(7) M. S. Gordon and J. A. Pople, Program MBLD, *QCPE*, **10**, 135 (1975).

(8) W. T. Wipke, in "Computer Representation and Manipulation of Chemical Information", W. T. Wipke, S. R. Heller, R. J. Feldmann, and F. Hyde, Eds., Wiley, New York, 1974, p 153.

(9) P. K. Weiner and S. Profeta, Jr., "Abstracts", 181st National Meeting of the American Chemical Society, Atlanta, GA, March 29–April 3, 1981, American Chemical Society, Washington, DC, 1981, COMP Division, Abstract No. 15.

(10) N. C. Cohen, P. Colin, and G. Lemoine, *Tetrahedron*, **37**, 1711 (1981).

(11) G. M. Crippen, *J. Comput. Phys.*, **24**, 96 (1977).

(12) G. M. Crippen, *J. Comput. Phys.*, **26**, 449 (1978).

(13) G. M. Crippen and T. F. Havel, *Acta Crystallogr., Sect A*, **A34**, 282 (1978).

(14) J. G. Nourse, *J. Am. Chem. Soc.*, **101**, 1210 (1979).

(15) J. G. Nourse, R. E. Carhart, D. H. Smith, and C. Djerassi, *J. Am. Chem. Soc.*, **101**, 1216 (1979).

(16) *Spec. Publ.—Chem. Soc.*, **No. 11** (1958); **No. 18** (1965).

(17) This formula has been used for test runs of the EMBED program. See Acknowledgment for further information.

(18) T. F. Havel, G. M. Crippen, and I. D. Kuntz, *Biopolymers*, **18**, 73 (1979).

(19) T. M. Dyott, "Utilization of Stereochemistry and Other Aspects of Computer-Assisted Synthetic Design", Ph.D. Thesis, Princeton University, 1973.

(20) R. Fletcher and C. M. Reeves, Comput. J., 7, 149 (1964).

(21) W. Braun, C. Bosch, L. R. Brown, N. Go, and K. Wuthrich, *Biochim. Biophys. Acta*, **67**, 377 (1981).

(22) Y. Beppu, *QCPE*, **11**, 370 (1978).

Exhaustive Generation of Structural Isomers for a Given Empirical Formula—A New Algorithm

SI-YU ZHU* and JIN-PEI ZHANG

Institute of Elemento-Organic Chemistry, Nankai University, Tianjin, People's Republic of China

Received July 20, 1981

A new algorithm is presented for generating an exhaustive, irredundant list of structural isomers that are consistent with a given empirical formula. The new algorithm is simpler than existing ones, and the required program can be executed on many mini/microcomputers.

INTRODUCTION

Structural isomerism is a classical problem in organic chemistry, and the enumeration of such isomers has received considerable attention.[1,2] Yet, despite these efforts, it is still not possible to determine the number of structural isomers that are possible for an arbitrary empirical formula C_nH_m.

In recent years Smith and co-workers[3] have developed a computer program which aids considerably in this task for molecules which are not too large. However, their algorithm is rather complex and their program, CONGEN, is written in a generally unfamiliar language, INTERLISP. In this paper we described an alternative algorithm. The resulting program, ISOGEN, which is written in FORTRAN-IV can be executed on many mini/microcomputers.

RESULTS

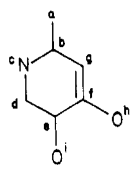
Representation of Structures. Following previous precedent, we have elected to represent molecular structures through the convenience of a connection matrix. Thus, a structure having n nonhydrogen atoms can be represented by a $n \times n$ connection matrix (CM) in which the ith row and the ith column correspond to the ith atom, and the entries CM(i,j) and CM(j,i) represent the bond connecting the ith and jth atoms, while single, double, and triple bonds are represented, respectively, by the digits 1, 2, and 3 and no bond by 0. For example, structure **1b** is represented by the following CM matrix:

$$\begin{array}{c|ccccccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
\hline
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
3 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
4 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
5 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
6 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
7 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\
8 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 2 \\
9 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 2 & 0
\end{array}$$

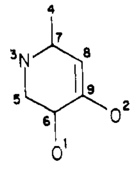
Since it is well-known that there are n! permutations for a structure containing n nonhydrogen atoms, a unique representation under such a system depends upon the ability of the algorithm to designate individual atoms uniquely. Following the examples of several previous authors who employed numbering schemes to effect this end,[4,5] we have chosen as our defining algorithm a numbering sequence based in part upon the earlier algorithm of Ugi and Shubert.[6] However, inasmuch as their method fails to provide a unique numbering for some molecules,[6] we have carried out several modifications. The resulting algorithm and its application to the generation of structural isomers is described below.

