in the Introduction. The total number of rings generated in this way would generally be far fewer than if all the atoms of the structure were included; this reduction is possible because the first two phases of the algorithm create a partial SSSR which could be used to select the missing rings.

CONCLUSION

This paper has presented a fairly comprehensive algorithm for finding the smallest set of smallest rings. The speed with which the smallest rings are found is dependent on the sequence in which the paths are explored. This, in turn, depends on the way in which the atoms are numbered when input to the algorithm. Although the algorithm can be shown to fail in some cases, the ring systems for which it fails constitute a very minute portion of those which are chemically possible.

The algorithm was programmed in PL/1 and required 258 statements. It should be noticed that the basic procedure for finding the smallest set of smallest rings is independent of the technique used to implement the ring-finding algorithm. Although the ring-finding algorithm illustrated in this paper uses a path-tracing technique, other techniques such as growing a tree from the selected atom or atoms might offer advantages in particular situations.

ACKNOWLEDGMENT

The author thanks T. Ebe and J. Mockus for their encouragement and stimulating discussions.

REFERENCES AND NOTES

(1) A. M. Patterson, L. T. Capell, and D. F. Walker, "The Ring Index", 2nd ed, American Chemical Society, Washington, D.C., 1960.
(2) E. G. Smith, "The Wiswesser Line-Formula Chemical Notation", McGraw-Hill, New York, N.Y., 1968.
(3) J. T. Welch, Jr., "A Mechanical Analysis of the Cyclic Structure of Undirected Linear Graphs", *J. Assoc. Comput. Mach.*, **13**, 205–10 (1966).
(4) N. E. Gibbs, "A Cycle Generation Algorithm for Finite Undirected Linear Graphs", *J. Assoc. Comput. Mach.*, **16**, 564–8 (1969).
(5) C. C. Gotlieb and D. G. Corneil, "Algorithms for Finding a Fundamental Set of Cycles for an Undirected Linear Graph", *Commun. Assoc. Comput. Mach.*, **10**, 780–3 (1967).
(6) K. Paton, "An Algorithm for Finding a Fundamental Set of Cycles of a Graph", *Commun. Assoc. Comput. Mach.*, **12**, 514–8 (1969).
(7) J. C. Tiernan, "An Efficient Search Algorithm to Find the Elementary Circuits of a Graph", *Commun. Assoc. Comput. Mach.*, **13**, 722–726 (1970).
(8) M. Plotkin, "Mathematical Basis of Ring-Finding Algorithms at CIDS", *J. Chem. Doc.*, **11**, 60–63 (1971).
(9) A. Zamora and T. Ebe, "PATHFINDER II. A Computer Program That Generates Wiswesser Line Notations for Complex Polycyclic Structures", *J. Chem. Inf. Comput. Sci.*, preceeding paper in this issue.

Principle for Exhaustive Enumeration of Unique Structures Consistent with Structural Information

YOSHIHIRO KUDO and SHIN-ICHI SASAKI*
Miyagi University of Education, Sendai, 980 Japan

Received March 26, 1975

Unique structures consistent with structural information are enumerated by means of the "connectivity stack", the proper situation to provide an effective examination of the correct estimation of each structure, complete or even under construction, as one of the members of the "informational homologues". Both cyclic and acyclic structures are treated.

We have developed an integrated system for structure elucidation of organic compounds,[1-4] and called it CHEMICS.[5] The generic acronym CHEMICS stands for Combined Handling of Elucidation Methods for Interpretable Chemical Structures. It is a system for deducing all logically valid structures,[6,7] acyclic and cyclic, on the basis of previously settled propositions according to input information concerned with the structure of a given compound. Each logically valid structure is defined as an *informational homologue*[5,8] of provided structural information. If the information consists of only a molecular formula, the informational homologues are identical with structural isomers, whose members may even exceed millions.[9] Their composition depends on only the nature of the provided information; that is, the richer the information, the fewer informational homologues there are. In order to enumerate them not only completely and uniquely but also as quickly as possible,[16] a new principle of enumeration has been devised and has yielded many results for CHEMICS,[1-4] though most are not published in the literature. It was recently known that the principle in the heuristic DENDRAL[10,11] is very similar to ours because of mathematical permutation, though the object and order of application are different from each other. Mathematical permutation is one of the best ways for exhaustive enumeration, but really has practical value when hopeless branches of a logical tree are eliminated as early as possible. Balaban's report[12] directly stimulated publication of the original principle of our enumeration methods.

REPRESENTATION OF THE STRUCTURES

The enumeration part of CHEMICS combines static features with dynamic ones. The former is to carry out correct enumeration and the latter is to decrease execution time. How to represent structures goes along with both features.

Component and Segment. Most chemical systems, e.g., DENDRAL,[11] CAS/Morgan,[13] IUPAC/Dyson,[14] WLN,[15] represent a structure with canonical connectivities and after this with segments under constraint of hierarchical orders, in their own peculiar ways. On the other hand, CHEMICS considers segments in a hierarchical order by their parent components first and secondly constructs a suitable connectivity representation according to the order. The two concepts, component and segment, correspond to chemical element and atom in general chemistry, respectively. That is, the component is a logical division of partial structures, and the segment is an entity with the component as property. After setting the components, each part of a whole molecule is always specified with exactly one component. Two conditions, (1) and (2), define the concept of the component, C_i:

$$\bigcup_i C_i = \text{all whole structures} \quad (1)$$

$$C_i \cap C_j = 0 \quad (i \neq j) \quad (2)$$

There are many possible ways to set up components under the two conditions. There is no natural component set and a

Sequence number	1	2	3	4	5	6	7	8	9
(A,B,C)	A	A	B	B	B	C	C	C	C
(A,C,B)	A	A	C	C	C	C	B	B	B
(B,A,C)	B	B	B	A	A	C	C	C	C
(B,C,A)	B	B	B	C	C	C	C	A	A
(C,A,B)	C	C	C	C	A	A	B	B	B
(C,B,A)	C	C	C	C	B	B	B	A	A

