Mathematical Basis of Ring-Finding Algorithms in CIDS*

MORRIS PLOTKIN
Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia, Pa. 19104

Received May 7, 1970

CIDS is an information storage and retrieval system that performs on-line chemical structural searches. Among its retrieval screens are the atom-population of individual rings. This paper defines a class of rings for which screens are assigned and describes briefly the algorithms by which computer programs analyze the structures for automatic screen assignment. The class of rings is independent of structure orientation, and may be defined equivalently as either the rings that are not the envelope of shorter rings or the rings that can be a member of a smallest set of smallest rings.

CIDS (Chemical Information and Data System) is an information storage and retrieval system designed by the University of Pennsylvania for the U.S. Army. CIDS performs on-line retrieval of chemical compounds in response to queries that may specify structural characteristics. In particular the queries may specify, among many other structural characteristics, atom-population of rings. For example, a query might include the specification that each retrieved compound contain a ring consisting of five carbon atoms, one nitrogen atom, and one oxygen atom. As each new compound is registered by the system, it is analyzed by computer programs for certain structural characteristics, including the atom-population of rings. Because in a large fused-ring system many dozens of rings can be traced, few of possible chemical interest, not all rings are found and recorded. This paper defines the class of rings found by the computer programs and describes briefly the algorithm used.

The rings found are the rings that contain eight or fewer atoms, or belong to a certain class K, or both. The eight-or-fewer-atoms property is of little interest and is not discussed further. The class K is defined later.

WORK BY OTHERS

The work of Welch[1] and Gibbs[2] provides a method for finding all the rings in a given compound. The method begins with finding an arbitrary set of fundamental cycles, a problem to which contributions have been made by Gotlieb and Corneil[3] and by Paton.[4] The work described below finds only the rings of a limited class, selected for its interest to chemists. The algorithm does not make use of an arbitrary set of fundamental cycles.

The work of Fugmann, Dölling, and Nickelsen[5] is directed at the same problem as the present paper. The class of rings found by their algorithm is too complicated to define here. The author believes that the definition of the class of rings found here is more meaningful to chemists than the definition of the class of rings found by Fugmann, Dölling, and Nickelsen.

A by-product of the algorithms described below is that for each ring system in the given compound there is found a smallest set of smallest rings, of interest in its own right. The number of rings and the number of atoms in each ring in the smallest set of smallest rings together constitute one of the CIDS retrieval screens.

THE CLASS K OF RINGS

The structure of a chemical compound is commonly represented by a linear graph in which the vertices are the atoms, and the edges are the chemical bonds. The rings of the structure are, to a chemist, what mathematicians call the cycles of the linear graph.

Mathematicians define addition of cycles as follows.

Let $(R_1, R_2 \ldots, R_m)$ be a set of cycles in a linear graph. Let $(E_a, E_b, \ldots)$ be the set of edges each of which occurs in an odd number of the cycles $R_1, \ldots, R_m$. If the edges $E_a, E_b, \ldots$ constitute a cycle R, then $R = R_1 + \ldots + R_m$. Chemists call such a ring R the envelope ring of the set $(R_1, \ldots, R_m)$ of rings.

If R is a ring of m bonds, the length of R is denoted by $|R| = m$.

Now the class K may be defined. K consists of all the rings that are not expressible as the sum or envelope of shorter rings—i.e., if there exists a relation $R = R_1 + \ldots + R_n$ where $|R_1| < |R|, \ldots, |R_n| < |R|$, then R is not in K; otherwise R is in K. After the introduction of additional concepts that will be useful later for describing the algorithm, another characterization of the set K is given below. The algorithm for finding the rings of K is based on the latter characterization.

LINEAR INDEPENDENCE OF SETS OF RINGS

Let B be the number of bonds in a structure. Each ring R can be regarded as a vector in a space of B dimensions, the jth coordinate being 1 or 0 depending on whether the jth bond does or does not occur in R. Each set of bonds corresponds to a vector in this space, but only those sets that constitute a ring are of interest at the moment. Addition of rings is vector addition modulo 2.