I. NUMBERING ALGORITHM

First Level. Sort all the nonhydrogen atoms of a given structure according to their atomic *type* (i.e., C, N, O, S, etc.). Although the ordering of the types of atoms is arbitrary, in this paper we have assigned precedent to atoms according to their respective valence, R. Thus, for example, O (R = 2) and S (R = 2) precede N (R = 3) and P (R = 3) which in turn precede C and Si (R = 4) and so on. Atoms which are dissimilar in type but which have equivalent valences (for example, O and S) are ordered arbitrarily. Application of these first-level criteria to the numbering of structure **1a** is shown in Table I.



1a



1b

Second Level. Atoms which are the same in type (e.g., all carbon atoms) require a second level of numbering, the criteria for which is based upon the *number* and *type* of nonhydrogen atom bonds connected to such atoms. If we allow Vnh to represent the sum of orders of bonds connected to adjacent nonhydrogen atoms, employing the integral values 1 for single bond, 2 for double, and 3 for triple, the resulting index can be used to order these atoms. Atoms with the smaller Vnh index precedes that with a larger one; thus, according to this criterion, atom g in **1a** precedes atom f [Vnh(g) = 3 < Vnh(f) = 4]. However, this still results in some ambiguity.

To further distinguish those atoms which have identical R and Vnh indexes (e.g., atom b and g), it is necessary only to

0095-2338/82/1622-0034$01.25/0

Table I. Results of Different Levels of Numbering

	node								
level	a	b	c	d	e	f	g	h	i
1	3	3	2	3	3	3	3	1	1
2	3	5	2	4	5	7	6	1	1
3	4	7	3	5	6	9	8	2	1

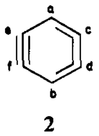
examine the bond order of each nonhydrogen atom bond, arranged in decreasing bond order. Thus, an examination of atom b in **1a** reveals three single, nonhydrogen atom bonds. This fact can be represented by a number containing R digits (in this case) where any undesignated digits (specifically those representing a bond to a hydrogen atom) are assigned a value of 0. Thus, atom g in **1a** has an R-digit representation of 2100, while atom e is represented by 1110, and so on. The atoms are then arranged according to the magnitude of the R-digit number—the lower this number, the higher the priority of the atom. Thus, atoms e and b (1110) precede g (2100), etc. The atoms are then renumbered according to this ranking; the results of application of this second level of numbering to **1a** are presented in Table I.

Third Level. Those atoms which remain undistinguished by applications of the level 2 criterion can be further distinguished by the following procedure. For a given atom, examine each of its adjacent atoms and arrange *their* level 2 numbers (m, n, etc.) into a numerical array such that the magnitude of the resulting number is minimal, i.e., for any numerical array $mnop$ to satisfy this procedure, $m < n < o < p$. The resulting number, $mnop$, provides a basis for ordering those atoms whose designations remain undistinguished after application of levels one and two described above. As in the case of level 2, the atoms are reranked numerically according to increasing magnitude of the numbers $mnop$, the atom whose $mnop$ value is the lowest being assigned the number 1, etc.

As an example, consider b in **1a**. Arrangement of the level 2 ordinal numbers of all adjacent atoms as described above generates the number 236. Similar treatment of atom e leads to the number 147. Since 147 < 236, atom e precedes atom b in ranking. A summary of level 3 numbering of the structure **1a** is seen in Table I and in **1b**.

Finally, it should be noted that under certain conditions, application of numbering levels 1–3 may still fail to define uniquely some atoms. Specifically, such instances occur when (1) all undistinguished atoms are exactly equivalent in which case it does not matter which atom is given precedent since the connection matrix is unique or (2) in those instances where the undistinguished atoms are not equivalent. This last instance results in duplicate structure generation and is discussed in detail below.