Figure 1.

designer/operator may arbitrarily adopt other conditions to fix their set, even temporarily according to requirements. For example, in one of the realized systems, CHEMICS-F,[4] which treats CHO compounds, four hierarchical sets of components (basic, primary, secondary, and tertiary) are provided to describe analytical results from quite different information sources (NMR, ir, etc.) effectively and without contradiction. Basic components are chemical elements (C, H, and O), primary ones are CH_3, OH, and so on, and secondary and tertiary ones are bigger units of partial structures, e.g., *tert*-butyl, *gem*-dimethyl carbon, without and with afferent natures (that is, which are efferent ones of neighboring segments), respectively. The sets can sufficiently play the roles, though alone. Again, evidently, the component is not fixedly assigned to any chemical term. If wanted, for example, even all *n*-alkyl radicals can become components. In this case, duplication of longer alkyl radicals may be avoided by prohibiting connections of shorter alkyl radicals with methylene groups. In each, components are always arranged in a hierarchical order. For *n* kinds of components, the possible numbers of orders are *n*! ways; e.g., for three components, A, B, and C, the orders are six ways (6 = 3!): (A,B,C), (A,C,B), (B,A,C), (B,C,A), (C,A,B), and (C,B,A). Either of these orders also may be arbitrarily adopted; however, in fact, it is usually done during consideration of effectivity. For example, when carbon and hydrogen are assumed as components, an order (H,C) would never be selected because of the connectivity stack's[8] nature (see later).

Numbering of Segments. According to a hierarchical order, all segments in a set concerned with construction of reasonable structures are numbered. Figure 1 shows all numberings in a case of composition, $A_2B_3C_4$, where A, B, and C are components arbitrarily defined and numerals, 2, 3, and 4, are numbers of segments of the components, respectively. The first line expresses the sequence numbers of nine (9 = 2 + 3 + 4) segments. The second to the seventh lines are six possibilities of the numbering. All of these are logically reasonable, but not always efficient. The effects of the hierarchical orders are shown in Table I.

The Connectivity Stack. There are several ways to describe the connectivities in the form of a connection: a connection matrix, a connection table, a linear notation, and a structural formula. CHEMICS added a new one, a *connectivity stack*.[8] All of these are equivalent from the viewpoint of connectivity, and are suitably handled according to the local requirements in CHEMICS. The connectivity stack, say S, is a string of local connectivities, b_k, whose order is important: $S = b_1, b_2, \ldots, b_k \ldots$; the k and segment numbers, i and j, should be related to each other through a proper correspondence rule. The CHEMICS' correspondence rule decides the canonical form easily by partial examination of local connectivities, unlike, except for the worst cases, Morgan's method, so enumeration of structures is very easily and quickly carried out together with other tactics. Generally, the relation between a connectivity stack and either of the connection matrices is never self-evident. Among many alternative possibilities, CHEMICS adopts the following correspondence rule as the best:

$$k = i + (j-1)(j-2)/2 \qquad (i<j) \tag{3}$$

This rule connects a stack with a specific matrix (a_{ij}), making the latter symmetric by defining a_{ji} to be equal to a_{ij}. The diagonal elements, a_{ii}, do not need to be defined here and are not always zero. They should be arbitrarily[16] defined according to local requirements (in later parts, they are assigned to have certain values).

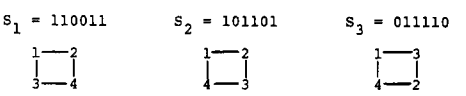

Figure 2.

In the same form as (3), the rule in Morgan's method, which does not always intend to enumerate, may be expressed by the equation

$$k = N(i-1) - i(i+1)/2 + j \qquad (i<j) \tag{4}$$

A remarkable difference between (3) and (4) is whether k is or is not a function of the total number of segments, N. That k is a function of only two concerned segments means using also a concept of infinition. Its effect qualitatively may be explained. In the examination of the canonical form, required iterations are about $j!$ (j represents the greater segment number of the two). Then during construction of one structure, their total would roughly become $\sum_{n=1}^{N} n!$. On the other hand, when k is a function of N as well as i and j, required iterations are $N!$, resulting in a total of $N(N!)$ iterations.

Canonical Stacks. One structure has usually more than one stack, being isomorphic. If there is a relation in the form of $S_i\Pi = S_j$ for either of the permutations

$$\Pi = \begin{pmatrix} p_1 p_2 \cdots p_k \cdots p_m \\ 1\ 2 \cdots k \cdots m \end{pmatrix}$$

the two stacks are isomorphic; otherwise not isomorphic. For example, when four segments of the highest bivalent component (assuming there is no other component) give a four-membered cyclic structure (imagine cyclobutane made from methylene groups, 1,1,2,2,3,3,4,4-octaethylcyclobutane, from *gem*-diethyl carbon groups, etc.) as only one informational homologue, three stacks represent an identical structure as shown in Figure 2. For example,

$$S_2 = S_1 \begin{pmatrix} 1243 \\ 1234 \end{pmatrix} \text{ and } S_3 = S_1 \begin{pmatrix} 1423 \\ 1234 \end{pmatrix} = S_2 \begin{pmatrix} 1324 \\ 1234 \end{pmatrix}$$

Among all isomorphic stacks, the greatest is a canonical form. In another words, if S is never less than $S\Pi$ for all possible Π's, the S is canonical, and vice versa. In the above case, S_1 is judged to be canonical, because it is the greatest of three (all). Checking the first and then the second elements, S_3 and S_2 will be denied, respectively, and the fact suggests that in the comparisons, not all elements in the stacks are always examined. The principle of the greatest values corresponds to the one of the smallest values in the heuristic DENDRAL,[11] though their objects to be applied to are different.