From well-known[6] properties of vector spaces it follows that for any set C of rings in a structure there is a number $d(C)$ that is equal to the number of rings in any maximal linearly independent subset of C. A subset C' of C is linearly independent if no ring of C' is expressible

* This work was supported by the U.S. Army, Edgewood Arsenal, under Contract No. DAAA15-69-C-0140 and by the National Science Foundation on Contract C-467.

as the sum of other rings of C'; it is maximal if no other ring of C can be adjoined to C' without destroying the linear-independence property. It follows that if C' is a maximal linearly independent subset of C, every ring of C can be represented as the sum of rings in C'. The latter property is expressed by calling C' a basis for C; a basis, by definition, must be linearly independent.

THE PROCEDURE P

Let all the rings of a given structure be ordered by length R_a, R_b, R_c, ... where $|R_a| \leqq |R_b| \leqq |R_c| \leqq \ldots$. If any two rings of the structure are of the same length, the ordering is not unique. The set S is defined as follows. R_a is in S. R_b is in S. Beyond R_b, each R_i is examined in sequence, and it is admitted to S if and only if it is not expressible as the sum of rings previously admitted to S. The procedure whereby S is constructed from an ordering-by-length of all the rings of a structure is called Procedure P. If C is the set of all rings of the structure, the set S is clearly a basis for—i.e. a maximal linearly independent set in—C. S therefore contains $d(C)$ rings, regardless of what ordering-by-length was used. Procedure P is not part of any algorithm used in CIDS; it is introduced here as an aid in the proof of statements.

THE SMALLEST SET OF SMALLEST RINGS

Let C_r be the set of all rings of length r or less, to continue the discussion about Procedure P. C_r is empty for $r = 1$ or 2, and also for some larger r in most chemical structures. Let S_r be the rings of S that are also in C_r—i.e., the rings of the set S constructed by Procedure P that contain r or fewer atoms.

The nature of Procedure P is such that S_r is a basis for the rings of C_r. The number of rings in the set S_r is therefore $d(C_r)$, regardless of the ordering-by-length that was used for constructing the set S.

Further, if S' is any other basis for the set C of all rings, the number of rings in S' of length r or less cannot exceed $d(C_r)$, the number of rings in S_r, because $d(C_r)$ is the maximal number of linearly independent rings in C_r. Therefore S' has at least as many rings of length greater than r, for each r, as does S. In this sense the rings of S are as small as possible, and the set S is called by chemists a smallest set of smallest rings, abbreviated SSSR. For any ordering-by-length of the rings of a structure, Procedure P produces an SSSR.

Conversely, given an SSSR—i.e., given a basis for the set C of all rings such that each ring of the basis can be paired off with a corresponding ring of equal length in the set S—there is an ordering-by-length of the rings of C that will, by application of Procedure P, produce the given SSSR. All that is necessary is to place the rings of length r, of the given SSSR, first among the rings of length r of the structure, in the ordering-by-length.

ALTERNATE CHARACTERIZATION OF CLASS K

It can now be shown that the class K might equivalently be defined as the set of all rings that appear in one or more SSSR for the structure. The proof is in two parts.

If a ring R is in any SSSR, it is a member of the set S produced by the application of Procedure P to some ordering-by-length of the rings of the structure. Let R be of length r. The selection of R by Procedure P means that R is not expressible as the sum of rings of a basis of C_{r-1}. Therefore R is not expressible as the sum of rings in C_{r-1}—i.e., R is in K, by the definition of K.

Conversely, if R is in K, it will be selected by Procedure P for any ordering-by-length in which R is first among the rings of its length, and is therefore in at least one SSSR. The proof is complete.

PATHS

The description of the algorithm for finding all the rings of class K requires the use of paths. A path of length L is a set of L bonds joining two distinct atoms A and B as follows: the first bond joins A to an atom X_1, the second joins X_1 to an atom $X_2, \ldots$, and the Lth joins atom X_{L-1} to atom B. The atoms $X_1, \ldots, X_{L-1}$, called the interior atoms of the path, must all be distinct, and distinct from A and B; the case $L = 1$, in which there are no interior atoms, is admissible. The length of the path P is denoted by $|P|$. If each interior atom is the terminus of only two bonds—namely, the bonds of the path—the path is said to be unforked. The useful property of unforked paths is that if any bond of an unforked path is in some ring, all of it is in that ring.

