ison, because the two work processes—index operation and output scanning—can be worked on separately. The frequent confounding of these operations by the user is because the index-couples, which are at the heart of index using, frequently carry the display also, as with card catalogs or subject heading indexes such as *Chemical Abstracts*. Thus, the user subconsciously proceeds from using the index-couples to using the display, to which the former have guided him. This intertwining of the operation in use should not blind one to the fact that the operations can be separately designed in system building.

To give a further example, a good display in a small system with a very shallow index can lull people into a false sense of security. The shallow index gives the user a large proportion of irrelevant material, but the system is small and he has a good display. If the act of perusing the display is erroneously considered part of the act of using the index, the following error could easily be encountered. A user poll might indicate high satisfaction with the index, with practically no false drop. But the users eliminated the false drop during the display scanning, thinking this was part of using the index. On the basis of the low false-drop report, management might expand into a large mechanized system, with the indexing depth frozen at the point of the user survey. At large volume a good display ceases to be a substitute for a deep index. This is not to argue against high-level display; it simply warns against confusing display with index.

Confusion of another sort has resulted when an index entry leads to a pertinent article, but the abstract serving as a display is misleading causing the inquirer to eliminate the item. This loss of a pertinent reference may be erroneously attributed to a weakness in the index, but is actually a weakness or misuse of the display.

In summary, the two steps of index using and display scanning are often confused, particularly in subordinate indexes. It is submitted that distinguishing between these two work processes results in better system design.

REFERENCES

(1) J. C. Costello, Jr., *J. Chem. Doc.*, **4**, 12 (1964).
(2) R. J. Runck, *ibid.*, **2**, 129 (1962).
(3) M. Taube, *et. al.*, "Studies in Coordinate Indexing," Vol. 1, Documentation Inc., 1953.
(4) J. W. Perry, and A. Kent, "Tools for Machine Literature Searching," Interscience Publishers, Inc., New York, N. Y., 1958.
(5) C. W. Brenner, and C. N. Mooers, in "Punched Cards," 2nd Ed., R. S. Casey, *et al.*, Ed., Reinhold Publishing Corp., New York, N. Y., 1958.
(6) C. K. Schultz, in "Punched Cards," ref. 5.
(7) J. P. McMurray, *Am. Doc.*, **13**, 66 (1962).
(8) J. C. Costello, Jr., *ibid.*, **12**, 20 (1961).
(9) W. E. Batten, in "Punched Cards," 1st Ed., R. S. Casey and J. W. Perry, Ed., Reinhold Publishing Corp., New York, N. Y., 1951, pp. 169–181.
(10) F. R. Whaley, in "Information Systems in Documentation," Interscience Publishers, Inc., New York, N. Y., 1957.
(11) J. J. Nolan, *Am. Doc.*, **10**, 27 (1959).
(12) P. D. Bradshaw, *ibid.*, **13**, 270 (1962).
(13) J. W. Kuipers, A. W. Tyler, and W. L. Myers, *ibid.*, **8**, 246 (1957).
(14) F. R. Whaley, *ibid.*, **12**, 101 (1961).
(15) C. M. Lauer, and F. R. Whaley, *J. Chem. Doc.*, **3**, 150 (1963).
(16) F. R. Whaley, *Special Libraries*, **53**, 65 (1962).

The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service

H. L. MORGAN
Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210
Received January 15, 1965

I. INTRODUCTION

As part of the development of a computer-based chemical information system at CAS, it has been necessary to devise techniques for the registration of drawings of chemical structures. A major purpose of the CAS registration process is to determine whether a particular structure has already been stored in the system. The ability to make this determination makes it possible to utilize a computer to assign to every chemical structure a unique identifying label. This identifying label, referred to as a registry number, is the thread that ties together all information associated with a particular compound throughout the developing CAS computer system. It is because of this association, made possible by the registration process, that CAS will be able to provide multiple-file correlative searches with assurance that all information on file for a particular compound has been located.

