In computer storage the unique table appears as follows:

From list	Ring closure
001001001002002003003004004005006	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

that specific chemical element associated with the node. The edges indicate the connectivity or bonding between the atoms. In the case of double or triple bonding, the degree of bonding between two atoms can be represented either as an attribute of a single edge (an assigned multiplicity) or by the presence of multiple edges.

Generally, identification or recognition of a chemical compound, when expressed in a structural form, is determined by the graphic connectivity between atoms, the chemical identity of these atoms, and their degree of bonding. Each of these three factors forms a set of differentiating criteria. However, the extent of differentiation varies considerably among the three sets. The degree of bonding covers but a small set of possible values. Therefore the information content and resultant differentiating power of this set is strictly limited. Much the same is true of the set designating chemical identity because of the predominance of but four elements (C, H, N, and O) in organic compounds.

On the other hand, the graphical representation of a compound exhibits a wide range of variation much on the same order as the possible sequence of moves of a chess piece. It would therefore appear that the development of a convenient means (notation) for describing graphical relationships between atoms will provide an information set with a very high level of differentiating power.

Connectivity Code.—The connectivity code describes in linear notation the graphic structure about an individual atom through three levels of connectivity,[1] and at the same time is reducible in the sense of indicating the possibility of either complete or partial mapping (embedding) between two structures. As is indicated, the code is not meant in anyway to describe a complete structure. It is a unique expression of the atomic network within the immediate neighborhood of the subject atom and is an *attribute* of the atom as much as is its chemical identity. In order to illustrate the generation of this code, the morphine structure in Figure 1 is taken as an example. For sake of simplicity, hydrogen atoms have been neglected.[2] Each nonhydrogen atom has been assigned an arbitrary identification number.

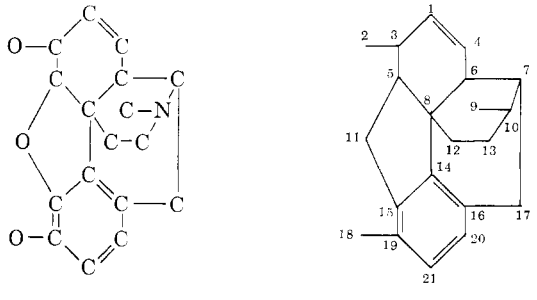

Figure 1.

(1) Use of three levels of connectivity is based upon an intuitive deduction and a superficial survey of the organic compounds. Only an exhaustive statistical analysis would actually determine an optimum level. The principle of the code remains the same; an alternate limiting level only curtails or extends the notational scheme.

(2) In most chemical recognition problems the indication or nonindication of hydrogen atoms is insignificant. When significant, their presence can be inferred on the basis of the bonding requirements.

Diagramming the connectivity about the carbon atom at (8), it is noted that (8) connects directly with the atoms at (5), (6), (12), and (14). This is pictured in Figure 2 as points on the circumference of a circle about (8).

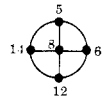

Figure 2.

Tracing further, the atom at (5) is connected at the next level to (3) and (11). They in turn are connected to (1) and (2), and (15), respectively, at the third level. The second and third levels of connectivity are indicated in Figure 3 as points connected through two additional circumscribed circles.

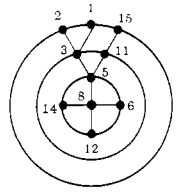

Figure 3.

The complete tracing of connectivity about (8) through three levels is shown in Figure 4. This, of course, could also be looked at as a tree graph which may be extended to whatever connectivity level appropriate.

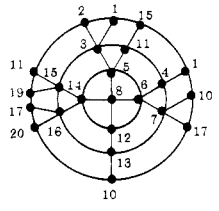

Figure 4.

Two things will be noted in Figure 4. First the diagram (and consequently the connectivity code) for atom (8) actually embraces seventeen additional atoms. Only the atom at (9), (18), and (21) are not represented. Also, the atoms at (1), (10), (11), (15), and (17) are represented twice. The significance of this latter situation will become evident in the section on the "Recognition of Ring Formations."

