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

Received March 24, 1980; Revised Manuscript Received March 19, 1981

A computer-assisted mass spectral interpretation system with a learning mechanism 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 self-organization of CSSC is "learning". Chemical structures are represented in terms of blocks, which facilitates the learning process.

INTRODUCTION

Many kinds of structural analysis methods of mass spectra have been represented 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 the analyzing procedure and saves processing time. 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–cutpoint tree (BCT) is defined as a graph whose vertex set consists of blocks and cutpoints, and an edge is defined between a block and a cutpoint, provided that the cutpoint is an element of the vertex set of the block. This BCT is used for the intermediate representation of chemical structures. For structural analysis, a block is treated as a constituent unit of an ion giving some spectral pattern. Substructure inference in the structural analysis is performed by using spectral patterns corresponding to blocks/superblocks. This brings about advantages in the form of simplicity and time saved in the analyzing process. The structural analysis describes a sample spectrum in terms of couples of substructure and spectral component. The substructure/spectral component of the couple are represented by blocks/spectral patterns (bit pattern). A set of these couples is organized into a file, CSSC (a table of correspondences between substructures and spectral components), and an element of this set is considered as knowledge to analyze spectra. CSSC is not fixed (i.e., not a ready-made data set), but is automatically generated and renewed by the system: the substructure might be articulated and the spectral component might be refined. This process of self-organization of knowledge is "learning". The advantages of the structural analysis method described here are (1) the unit of structure representation is a block and (2) the system has a learning mechanism.

ANALYZING SYSTEM

A block diagram of the system is shown in Figure 1. Structural data of chemical compounds, mass spectral data, and 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 activiated by program ANALYSIS if necessary, but usually the two programs work independently.

(1) Data Sets. Structural and spectral data sets are shown in Figure 2. Structural data are stored in two files, FCF and VCF. The FCF record is a bit sequence of fixed length which gives the block constitution of a compound. The length of the bit sequence is 540 bits and is divided into two fields: the unmodified and the modified. The ith bit value (1/0) of the unmodified field specifies the presence/absence of the block whose identification number is i, and this field is used for the blocks which occur most frequently in compounds. The modified field consists of two subfields: modifier and number subfields. The value of the modified field is computed from a pair of values: (modifier, number) = $100m + j$, where m is the value of the modifier subfield and j is the on-bit position of the number subfield. The modifier corresponds to a pagination physically: one page contains 100 records, and these block records are numbered from 1 to 100, corresponding to bit positions of the number subfield. Further details of block constitutions are described in VCF records which consist of the number of kinds of blocks, the number of each block, the degree of each block in BCT, and so on. Blocks are identified by block file BF.

The spectral data set consists of IF and SF. IF is an information file which contains items such as compound names, molecular weights, and measurements conditions. 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 characteristic data of compounds.

CSSC consists of two kinds of records—records whose substructure item is known and records whose substructure item is unknown. The latter is detected as a spectral component and unknown substructure pair in the structural analyzing process and is registered in CSSC. These records are also used for analyzing 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 the 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 substructure and spectral component pair is presented 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. CSSP is a 958-bit sequence whose ith 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 structure is 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 are represented in terms of blocks as an intermediate concept, and files are organized by using the block as a processing unit.

(2) Learning. Any sample spectra can be retrieved and identified from the spectral data set if spectral data of all compounds are prepared. But it is difficult to prepare spectra of all compounds in a form suitable for a quick search. Another way to analyze spectra is, firstly, to infer substructures

0095-2338/81/1621-0142$01.25/0

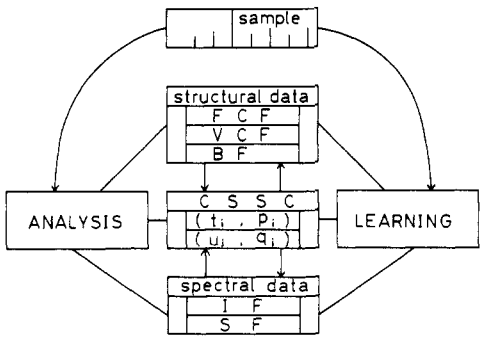


Figure 1. Block diagram of the system.