II. THE REGISTRATION PROCESS

The registration technique that has been selected by CAS requires computer generation of an alphanumeric description for each chemical structure that is unique for that structure. The machine technique has not yet been extended to all types of structural detail, but techniques and computer programs are complete for generating the unique description for the two-dimensional projection of fully known nonpolymeric chemical structures. The third dimension is presently handled by the addition of conventional stereochemical descriptors which are supplied by the chemist who prepares the structural diagram for input to the system.[1]

In the coming months present basic machine techniques will be extended to handle partially unknown and polymeric structures. Work is also progressing toward the inclusion of the third dimension directly in the graphic record so that in time the full steric picture will be in the form of a single detailed coherent record of each structure. The initial approach, however, permits CAS to provide an operable registry system that will accommodate all compounds without awaiting the utopia of a complete set of machine techniques and computer programs that will handle all chemical substances automatically.

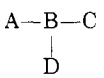
Once the unique descriptions for a set of input structures are obtained, the remainder of the registration process is simple and very fast. Since the description of each compound is in itself unique it is possible to organize both the input and registry files into a unique sequence. The use of this unique sequence reduces the actual registration process to a merging and updating of two serial files; therefore, it is the uniqueness of the machine representation of a chemical compound that is the key to an effective, efficient, reliable registration system.

III. CHARACTERISTICS OF THE STRUCTURE DESCRIPTION

The structure description employed in the CAS registration process is a uniquely ordered list of the node symbols of the structure (or graph) in which the value (atomic symbol) of each node and its attachment (bonding) to the other nodes of the total structure are described. Such a list and description is called a "connection table." Since this paper is not concerned with structure input, the connection table which is described is that stored and manipulated by the computer. The form of the table which is used within the computer is not the most convenient form for input to the system; thus the input form is translated by the computer into the "compact connection table" developed by D. J. Gluck of du Pont.[2] In this form of the table, the nonhydrogen nodes of the structure are listed according to an exact set of rules. The application of these rules alone does not produce a unique table; it does, however, produce a partial ordering among the nodes of the structure. "Partial ordering" in this context means that at certain stages in the formation of the table certain nodes will receive preference for earlier listing in the table. This is important since the generation of the unique table is based in part on a process of successive partial orderings as will be seen later in this paper.

After establishing a structure representation in the computer memory, the compact connection table is formed by first numbering the nonhydrogen nodes of the structure. This numbering proceeds from 1 using only the ordinal numbers. The numbers are assigned to the nodes of the structure according to the following rules: (1) a node is arbitrarily selected and assigned the locant, node number, 1; (2) the nodes attached to node 1 are numbered 2, 3, etc. When all the nodes directly attached to node 1 have been numbered, those which have not yet been numbered but which are attached to node 2 are numbered, and so on. This procedure is followed until all nodes have been numbered, or as in the instance of disconnected graphs such as represent ions, until the process leads to a point where not all nodes have been numbered, yet none of the unnumbered nodes is attached to a previously numbered node. Under such conditions another arbitrary choice is made among the unnumbered nodes for the next node to be numbered and the process of numbering is continued.

Example I.—Assume the structure



For this structure the following table shows the numberings that result from application of the above rules.

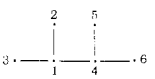
Locant	Possible node assignments											
1	A	A	B	B	B	B	B	B	C	C	D	D
2	B	B	A	A	C	C	D	D	B	B	B	B
3	C	D	C	D	A	D	A	C	A	D	A	C
4	D	C	D	C	D	A	C	A	D	A	C	A

For the structure in example I there are 24 possible numberings using the numbers 1–4; however, only 12 of the possible numberings comply with the rules cited above. This reduction is a characteristic of the numbering rules and becomes more significant as the size and complexity of the structure being treated increase.

When the entire structure has been numbered according to the preceding rules the connection table is formed by recording the structural relationships in the five lists which compose the connection table, as follows:

1. The "FROM ATTACHMENT" List.—This list is composed of X fixed length ranks where X is equal to the number of nonhydrogen nodes in the structure. In this list the ith rank is used to describe not more than one attachment between the ith node and one other node of the structure. At the ith rank is recorded the rank number of the lowest numbered atom attached to the ith node. If, however, the rank number which would be recorded at the ith rank is numerically greater than i, the ith rank is left blank.

Example II.—Assume the following structure with the numbering shown



(1) The system is now an operational element of the publication process of *CBAC*, a new CAS computer-based publication.

(2) D. J. Gluck, *J. Chem. Doc.*, 5, 43 (1965).