Construction of the Code.—The connectivity code as developed here (through three levels of connectivity) is made up of three parts: groups, the number of characters within a group, and the numerical value of these characters.

(1) Each branch from the subject atom to the first connection level corresponds to a group. The code is preceded by a slash (/) and each group is terminated by a slash. Returning to the morphine example, the four connections from atom (8) represent four groups and would appear in the code as ///// (note that there will always be one more slash than there are groups).

(2) The number of branches to the second level from each first level node is designated by the number of characters within each respective group. In our example, (5), (6), and (14) each have two connections into the second level, while (12) has but one. This is indicated in the code by */nn/nn/nn/n/*.

(3) The number of branches from each second level node to the third level is indicated by the numerical value of the characters described above. In the example of morphine, the complete connectivity code about (8) would be /22/21/21/1 (the full set of connectivity codes for morphine is listed in Figure 12).

As can be seen from the code, there are four groups (four connections to the first level). Three of these first level connections, /22/, /21/, and /21/, have double connections to the second level. The other, /1/, has but a single connection. Each of these second level connections has 2, 2, 2, 1, 2, 1, and 1 connection(s), respectively, to the third level.

Two cases not illustrated in the above example are worth noting: /0/ indicates a connection terminating at the second level; //, a connection terminating at the first level. This situation can be seen in Figure 5, the connectivity code being /310/0//.

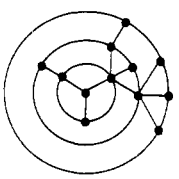

Figure 5.

A chemical structure can for many purposes be conveniently represented by a connection matrix. It is a square matrix, the number of rows and columns each being equal to the number of atoms in the structure. Each row and its corresponding column is associated with a specific atom. A connectivity between two atoms, (i) and (j), is indicated by a "1" at the intersection of their columns and rows, *i.e.*, the matrix elements a_{ij} and a_{ji} are each set to 1. Matrix elements representing no connectivity are zero.

Generation of the connectivity code from a metatable[3] can be readily accomplished by direct use of a connection matrix together with the sum (Σ) of the elements in the individual rows (or columns). The connection matrix for morphine, using the labeling of Figure 1, is pictured in Figure 6.

In generating the connectivity code about (8), it is first noted that the sum of the elements in row 8 is equal to 4, the number of first level connections. Therefore the code will be made up of four groups. Inspection of row 8 discloses nonzero elements in columns 5, 6, 12, and 14 which correspond to the four groups. The sum of the elements in each of these individual columns (or rows) will be one greater than the number of *characters* in their respective groups.

(3) "A metatable is a term often used to refer to a hypothetical 'standard' computer representation of structure data regardless of the mode or medium of input. It implies a tabular or matrix array which enumerates each atom, its connectivity, and other pertinent data as may be required such as valence, notational syntax, etc." J. Burger and W. Wilson, "A Review of Some Methods for Machine Manipulation of Chemical Structures." Rpt/RSIC-287, Sept. 1, 1964.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	Σ
1			1	1																		2
2			1																			1
3	1	1			1																	3
4	1					1																2
5			1					1			1											3
6				1			1	1														3
7						1				1								1				3
8					1	1						1		1								4
9										1												1
10							1		1				1									3
11					1									1								2
12								1					1									2
13										1		1										2
14								1							1	1						3
15											1			1					1			3
16														1			1			1		3
17							1									1						2
18																			1			1
19															1			1			1	3
20																1					1	2
21																			1	1		2

Figure 6.

When a nonzero element is encountered in row 8, that particular column (a group) is checked to determine which, if any, additional rows contain a nonzero element. If one or more such rows is found, each row represents a character within the current group and the numerical value of the character is equal to one less than the Σ for that row.

In our example, the generated connectivity code would initially be represented as /21/12/1/22/. This of course would be an arbitrary arrangement dependent upon the assignment of row and column sequence. It is necessary, however, that the code be expressed in a canonical form for uniqueness and to facilitate identity comparisons. The specific rules for specifying format are of secondary importance[4] as long as they are consistent throughout.