The methodology present in levels 1–3 is similar to that of Ugi and Schubert except for the following significant differences. First, the procedure presented here does not consider the ordering of hydrogen atoms. Second, only immediately adjacent atoms are taken into consideration here. This modification is equivalent to omitting all $r > 2$ entries in the calculation of the quantity $P(v,n)$ as described by Schubert and Ugi.[6] It is furthermore noteworthy that according to the algorithm employed by these authors, which does not consider bonds, some nonequivalent nodes [e.g., nodes (c, d) and (e, f) in structure 2] cannot be distinguished. By contrast, the algorithm we have presented readily provides that (c, d) precedes (e, f). Thus we can obtain two equivalent numberings: a b c d e f or b a d c f e. Finally, we note that although the algorithm of isomer generation described here does not itself require the operation of numbering, the numbering algorithm is nonetheless the basis of the generation algorithm (vide infra).



2

II. PROPERTIES OF THE CM MATRIX

The connection matrix CM, formed by the numbering procedure outlined above, has the following properties.

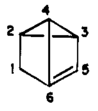
(1) Atoms are ordered according to their valences from low to high; atoms which have identical valences but which are typed differently are ordered arbitrarily.

(2) Within groupings of atoms of the same type, the atom with smaller Vnh index is assigned precedent. For atoms with equal Vnh indexes, precedent is assigned on the basis of bond order, i.e., those having only single bonds precede, followed by those having a double bond, and so on.

(3) If two arbitrary atoms numbered i and j ($i > j$) are of the same type, have equal Vnh indexes, and the same bonds, it can be shown from the criteria for numbering that when the entries of the ith row are compared with those of the jth row in the corresponding matrix, the following conditions must be satisfied.

(a) If the ith atom has T adjacent atoms, the ordinal numbers of the first level numbering of these atoms, i.e., the ordinal numbers of the types of the atoms corresponding to the nonzero entries of this row, can form a T-digit number TA_i. The first condition is that the TA of row i must be greater than or equal to TA of row j, i.e., $TA_i \geq TA_j$. In structure **1a**, O ranks as the first type of atom, N as the second type, and C as the third type; therefore, TA of b is 233, while TA of e is 133; since atom e falls in the 6th row and atom b in the 7th row of the corresponding CM matrix, $TA_7 > TA_6$.

(b) The Vnh index of the atoms corresponding to each nonzero entry of a row constitute a T-digit number VA; it follows that $VA_i \geq VA_j$. Thus, for example, in structure **3**, $VA_2 = 233 < VA_3 = 334$.



3

(c) The level 3 ordinal numbers of the atoms corresponding to each nonzero entry of ith row also constitute a T-digit number, NA_i, for which it may also be shown that $NA_i \geq NA_j$. Thus, for example, in structure **1b**, $NA_6 = 159 < NA_7 = 348$. It should be pointed out that when $CM(i,j) \neq 0$, the two digits corresponding to the entry $CM(i,j)$ of row i and $CM(j,i)$ of row j are considered as equal. This result is a consequence of the fact that the entries $CM(i,j)$ and $CM(j,i)$, representing the same bond connecting these two atoms, are equivalent. Thus, consider, for example, structure **3** in which $NA_3 = 245$ and $NA_4 = 236$; since $CM(3,4) \neq 0$, the middle digit of the NA number representing each of the two atoms should be considered as equal. If both are considered to be **3**, then $NA_3 = 235 < NA_4 = 236$. This property can be obtained directly from level three of the numbering procedure, while properties a and b can be obtained from the application of levels 1–3. The specific instance where $j = i - 1$ is very important to the structure generation algorithm discussed below.

III. ALGORITHM OF STRUCTURE GENERATION

A simple algorithm of structure generation may be abstracted from the above properties. In order to compare it with the existing algorithm,[3] the same example, C_6H_8, is used in this section.

