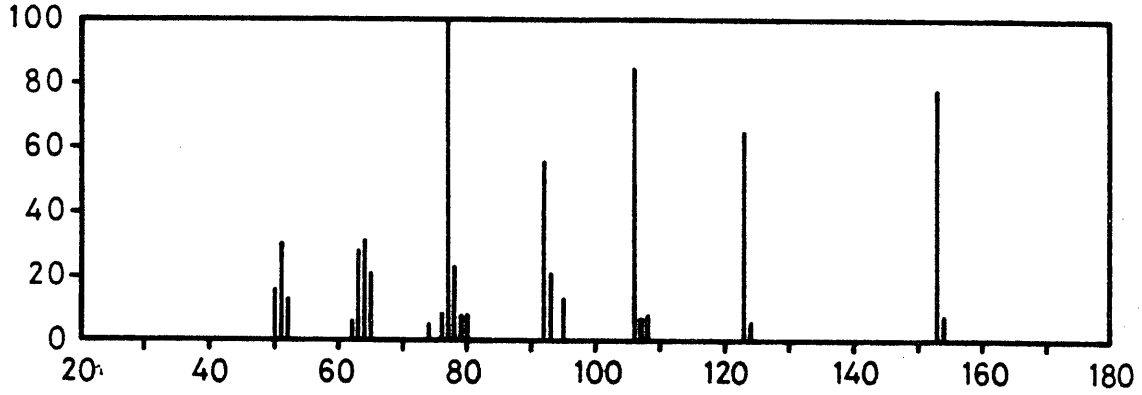
CSSP = (77)



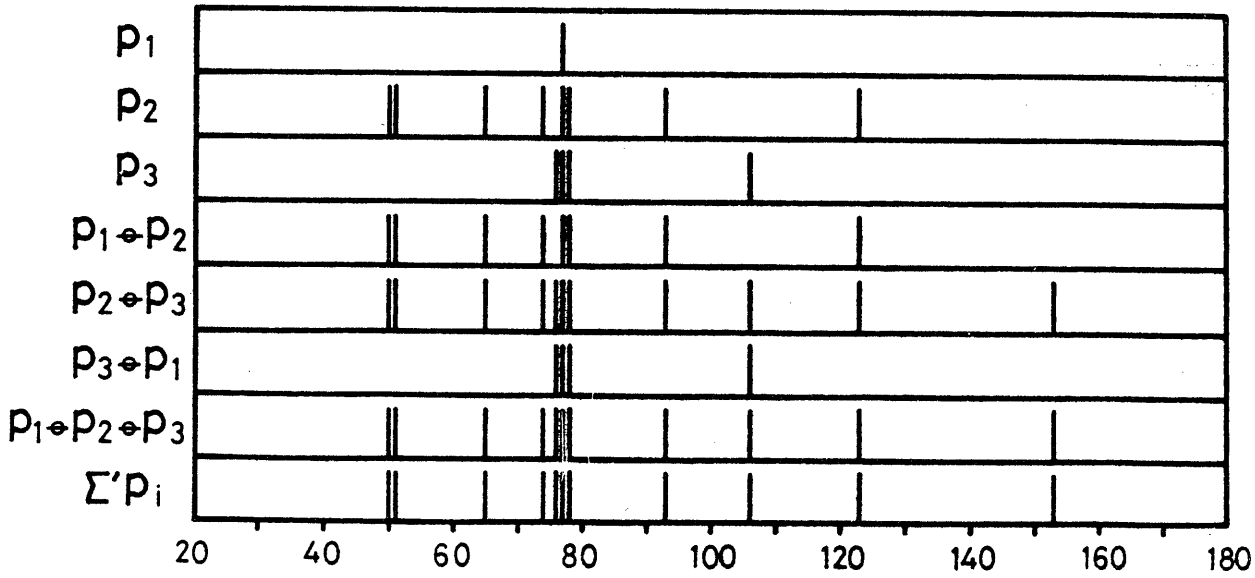
(b) similarity threshold = 0.75



(a) input sample spectrum



(b) noise cut spectrum



(c) synthesized spectrum

INSTITUTE OF INFORMATION SCIENCE AND ELECTRONICS
UNIVERSITY OF TSUKUBA
SAKURA-MURA, NIIHARI-GUN, IBARAKI, JAPAN

REPORT DOCUMENTATION PAGE	REPORT NUMBER ISE-TR-81-25

TITLE

LEARNING SYSTEM FOR AUTOMATIC STRUCTURAL ANALYSIS OF MASS SPECTRA

AUTHOR(S)

Takashi Nakayama [Central Research Laboratory, Kuraray Co. Ltd.]

Yuzuru Fujiwara [Institute of Information Science and Electronics, University of Tsukuba]

REPORT DATE	NUMBER OF PAGES
April 1, 1981	23

MAIN CATEGORY	CR CATEGORIES
Learning System Graph database	3.13 Applications, Chemistry 3.62 Artificial Intelligence, Learning and Adaptive Systems

KEY WORDS

Learning System/Adaptive data structure/Graph database
Automatic Analysis of Mass Spectra/Structure Elucidation

ABSTRACT

A computer-assisted mass spectral interpretation system which has the capability of learning is described. The set of correspondences between substructure and spectral component (CSSC) is used for interpreting mass spectra. CSSC is generated, renewed and improved automatically in the system. This automatic generation is learning. Chemical structures are represented in terms of blocks.

SUPPLEMENTARY NOTES