For our use, the characters within groups will be arranged in descending numerical order and the groups themselves in descending order by the most significant character. The resultant format is therefore /22/21/21/1/.

Recognition of Ring Formations.—The two structures in Figure 7 differ in that one consists of two hexagonal ring

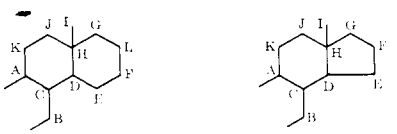

Figure 7.

formations while the other consists of one hexagonal and one pentagonal ring. The connectivity code about atom (D) in each case is identical: /110/10/1/. This is due to the tree nature of the structural description. Within the code, there is no indication of a distinction between position L in the first structure and the overlapping of positions F and G in the second. While generating the connectivity code, it can be determined whether or not the subject atom is part of one or more ring formations and the configuration of these rings.

Pentagonal Ring Identification.—The rule for determining a pentagonal ring is as follows: if there is direct connectivity between any two atoms in the second level, a pentagonal ring has been identified; each pair of such atoms determines a distinct ring. In the above example,

(4) The rules of format should depend upon the technique used by the programmer and be a material aid to him in checking for the embedding of fragments (see "Recognition of Fragments").

Figure 8b exhibits such a connectivity between (F) and (G). Therefore, the subject atom (D) is part of a pentagonal ring, extending through (E) and (H) to (F) and (G).

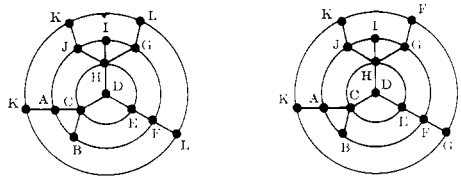

Figure 8.

Hexagonal Ring Identification.—The rule for determining a hexagonal ring is but an extension of the above: if an individual atom is common to two points on the third level, or expressed differently, if two atoms at the second level have a common connectivity at the third level, a hexagonal ring has been identified; each such incident determines a distinct ring. In the above example, the atom at (K) in both Figures 8a and b and the atom at (L) in Figure 8a exhibit this property. Therefore, the subject atom (D), in both cases, is part of a hexagonal ring through (H) to (J), and (C) to (A), converging at (K). In Figure 8a, (D) is likewise part of the hexagonal ring converging at (L).

The General Case.—There is obviously a basic underlying algebraic expression for identifying ring structures in terms of the connectivity levels. In the case of rings with an odd number of members, there is a direct connectivity between two elements in the $\frac{1}{2}(m-1)$th connectivity level (where m is the number of members in the ring). For rings with an even number of members, there is an element in common with two branches in the $\frac{1}{2}(m)$th connectivity level. This is illustrated in Figure 9 for $m = 3$, 4, and 7.

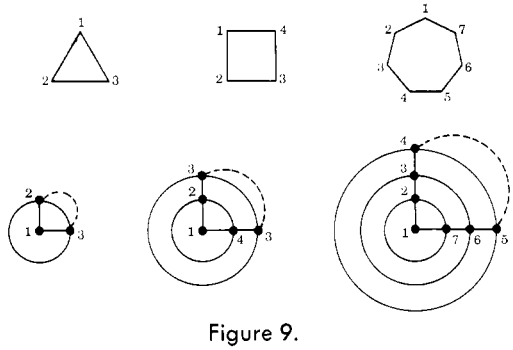

Figure 9.