For this structure with the numbering shown the "FROM" list is shown below. The rank numbers to the left in this and following examples are for the reader's convenience and do not appear in the actual list.

Rank no.	From attachments
1	Blank
2	001
3	001
4	001
5	004
6	004

2. The "RING CLOSURE" List.—This list is composed of X fixed length ranks where X is equal to the number of cycles (rings) in the structure. Structures containing no cycles have no such list.

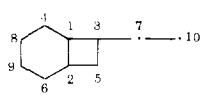
After the formation of the FROM list there will remain one connection, not described in the FROM list, for each cycle in the structure. These additional bonds or ring closures are defined in the RING CLOSURE list as follows:

(a) For each ring closure, record in a rank of the RING CLOSURE list the locants of the two atoms involved.

(b) In each rank of the RING CLOSURE list, order the two locants so that the first is numerically less than the second.

(c) Order the ranks of the RING CLOSURE list so that the locant pair of the first rank is numerically less than the second, which is less than the third, etc. Thus, 002 007 < 003 005 < 003 006.

Example III.—Assume the following structure with the numbering shown.



For this structure with the numbering shown, the FROM list and the RING CLOSURE list are as follows:

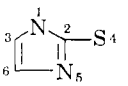
Rank no.	From attachment	Ring closure	
1	Blank		
2	001		
3	001		
4	001		
5	002		
6	002		
7	003		
8	004		
9	006		
10	007		
		003	005
		008	009

It should be noted at this point that the FROM list and the RING CLOSURE list are sufficient to completely describe the interconnections of the graph for the two-dimensional projection of the compound.

3. The "NODE VALUE" List.—This list is composed of X fixed length ranks where X is equal to the number of nonhydrogen nodes of the structure. In this list the ith rank is used to describe the node value (atomic symbol) of the ith node (see example IV below).

4. The "LINE VALUE" List.—This list is composed of X fixed length ranks where X is equal to the number of bonds in the structure between two nonhydrogen nodes. In this list the ith rank is used to describe the line value or bond for the attachment defined at the ith rank of the FROM list or RING CLOSURE list. For purposes of definition, the ranks of the RING CLOSURE list are assumed to be numbered consecutively after the FROM list. The bonds (*i.e.*, line values) are described by assigned code.

Example IV.—Assume the following structure with the numbering shown.



Rank no.	From attachment	Ring closure		Node value	Line value
1	Blank			N	Blank
2	001			C	1
3	001			C	1
4	002			S	1
5	002			N	2
6	003			C	2
		005	006		1

5. The "MODIFICATIONS" List.—This list is used to describe any other modifications of the nodes and lines as listed, such as the charges of ions, isotopic mass, and citation of unusual valence.[3] Such modifications are described by citing the type of modification in coded fashion, followed by the node number or line number being modified, followed by a description of the modification in coded form. Since the techniques for treating this list are merely an extension of the techniques applied to the previous four lists, discussion of the MODIFICATIONS list will be omitted from the remainder of this paper.

The compact connection table is at this stage an unambiguous description of the two-dimensional projection of a chemical structure drawing. Where necessary it is made unambiguous for three-dimensional structures by the addition of conventional stereochemical descriptors as previously mentioned. Thus, the table is at this stage an unambiguous but non-unique machine representation of the chemical structure. It is one of a family of unambiguous descriptions of the structure. The exact table selected for use in the CAS Registry System is a member of this family and is selected by further computer processing.

In the following pages the techniques for selecting the unique table from among the family of unambiguous tables will be shown to be completely independent of the order of the nodes in the input table. Since the ordering process is independent of the order of the nodes in the input table, it follows that the unique table is also independent of both the orientation and the projection of the drawn structure. It also follows that the ordering process is independent of the means by which the drawn structure is converted to a machine representation, *e.g.*, Army Chemical Type-

(3) The editing routines include a check for normal valence. Thus, for example, a trisubstituted methyl free radical requires the specification that only three groups are directly bonded to the methyl carbon instead of the usual four.

writer,[4] optical scanning,[5] clerically generated connection table,[6] grid structure,[7,8] or linear notations,[9] so long as the resulting machine representation is in fact a representation of the structure in question. The points expressed in this paragraph are very important since the CAS Registry System is based on the premise that a unique structure will be stored once and only once, thus making the registry number a unique and unambiguous identification of a chemical substance.

