A Procedure for Rapid Recognition of the Rings of a Molecule

By **Ashmeed Esack**, Department of Chemistry, University of Toronto, Toronto, Canada M5S 1A1

A procedure ideally suited for ring recognition by computer programs that plan multi-step organic syntheses is proposed. The procedure consists of a modified ring algorithm, together with strategies for deducing the ring systems of the product (reactant) of a reaction from a knowledge of the reactant (product) ring systems and the nature of the reaction.

In multi-step synthesis programs [1,2] the entire task of generating optimal synthetic routes is delegated to a computer program. The program can operate in two possible modes (A) and (B). (A) The molecule to be synthesized, the goal molecule G, is examined and all possible predecessors (subgoals) L1, L2, . . ., L*n* (see Figure) are generated. Next all possible subgoals K*i*1, K*i*2, . . ., K*im* from each possible subgoal L*i* are generated and so on until a subgoal is found which is a readily available substance. Then the pathway from the available substance to the goal molecule G is a possible synthetic route provided that all co-reactants along the route can also be

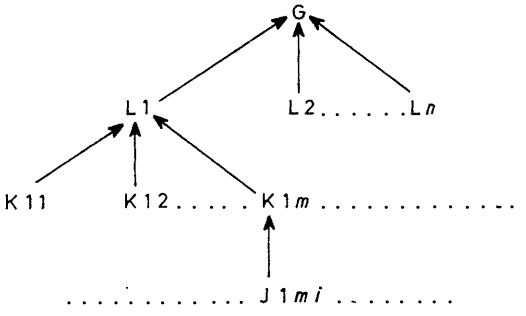
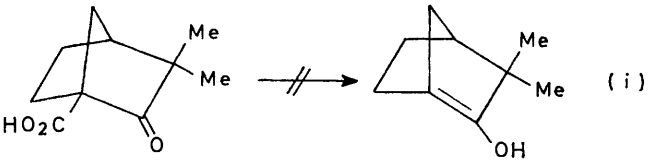


FIGURE A synthetic tree; G is the goal molecule

synthesized. This procedure is called the Backward Search method. (B) Alternatively, one can select a readily available material which resembles the goal molecule and, proceeding in a manner analogous to the Backward Search method, arrive at the goal molecule. This procedure is called the Forward Search method.

At each cycle of subgoal generation, information concerning the nature and position of functional groups and ring systems must be known to the program in order to facilitate synthetic decisions. Let us assume that each molecule in the Figure has an average of 20 chemically reasonable predecessors. Then, if we desire a five-step synthetic sequence to the goal molecule, the multi-step program must examine 20^5 or 3,200,000 intermediates. Furthermore, if we assume that 10 ms is a typical subgoal generation time, then the program would require 8·88 h of continuous computing to traverse the synthetic tree. If multi-step synthesis programs are to contribute significantly to organic chemistry, strategies must be devised to reduce the number of subgoals and the time required to generate each subgoal. It is the latter problem which is considered here.

At each cycle of subgoal generation the program must possess knowledge about the position and size of all synthetically relevant ring systems if it is to simulate reactions (*e.g.* the Diels–Alder reaction) which generate or manipulate ring systems. The generation of excessively strained molecules must also be avoided. In reaction (i) the decarboxylation either fails or takes a



(i)

different reaction path, in accordance with Bredt's rule. In addition, reactions of functional groups which are cyclic exhibit yields which can be appreciably different from those of acyclic groups. Finally, the perception of stereorelationships depends on the discovery of rings. The definition of a synthetically relevant ring is that of Corey and Petersson.[3] A ring is synthetically important if it contains six or fewer atoms or if it is not the envelope of other rings. The following ring algorithm can reduce the average subgoal generation time by reducing the time required for ring discovery.

Description of the Ring Algorithm.—Bersohn [4] has reported an algorithm for finding the chemically important rings in a connection table description of molecular structures. The efficiency of the procedure can be significantly enhanced if restrictions are imposed on (*a*) the choice of the particular atom in the pruned structure used as a starting atom in ring discovery, and (*b*) the choice of the particular neighbouring atom of the starting atom to which we subsequently advance. The modified algorithm is now described.