Resolution of Isomorphic Ambiguity.—More important in a graphical sense (and in practical application), the connectivity code handles the problem of recognition of inverted or rotated configurations. A structure and its mirror image will have identical codes. No matter how two or more structures may originally be "pictured," if there is a one-to-one relationship (mapping) between the connectivities, the codes will agree and any isomorphic ambiguity is resolved. This is illustrated by the three examples in Figure 10. In each case the connectivity code

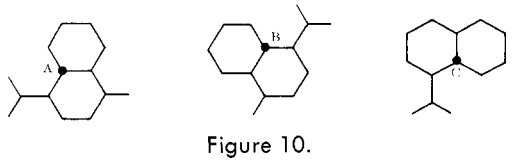

Figure 10.

about (A), (B), and (C) is identical: /21/21/1/. This property of the code is a result of the format rules. The relative position of the groups and the sequencing of characters within groups does not necessarily reflect a graphical configuration relative to any preconceived coordinate or geometric system; rather, the code is ordered on an *arithmetic* basis.

Application of Sets.—Both Sussenguth[5] and Unger[6] have developed algorithms for determining isomorphism based upon graph theory and set theory. The graphs Unger has analyzed are of a general geometric nature, while Sussenguth has been concerned primarily with the graphical representation of chemical structures.

The technique employed in each case is quite similar. The structural nodes are separated into pairs of sets according to their characteristics: (1) "node value," chemical element; (2) "degree," number of connections into the first level; (3) "branch value," whether the atom is connected to its neighbors by a single, double, and/or triple bond; and (4) "order," the minimum number of structural segments that must be traced to produce a circuit through a node.

The pairs of sets are analyzed and "partitioned" into a finer subdivision of sets in an effort to determine a unique correspondence between nodes. When a unique correspondence cannot be generated by further partitioning of the sets, the ambiguous matchings are tested by an attempted mapping of one structure upon the other.

This method, or variations thereof, provides a much more powerful tool for attacking the structural recognition problem than does a strictly atom-by-atom search technique. The value of the graph and set theory approach depends to a large extent on maximizing the initial division of sets and minimizing the number of subsequent partitioning operations.

The connectivity code can be used as a formidable basis for initial set generation. As an example, two diagrams of morphine are pictured in Figure 11 indicating the nonhydrogen atoms, their connectivity, bonding, and identifiers for each node.

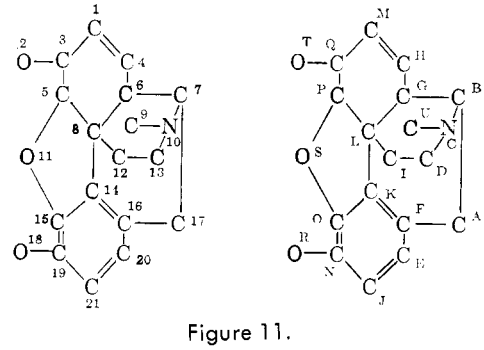

Figure 11.

The set generation and corresponding connectivity codes are listed in Figure 12. As can be seen, a majority of the nodes are already uniquely determined by this initial set subdivision. The sets (10, 19) and (C, N) can be resolved either on the basis of chemical identity or bonding. The (9) and (U) can also be matched by chemical identity. The need for extensive partitioning operations is unneces-

(5) E. H. Sussenguth, Jr., *J. Chem. Doc.*, 5, 36 (1965).
(6) S. H. Unger, *Comm. ACM*, 7, No. 1, 26 (1964).

sary. What few ambiguities still exist can be resolved by comparing the actual connectivities.

(1, 21)	/20/2/	(J,M)
(2, 9, 18)	/21/	(R,T,U)
(3)	/31/1//	(Q)
(4)	/32/2/	(H)
(5)	/221/2/10/	(P)
(6)	/221/21/1/	(G)
(7)	/31/2/10/	(B)
(8)	/22/21/21/1/	(L)
(10, 19)	/21/1//	(C,N)
(11)	/32/22/	(S)
(12)	/222/2/	(I)
(13)	/3/20/	(D)
(14)	/221/221/11/	(K)
(15)	/32/2/10/	(O)
(16)	/32/2/1/	(F)
(17)	/22/21/	(A)
(20)	/21/2/	(E)

Figure 12.

Recognition of Fragments and Generic Searches.—The connectivity code has been discussed so far only in connection with determining complete isomorphism between two structures. The code can also be of considerable aid when the problem is to determine if a particular fragmentary structure is contained within a more complex network. In Figure 13, the left-hand structure can be embedded within the network of the right-hand structure.