Some examples on comparison of canonical forms by several methods are shown in Figure 3. To make clear different points, chemical elements themselves are adopted as components in CHEMICS. In the real WLN, not numerals but alphabetical characters are used as the "locants".

This method may be expanded to incomplete structures. For example, the four-membered cyclic structure may also be formed with the first four segments of the highest component which can be trivalent, tetravalent, or higher, using the same stack. For an incomplete structure consisting of more than one fragment (maybe the most frequently encountered cases during real construction of structures), e.g., (–A–A–) together with (–A–), where A is the highest component, the three segments should be numbered, #1, #2, and #3. Possible stacks are 100 (S_1), 010 (S_2), and 001 (S_3), corresponding to numberings (–1–2–) and (–3–); (–1–3–) and (–2–); and (–2–3–) and (–1–). Of course, the canonical stack is always S_1.

Table I. Examples of the Implementation. Numbers of the Informational Homologues (IH) and Times Required under Various Hierarchical Orders, in the Run of a Certain System

						X = I(I), O(II), N(III) or P(V)					
	IH	CH sec	HC sec		IH	CXH sec	XCH sec	CHX sec	XHC sec	HCX sec	HXC sec
CH_4	1[9]	0.0	0.0								
C_2H_6	1[9]	0.0	0.4	C_3H_7I	2	0.00	0.00	0.8	375	397	574
C_3H_8	1[9]	0.00	2128	C_4H_9I	4	0.16	0.34	9.5			
C_4H_{10}	2[9]	0.1									
C_5H_{12}	3[9]	0.4		C_4H_4O	62[11]	3	4	7	34	74	94
C_6H_{14}	5[9]	1.1									
C_7H_{16}	9[9]	4		C_6H_6O	2237[11]	301	367				
C_8H_{18}	18[9]	12		C_6H_8O	1623	235	271				
C_9H_{20}	35[9]	41		$C_6H_{10}O$	747[11]	118	130				
C_3	3	0.0	(0.0)	$C_6H_{12}O$	211[11]	40	40				
C_3H_2	2	0.0	0.0	$C_6H_{14}O$	32[9]	7	7				
C_3H_4	3	0.0	0.6								
C_3H_6	2	0.00	2.1	$C_3H_2N_2$	86	4	5	5	9	6	9
C_3H_8	1[9]	0.00	2128	$C_3H_4N_2$	155[11]	8	8	14	54	91	182
C_3H_{10}	0	0.0	0.0	$C_3H_6N_2$	136[11]	8	8	25	695	3194	7826
C_6	19	8	(8)								
C_6H_2	85	36	36	$C_4H_9P^V$	110[11]	9	9	118	*b*		
C_6H_4	185	55	980								
C_6H_6	217[11,12]	52	*a*								
C_6H_8	159[11]	35									
C_6H_{10}	77[11]	22									
C_6H_{12}	25[11]	5									
C_6H_{14}	5[9,11]	1									
C_6H_{16}	0	0.0	0.0								

[a] 480/the first three, 1000/the first six, and 1670/the first nine. [b] 230/~1; 458/~2; 738/~3; 1139/~4 and 1416/~5.

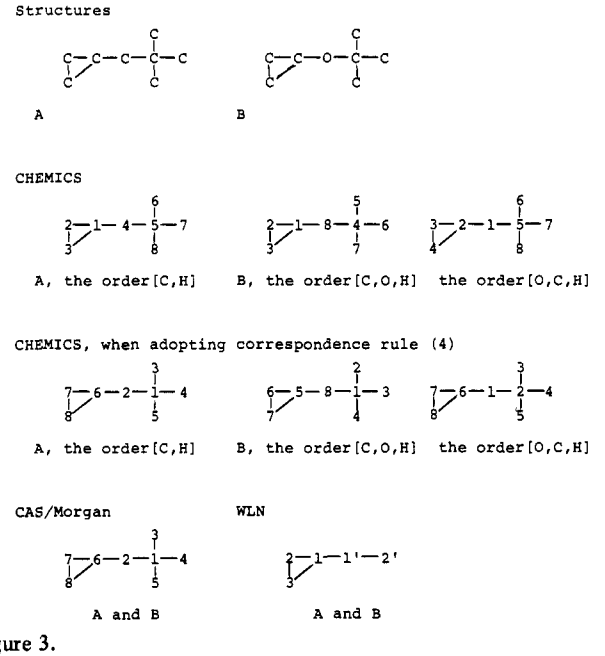

Figure 3.

Independent of how many components are concerned with construction, the canonical stack depends on connectivities between segments of the highest component first, and, after this, segments of the next highest component. Therefore as already mentioned, the order (H,C), instead of (C,H), results in zero elements (because connection of two hydrogen atoms is automatically prohibited, leading to nonconnective structures) in important places for judgment of the canonical form; e.g., in the case of $X^{IV}M^{I}_4$ (e.g., CH_4), *S*'s for (X,M) and (M,X) are 1101001000 and 0000001111, respectively. In the case of X_2M_6 (e.g., C_2H_6) in both situations, (X,M) and (M,X), the canonical stacks and some noncanonical stacks, in each case, of the 20 possible stacks, are shown in Figure 4.

Clearly the difference of effects would be appreciated by considering that comparison of two isomorphic stacks starts from the first elements. In the case of examination of whether the stack itself is canonical or not, the difference would increase because of permutation. That is, in the examination of the *k*th place, *k*! times of iterative calculations are required, qualitatively. In the latter situation, the value of *k* is usually greater than that in the former, and, further, the judgment is impossible until nonzero elements are detected.

The method for detection of isomorphism may accommodate only components which have bonding sites of single nature. On the other hand, the real chemistry demands ways which accommodate components which have bonding of plural nature (e.g., –CO–O–, =N–). The realized systems devised several ways connecting demand and ability. One of them is a set of tertiary components in CHEMICS-F. They are provided with efferent and afferent natures; e.g., –CO– (O,D) can be connected only with ether oxygen segments at the first bond and with olefinic carbon segments at the second, and cannot but form vinyl esters. Also in the other system, the same ideas are used.[2,3] The second way is the matching method in SIEDS;[7] any desired partial structure is examined whether or not it is contained in a whole structure after construction. The third way, which is not always good, is to select an easy operation rather than to avoid duplication in cases of cyclic partial structures in SIEDS. In the last case, if the partial structure is not formed from the remainder of the segments, no duplication occurs.