IV. THE GENERATION OF THE UNIQUE DESCRIPTION

As has been stated, the unique table used in the CAS system is a member of a family or set of tables, all of which describe the same structure equally well. It is unimportant, therefore, which member of the set is labeled unique so long as the same table is always selected for the same structure. Since it can be shown that the set is finite for any graph composed of a finite number of nodes, it is possible to select the unique table by generating all members of the set, lexicographically ordering the members of the set based on the characters involved in the description, and then selecting the first member of the resulting list as the unique table. This concept is a restatement of a technique proposed by C. N. Mooers for generating a unique cipher based on a process of making all possible "cuts" and comparing the resulting ciphers.[10,11]

The generation of all possible tables of the type described would, in the case of large molecules, be prohibitively expensive. It is necessary, therefore, to devise techniques to limit the number of tables that must actually be generated to some invariant subset of tables which is small enough to make the process economically feasible. Having generated this invariant subset, the unique table is selected in exactly the same way as if the entire set had been generated. It does not necessarily follow that the same table would be selected from the subset as would be selected from the entire set, but that fact is not important so long as only a single subset is generated for a given compound regardless of the order of the nodes in the input table for that compound.

In order to generate only an invariant subset of the possible set of tables, the computer program first employs the rules for numbering the structure and forming the table as described earlier. This procedure reduces what would be a factorial expression to a number which is almost always significantly smaller. For instance, in a simple six-membered ring there are 720 possible numberings; however, only 12 comply with the rules for numbering. Thus, the numbering rules have created an invariant subset. In addition to the rules of numbering, the computer program employs certain invariant properties of the graph to reduce further the size of the subset. These properties are the "connectivity value" of each node, the node value (atomic symbol), and the line value (bond).

The second means by which the subset is reduced in size is by introducting a partial ordering among the nodes of the graph. The selection of the next node to be listed, where a choice is possible, can then in many cases be resolved on the basis of a preference implicit in this partial ordering. A simple illustration of such a partial ordering is shown in example V where preference is given to the nodes with the greater number of attachments at each point of choice.

Example V.

Structure	Possibilities for the order of citation of the nodes	
A—B—C—D	B	C
	C	B
	A	D
	D	A

In example V only nodes B and C were considered for node 1 because of the preference introduced by the partial ordering. Having selected one, the other is given preference for node 2 again because of the partial ordering. Having listed nodes 1 and 2, nodes 3 and 4 are fixed because of the rules for numbering. Thus, in this example the subset generated will consist of only two tables, whereas without the use of the partial ordering six tables would have been generated, and without both the partial ordering and the rules for numbering twenty-four tables would have been required.

Although the partial ordering of the nodes based on the number of attachments will usually greatly reduce the number of tables in the subset, it is not sufficient to adequately partition the set. The reason for this is that in organic chemistry the number of bonds to any given atom rarely exceeds four or five. In order to increase the effectiveness of the partial ordering, a technique has been devised for computing a "connectivity value" for each node based on the invariant properties of the graph. These values are then used to introduce a partial ordering among the nodes in the same fashion as the number of connections were used in example V.

The "connectivity values" are computed by first assigning to each node an initial "connectivity value" equal to the number of nonhydrogen atoms attached to that node. This number is clearly an invariant property of the graph. The computer then calculates the number (k) of different "connectivity values" which had been assigned. An iterative process is then established which calculates a new "connectivity value" for each node. This new value is the sum of the assigned values for the nodes connected to the one under consideration. Having computed a new value for each node based on the previous values, the computer calculates the number (k') of different values in the set of new values. If $k' > k$, the new values are assigned to the corresponding nodes, k is set equal to k', and the summation process is repeated. If, however, $k' \leqq k$ the process is terminated, and the last set of values assigned to the nodes is used to induce a partial ordering among the nodes. Using this partial ordering, the size of the subset is reduced by giving preference to the