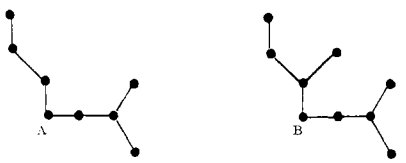

Figure 13.

The connectivity codes about (A) and (B) are /2/1/ and 2/10/, respectively. First, a test is made for congruity by by comparing the two complete codes. If, as in this case, equality is not found, a more detailed examination is required. The number of groups comprising the fragment's code must be compatible with (less than or equal to) the number of groups in the test structure's code. In like manner, the number of characters within the groups and the numerical value of these characters must fulfill the same criteria, *i.e.*, less than or equal to. As can be seen from the two connectivity codes in our example, these criteria are met and at (A), the fragment can be embedded within the structure at (B).

Determination of the above conditions is easily made by "mere" visual inspection, but this determination is not as easily accomplished by a computer. The value of the connectivity code depends upon the speed its utilization will impart to a computer program. Without going into the details of the required machine operations, it is sufficient to say that the ultimate rapidity of search will be a function of the ingenuity of the computer programmer, and poses an interesting application for an experienced "bit-juggler."

Generally, fragment recognition and generic search cannot be accomplished by set generation alone. Because of the incomplete nature of the information, the combination of set generation and atom-by-atom search techniques is required. In Figure 14 the three-ringed fragment is to be tested for inclusion within the morphine structure.

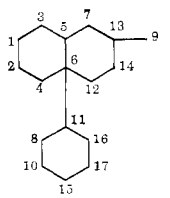

Figure 14.

The connectivity codes and corresponding sets for the fragment are indicated in Figure 15. To the right are listed the sets applicable to the morphine structure. It should be noted that the latter sets are not based on the equality of the codes, but on the ability to embed the fragment's code within those of the more complex structure.

(1, 10, 17)	/2/1/	(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,S)
(2)	/3/1/	(B,D,F,H,O,Q,S)
(3, 8, 16)	/31/1/	(B,F,H,O,Q,S)
(4)	/221/1/	(G,I,K,P)
(5)	/211/2/1/	(G,K,P)
(6)	/11/11/1/1/	(L)
(7)	/31/10/	(B,O,S)
(9)	/11/	(A,B,C,E,F,G,H,I,K,L,N,O,P.Q.R,S,T,U)
(11)	/211/1/1/	(G,K,P)
(12)	/221/2/	(G,I,K,P)
(13)	/2/1//	(B,C,F,G,K,L,N,O,P,Q)
(14)	/3/10/	(B,D,O,S)
(15)	/1/1/	(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,S)

Figure 15.

The sets can be further partitioned either on the basis of their inherent mathematical properties or by applying additional criteria such as node values, bonding, etc. In any event, an atom-by-atom search can now be undertaken to determine a one-to-one correspondence between the fragment nodes and a set of those from the compound. The various atom-by-atom search techniques will not be discussed here other than to indicate that the only paths traced should be those conforming to the set limitations.

Conclusion.—The connectivity code, as expressed, does not reflect any unique graphic arrangement of the connectivities, but has a universal character regardless of how the structure is rotated or inverted. Provided an appropriate input convention is established, additional information bits can be employed to indicate if a connection is above or below the planar surface. The problem here is not how stereo information may be represented or manipulated once within a computer, but how best to convey this information to the computer initially.

The connectivity code developed above is readily adaptable to computer methodology and provides a short cut to both atom-by-atom search and set generation procedures. When used with the atom-by-atom search technique, it gives the the computer program a "look-ahead" feature through three levels of connectivity and can thereby eliminate, with but one comparison, the detailed tracing of countless nonapplicable paths. More important, the connectivity code can form the basis for generating a powerful information set for graphical differentiation between atoms. The code, although not a solution in itself, becomes a valuable tool when used in conjunction with these general techniques for solving problems in chemical isomorphism.