Table I shows several examples of the required execution time for enumeration, in which the time for printing individual structures is excluded. Numerals in the table are the time (sec) required to run the program on the JEC-6 spectrum computer (JEOL; core 4 kwords, 16 bits/word, and execution time 5 μsec). The program, if it is designed to treat less than 31 segments at the same time, may be stored in 4 kwords of the memory size, containing all necessary subroutines. As shown in the table, the times depend on hierarchical order, number of concerned components, valences of the components, number of segments of each component, and others. Clearly, it increases greatly with *M*, the number of segments of the highest and monovalent component (represented by hydrogen here). Qualitatively, it has been predicted that the time is about

```
(A) [X,M]

      XMMM-XMMM
# 1   1345-2678    1 10 100 1000 01000 010000 0100000
# 2   1346-2578    1 10 100 0100 10000 010000 0100000
# 3   1347-2568    1 10 100 0100 01000 100000 0100000
--------------
#20   1678-2345    1 01 010 0100 10000 100000 1000000


(B) [M,X]

      XMMM-XMMM
# 1   7123-8456    0 00 000 0000 00000 111000 0001111
# 2   7124-8356    0 00 000 0000 00000 110100 0010111
# 3   7125-8346    0 00 000 0000 00000 110010 0011011
--------------
#20   7456-8123    0 00 000 0000 00000 000111 1110001
```

Figure 4.

directly proportional to $M!$ because of the nature of canonicalization rule and construction method of the stack in CHEMICS. Confirmation of exactly one isomer of C_3H_8 requires 0.00 and 2128 sec in the order (C,H) and (H,C), respectively. Of course, the program itself does not care about chemical meaning. In empty spaces in the table, very large values, which are at least larger than any of the neighbors, are expected, some of which were too large to finish the measurement because even the first structure of some hundreds is not constructed in the first 15 min.

CONSTRUCTION OF STRUCTURES

Accomplishment of All Unique Connectivity Stacks. The informational homologues are enumerated by construction of all possible unique structures, after obtaining a proper set of segments. The enumeration starts with arrangement of the component segments in the order of the component hierarchy. If there are n segments of the highest hierarchical component of components concerned with, they are sequencially numbered from the 1st to the nth. Further there are m segments of the next highest component, which are numbered from the $(n+1)$th to the $(n+m)$th. They must neither become earlier than the $(n+1)$th nor later than the $(n+m)$th. First, as a target, the second segment (#2) is picked up and the connectivity with the first segment (#1) is examined as to whether it obeys the several limitations mentioned below. Generally, a target, say the jth segment (#j), is picked up and the connectivities with the segments from #1 to #$(j-1)$ are examined in turn. The value of j varies between two and the total number of the segments, always by one.

This means constructing the stacks from the first place to the last place. To a connectivity between #i and #j b_k, the bond degree between the two, is assigned (3 for triple, 2 for double bond, etc.; hypothetical or conventional values more than three are also possible; otherwise b_k is zero). When the element of the last $[(N-1)(N-2)/2]$ place (N: total number of segments) is reasonably established, a new informational homologue appears. After that or after any failure of assignment during construction, the last element at that time is examined. If it can be decreased by one, the value is substituted with a value smaller by one and stack formation is continued forward to the $[(N-1)(N-2)/2]$th place; otherwise, by retrogradation of the stack, the last element, of which decrement by one is possible, is searched for. As a result, in the retrogradation, the stack is shortened one by one. Nonzero elements are increasingly shifted to higher places, so that if the retrogradation, at a time when the lower places are all zero, erases all elements of the stack, the enumeration is complete. (This aspect is shown in Figure 8).

Examination of Duplication. An obvious way to avoid duplication, like comparison of a structure with complied canonical representations, is impractical. As above mentioned, a permutation method was created for CHEMICS. Some of its merits are that a structure can examine by itself whether

```
Π                          sΠ
1 2 3 4      -3-1-2-4-     S = 1 10 010
1 2 4           1 2 3          1 0
1 3           2 1              1
1 3 2         2 1 3            1 10
1 3 2 4       2 1 3 4          1 10 00
1 3 4         2 1   3          1 0
1 4             1   2          0
2 1             2 1            1
2 1 3         3 2 1            1 0
2 1 4           2 1 3          1 10
2 1 4 3       4 2 1 3          1 10 010
2 3           2   1            0
2 4               1 2          1
2 4 1             3 1 2        1 10
2 4 1 3       4 3 1 2          1 10 00
2 4 3         3   1 2          1 0
3 1           1 2              1
3 1 2         1 2 3            1 0
3 1 4         1 2   3          1 0
3 2           1   2            0
3 4           1     2          0
4 1             2   1          0
4 2               2 1          1
4 2 1           3 2 1          1 0
4 2 3         3   2 1          1 0
4 3           2     1          0
END
There is no sΠ greater than the original S.
```

Figure 5. Examination of S = 1 10 010.

```
Π                          sΠ
1 2 3 4      -1-2-3-4-     S = 1 01 001
1 2 4          1 2   3         1 00
1 3            1   2           0
1 4            1     2         0
2 1            2 1              1
2 1 3          2 1 3            1 1        >S
There is at least one sΠ greater than the original S.
```

Figure 6. Examination of S = 1 01 001.

it is canonical or not, that the examination is applicable also to incomplete structures during construction, and that it is not necessary to check all elements. In a previous explanation, it was considered that $n!$ iterative calculations are required for a set of n segments. However, the number is limited when all segments belong to one component. When we are concerned with plural components, for instance, in the case of $A_2B_3C_4$, the required number is roughly at most not 3×10^5 (9!) but 3×10^2 (2! 3! 4!), because, in the hierarchical order (A,B,C), numbering two, three, and four segments of components A, B, and C is limited within (#1 to #2), (#3 to #5), and (#6 to #9), respectively.