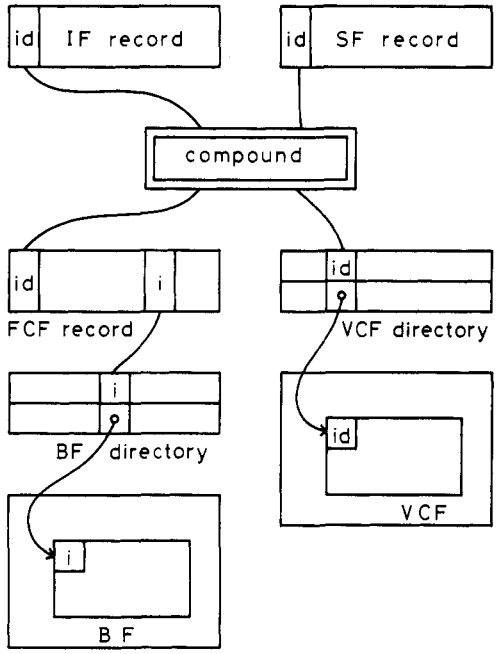


Figure 2. Organization of structural and spectral data files.

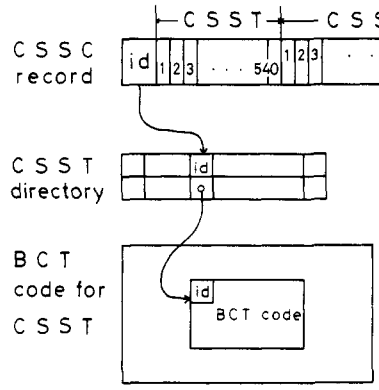


Figure 3. CSSC record format.

of a sample by using empirical laws and theories and, secondly, synthesize a structure from those substructures. CSSC is a set of pairs of correspondences between substructures and spectral components. If these correspondences are precise and sufficient, any mass spectra can be described by CSSC. The method described here identifies a compound constructively by CSSC; so it is possible to analyze as many 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 used to organize good CSSC; 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 fixed input data but is self-organized from spectral data or CSSC records.

Generation of Initial CSSC. The initial CSSC is generated from the sample spectra by the program LEARNING as 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) and then to extract a common substructure and a common spectral component from the set. Using the similarity between two spectra as a criterion, the generation procedure is outlined 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. The variable x refers to 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 in SF (P_2) is computed. The similarity is defined as

$$S(P_1,P_2) = \frac{(P_1,P_2)}{|P_1||P_2|} \qquad (1)$$

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 overflows the size of SF is stored in additional SF. The spectrum P_2 in expression 1 is often filtered. The filtering vector $F = (f_1,\ldots,f_{958})$ is determined 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, $P_2 = (p_1^2,\ldots,p_{958}^2)$ is renewed by $p_i^2 = f_i p_i^2$.

(3) Extraction of common substructures and common spectral components—a set of similar compounds $C = \{c_1,\ldots,c_n\}$ is obtained through preprocessing (1) and (2) described above, where c_i ($i = 1,\ldots,n$) is selected for a member of the set if the similarity between a sample spectrum and c_i's spectrum is greater than some standard value. The structural data of $c_1,\ldots,c_n$ are consulted in data files shown in Figure 2. The common substructure is extracted in the form of a set of blocks. The extraction procedure is performed rapidly by using a file FCF whose record is a bit sequence. The spectral data of $c_1,\ldots,c_n$ are obtained from SF, and 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) at which compounds $c_1,\ldots,c_n$ give significant peaks in common (noise peaks are already eliminated at this stage).

(4) Check of extracted correspondence—actually, there occur many cases in which a similar compound set is not generated or a common substructure/spectral component 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 between a substructure and a spectral component is obtained, it is checked to see if it is a proper CSSC record. The check points are (a) the common substructure should be a connected subgraph for all members of the similar compound set, (b) the common substructure should contain at least one terminal block, and (c) the maximal mass number of the common spectral component should not exceed the mass of the corresponding substructure. Conditions a and b stem from the supposition that the mass spectra reflect mainly the simple fragmentation. However, these conditions do not mean that fragmentations of other types are eliminated.