(4) A. Feldman, D. B. Holland, and D. P. Jacobus, *J. Chem. Doc.*, 3, 187 (1963).
(5) W. E. Cossum, M. E. Hardenbrook, and R. N. Wolfe, *Proc. Am. Doc. Inst.*, 269 (1964).
(6) G. M. Dyson, W. E. Cossum, M. F. Lynch, and H. L. Morgan, *Inform. Storage Retrieval*, 66 (1963).
(7) P. Horowitz, and E. M. Crane, "HECSAGON: A System for Computer Storage and Retrieval of Chemical Structure," Eastman Kodak Co., Rochester 4, N. Y., 1961.
(8) W. H. Waldo, and M. DeBacker, "Proceeding of the International Conference on Scientific Information, Washington, D. C., Nov. 16-21," Washington, D. C., 1959, pp. 711-730.
(9) H. T. Bonnett, *J. Chem. Doc.*, 3, 235 (1963).
(10) C. N. Mooers, "Ciphering Structural Formulas—The Zatopleg System," Zator Technical Bulletin No. 59, Zator Co., 79 Milk St., Boston 9, Mass.
(11) C. N. Mooers, "Generation of Unique Ciphers for a Finite Network," Zator Technical Bulletin No. 49, Zator Co., 79 Milk St., Boston 9, Mass.

node associated with the higher "connectivity value" at each point of choice in the numbering process described earlier. It is important to note that the iterative process is finite for any graph of X nodes where X is a finite number. The process will terminate, under the conditions cited, after no more than $(X + 1)$ iterations since there are at most X values that can be assumed by k which will cause the process to continue. Examples 1 and 2 of Appendix I illustrate the application of this technique for introducing a partial ordering among the nodes of the graph.

After introducing the partial ordering among the nodes, the generation of the subset of tables defined by this ordering is begun. Even at this stage, however, the entire subset is not always generated since in practice it is often possible to eliminate large blocks of potential tables. To describe the means by which potential tables are eliminated during the generation process, it is necessary to describe the means by which the unique table is ultimately selected.

After generating any two of the tables of the subset, a preference between them is introduced by "alphabetizing" on the basis of the collating sequence of the machine symbols involved in the tables. The table which "sorts" to the top of the list is then selected as preferred over the other. If the two tables are identical, one is arbitrarily selected as preferred over the other. For purposes of this "alphabetization," the tables are treated as a string of symbols in the following order (see example 3 of Appendix I):

A. The "FROM ATTACHMENT" list
B. The "RING CLOSURE" list
C. The "NODE VALUE" list
D. The "LINE VALUE" list
E. The "MODIFICATION" list

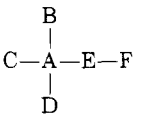
Since a preference or a lack of preference is introduced each time two tables are completed, it is never necessary to have more than two complete tables in memory at any given time.

During the table generation process, when a complete table is in the computer memory and a second table is being generated, a determination is made after each step in the generation process whether the first, completed table is already preferable to the second, partially generated one. This determination is accomplished by comparing the two FROM lists up to the point of completion of the second and selecting, as preferred, the one which "sorts" first. If it is determined that the completed table is already preferred to the second, further generation of the second table is stopped, and all tables based on the fragment thus far generated are eliminated.

Another means of eliminating potential tables during the generation process is the provision of performing a simple look-ahead to determine a preference or lack of preference.

In chemical structure drawings it is quite common to have two or more terminal atoms attached to the same atom. (A terminal atom is here defined as an atom attached to only one nonhydrogen atom.) The partial ordering of the nodes as described above does not resolve the order of selection of these terminal atoms. Thus, without provision for a simple look-ahead, the alternatives would need to be generated and the tables compared to determine a preference.

Example VI.



Computer connectivity values

	A	B	C	D	E	F	
i =	0	4	1	1	1	2	1 k = 3*
i =	1	5	4	4	4	5	2 k = 3

Partial ordering of nodes
$\{A\} > \{E\} > \{B,C,D,F\}$

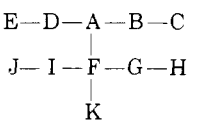
* These values used for partial ordering

In example VI the partial ordering will cause nodes A and E to be selected as the first and second nodes, respectively. At this point, however, B, C, and D are equal candidates for selection as the third, fourth, and fifth nodes, thereby giving rise to the generation of six tables. To prevent this common situation the computer detects the condition and performs a look-ahead to determine the effect of the possible choices on the next levels of the table. This look-ahead can be done since the choice cannot affect the FROM or RING CLOSURE lists. Therefore, the node values are examined and any preference implied by them is introduced. If, however, the node values are equal, the determination of a preference falls next to the line values and finally to the node and line modifications. If the choices are equal at every level then it makes no difference which is selected next since the choices give rise to identical tables. By application of this simple look-ahead the program is able to eliminate the generation of the possible alternatives and the selection from among them. In example VI, for instance, only one table will be generated instead of the six which would have been generated without the look-ahead technique.