Moreover, in fact, in some cases even one permutation is not necessary to be established. During the establishment, judgment is involved, and if a given permutation is unsuitable, another permutation is searched for in an as short a way as possible. Two stacks of many possibilities for a series of four segments of a bivalent component, (-B-B-B-B-), S = 1 10 010 to (-3-1-2-4-) and S = 1 01 001 to (-1-2-3-4-), are examined as shown in Figures 5 and 6, respectively. Figure 5 shows that there is no $S\Pi$ greater than the original S; S = 1 10 010 is canonical. Figure 6 shows that there is at least one $S\Pi$ greater than the original S; S = 1 01 001 is not canonical. Both figures, moreover, suggest that almost all permutations may play their roles before complete formation; this is very important for quick enumeration.

Examination of Valency. Any connectivity b_k has two corresponding two-dimensional elements, a_{ij} and a_{ji} ($a_{ij} = a_{ji}$). A connectable number of each segment of component n must not exceed its peculiar valence, so condition 5 must be satisfied for all n's.

$$\sum_{k \neq n} a_{nk} \leq v_n \tag{5}$$

Diagonal elements, a_{ii}'s, are unnecessary to be defined, but

in the algorithm they are all temporarily assigned zero, for simplicity. (For different purposes, different values are given.) During the buildup of the structures, the values of all a_{ij}'s are not simultaneously determined, so the condition is checked whenever each a_{ij} is examined. If equalities are attained for all n segments, the stack represents a complete structure which has no unused bond; otherwise, it represents an incomplete structure. The second condition comes from the fact that real structures are finite though partially under the concept of infinity. When an element of the kth place of a stack is established, with k, defined in the relation

$$k = (j-1) + (j-1)(j-2)/2 = (j-1)j/2 \quad (i \equiv j - i) \tag{6}$$

Summation of unused bonds of segments in higher numbers must not be smaller than summation in lower numbers

$$\sum_{n \leqslant j} (\text{unused bond})_n = \sum_{n>j} (\text{unused bonds})_n \tag{7}$$

Examination of Connectivity. Clathrates, catenanes, rotaxanes, tanglanes,[3] and so on, which are nonconnected structures, may be neglected intentionally, if disliked. The unconnected structures are differentiated from the connected ones by the following method, which was empirically developed. The two contradictory concepts came from the graph theory. A nonconnected structure consists of more than one connected structure as a component (which is a technical term in the theory). For simplicity, a connection matrix which is equivalent to a connectivity stack is used. However, this time, diagonal elements are not defined, so, as if they are eigenvalues, each a_{ii} is assigned minus values of valences ($-v_i$). The determinant D is derived from the matrix M.

For a complete structure,

$$D = |M| = 0 \tag{8}$$

because sums of every row or column become zero by definition on diagonal elements (cf. eq 5).

For an incomplete structure, which is obtained by elimination of any segment or bond from the connected structure, the determinant D_- is derived from the corresponding matrix M_-, leading to the following empirical relation.

$$D_- = |M_-| \neq 0 \tag{9}$$

This is a working hypothesis which neither has mathematical proof nor disproof.

Another type of incomplete structure, which is obtained by addition of some segments to a complete and connected structure, is represented by a matrix $M+$, from which the determinant $D+$ is derived. This structure cannot but form nonconnected structures; i.e., it is a precursor of nonconnected structures. In this case, $D+$ is always zero regardless of added segments, as clarified by the equation

$$D_+ = |M_+| = \begin{pmatrix} M & 0 \\ 0 & A \end{pmatrix} = |M| \times |A| = 0 \times |A| = 0 \tag{10}$$

The fact that the determinant becomes zero means construction accomplishment of a connected structure or appearance of a precursor of nonconnected structures, and, as mentioned in the section of structure construction, this is one of the occasions for which a connectivity stack has to be remade by retrogradation.

The real algorithm examines whether a new family, which is formed from two families of segments, i and j, by combining them, has at least one unused bond, where family means a group of segments connected with bond–segment–bond fashion.

Examination of Number of Rings. The maximum number of rings, R, in a structure is automatically derived from a set of segments composing the structure according to the equation

$$R = B - E + F \tag{11}$$

where B, E, and F are numbers of bonds, segments, and components (in the graph theory), respectively. The former two are calculated from numbers of segments, T_v, with valences in which v is a valence, as follows:

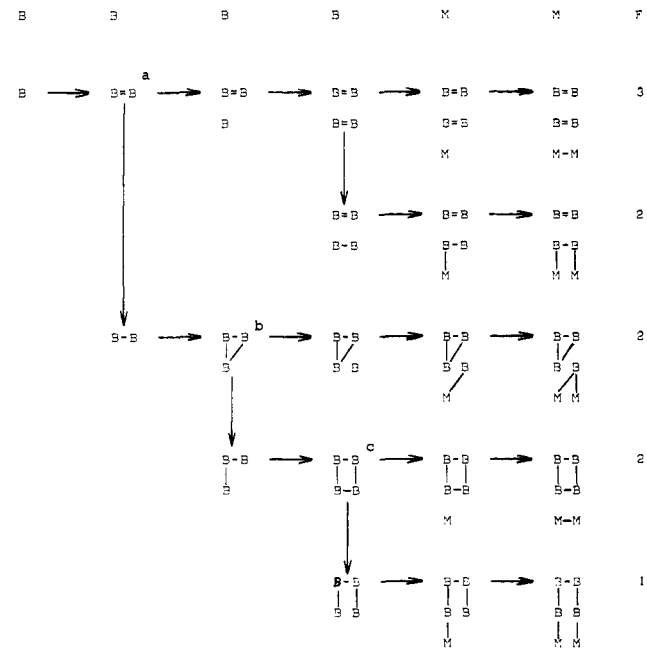

Figure 7. Roles of the examination on connectivity of structures and number of rings in the construction of the informational homologues of $B^{II}{}_4M^{I}{}_2$.