(1) Considering the molecule as a graph wherein each node is an atom and each edge connects pairs of nearest neighbour atoms, calculate the number of rings, R, in the molecule by using the equation $R = e - n + 1$, where e is the number of edges and n the number of nodes in the

[1] M. Bersohn, *Bull. Chem. Soc. Japan*, 1972, **45**, 1897.
[2] H. Gelernter, N. S. Sridharan, A. J. Hart, S. C. Yen, F. W. Fowler, and H. Shue, *Topics Current Chem.*, 1973, **41**, 113.
[3] E. J. Corey and G. A. Petersson, *J. Amer. Chem. Soc.*, 1972, **94**, 460.
[4] M. Bersohn, *J.C.S. Perkin I*, 1973, 1239.

graph. If $R = 0$, the procedure terminates as no rings are present in the molecule.

(2) Prune the structure by removing all free chains from a copy of the molecular structure. Chains of atoms connecting two rings but not a part of any ring, are not removed in this step.

(3) If $R = 1$, the pruned structure is a single ring and we store the size of the ring and its members suitably. We exit as the prodecure terminates.

(4) The pruned structure is examined and a list of all polyvalent nodes (atoms with three or more neighbours) is constructed.

(5) If there exists in the pruned structure a polyvalent node, such that three or more of its neighbouring atoms are each divalent (*i.e.* having two neighbouring atoms) or each polyvalent, then we go to step (10) of this algorithm.

(6) From the list of polyvalent nodes a starting atom, named A, is selected. Thus the condition we impose on the starting point in ring discovery is that A must be polyvalent. That is, the search for rings must begin at a ring junction atom or at the junction between a ring and the chain of atoms connecting that ring to some other ring system. Atom A may reside on a previously found ring.

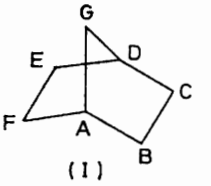
(7) Trace out all paths of length k which begin at A and see if any returns to the atom A. The restrictions on A imposed in step (5) imply that at least one and at most two neighbouring atoms of A can be polyvalent. We now impose the further restriction that the first advance from A must preferentially be to a neighbouring atom of A which is divalent and which does not reside on a previously found ring. This restriction allows us to discover each ring only once; that is, we do not traverse the same cyclic path more than once. Initially k is set equal to 3. If no cyclic paths are discovered, then k is increased by 1 and we continue. Rings of more than six atoms which are the envelope of two or more smaller rings are not generated, in accord with the definition of synthetically relevant rings. When we are successful in finding a cyclic path, we store the size of the ring and its component atoms suitable. A table which records the number of rings that each atom is in and the number of rings that each edge is in, is updated. The second, third, *etc.*, advance from A can be to any neighbouring atom of A that is not on previously found rings. No advance from A is possible if each of the neighbouring atoms of A resides on previously found rings. When this condition is encountered we proceed to step (8).

(8) Select from the remaining atoms on the list of polyvalent nodes another atom A and proceed to step (7). If there are no remaining starting atoms we go to step (9).

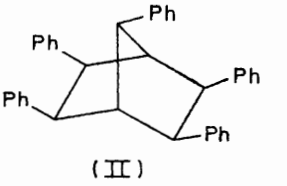
(9) Finally we examine the table that records the number of rings in which each edge is present. We impose the last restriction that all edges which reside in more than one ring must have polyvalent nodes at both ends. If this condition is satisfied, the procedure terminates normally and we have found all the synthetically relevant rings in the molecule. If this condition is not satisfied we proceed to step (10).

(10) All restrictions on the nature of the starting atom and on the nature of the neighbouring atom to which we advance are relaxed. Using the pruned structure as its argument we execute the Bersohn algorithm as originally described.[4]

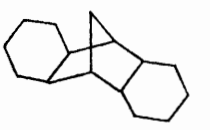
The algorithm of steps (1)—(9) is incapable of discovering all the synthetically relevant rings in molecules in which overlapping rings occur. The molecule bicyclo-[2.2.1]heptane (I) is such an example. This molecule

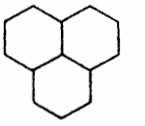


(I)



(II)


(III)