The definition of addition of rings in terms of addition modulo 2 of the bond-set vectors extends naturally to the addition of paths: the sum of two paths P and Q is the set of bonds that occur in P or in Q but not in both. If P and Q each join the same pair of atoms A and B, and if P and Q are disjoint—i.e., have no bonds or interior atoms in common—their sum constitutes the bonds of a ring R, and the fact is expressed by writing $P + Q = R$. In this case, $|P| + |Q| = |R|$. If paths P and Q each join the same pair of atoms A and B but the paths are not disjoint, then (a) their sum constitutes the bonds of two or more rings and (b) the sum of the lengths of the rings is at most $|P| + |Q|$. Fact (a) follows from a theorem[7] in graph theory that a finite linear graph with an even number of edges terminating at each vertex is decomposable into cycles; fact (b) follows from the fact that any bond in $P + Q$ is a bond in either P or Q.

FINDING AN SSSR

The first step in the algorithm for finding all the rings of class K in a structure is finding an SSSR for the structure. The principal tool is the following theorem.

Theorem 1. If P is an unforked path in a structure G and there is a shortest ring R through P such that $|R| \leqq 2|P|$, there is an SSSR of G that contains R and contains no other ring in which P occurs.

Proof. Consider the SSSR produced by the application of Procedure P to an ordering-by-length in which R is first among the rings of length $|R|$. Since no ring before R in the ordering could have contained the path P, the SSSR produced will contain R. Let R' be any other ring containing the path P. Let Q and Q' be the paths

respectively that combine with P to make R and R'—i.e., $P + Q = R$ and $P + Q' = R'$. There are two possibilities, depending on whether Q and Q' are disjoint.

If Q and Q' are disjoint, then their sum $Q + Q' = R''$ is a ring of length $|R''| = |Q| + |Q'|$. By hypothesis $|Q| + |P| = |R| \leqq 2|P|$, so that $|Q| \leqq |P|$ and $|Q| + |Q'| = |R''| \leqq |R'| = |P| + |Q'|$. That is, the ring R'' consisting of the paths Q and Q' is no longer than the ring R'. It follows that if R' is in an SSSR, it can be replaced in the SSSR by R'' which does not contain the path P, since R'' is no longer than R', and the new SSSR will be a basis for the set of all rings if the old SSSR is (and of course it is) by virtue of the relation $R' = R'' + R$. Therefore, the SSSR need not contain R' if Q and Q' are disjoint.

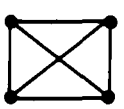
If Q and Q' are not disjoint, then by the discussion under Paths their sum is equal to the sum $\sum R_a = Q + Q'$ of smaller rings. But then $R' = R + \sum R_a$ and so would not appear in the SSSR produced by Procedure P on the ordering-by-length in which by hypothesis R is first among the rings of its length.

To summarize, an arbitrary ring R' other than R containing the path P either does not appear in the SSSR constructed as described above or can be replaced in it by a ring R'' that does not contain P. The proof is complete.

To begin with, the algorithm strips off—i.e., deletes—any hanging chains that are clearly not in any rings. Then it seeks to apply Theorem 1. The search for a suitable unforked path P and ring R is exhaustive—each bond B of the structure is tested in turn. First there are determined the longest unforked path P containing B, and the end-point atoms X and Y joined by P. Then the shortest rings through P are found by tracing out from X all paths of length 1,2,... that do not intersect P. If the atom Y is not encountered by the time all such paths of length $|P|$ have been built out from X, P does not lie in any ring R for which $|R| \leqq 2|P|$. If the point Y is first encountered at some distance $L \leqq |P|$ from X, each path of length L which joins X and Y, together with P, gives a shortest ring R through P of length $|R| = L + |P| \leqq 2|P|$.

Each time a suitable path P is found contained in one or more suitable rings R, the algorithm records one of the shortest rings R and deletes P from the graph. Theorem 1 makes the deletion permissible. The procedure is repeated. After $d(C)$ iterations, where C is the set of all rings, there are no rings left and the set of rings R that has been found is an SSSR for the original structure. The value of $d(C)$ is known beforehand from the formula, long familiar from electric circuit theory, $d(C) = B - A + n$ where B is the number of bonds, A the number of atoms, and n the number of disjoint pieces in the original graph. This formula is used in the algorithm as an indicator of whether any rings remain.