$$B = (\sum_v v T_v)/2 \tag{12}$$

$$E = \sum_v T_v \tag{13}$$

With (11) this results in

$$R = \sum \left(\frac{v}{2} - 1 \right) T_v + F \tag{14}$$

In cases of connected structures, F is one. During construction, if more than one ring and multibond is formed beyond the maximum number, the construction cannot succeed. For example, there is a set of four and two segments of a bivalent component, B, and a monovalent component, M, respectively. In that case T_1 equals 2, and T_2 equals 4, resulting in $R = F - 1$. If B is higher than M in hierarchical order, construction of all possible structures is carried out as shown in Figure 7.

In the figure arrows to the right and the lower mean bond formation and bond order decrement, respectively. In this case formation of any multibond and any ring results in nonconnected structures ($F > 1$), and only one connected structure ($F = 1$) is constructed. In the algorithm, three unsuitable structures (a, b, and c in the figure) of eight precursors are not constructed, but only imagined for the examinations.

Thus, the examination of the number of rings during structure construction makes prediction of nonconnected structures. The algorithm examines whether two segments, which determine the possibility of bond formation, have any path through other segment(s).

In order to examine whether a segment j is one of ring segments in any n-membered ring, the diagonal elements m_{ii}'s in a product of the connection matrix are calculated. Let $(M)_n$ be the nth power of (M); thus

$$(M)_n = (M)^n \quad (n \geqslant 2) \tag{15}$$

In this multiplication, the conditions in (16) have to be satisfied:

(a) $a_{ij}{}^2 = 0$
(b) $a_{ij} \times a_{ji} = 0$
(c) In all other cases, a simple scalar multiplication is carried out. (16)

Then, if a segment j is not one of ring segments of any n-

```
 J I   J I   BOND V C D  J-I  J/I     STACK

 2 1         =   G B
             -   G G G  2-1         1
 3 1         -   G G G  3-1         1 1
 3 2         -   G B                1 10
 4 1         -   B                  1 10 0          1-2
 4 2         -   G G G  4-2         1 10 01
 4 3         -   G G G  4-3         1 10 011        3-4
             -   G B          4/3   1 10 010
>4                                  1 10 010***    -3-1-2-4-
       4 3                          1 10 010
       4 2                          1 10 01
 4 2                          4/2   1 10 00
 4 3         -   G G B              1 10 000
>4                                  1 10 000***    -2-1-3-, & -4-
       4 3                          1 10 000
       4 2                          1 10 00
       4 1                          1 10 0
       3 2                          1 10
       3 1                          1 1
 3 1                          3/1   1 0
 3 2         -   G G B              1 00
 4 1         -   G G B              1 00 0
 4 2         -   G G B              1 00 00
 4 3         =   G B
             -   G G G  4-3         1 00 001
>4                                  1 00 001***    -2-1-, & -4-3-
       4 3                          1 00 001
 4 3                          4/3   1 00 000
>4                                  1 00 000***    -2-1-, -3-, & -4-
       4 3                          1 00 000
       4 2                          1 00 00
       4 1                          1 00 0
       3 2                          1 00
       3 1                          1 0
       2 1                          1
 2 1                          2/1   0
 3 1         =   G B
             -   G G B              0 0
 3 2         =   G B
             -   G G B              0 00
 4 1         =   G B
             -   G G B              0 00 0
 4 2         =   G B
             -   G G B              0 00 00
 4 3         =   G B
             -   G G B              0 00 000
>4                                  0 00 000***    -1-, -2-, -3-, & -4-
       4 3                          0 00 000
       4 2                          0 00 00
       4 1                          0 00 0
       3 2                          0 00
       3 1                          0 0
       2 1                          0
       END                          VANISHED
```

Figure 8.

membered ring, m_{ii} becomes zero, and vice versa. A history of a each diagonal element of $(M)_n$ is implied with all factors, which are elements of M. For example, $(a_3)_{11}$ of $(M)_3$ consists of three connectivities, $2(a_{12}a_{13}a_{23} + a_{12}a_{14}a_{24} + \ldots + \alpha_{1(n-1)}\alpha_{1n}\alpha_{(n-1)n})$. A term, say $a_{ij}a_{ik}a_{jk}$, denotes whether or not three segments i, j, and k, form a three-membered ring. This examination would be useful not only for control of numbers of rings of desired size, but also for detection of ring closure.

Figure 8 shows aspects of formation of canonical connectivity stacks with several kinds of examinations. A composition of the model is A_4B_b, where component A is the highest bivalent component and B represents all other components. Though only connectivities between four segments of component A are discussed, the figure would make the functions of many examinations deducible. Evidently, these connectivities are the most important for formation of the canonical stacks. There are six states in the relation between the four segments: (i) four-membered ring; (ii) (–A–A–A–A–); (iii) (–A–A–A–) and (–A–); (iv) (–A–A–) and (–A–A–); (v) (–A–A–), (–A–), and (–A–); and (vi) (–A–), (–A–), (–A–), and (–A–). Which of them are most reasonable depends on the content of B_b. For example, if "b" is equal to zero, only (i) is possible; otherwise (i) forms a component of any nonconnected structures.

In the first line of Figure 8, J and I are segment numbers we are concerned with, BOND is the bond order to be examined, V, C, and D are the examinations on valence, connectivity, and duplication, respectively, and their results are expressed with G (good) and B (bad), J-I and J/I stand for bond formation and cleavage, and STACK is a connectivity stack, whether complete or incomplete structure. The aspects of the connection correspond to six states of A_4 in the text. The contents of the stacks which are implied in the expression of three asterisks (***) depend on B_b of the composition of A_4B_b.