(IV)

contains three rings: the six-membered ring (A,B,C,D,E,F) and two five-membered rings (A,B,C,D,G) and (A,F,E,D,G). The algorithm of steps (1)—(9) is capable of discovering only two rings: (A,B,C,D,E,F) and (A,B,C,D,G) if atom A is chosen as a starting atom and the first advance is from A to B. It is for this reason that steps (5) and (9) have been inserted in the algorithm. Step (9) describes a test by which systems with overlapping rings are detected. For example, since the edge BC belongs to two rings and its end nodes B and C are both divalent then we would abandon the ring discovery process in step (9) and find the required three rings by execution of the more time-consuming but rigorous Bersohn algorithm, in step (10). Step (5) is used as a screen for early detection of systems of overlapping rings, thereby reducing the number of times the routine is needlessly executed. For example, structure (II) would pass the test in step (9) but is detected in step (5), and structure (III) passes the test in step (5) but is detected in step (9). There are systems which do not possess overlapping rings but which are rejected in step (5). Structure (IV) is such an example. The inclusion of step (5) also implies that the rings of spiro systems and systems with bridgehead atoms are not amenable to discovery by the algorithm of steps (1)—(9). The algorithm is however particularly fast for fused ring systems such as steroids, alkaloids, triterpenes, and other natural products.

This algorithm was written in the IBM 370 assembly language and executed on the IBM 370/165 computer. The average execution time for a single store instruction on this machine is 0·32 μs. The algorithm required 0·95 ms to discover the four rings of cholesterol; *cf.* with the previously reported time of 4·31 ms.[4] The increased efficiency is clear.

Strategies for Ring Deduction in the Backward Search Method.—As previously indicated in the Backward Search

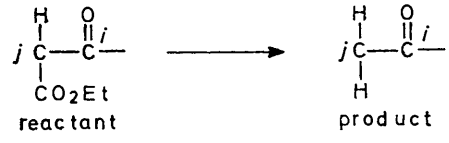
method we start at the root of the synthetic tree, with the goal molecule in hand, and work our way down the tree to readily available materials. At every node in the tree, information concerning the functional groups, ring systems, and stereochemical features of the product molecule is known to the program. This information is subsequently utilized in simulating a chemical reaction to generate a reactant (or reactants) that can reasonably be expected to give rise to the product in question. Having performed the reaction, we must once again gather information concerning functional groups, rings, and stereochemical features of the newly generated molecule. In certain cases the ring systems of the reactant molecule (the newly generated molecule) can be unambiguously deduced from a knowledge of the ring systems of the product molecule and the nature of the reaction performed. The cases where this is possible are systematized below and for pedagogic reasons are classified according to the number of non-hydrogen atoms (termed reaction sites) involved in the reaction.

All reactions are defined in the synthetic sense.

I. Reactions involving one reaction site. In the notation of Hendrickson [5] these are reactions which change the functionality of the reaction site. Some examples are oxidation of an alcohol to a ketone, reduction of a nitro-group to an amine, and halogenation of an aromatic carbon. Since the ring systems of the product are unaffected by these reactions, the ring systems of the reactant are exactly those of the product. Included in this category are reactions which add or remove blocking groups.

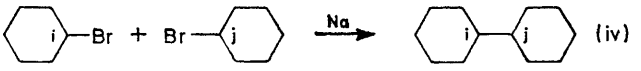
II. Reactions involving two reaction sites, i and j. These reactions can be subdivided into three classes:

(1) Reactions in which the σ-bond between i and j remains intact. In such reactions, the ring systems of the reactant are those of the product, *e.g.* (ii) and (iii).

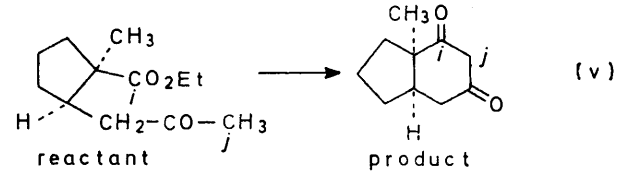


(2) Reactions in which a σ-bond is made between i and j. In these reactions the nature of the product edge ij, that is the number of rings the edge is in and the number of neighbouring atoms of i and j, becomes important. The following reactions are classified according to the nature of the edge ij.