First, the CM matrices, i.e., the isomers, must be ordered. As the CM matrices are symmetric, only the upper triangular matrix need be considered. As indicated above, the order of atoms is fixed, and the CM matrices can be ordered by comparing the matrices row by row from top to bottom so that a rule is needed to decide the relative order of rows. As has been pointed out, rows having *only* single bonds are given precedent followed by those having double bonds, and so on. But this is not enough since the order of bonds must be taken into account. For the convenience of programming, the order of the partitions of bonds is as follows: valence 2, 11,2; valence 3, 111,12,21,3; valence 4, 1111,112,121,211,22,13,31; and so forth. When the partitions of bonds are still the same, a row can be read from left to right as a number, and the row corresponding to the larger value is defined to be precedent. Thus, for example, for matrices a and c below, the first rows are the same. The second row of a can be read as 101000 while for c it is 100100, which is smaller than that of a. Therefore, matrix a precedes c.

$$\underset{\text{a}}{\begin{array}{cccccc} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 1 & 2 & 0 \end{array}} \qquad \underset{\text{b}}{\begin{array}{cccccc} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{array}} \qquad \underset{\text{c}}{\begin{array}{cccccc} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 1 & 2 & 0 \end{array}}$$

For the sequence of the CM matrices, the algorithm for a given molecular formula can be described as follows. Calculate the total valences of all nonhydrogen bonds. Then partition them among all atoms in all possible ways. For example, there are eight such unique partitions for C_6H_8 (Table II). When there exist heteroatoms, the partition becomes more complex, but it is still a common combinatory problem and will not be discussed in detail here.

For each partition, all possible isomers can be generated as follows:

(A) The CM matrix of lowest order is generated first for a given partition of Vnh.

(1) For row 1 ($i = 1$), if Vnh equals M, distribute bonds according to the highest possible priority, that is, align M single bonds beginning from CM(1,2). For example, if the empirical formula is C_6H_8 and the partition of Vnh is 223333, the distribution of highest possible priority for bonds of row 1 is 011000.

(2) For the next row ($i = i + 1$), distribute the bonds according to their priority, subject to the condition that the sum of the entries of ith row equals Vnh(i) and as row i is compared with row $i - 1$, the properties of CM matrix described above must be satisfied. If this requirement cannot be met, then proceed to (D) below.

(3) Repeat step 2 until $i = n - 1$ (n is the number of nonhydrogen atoms). At this point the matrix is completed, since row n = column n. For the partition of Vnh 223333, the first matrix obtained is matrix a below.

(B) Examine the entries of row n of the generated matrix to see if the sum of the entries of that row equals Vnh(n); if it does not, then proceed to (D).

(C) If the sum equals Vnh(n), examine the structure to determine if it is connected, i.e., it cannot be separated into two or more fragments without breaking bonds. If it is not connected proceed to (D); however, if it is connected, then a legal isomer has been generated, and one then proceeds to (D) to generate the next structure. For example, in matrix a the sum of entries in row 6 equals 3, i.e., equals Vnh(6), and the structure is connected; it is, therefore a legal structure. This structure is shown in **4a**.

(D) For the preceding row ($i = i - 1$), distribute bonds according to the next highest possible priority, then return to (A2). If this requirement is impossible, repeat (D).

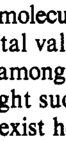
For example, after the generation of matrix a, examine row 4 ($i = 4$) for which the next possible partition of bonds is 001020; examination of row 5 reveals a value of 000201; thus, matrix b is obtained. In matrix b, the sum of the entries of row 6 equals 1, so it is not a legal structure; consequently, we proceed to step D, and so on. Iterating these steps yields the matrix c representing the structure **4b**. For the partition 223333 of C_6H_8, 14 isomers can be obtained in this way. The first five are shown in **4a**–**4e**.

Table II. Number of Isomers of C_6H_8

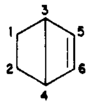
partition of Vnh	no. of isomers
112444	9
113344	19
122344	40
123334	34
133333	4
222244	10
222334	29
223333	14
	total = 159

Table III. Two Possible Numberings of Benzene

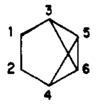
	a	b	c	d	e	f
1st	1	2	3	4	5	6
2nd	1	3	2	5	4	6



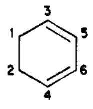
4a



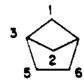
4b



4c



4d



4e

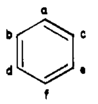
Table II shows the numbers of isomers of each partition of Vnh of C_6H_8. The total number is 159, which is consistent with the result of Smith et al.[3]

The algorithm described above is a rather simple one. Nonetheless, it is apparently capable of generating directly all isomers, including acyclic and cyclic structures, without requiring complex descriptions of molecular symmetry. Inasmuch as all possible permutations of the matrix have been considered in this algorithm, the results seem to be exhaustive. Moreover, it can be easily demonstrated that an interchange of any two nonequivalent rows of a CM matrix satisfying the conditions of section II will violate these conditions. Hence, as long as there exist no different atoms of the same type and with the same partition of Vnh, the matrix is unique; therefore no redundant structure may occur. Nevertheless, there are a few exceptions where the structures generated from above algorithm may be duplicate. This situation is discussed in the following section.

