An Algorithm for Finding the Synthetically Important Rings of a Molecule

By **Malcolm Bersohn,** Chemistry Department, University of Toronto, Toronto, Canada M5S 1A1

An algorithm for finding rings in the computer representation of molecular structures is presented. The method generates only chemically important rings. The difficulties of deciding on the best algorithm are discussed.

GRAPH theoreticians and those interested in chemical information processing by computer have both been concerned with the problem of counting and listing the rings of a structure. Their points of view differ; all rings can be equivalent to a graph theorist but not to a chemist.

The elementary ideas of rings in graph theory may be stated as follows.[1,2] We will speak only about undirected graphs and use the term graph as synonymous to undirected graph. A graph is a set of points (nodes) connected by lines (edges). A tree is a graph without rings (cycles). To construct a tree we start with a lone node. Then we add an edge which ends in a second node. Next we add a second edge which ends in a third node, *etc.* Thus trees have one more node than they have edges. In symbols, we have the formula

$$e = n - 1 \quad (1)$$

where e is the number of edges in the tree and n is the number of nodes. If, now, we add a new edge to a tree without introducing a new node then the new edge must connect two nodes, let us call them A and B, which are already in the tree. Since they already are part of the graph then there must exist a path from A to B. The newly introduced edge extends this path to complete a cyclic path from A to A, *i.e.* a ring. Since every time we add an edge without adding a new node we create a ring it follows that

$$e = n - 1 + R \quad (2)$$

where e and n are the numbers of edges and nodes in the graph and R is the number of its rings.

It we consider a molecule as a graph by taking each atom as a node and drawing an edge between all pairs of nearest neighbour atoms, we can translate equation (2) into an equivalent equation (3) for the number of rings in the molecule:

$$R = 1/2\{2N(\mathrm{C}) + 2 - \mathrm{UNS} - N(\mathrm{H}) + \sum_{i \neq \mathrm{C}} N(i)[\max(i) - 2]\} \quad (3)$$

where R = number of rings in the molecule, $N(\mathrm{C})$ = number of carbon atoms in the molecule, $N(\mathrm{H})$ = number of hydrogen atoms in the molecule, $N(i)$ = number of atoms of element i in the molecule, and $\max(i)$ = the greatest number of nearest neighbours that an atom of element i can have. This is 4 for a carbon atom, 3 for an uncharged nitrogen atom, 2 for oxygen, 1 for halogens adjacent to carbon atoms, *etc.* When the maximum number of nearest neighbours is present the atom is said to be saturated. The unsaturation of an atom can be defined as the difference between the actual number of nearest neighbours and

[1] O. Ore, 'Theory of Graphs,' American Mathematical Society Colloquium Publications, vol. XXXVIII, Providence, Rhode Island, U.S.A., 1962.

[2] C. Berge, 'Theorie des graphes et ses applications,' 2nd edn., Dunod, Paris, 1969.

the maximum number of nearest neighbours, and UNS is defined as the sum of the unsaturations of all of the atoms of the molecule. For example the values of UNS for acetone, acetylene, butadiene, and benzene are 2, 4, 4, and 6, respectively.

Use of either equation (2) or (3) will show that biphenyl has two rings; the chemist and the graph theorist would agree here and in all cases where the rings have at most one atom in common. For decalin, use of the equations again shows that two rings are present. From the point of view of graph theory we can construct decalin from any of a set of certain ten-node trees by adding two edges. Hence building any two of its three conceivable rings produces the molecule. For this reason, graph-theoretical algorithms [3-5] which find the rings of a graph will in a case like decalin either describe the molecule as being composed of a ten-membered ring and a six-membered ring or describe it as being composed of two six-membered rings. (The former result is obtained if the edge connecting the atoms at positions 4a and 8a happens to be one of the two edges omitted from the 'spanning tree' which these algorithms generate. If this edge is not one of those omitted from the tree then the latter result is obtained.) The synthetic chemist, however, needs to regard the decalin molecule as being built of two six-membered rings. The reason for this is that we have a variety of synthetic methods uniquely suited to making six atom rings,[6] but no unique methods for making ten-atom rings, especially ten-atom rings with functional groups appropriately placed so that the ten-atom rings can be joined at their middle to make, additionally, two six-membered rings. Hence, from the point of view of present-day synthetic chemistry, the description of decalin as two six-membered rings is the only reasonable one.