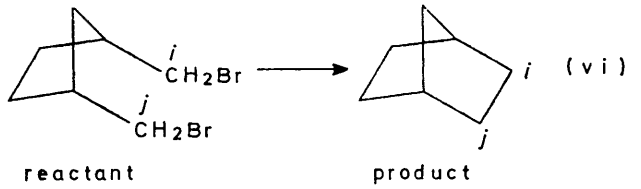
(2a) The edge ij is not a member of any ring [reaction (iv)]. Assuming that only two major reactants are all that is necessary to make the bond, the rings of reactant 1 are obtained by deleting all rings containing node j from the list of rings of the product molecule. Similarly, the rings of reactant 2 are obtained by removing all rings containing node i from the list of rings of the product molecule.



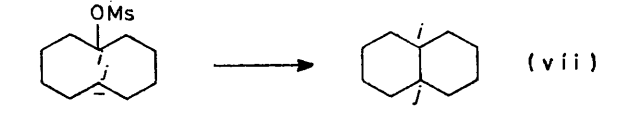
(2b) The edge ij is a member of only one ring [reaction (v)]. In such a case, the reactant rings are obtained by deleting all rings containing this edge from the list of product rings.



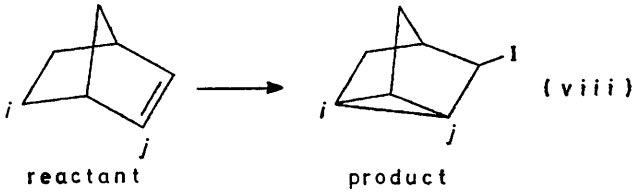
(2c) The edge ij is a member of more than one ring and in the pruned structure nodes i and j are either both divalent or node i is divalent and node j polyvalent [reaction (vi)]. The reactant rings are generated by deleting rings containing this edge from the list of product rings. An example of such a reaction is:



(2d) The edge ij is a member of two rings and in the pruned structure nodes i and j are both polyvalent [reaction (vii)]. The reactant rings are generated by deleting the two rings containing the edge ij from the list of product rings. In addition, suppose the two rings were $(i,m,n,\ ,\ ,j)$ and $(i,p,q,\ ,\ ,j)$, then the path $(i,m,n,\ ,\ ,j,\ ,\ ,q,p)$ is examined. If the path returns to itself but not to the starting point i, that is if a member atom appears more than once in the list $(i,m,n,\ ,\ ,j,\ ,\ ,q,\ ,\ ,\ ,p)$, then this path is deleted. Otherwise the path $(i,m,n,\ ,\ ,j,\ ,\ ,q,p)$ is a valid cyclic path in the reactant molecule.



(2e) The edge ij is a member of more than two rings and in the pruned structure nodes i and j are polyvalent [reaction (viii)]. The reactant rings are generated by

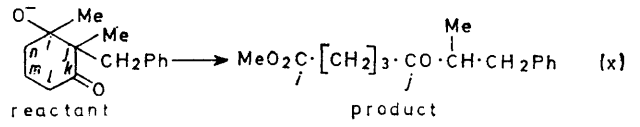


[5] J. B. Hendrickson, *J. Amer. Chem. Soc.*, 1971, **93**, 6847.

deleting from the list of product rings all rings containing the edge *ij*.

(3) Reactions in which the σ-bond between *i* and *j* is cleaved. Reactions in this category are of three kinds depending on the nature of the end nodes *i* and *j*. Notice that the edge *ij* does not exist in the product molecule.

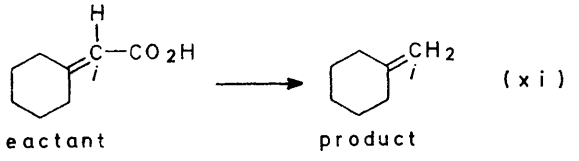
(3a) Both *i* and *j* are not members of any product rings [reaction (x)]. If the intermediate atoms (*k,l,m,n*) separating *i* and *j* in the product molecule are known to the



(x)

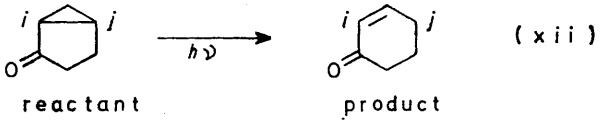
program, then the reactant rings are produced by the sum of the product rings and the new ring (*i,j,k,l,m,n*). If the intermediate atoms are unknown, then the ring systems of the reactant are discovered by execution of the ring algorithm previously described.