COMBINATION OF NUMBERS OF SEGMENTS

As shown in above sections, the fewer the segments of each component, the less the required calculations. For the purpose of reasonable distribution to a more precisely divided component set, combination of numbers of segments is carried out within numbers of segments at a parent level of components. In CHEMICS-F, four levels of component sets are provided. This means conversion of a considerable part of inevitably time-consuming checking for avoiding duplication to an easier arithmetic combination of simple values. For example, if chemical elements are replaced by methyl–methylene level components, the differentiation between this level of partial structures is accomplished not with connection states but with component numbers. Its compensation is a small increment of required memory size for storing a table of the new components and the algorithm for the combination. Of course, if a set of numbers of segments at the deepest level of components is directly input, as Setting 1 in Chart I, this function is unnecessary.

For instance, eq 17 must be solved to enumerate the informational homologues for an input molecular formula via a component set, e.g., consisting of six components, CH_3, CH_2, CH, C, OH, and O. The equations are equivalent to si-

$$\begin{array}{c} \quad CH_3\ CH_2\ CH\ C\ OH\ O \\ \begin{matrix} C \\ H \\ O \end{matrix}\begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}\begin{pmatrix} CH_3 \\ CH_2 \\ CH \\ C \\ OH \\ O \end{pmatrix} = \begin{pmatrix} 2 \\ 6 \\ 1 \end{pmatrix} \end{array} \qquad (17)$$

multaneous linear equations with six inequalities. In this case the number (three) of equations is less than that (six) of the unknown quantities and, though six inequalities are added, the unique solution is not obtained, but two solutions are derived:

$(CH_3\ CH_2\ CH\ C\ OH\ O) = (2\ 0\ 0\ 0\ 0\ 1)$ and $(1\ 1\ 0\ 0\ 1\ 0)$

Equation 17 is generalized by the following formula:

$$(CM)\cdot(DCV) = (PCV) \qquad (18)$$

where (PCV) and (DCV) are component vectors meaning a set of numbers of segments, of parent and daughter components, respectively; (CM) stands for a component matrix defining the properties of daughter components with parent components and is fixed when the daughter component set is established. In the above case, (PCV) is a molecular formula. Generally, (PCV) is directly input or calculated as (DCV) from its (PCV). In other words, CHEMICS obtains (DCV) from known (CM) and (PCV) with a substitution method: possible values are substituted for all elements of (DCV), and whether eq 18 is true or not is examined. In the results, the first component vector, (2 0 0 0 0 1), means two methyl groups and an ethereal oxygen atom, and the second one, (1 1 0 0 1 0), a methyl, a methylene, and a hydroxy group. Each component vector yields a subset of the informational homologues. Clearly, dimethyl ether and ethanol are derived from the first and the second vectors, respectively.

Examples of Enumeration. Three examples of the enumeration are shown in Chart I: settings 1, 2, and 3. In all cases, the informational homologues are enumerated according to the information of molecular formula C_6H_6. Their settings of components are near to basic, primary, and secondary ones in CHEMICS-F, irrespective of intention.

Since only a molecular formula is the input information, the informational homologues in all examples are identical with the 217 structural isomers from the chemical viewpoint. This fact also shows that the informational homologues depend on not a process of an algorithm which determines the efficiency,

Chart I

Setting 1: (a) Input information: molecular formula C_6H_6.

(b) Enumeration

Components	Component vectors
	#1
C	6
H	6
Number of informational homologues	217

Setting 2: (a) Input information: molecular formula C_6H_6.

(b) Enumeration

Components	Component vectors						
	#1	#2	#3	#4	#5	#6	#7
C	4	3	3	2	2	1	0
CH	0	1	0	3	2	4	6
CH_2	0	1	3	0	2	1	0
CH_3	2	1	0	1	0	0	0
The homologues	7	46	16	34	76	32	6
						Total	217
Corresponding	#1	#2	#4	#7	#8	#10	#11
Component vectors	#5	#3		#9			
In Setting 3		#6					

Setting 3: (a) Input information: molecular formula C_6H_6.

(b) Enumeration

Components	Component vectors										
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11
C	3	3	3	3	2	2	2	2	1	1	0
CCH_3	0	0	0	0	2	1	0	0	1	0	0
$C(CH_3)_2$	1	1	0	0	0	0	0	0	0	0	0
$C(CH_3)_3$	0	0	0	0	0	0	0	0	0	0	0
CH	0	1	0	0	0	1	2	2	3	4	6
$CHCH_3$	0	0	1	0	0	0	1	0	0	0	0
$CH(CH_3)_2$	0	0	0	0	0	0	0	0	0	0	0
CH_2	0	0	1	3	0	1	0	2	0	1	0
CH_2CH_3	0	1	0	0	0	0	0	0	0	0	0
The homologues	2	7	9	16	5	30	20	76	14	32	6
										Total	217

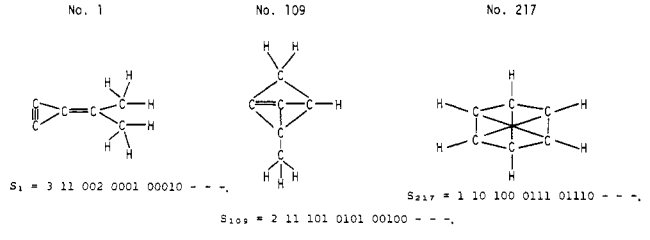

Figure 9. Three structures (the first, the middle, and the last) of 217 structural isomers of the molecular formula C_6H_6.

but on the input information. In setting 1 (Chart I), a necessary component vector is directly input as a molecular formula, so in this case the combination part is unnecessary. Figure 9 shows the first, the middle, and the last of 217 structures in the form of a structural formula with the corresponding canonical connectivity stacks. In setting 2, they are enumerated through a set of four components. In setting 3, the number of components increases to nine by subordinating the methyl group. Figure 10 shows some of the 217 structures by means of components in settings 2 and 3: (a) seven from #1 in setting 2, and the corresponding two and five from #1 and #5, respectively, in setting 3; (b) six structures from #7 in setting 2 or #11 in setting 3. These processes suggest a possibility for various types of manipulation methods of structural information. Really, in CHEMICS-F, four levels of components are properly utilized to describe the structural information according to contents of analytical results of the input data.