At present the look-ahead technique is used only to deal with the case of terminal atoms. The technique could be extended, however, without affecting the ultimate choice of the unique description. Determination of whether this extension is economically required will be made on the basis of operating experience in the coming months, but at this point it seems unlikely to be necessary.

The last technique for reducing the number of tables generated is the provision to recall, under certain conditions, preferences detected during the generation process. Because of the nature of the techniques thus far described, the size of the subset is a product function based on the number of choices arising; that is, the same preference or lack of preference is rediscovered several times.

Example VII.



For instance, in example VII the preference or lack of preference between B and D will be determined twice, once when G and I are listed third and fourth, respectively, and once when G and I are listed fourth and third, respectively. It would be more efficient to remember the preference once detected and to use this information should the same choice arise again. The problem is that

the preference can be recalled only when it is independent of any previous choices; therefore, the preference can be remembered only under certain conditions. These conditions are: first, the atom from which the choice arises, atom A in example VII, must be bonded to exactly three other nonhydrogen atoms, two of which are involved in the choice; and, second, the bond not involved in the choice, bond A–F in example VII, must not be part of a cycle.

If during the generation process a point of choice is reached which meets the conditions cited, the computer program divides the graph into two subgraphs by removing the bond which is not involved in the choice. In example VII the bond between A and F is temporarily removed. By virtue of the fact that the removed bond is not a member of a cycle (specified condition), the result of this removal is to divide the graph into two subgraphs. The program then operates on the subgraph involved in the choice so as to determine a preference or lack of preference between the two choices. This preference is determined by generating the set of tables which arises from the choices in the subgraph. Once such a preference is determined the graph is restored, and the preference is recalled when the same choice arises again. If there is no preference between the two choices, then it makes no difference which is selected since they are indistinguishable. In this case, the preference will be made arbitrarily and reused should the opportunity arise again.

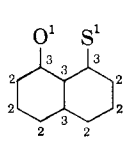
Of the several methods employed to reduce the number of tables generated the two most significant are (1) the partial ordering of the nodes by the computed "connectivity values" and (2) the rules for numbering the nodes for table generation. Together these two methods complemented by the other techniques reduce what would otherwise be a devastatingly time consuming task to one which requires only a trivial amount of time.

In order to demonstrate the presumed advantages of the techniques described, they were programmed for an IBM 1410 Data Processing System. Over 25,000 chemical structures from CAS files, selected solely on the basis of immediate availability, were processed. The description of these structures and the statistics resulting from this test are shown in Appendix II. Based on these statistics and the published timings of other techniques which have been described in the literature, it appears that the present technique offers significant economic advantage over other methods for accomplishing the same end.

APPENDIX I

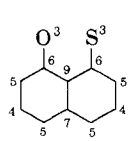
Example I.

$i = 0$



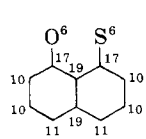
connectivity values
1, 2, 3
$k = 3$

$i = 1$



connectivity values
3, 4, 5, 6, 7, 9
$k = 6$

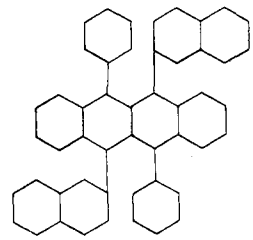
$i = 2$



connectivity values
6, 10, 11, 17, 19
$k = 5$

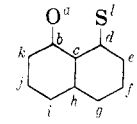
In this example the iterative process will be terminated after two iterations and the values assigned after the first will be used to introduce the partial ordering. In this example the subset of tables will consist of exactly four tables.

Example II.



	Connectivity values	Value of k	Size of the subset, tables
$i = 0$	2, 3	2	14,592
$i = 1$	4, 5, 6, 7, 8, 9	6	160
$i = 2$	8, 9, 11, 12, 14, 17, 18, 19, 20, 22, 24, 27	12	32
$i = 3$	18, 19, 20, 21, 26, 27, 28, 29, 31, 32, 38, 41, 42, 46, 50, 58, 68, 69, 75	19	8
$i = 4$	will also yield a value for k of 19; thus the process terminates and the values at $i = 3$ will be used to introduce the partial ordering of the nodes.		