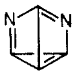
IV. AVOIDANCE OF REDUNDANT MATRICES

Occasionally a few of the CM matrices generated by the above sequence may prove to be redundant. We suggest that these instances can be divided into two categories.

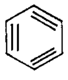
(i) Consider a structure which has an "aromatic ring" that has an axis or a center of symmetry that involves only vertices (i.e., the element of symmetry does *not* bisect any bonds). Such a structure may not be numbered uniquely under the considerations described above. Examples of such structures are shown in structures **5a**–**5d**. Here the term "aromatic ring" is used in broad sense, including those rings such as **5b**, **5c**, and **5d**. If node a in **5a** is numbered 1, either node b or c can be numbered 2; hence two numberings corresponding to unequal CM matrices can be obtained for the same structure,



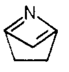
5a



5b



5c



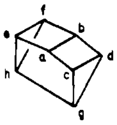
5d

Table IV. Several Possible Numberings of Structure 6

	node							
no.	a	b	c	d	e	f	g	h
1	1	2	3	5	4	6	7	8
2	2	5	1	3	6	8	4	7
3	4	5	1	2	7	8	3	6
4	7	8	3	4	5	6	1	2

as shown in Table III. Both matrices satisfy the conditions described in section II; thus redundant matrices exist.

(ii) In some circumstances the algorithm of Ugi and Schubert[6] fails to distinguish nonequivalent atoms. For example, structure **6** has eight nodes, all of which have three single bonds; intuitively, they can be divided into three equivalent groups—ab, cdef, and gh. However, nodes ab cannot be distinguished from gh by the algorithm of Ugi and Schubert.[6] By contrast, using the numbering method we have described, all eight nodes will be regarded as the same; thus there will exist several numberings corresponding to different CM matrices and all these CM matrices satisfy the conditions described above. Some of these numberings are shown in Table IV.



6

Among these duplicates only one can be regarded as legal. This one may be selected as that matrix corresponding to the largest valued number derived by reading the entries of its CM matrix row by row from left to right and from top to bottom. All other duplicate matrices should be discarded. Ideally, any algorithm capable of accomplishing this tasks should do so without retrospective checking of each generated structure with existing structures. One such algorithm, which can also further serve to order undistinguished atoms, is briefly described below.[8]

(1) The complete algorithm of Schubert and Ugi[6] can serve further to distinguish and order these atoms. Thus, by their algorithm one can determine that the nodes c, d, e, f of structure **6** are different from other nodes and should take precedence. In our program a simplified method is used which is essentially the same as that of Schubert and Ugi.

(2) For an atom and any pair of its adjacent, nonhydrogen bonds one can obtain one or more rings, the smallest of which is defined as the least ring. If a node (atom) has T adjacent edges (bonds), apparently it can have M least rings, $M \leq T(T-1)/2$. The numbers representing the members of these least rings may be arranged in increasing order to form a M-digit number. The resulting M-digit numbers can be used further to distinguish and order these nodes: the nodes corresponding to the smaller value should precede. For example, the nodes a, b of structure **6** have a number of 445, and g, h have a number of 355; thus g, h should precede a, b since smaller the number of an atom, the more closely its adjacent atoms connect to each other. Therefore, the entries of these rows will correspond to larger valued numbers when the entries are read row by row from left to right and from top to bottom.

Since the above algorithm considers only the adjacent rings of an atom, obviously it may not be sufficient for large, highly unsaturated molecules, but within the available size of problems which can be solved by this algorithm, it appears to be capable of distinguishing all unequal atoms.