The necessity of a 'description' of the rings may not be evident to a human being with sight, who can easily discern from a structural diagram what situation he is confronted with. To a computer program, which is, in a manner of speaking, blind, the decalin molecule must also appear as two six-membered rings, otherwise the program will not be as efficient in solving problems of chemical synthesis. Specifically, in this case the program should not miss either of the two six-atom rings because, depending on the substituents, either ring might be a good starting point for a total synthesis. In addition the program should not waste time synthesizing a ten-atom ring and then trying to build a connection at its middle.

Sometimes the chemist wants to find more rings than equation (3) predicts will exist in the molecule. For example, in norbornane there are a six-atom ring and two five-atom rings, all three being of synthetic importance. Similarly in bicyclo[2.2.2]octane, there are three six-membered rings, any one of which could be important synthetically, depending on the substituents. But equation (3) finds only two rings in these two molecules. In decalin the enveloping ring was unimportant; in the latter two molecules the enveloping ring is important. Clearly a computer program that is to find synthetically important rings must have a precise definition of important rings. Corey and Petersson [7] have produced the first explicit definition of a synthetically important ring, which I paraphrase as follows. A ring is synthetically important if it contains six or fewer atoms or if it is not the envelope of other rings.

In addition to a definition of synthetically important rings we need an algorithm, that is to say a stepwise procedure, which will find all rings of a molecule satisfying the definition and only such rings. Fugmann *et al.*,[8] Plotkin,[9] and Corey and Petersson [7] have presented algorithms for finding the important rings of molecules. Their algorithms differ markedly. The first mentioned algorithm forms rings by combining two different paths between two atoms, provided certain conditions are met. The second algorithm starts with each bond in a pruned structure (*i.e.* one with chains removed) and then determines the longest path including this bond which has no branching possibility. Such a path is used as a basis for finding non-enveloping rings. This algorithm requires that the search for non-enveloping rings and the search for rings of less than a certain size be carried on in two different stages. The last algorithm uses logical operations to discard various extra, enveloping rings that are generated. In the first algorithm the ring is a ring of atoms; in the last the ring is a ring of atomic connections. The purpose of the present paper is to present an alternative algorithm to the aforementioned three. The algorithm to be discussed adheres to Corey and Petersson's definition of chemically important rings but it is different internally from any of the three previously mentioned algorithms.

It is difficult to assert the superiority of one of these four algorithms because the speed of execution of such an algorithm depends on: (1) the speed of the hardware, (2) the excellence of the programming art, *i.e.* the efficiency of the sub-algorithms used, (3) the programming language used, and (4) the particular representation of molecular structure that is used. Factor (3) has an important influence on factor (2). As to factor (4), it should be noted that one representation of molecular structure may be less efficient than another for finding rings but more efficient than the other for different substructure searches or for being transformed in simulated chemical reactions. After all these caveats, I

[3] J. T. Welch, jun., Comm. ACM 12,9 (Sept. 1965), 514—518.
[4] G. C. Gotlieb and D. G. Corneil, Comm. ACM 10,12 (Dec. 1967), 780—783.
[5] K. Paton, JACM 13,2 (Apr. 1969), 205—210.
[6] *Cf.*, for example, C. A. Buehler and D. E. Pearson, 'Survey of Organic Syntheses,' Wiley, New York, 1970; H. O. House, 'Modern Synthetic Reactions,' Benjamin, New York, 1965.
[7] E. J. Corey and G. A. Petersson, *J. Amer. Chem. Soc.*, 1972, **94**, 460. The work of P. L. Long, R. F. Phares, J. E. Rush, and L. J. White, presented at the 160th National Meeting of the American Chemical Society, 1970, appears to be similar but not directed towards the particular needs of synthetic chemistry.
[8] R. Fugmann, V. Dolling, and H. Nickelsen, *Angew. Chem. Internat. Edn.*, 1967, **6**, 723.
[9] M. Plotkin, *J. Chem. Documentation*, 1971, **11**, 60.

report for engineering benchmark purposes that a computer program using the algorithm of this paper required 4·31 ms to find the four rings of cholesterol. An IBM 370/165 computer was used and the program was written in the IBM 370 assembly language. This execution time is 13,470 times the execution time of one store instruction of the same computer. A connection table type of representation of molecular structure was used. (M. F. Lynch *et al.*[10] give a thorough discussion of computer representations of molecular structure in general and connection tables in particular.) The procedure is now described.