Example III.



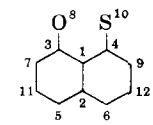
Using the connectivity values shown in example I of this appendix a partial ordering among the nodes is introduced.

$$\{c\} > \{h\} > \{b, d\} > \{k, e, i, g\} > \{j, f\}$$

Using this partial ordering the possible tables are generated and compared giving preference for lower numbering to the node or nodes which are between the leftmost pair of braces at each point of choice. For this example four tables must be generated and the unique table selected from among the set.

The preferred numbering and the corresponding unique table are shown below:

1	C	–	–
2	C	1	1
3	C	1	1
4	C	1	1
5	C	2	1
6	C	2	1
7	C	3	1
8	O	3	1
9	C	4	1
10	S	4	1
11	C	5	1
12	C	6	1
Rings	7	11	1
	9	12	1



In computer storage the unique table appears as follows:

From list	Ring closure
00100100100200200300300400400500600	007011009012

Node values	Line values
C C C C C C C O C S C C	1 1 1 1 1 1 1 1 1 1 1 1 1

APPENDIX II

In order to test the presumed economic advantages of the technique described in this paper, over 25,000 chemical structures were selected from the CAS files. These structures were selected solely on the basis of immediate availability and consisted of the following:

A	The "Ring Index" structures including the first supplement	9,568
B	A CAS File of commercial compounds	7,154
C	The structures from Lange's "Handbook"	4,596
D	The CAS File of compounds containing only carbon, hydrogen and sulfur	4,287
	Total	25,605

The following is a table of statistics resulting from the testing of these techniques using the file described above:

A	Sample size	25,605 structures
B	Total 1401 computer time for the generation of the unique description	4.93 hr.
C	Average number of compounds per minute for the generation of the unique description	92.8/min.
D	Average cost per compound for the generation of the unique description	2.2 cents
E	Average number of tables generated per compound	4.3

A Connectivity Code for Use in Describing Chemical Structures

ROBERT H. PENNY
General Electric Company, Computer Department, Falls Church, Virginia
Received June 17, 1964

This paper discusses a technique for efficient utilization of a computer to search a file of chemical compounds stored in structural form. The object of the search is to recognize in the file: (1) a compound identical with a given structure; (2) those compounds containing a given chemical fragment within their more complex structure; or (3) those compounds generic to a given structure.

A wide variety of codes and notations have been developed for describing chemical structures. Most of these notations were designed with a specific purpose or application in mind. As a result, the techniques employed for analyzing chemical structures expressed in these notations, although adequate from either a chemical or mathematical point of view, seldom lend themselves to efficient computer processing.

Three different computer techniques are currently being investigated as approaches to handling the structural recognition problem. One method is based on a division of the structure into basic groups with links to indicate how these groups are connected. This method has been used extensively and successfully in the past, but probably has the least promising future as far as computer application is concerned because of a lack of agreement as to what constitutes a "basic group."

A second method is the atom-by-atom comparison and search technique. Even with a large-scale digital computer this method can become time consuming unless extensive screening devices and short cuts can be formulated to minimize nonproductive path tracing and backtracking.

A third, and more recently proposed, method is based on set theory whereby sets are generated from graph theoretic and chemical characteristics of the structure. This particular method is discussed in more detail below.

The evaluation and implementation of any methodology must be based on how well it provides the chemist with the *best information* for the *least cost*. This can only be done if the problem is approached, not only as a chemical problem, but also as a computer problem since the computer is the medium through which this information must be processed. Chemical data representation and its subsequent processing must always be considered in terms of computer adaptability and efficiency.

With this in mind, a numerical code indicating the connectivity about an atom is presented which can be effectively used with a computer either as a tool in the atom-by-atom search technique or as a prime criterion for set generation.

Graphical Representation of Chemical Structures.—A chemical structure can be thought of as a graph, *i.e.*, a geometric figure consisting of points (nodes) and lines (edges) connecting these points. A node represents an individual atom in the structure and its "node value" is