The initial CSSC is generated in another way: first a common substructure is specified, then a set of compounds are

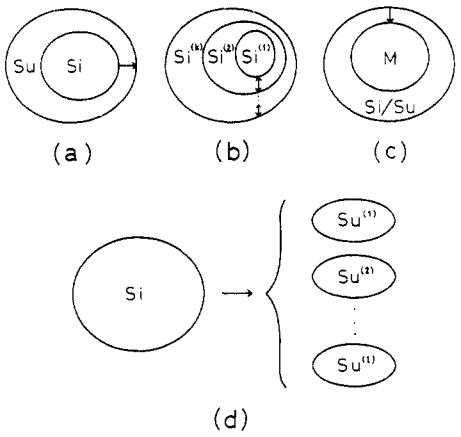


Figure 4. Reconstruction of compound sets according to parameters: spectral similarity (S_i), substructure (S_u), and molecular weight (M).

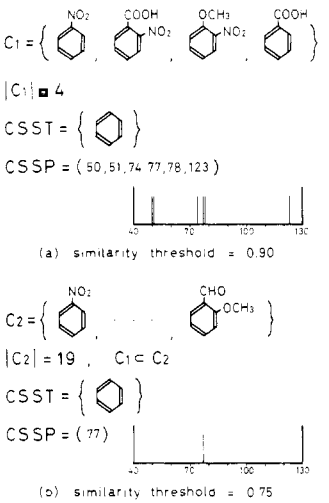


Figure 5. Improvement of a CSSC record by varying parameter, namely, spectral similarity.

chosen which contain the substructure in common, and finally the common spectral component from the set is extracted. This common spectral component should be checked by the condition c described above.

Improvement of CSSC. The accuracy of the correspondence of the CSSC record is checked when it is initially generated, but it is not always satisfactory. Even when the substructure is accurately specified, it is still probable that the corresponding spectral component may contain noise peaks or lack some significant peaks. The accuracy of the correspondence 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) the distance between two spectra—we adopted the Euclidean distance defined by expression 1, which is often used in classifying spectra as one of the pattern recognition techniques, K-nearest-neighbor classifier;[2] this measure is called spectral similarity; (2) existence of a common substructure; (3) molecular weights. Figure 4 shows that a compound set varies when these three measures vary. There are two ways of reconstructing compound sets: reduction and expansion. The reconstructing method shown in Figure 4a 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 4b. This construction technique 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 4c 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 are 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 differ widely (this implies that the similarity measure other than the existence of a common substructure should be excluded as much as possible). Figure 4d shows a kind of partitioning of S_i. When no common substructure can be found in S_i, the following partitioning procedure is applied: Suppose SS(1),. . ., SS(l) presents all the substructures contained in the compound set S_i, then the subset $S_u^{(k)}$ of S_i which consists of the members containing SS(k) is constructed for k = 1,2,. . ., l ($S_i = S_u^{(1)} \cup \ldots \cup S_u^{(l)}$, $S_u^{(p)} \cap S_u^{(q)} = \phi$ does not always hold). 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 common substructure and spectral component pair when the spectral similarity is varied. Figure 5a shows that the size of compound set C_1 is 4, CSST (common substructure) is the benzene ring, and CSSP (common spectral component) is given as a mass number set {50,51,74,77,78,123} (this is treated as a vector of 958-bit sequence in the system) for reference similarity 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 5b. This showns 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 4a, and this set gives the same result as described above. Figure 6 shows some elements of CSSC.

(3) Analysis. If CSSC is provided with records that are adequate in both quality and quantity, it is possible to analyze 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 $S' = 0$,

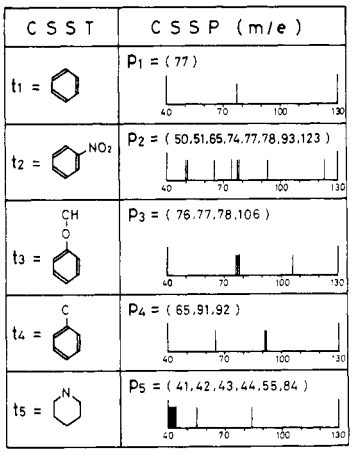


Figure 6. Example of CSSC.