(3b) Only atom *i* (or *j*) is present in the product molecule. Such reactions are common, for example the cleavage of a hydrocarbon side-chain [reaction (xi)]. The ring systems of the reactant are those of the product molecule.



(xi)

(3c) Both *i* and *j* are members of one or more rings in the product. Consider reaction (xii). The rings of the



(xii)

reactant are those of the product plus all synthetically relevant rings of the type (*i* , , , *j*). The latter are generated by execution of the following procedure.

(*a*) Examine, one at a time, each ring on the list of product rings. A ring description is a linear array of numbers (atom labels). Each ring is described in a clockwise or anticlockwise manner proceeding from the starting node.

(*b*) Examine the first node in the ring description. If the node is *i* (or *j*) proceed to step (*c*). Otherwise go to step (*d*).

(*c*) Proceeding left to right along the ring description, add to node *i* all nodes which follow *j* sequentially in the array. If the ring so produced has not been previously found (*i.e.* does not appear on the list of reactant rings), then store the size and members of the ring (*i,j* , , ,) suitably. Go to step (*e*).

(*d*) Proceeding left to right along the ring description, sequentially collect all nodes until node *i* (or *j*) is encountered. Add *j* to the path so produced and, if the resulting path has not been previously reported, then store its size and members (, , , *i,j*) suitably. Go to step (*e*).

(*e*) Examine the last node in this same product ring. If the node is *i* (or *j*) go to step (*f*). Otherwise go to step (*g*).

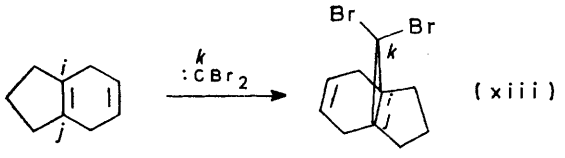
(*f*) Proceeding right to left along the ring description, sequentially collect nodes until node *j* is encountered. If the ring so generated has not been previously found, then store its size and members (*i*, , , *j*) suitably. Go to step (*h*).

(*g*) Proceeding right to left along the ring description, sequentially collect nodes until node *i* (or *j*) is encountered. Add *j* to the found path so produced and if the resulting ring has not been previously reported, then store its size and members (, , , *i,j*) suitably. Go to step (*h*).

(*h*) Steps (*b*)—(*g*) are repeated for each product ring. When this has been completed, the reactant ring systems are the new rings discovered by the above procedure plus the ring systems of the product molecule. The above procedure is independent of the number of neighbouring atoms of both *i* and *j*.

III. Reactions involving three reaction sites *i,j,k* and where the edge *ij* is present in the reactant and product. Such reactions fall into two categories.

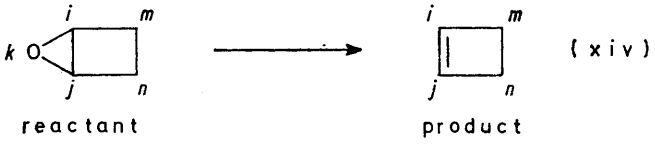
(1) Reactions in which the bonds *ik* and *jk* are made



(xiii)

[*e.g.* reactions (xiii)]. The ring system of the reactant are produced by removing all rings containing the edge *ik* (or *jk*) from the list of product rings.