Steps of the Algorithm.—1. From equation (3), calculate R. If $R = 0$ then exit and report that there are no rings present.

2. Delete the chains with a free end from a copy of the molecular structure in the following fashion. Remove all terminal groups such as methyl, amino, hydroxy, mercapto, halogen, *etc.* If the adjacent atoms are made terminal by the first removal, *e.g.* methylene adjacent to methyl, then remove them also, and so forth. Eventually a scan through the atoms that remain in the structure will find no more terminal atoms, indicating that all free chains have been removed. Chains of atoms connecting two rings, but not a part of any ring, are not removed in this step; this does not affect the final result. Step 2 is not necessary; it is only a convenience for shortening the running time of the time consuming step 5.

3. If $R = 1$, the pruned structure is a single ring. Store the size of the ring (the number of its atoms) and the atoms of the ring suitably and then exit.

4. Choose the ' first ' atom of the remaining atoms and call it A. Here and in what follows we use the word ' atom ' to mean an atom of an element other than hydrogen.

5. Trace out all paths of length k (chains of k atoms) which begin at A and see if any of them returns to the atom A. Initially k is set equal to three. If no cyclic paths are discovered then increase k by one and continue the process, *i.e.* increase the length of each path by one atom. If a path returns to itself but not to its starting point, *e.g.* the sequence of atoms ABCDEFC, then the path is deleted from our list of possible cyclic paths starting from A. Suppose that D has neighbours E and G as well as C, then the original path ABCD becomes two paths, ABCDE and ABCDG. When a cyclic path is found we check to see if it has been found previously. If it has been found previously this path is deleted from our list of possible cyclic paths starting from A. Suppose that the name of the second atom in this previously discovered ring is B, so that the path is AB . . . A, then all paths beginning with AB are deleted if the length of the discovered ring is greater than six. If the length of the previously discovered ring is not greater than six, then the paths beginning AB . . . *etc.* are only deleted when and if their length becomes greater than six. This deletion guarantees that no ring of more than six atoms will be generated if it is the envelope of two or more smaller rings.

If the cyclic path has not been found previously then we store it and its length in suitable places. We continue to generate possible cyclic paths from A until all the possible paths have been eliminated by success, in becoming a newly found cycle, or by failure. Failure can come in three ways, *i.e.* by becoming a previously discovered cycle, by meeting itself at some atom other than A, or as above when it is eliminated to avoid the possibility of generating an enveloping ring of more than six atoms.

6. Select the next atom in the list of atoms of the structure which has been pruned of free chains. If there is no next atom, *i.e.* we have reached the end of the list, then we exit and the algorithm terminates. If the next atom has only two neighbours in the pruned structure and it lies on a known ring then we return to the beginning of step 6.

7. Use this next atom as A and go to step 5.

An earlier, simpler version of this algorithm is used in a program concerned with synthetic organic chemistry, which generates lowest cost synthetic pathways.[11] It is clearly important for such a program to recognize, for example, whether or not a ketone is cyclic, since cyclic ketones undergo certain reactions with yields different from those from acyclic ketones. For the same reason the computer program must be able to recognize a ketone as belonging, let us say, to a five-membered ring rather than to a six-membered ring.

I thank the National Research Council of Canada and the John Simon Guggenheim Memorial Foundation, New York, for support.

[3/002 *Received, 1st January, 1973*]

[10] M. F. Lynch, J. M. Harrison, W. G. Town, and J. E. Ash, 'Computer Handling of Chemical Structure Information,' Macdonald, London, 1971.
[11] M. Bersohn, *Bull. Chem. Soc. Japan*, 1972, **45**, 1897.