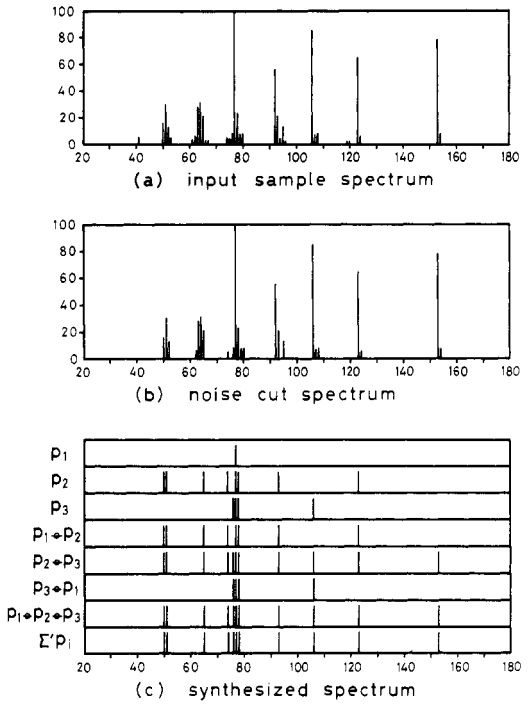


Figure 7. Example of spectral analysis.

the description of a sample spectrum is considered complete. $\sum'$ is interpreted as follows:

$$\sum_{i=1}^{n}{}'P_i = \sum_{i=1}^{n} P_i + \sum_j \sum_k P_j \ominus P_k + \ldots + P_1 \ominus P_2 \ominus \ldots \ominus P_n \quad (3)$$

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 $\ominus$ represents the composition of substructures; so $P_j \ominus P_k$ represents the spectral component corresponding to the substructure $t_j \ominus t_k$ composed of substructures t_j and t_k. Therefore, $\sum_j \sum_k P_j \ominus P_k$ represents the vector summation of spectral components corresponding to all the possible structures composed of two CSST's. Similarly, spectral components up to $P_1 \ominus P_2 \ominus \ldots \ominus P_n$ (this corresponds to the total structure) are computed, and the total vector summation of these components gives $\sum' P_i$.

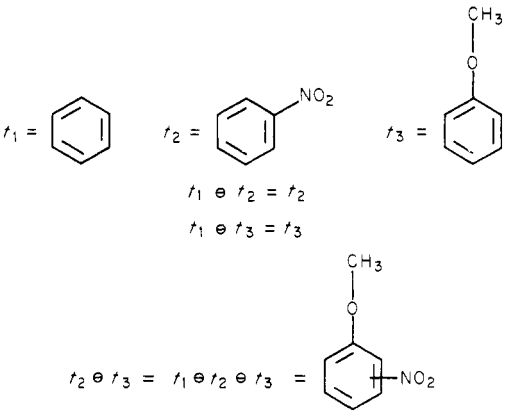
When $S = \sum' P_i$ (i.e., $\sum' Q_i = S' = 0$ in expression 2), the chemical structure of the sample spectrum contains substructures $t_1, \ldots, t_n$ (t_i corresponds to P_i), composed substructures $t_j \ominus t_k, \ldots$, and $t_1 \ominus \ldots \ominus t_n$ as a total structure. An example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

$$S = \sum_{i=1}^{3}{}' P_i \quad (4)$$

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

$$S = \sum P_i + \sum\sum P_j \ominus P_k + P_1 \ominus P_2 \ominus P_3 = P_1 + P_2 + P_3 + P_1 \ominus P_2 + P_2 \ominus P_3 + P_3 \ominus P_1 + P_1 \ominus P_2 \ominus P_3 \quad (5)$$

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



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-hand side means that $P_1 \ominus P_2 \ominus P_3$ include the mass position m/e 153. The synthesized spectrum is shown in Figure 7c.

CONCLUSION

As the argument 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 analyze any sample spectra. That is to say, the more the data increase in quantity, the more the available substructures (quantity of knowledge) increase, and the more spectral noises become reduced, the faster the learning speed becomes (quality of knowledge). The experimental CSSC was generated by using EPA/NIH Mass Spectral Database (1975 edition). The mass spectral data of this database are not always suitable for CSSC generation, because they include systematic noises such as spectral patterns of solvents, air, etc.; so they are preprocessed for practical use.

Chemical structures, therefore CSST (an entry of CSSC for substructures), are represented in terms of BCT; so the processing efficiency has been improved largely for CSSC generation.