(2) Reactions in which the bonds *ik* and *jk* are cleaved [*e.g.* reaction (xiv)]. The ring systems of the reactant

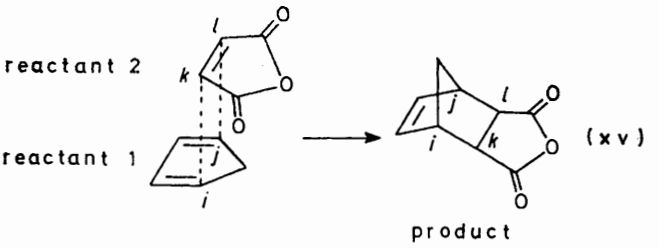


(xiv)

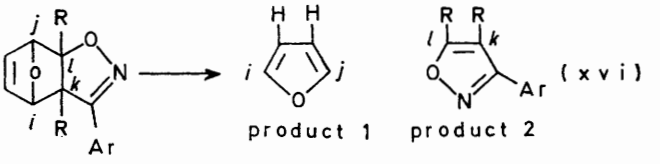
are produced by adding the ring (*i,j,k*) to the list of product rings. In addition, cyclic paths of the type (*i,m,n,* , , *j,k*) are generated, where (*m,n,* , ,) are the intermediate atoms in rings which contain the edge *ij*. Such rings are synthetically important if they are of size six-atoms or less.

IV. Reactions involving four reaction sites *i,j,k,l*. The reactions for which we can unambiguously deduce the reactant rings knowing the product rings are of two kinds.

(1) Reactions in which two edges *ik* and *jl* are made. Examples of such reactions are cycloadditions, the Diels–Alder, and 1,3-dipolar addition reactions. The ring systems of reactants 1 and 2 are generated by deleting from the list of product rings those rings containing the node *k* (or *l*) and *i* (or *j*), respectively. Reaction (xv) is an example of the Diels–Alder type.



(2) Reactions in which two edges *ik* and *jl* are cleaved. An example is the retro-Diels–Alder reaction, *e.g.* (xvi). The rings of the reactant are deduced as follows.



All the rings in the products are also present in the reactant. In addition, the reactant possesses those cyclic paths of the type $(i,,,j,l,k)$ and $(i,,,j,l,,,k)$ which are synthetically relevant.

For all other reactions, the ring systems of the reactant are determined by executing the ring algorithm previously described.

Strategies for Ring Deduction in the Forward Search Method.—A multi-step synthesis program operating in the Forward Search mode starts with a readily available material and traverses the synthetic tree until a molecule which matches identically the goal molecule is generated. At any given time, attention is focused on the reactant and all information of synthetic interest about the reactant is used to simulate a reaction and generate a product. In the next cycle, the newly generated product becomes the reactant and we proceed as before. It is possible to deduce unambiguously the ring systems in the following reactions. Reactions are defined in the synthetic sense and are classified according to the number of reaction sites involved in the reaction. Examples of each reaction type can be found in the section of this paper dealing with the Backward Search method.

I. Reactions involving one reaction site. The product rings are those of the reactant molecule.

II. Reactions involving two reaction sites *i* and *j*. These reactions can be divided into three classes.

(1) Reactions in which the σ-bond between *i* and *j* remains intact. In such reactions, the ring systems of the product are those of the reactant molecule.

(2) Reactions in which a σ-bond is made between *i* and *j*. Here the edge *ij* does not exist in the reactant and two cases are distinguishable.

(2a) Only node *i* (or *j*) is present in the reactant molecule. Such reactions usually require two reactants and the product ring system is the sum of the ring systems of each reactant.

(2b) Both *i* and *j* are present in the reactant molecule. The product ring systems are those of the reactant. In addition all cyclic paths of the type $(i,,,j)$ are added to the list of product rings.

(3) Reactions in which the σ-bond between *i* and *j* is cleaved. The product rings are obtained by deleting all rings containing the edge *ij* from the list of reactant rings.

The strategies for handling the reactions belonging to categories III and IV are closely analogous (but in reverse) to those of the Backward Search method.

For all other reactions, the ring systems of the product are determined by executing the ring algorithm previously described.

The strategies outlined above for ring deduction in the Backward and Forward Search methods effectively eliminate the necessity of executing the ring algorithm at each cycle of subgoal generation. The reactions which are amenable to these strategies are numerous and arise frequently in a typical synthesis. Thus, a procedure that combines these strategies for ring deduction with the modified ring algorithm will certainly reduce the average subgoal generation time by reducing the time required for ring recognition. The foregoing strategies are now being incorporated into Backward and Forward synthetic programs.

I thank the National Research Council of Canada for a scholarship, and M. Bersohn and P. Yates for discussions.

[4/1733 *Received*, 19*th August*, 1974]