In the accompanying paper, we report (i) a Structural Isomers Enumeration and Display System (SIEDS)[7] which enumerates the informational homologues and displays each structure in the form of the structural formula, and (ii) in the

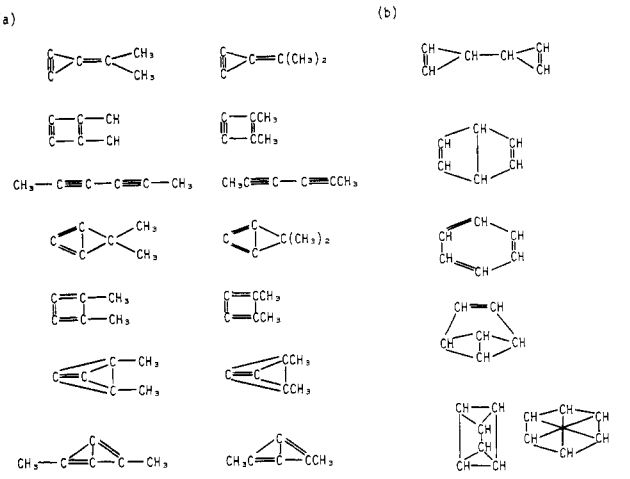

Figure 10. (a) Seven structures of the formula $[C]_4[CH_3]_2$ and their corresponding structures of $[C]_2[CCH_3]_2$ and $[C]_3[C(CH_3)_2]$; (b) six structures of the formula $[CH]_6$.

near future we shall report a system for structural elucidation of organic compounds consisting of carbon, hydrogen, and oxygen, (CHEMICS-F),[4] which enumerates the informational homologues on the basis of a molecular formula and chemical spectral information. In both systems, the enumeration is performed with the same principles as mentioned above, while more complicated conditions and many limitations are added.

REFERENCES AND NOTES

(1) S. Sasaki, Y. Kudo, S. Ochiai, and H. Abe, *Mikrochim. Acta*, 726 (1971).
(2) The 1011 System: S. Sasaki, Y. Kudo, S. Ochiai, and I. Fujioka, *Jpn. Anal.*, **22**, 25 (1973).
(3) Y. Kudo, *Kagaku no Ryoiki Zokan*, **98**, 115 (1972).
(4) CHEMICS-F, unpublished.
(5) S. Sasaki, H. Abe, Y. Kudo, S. Ochiai, and Y. Ishida, *Kagaku No Ryoiki*, **26**, 981 (1972).
(6) Y. Kudo, and S. Ochiai, *Anal. Instrum. (Bunseki-Kiki)*, **11**, 654 (1973).
(7) A Structural Isomers Enumeration and Display System (SIEDS); Y. Kudo, Y. Hirota, S. Aoki, Y. Takada, T. Taji, I. Fujioka, K. Higashino, H. Fujishima, and S. Sasaki, *J. Chem. Inf. Comput. Sci.*, following paper in this issue.
(8) Y. Kudo and S. Sasaki, *J. Chem. Doc.*, **14**, 200 (1974).
(9) H. R. Henze and C. M. Blair, *J. Am. Chem. Soc.*, **53**, 3042, 3077 (1931); **56**, 157 (1934).
(10) L. M. Masinter, N. S. Sridharan, J. Lederberg, and D. H. Smith, *J. Am. Chem. Soc.*, **96**, 7714 (1974).
(11) L. M. Masinter, N. S. Sridharan, R. E. Carhart, D. H. Smith, *J. Am. Chem. Soc.*, **96**, 7714 (1974).
(12) A. T. Balaban, *Rev. Roum. Chim.*, **18**, 635 (1973).
(13) H. L. Morgan, *J. Chem. Doc.*, **5**, 107 (1965).
(14) IUPAC, "Rules for IUPAC Notation for Organic Compounds", Commission on Codification, Ciphering and Punched Card Techniques of the IUPAC, Wiley, New York, N.Y., and Longmans, Green, London, 1961.
(15) E. G. Smith, "The Wiswesser Line-Formula Chemical Notation", McGraw-Hill, New York, N.Y., 1968.
(16) In connection with the theme, the concepts, Atom Connection Matrix (ACM) and its Characteristic Polynomial (ACMCP), have been discussed. The ACM is a connection matrix whose diagonal elements are not numeral values but the attributes of atoms. There was a proposition that ACMCP uniquely presents the topology of a molecule. "Our" article on the proposition [*J. Chem. Doc.*, **13**, 225 (1973)] was vigorously attacked by W. Herndon [*ibid.*, **14**, 150 (1974)]. He is quite correct because "our" article is logically funny considering the way "we" announced a conclusion without any proof, regardless of his proof that the proposition is false. Accidently "our" conclusion was that the proposition is true. Fortunately, the conclusion on the proposition does not have influence on the present paper.

After Spialter's conjecture and Balaban's and Hosoya's proofs, we wrote an article, "Does not the Characteristic Polynomial Uniquely Determine the Topology of a Molecule?", to this Journal, in which we indicated that Balaban's and Hosoya's proofs were not correct because their "ACMCP" was a special one as a result of a certain modification, and that no example denying the conjecture, that the proposition is true, had up to that time been found (i.e., it was not clear at that time whether the proposition was true or not). To our surprise, we received a galley proof, "The Characteristic Polynomial Uniquely Represents the Topology of a Molecule". We then claimed that our original title and content had to be maintained. We found, however, only "our" article in the same form as the galley, whose author is unknown, in the Journal.