After the ordering of all similar nodes, one can verify if a generated CM matrix is legal or should be discarded [cf. (2) above] by determining whether the order of the atoms of generated matrix is legal. For example, in structure **6**, the first atom must be either d, c, e, or f. If we select c, the order of the adjacent atoms should be d, g, a. Therefore, the only legal numbering is the third row in Table IV, and all matrices corresponding to other numberings should be discarded.

The differences between the numbering of Ugi and Schubert and that proposed in this paper are readily apparent from a further examination of this example. According to their algorithm, the nodes c, d, e, f in **6** must be the first four nodes, which of course, is different from the numbering we have obtained. In addition, their algorithm also fails to number those structures as shown in **2**, **5**, and **6**. Hence, the numbering algorithm presented in this work appears to be a superior method.

For case 1 mentioned above, one can look for vertex-symmetric aromatic rings throughout a generated matrix. If such a ring can be found and the order of nodes is not legal, this matrix should be discarded. Thus, for example, consider the two numberings of benzene in Table III. The second numbering is legal, and the matrix corresponding to the first numbering should be discarded.

Although the procedures as outlined above for preventing redundant structures is rather lengthy, it should be recognized that only a small part of generated CM matrices need this verification. For example, there is only one redundant structure in 218 generated matrices for C_6H_6; hence it does not affect the rate of generation seriously.

V. THE PROGRAM

The program based on the algorithm suggested here is written in FORTRAN-IV. The resulting source program contains about 800 FORTRAN lines, in which more than half is occupied by the subroutine for preventing redundancies; hence, this program can be executed on many mini/microcomputers. On our microcomputer system having a Z-80 CPU, the object program occupies 22K bytes of memory and generates isomers at the rate of tens to hundreds of structures per minute, depending on the empirical formulas. Obviously, the more the types of atoms, the more efficient this program will be. For example, it takes about 2 min to generate 1371 isomers consistent with the empirical formula $C_3H_4N_2O$.

Perhaps the best verification of the validity of such an algorithm is the numbers of isomers obtained from the corresponding program. There exist no theoretical values for arbitrarily given empirical formulas; however, Smith et al. have obtained the number of isomers for about 300 empirical formulas by CONGEN,[7] and this can serve as a base of comparison. Using our program we have obtained the same results. For larger molecules, however, the validity of this algorithm and corresponding program remains to be proved.

In summary, the algorithm presented in this paper is a simple, exhaustive one. However, for large, highly unsaturated molecules its results may contain a few duplicates; hence theoretically it is not as complete as the algorithm of CONGEN.[3,9] But, due to the exponential nature of the increase in numbers of isomers, for even slightly large molecular formulas, no one algorithm is applicable for the exhaustive generation of isomers because of too long execution time. Hence, from a practical point of view this algorithm is sufficient for a structure generator.

In some existing structure-elucidation programs, the problem of generating isomers of large molecules has been solved by extending already proven algorithms to deal with polyatomic building blocks.[10-12] For instance, by introducing a superatom and by embedding successively, program CONGEN can deal with realistic molecular formulas.[10] We would like to point out that the algorithm of this paper can also be extended to deal with superatoms. Using the procedure described in the section about redundant structures (vide supra), the symmetry of a superatom (in the form of a submatrix) can be determined, and the superatoms can be properly embedded into a structure without duplicates. Also by introducing some constraints the generated structures can be pruned to a few candidates. Therefore our algorithm can serve as a practical structure generator, but has the advantage of being simple and can even be executed on many mini/microcomputers.

ACKNOWLEDGMENT

We are particularly indebted to Professor Joseph San Filippo, Jr. of Rutgers University and Dr. Dennis H. Smith of Standford University for their comments in preparing this manuscript. We acknowledge partial financial support by NSF (Grant 80-17045) and DOE (Contract DE-AS05-80ER-10662).

REFERENCES AND NOTES

(1) Rouvray, D. H. *Chem. Br.* **1974**, *10*, 11.

(2) "Chemical Applications of Graph Theory"; Balaban, A. T., Ed.; Academic Press: New York, 1976.

(3) Masinter, L. M.; Sridharan, N. S.; Carhart, R. E.; Smith, D. H. *J. Am. Chem. Soc.* **1974**, *96*, 7702.

(4) Morgan, H. L. *J. Chem. Doc.* **1965**, *5*, 107.