The description of a sample spectrum by CSSC is to infer the constituent substructures of the compound. Structure generation based on BCT representation of chemical structures is the subsequent step of structural analysis.[9]

REFERENCES AND NOTES

(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 Neighbor Voting Schemes with Self-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, MA, 1969.
(8) Takashi Nakayama and Yuzuru Fujiwara, "BCT Representation of Chemical Structures", *J. Chem. Inf. Comput. Sci.*, **20**, 23 (1980).
(9) The method of structure generation based on BCT representation of chemical structures will be dealt in our next paper.

Realistic vs. Systematic Nomenclature

JOHN A. SILK*

Imperial Chemical Industries Ltd., Plant Protection Division, Jealott's Hill Research Station, Bracknell, Berkshire, United Kingdom

The place of systematic nomenclature is appraised by relating its functions to recent developments.

The recent papers by Goodson and others on graph-based chemical nomenclature[1,2] raise afresh in my mind the difficult question of the value of truly systematic nomenclature. We already have systems capable of meeting the great majority of needs, and the fact that some of them are specially designed to meet particular requirements reflects the manifold variety of molecular architecture.

This seems to be an opportune time to reappraise the place of systematic nomenclature by relating the functions it is required to serve to recent developments. For what applications would a completely systematic and comprehensive nomenclature system be useful?

(1) Among chemists the primary means of communication is the structural formula, and the role of nomenclature is secondary in providing linear descriptions of structures in forms which can be both written and spoken. For communication *descriptiveness at an appropriate cognitive level* is required. Current systems of nomenclature reflect this need by employing a suitably rich vocabulary with characteristic names for important ring systems and functional groups. The variety speeds communication, at least among the knowledgeable, by enabling larger entities to be designated by a few syllables. This psychological aspect is important: the names are fit for their purpose.

With a fully systematic nomenclature, by contrast, the vocabulary is deliberately limited to a basic set of terms. While in principle this simplifies the construction and interpretation of names, in practice it has the obvious disadvantage of a higher degree of fragmentation (which lengthens names and slows comprehension) and the less-recognized disadvantages that complex sets of numerals, punctuation, parentheses, and other symbols are required to specify syntactical relations among components and that the names have a greater overall sameness and lack of distinctiveness. Consequently, they become more liable to errors in transcription (direct copying) and translation (to a structural diagram).

Whatever their logical attractions, such names are not generally of a type which chemists would willingly use in preference to established styles. The situation may be likened to that with a synthetic language, such as Esperanto, which has found few supporters as an international means of communication in comparison with two or three living languages. A new nomenclature system is likely to find applications only in areas where it can deal with new situations in a useful manner, for example, cyclophanes, where nodal nomenclature is clearly relevant.

(2) The major use of systematic names is in documenting the literature of chemistry in abstract journals and reference works. Chemical Abstracts Service has taken the lead in this important role by rationalizing current practices and providing valuable accessories, particularly the *Parent Compound Handbook* and the *CA Index Guides*. Despite all this, the role of systematic names is still mainly secondary: the molecular formula index is the primary tool for locating compounds, and the names in it serve merely to distinguish isomers of the same molform.

For this purpose it is not strictly necessary to derive a *unique* name for each compound; an *unambiguous* name suffices. This then accords with the situation of a searcher who is using formula indexes without having expert knowledge of nomenclature systems. He is able to interpret the alternatives which may be presented in a way which is heuristic rather than algorithmic.

Information scientists who have tried to use names for generic search, particularly in online mode, have come to realize the inherent limitations of even *9 CA* names as a basis for uniformly predictable descriptions of molecular structure. These arise from three general rules, which are designed to lead to unique names. They are, firstly, the priority rankings among functional groups, secondly, the precedence which is always given to the longest carbon chain, and thirdly, the alphabetical sequencing of substituents. The special methods used for naming symmetrical structures also cause substantial variations in styles of names (see Figure 1). Consequently, relatively small changes in structure can often lead to major changes in the forms of names for related compounds.

While CAS can be its own arbiter, the IUPAC rules for organic nomenclature[3] illustrate another facet of the problem. For many classes of structure two or even three alternative styles of name are permitted. Moreover, it has been my experience that the guidance provided is inadequate for many compounds encountered in practice and that even experts in IUPAC nomenclature differ over details of the name for a

* Address correspondence to 5 Albert Road, Wokingham, Berks RG11 2AL, United Kingdom

0095-2338/81/1621-0146$01.25/0