(5) Randic, M. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 171.

(6) Schubert, E.; Ugi, I. *J. Am. Chem. Soc.* **1978**, *100*, 37.

(7) Smith, D. H. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 203.

(8) This process can also be carried out by permutation, but the procedure is rather lengthy.

(9) A referee has pointed out that the most current (but as yet unpublished) version of the CONGEN program has incorporated a structure generator which functions in a manner very similar to that described in this paper: Carhart, R., unpublished results.

(10) Carhart, R. E.; Smith, D. H.; Brown, H.; Djerassi, C. *J. Am. Chem. Soc.* **1975**, *97*, 5755.

(11) Sasaki, Shin-Ichi; Abe, Hidetsugu; Hirota, Yuji; Ishida, Yoshiaki; Kudo, Yoshihiro; Ochiai, Shukichi; Saito, Keiji; Yamsaki, Tohru *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 211.

(12) Shelley, C. A.; Woodruff, H. B.; Snelling, C. R.; Munk, M. E. In "Computer-Assisted Structure Elucidation"; Smith, D. H., Ed.; American Chemical Society: Washington, DC, 1977; p 92.

(13) Copies of the program listing and appropriate documentation are available upon request.

The Development of an Environmental Fate Data Base

PHILIP H. HOWARD,* GLORIA W. SAGE, and A. LAMACCHIA
Syracuse Research Corporation, Syracuse, New York 13210

ANDREW COLB
U.S. Environmental Protection Agency, Washington, DC 20460

Received September 16, 1981

Three components (DATALOG, XREF, and CHEMFATE) of a new Environmental Fate Data Base are described. Environmental fate is a term describing the behavior (i.e., transport and degradation) of a chemical which is released to the environment. The system stores and retrieves data and provides sufficient flexibility to meet the needs of a variety of environmental fate data users. It is anticipated that both government and industry will find the information in these files useful as a source of data for estimations and modeling of environmental fate and exposure evaluations, as well as structure–reactivity, persistence, or transport correlations. These correlations are particularly desirable for predicting environmental behavior of chemicals for which only a limited amount of enivronmental fate data are available. The CHEMFATE data base will also be useful in determining where research effort is needed to supply missing data on physicochemical properties and environmental degradation and transport behavior.

INTRODUCTION

With the growing awareness of the health and environmental hazards associated with the commercial production, use, and disposal of industrial chemicals, risk assessment has become an area of increasing concern and activity.[1] There are two major factors that have to be considered for an overall environmental risk assessment:[2] (1) exposure and (2) toxicity. Numerous references are available containing tabulated biological effects data,[3-6] some of which are available online;[3,7-9] however, in contrast, few tabulations of data relevant to environmental exposure (e.g., environmental release and environmental fate) exist other than a limited number of monitoring data bases.[10] This is particularly true of the substantial amount of environmental fate (i.e., transport and degradation) information that is available. In November 1979, the development of an Environmental Fate Data Base was initiated in an attempt to fill this gap.

The knowledge of how a chemical will behave in the environment once it is released is particularly important in determining whether a chemical will come in contact with a critical species or with man in sufficient concentrations to cause a toxic effect or, in contrast, be rapidly degraded to innocuous products. The type of information pertinent to the fate of a chemical released into the environment is diverse and includes physical and chemical properties, transport and degradation studies, ambient monitoring data, and field studies.[11] Considerable amounts of time and money must be expended to extract this type of data from primary literature sources. Thus, once obtained, these data should be stored in a form that is readily accessible to other investigators. A data bank of environmental fate information serves the following purposes:

(1) Allows rapid access to all available fate data on a given chemical without having to resort to expensive, time-consuming, and inefficient searches of the primary literature.

(2) Identifies critical gaps in the available information to facilitate planning of research needs.

(3) Provides a source of data (training set) for constructing structure–activity correlations for degradability and transport of chemicals in the environment. Such correlation would be a tremendous aid in identifying persistent chemical classes as well as physical or chemical properties that may correlate to particular behavior in the environment.

SYSTEM OVERVIEW

The Environmental Fate Data Base is comprised of three interrelated files called DATALOG, XREF, and CHEM-

0095-2338/82/1622-0038$01.25/0 © 1982 American Chemical Society