For the great majority of compounds repeated application of Theorem 1 is enough to determine an SSSR. Occasionally a situation is encountered in which no suitable path P exists. An example is shown:

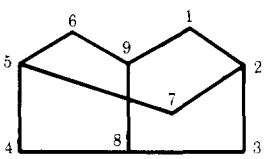


The algorithm then seeks a bond B to delete temporarily, to create the following situation: (*a*) the remaining structure can be handled—i.e., an SSSR found—by repeated application of the step of Theorem 1, and (*b*) the longest ring in the SSSR so found is not longer than the shortest ring R_B through B in the original structure. The proof is omitted here, but a straightforward use of Procedure P shows that if (*a*) and (*b*) hold, the ring R_B plus the SSSR for the reduced structure gives an SSSR for the original structure.

The choice of the bond B to be deleted is made as follows. A bond is selected at random and called B_1. A shortest ring R_1 is found through B_1, in the same way a shortest ring was found earlier through the path P of Theorem 1. For $j = 2,3,\ldots$, a bond B_j is selected at random not in any ring $R_1,\ldots,R_{j-1}$, and a shortest ring R_j containing B_j is found. This procedure continues until every bond is in at least one of the rings R_j. If there is a unique largest R_j, the corresponding B_j is the only bond tried as the bond B to be deleted; if there is a tie for largest R_j, the corresponding B_j's are tried in turn until either one results in determining an SSSR or all fail. If all fail, the algorithm concedes defeat and prints out a message of failure. The search for a bond B to be deleted takes place not only at the start of the algorithm, but after any number of applications of Theorem 1, if a suitable P cannot be found for the next application of Theorem 1. The search procedure just described does not ensure that a suitable B will be found if one exists, but it has been at least adequate for practical purposes.

EXAMPLES

Structure 2553 of the Ring Index[8] is used to demonstrate the working of the SSSR-finding algorithm. The longest unforked paths in the structure are of length 2; they



Ring Index # 2553

are 2, 1, 9; 2, 3, 8; 2, 7, 5; 5, 4, 8; and 5, 6, 9. None of these is part of a ring of four or fewer atoms. Therefore none of them is a suitable P for the purposes of Theorem 1. However if any bond is deleted, the remaining structure yields to Theorem 1. For example, if either the bond 2,7 or 5,7 is selected as bond B and deleted, the algorithm deletes the other (5,7 or 2,7 respectively) and applies Theorem 1 to the remaining structure, a fused pair of 5-atom rings. In the remaining structure the unforked path 9,1,2,3,8 is of length 4 and it is in a 5-atom ring; since for this path P

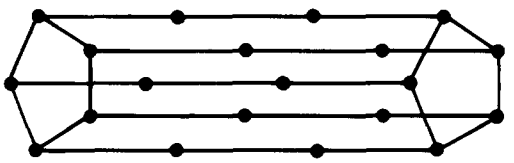
$$|R| = 5 \leqq 8 = 2|P|$$

is satisfied, the algorithm records the ring R: (9,1,2,3,8) and deletes the path P, leaving a single 5-atom ring (4,5,6,9,8) which it also records. The reduced structure left after the deletion of the bond B has been found to have an SSSR of two 5-atom rings.

It remains to verify that the shortest ring through B has at least five atoms. This is true: the algorithm finds that the shortest ring through B has six atoms. Therefore the algorithm decides that either of the two 6-atom rings through B, together with the two 5-atom rings (1,2,3,8,9) and (4,5,6,9,8), constitute an SSSR for the original structure.

If the bond selected for deletion as bond B had been the bond between atoms #8 and #9, the remaining structure would by Theorem 1 have been found to have the SSSR consisting of the 6-atom ring (1,2,7,5,6,9) and the 7-atom ring (1,2,3,8,4,5,6,9). The test that follows would have failed, because the shortest ring through B has only five atoms. The program would have sought another bond to delete. This situation demonstrates the need for the test involving the shortest ring through the deleted bond. Were this test not applied, the program would have reported an SSSR with rings of length 5, 6, and 7, instead of the correct values of 5, 5, and 6. The pivotal fact is that the correct SSSR may, and in this example does, have to contain two or more rings through some bond—in this case the bond between atoms #8 and #9—and such a bond may not be used as bond B. The test involving the shortest ring through B ensures that the B used is not such a bond.

Structure 9472A of the Ring Index is the only example found so far for which the algorithm fails. The algorithm selects as B one of the 15 bonds in the five parallel edges connecting the pentagons at the ends of the prism. Deletion of B leaves a structure with no suitable unforked



Ring Index #9472A

path P for applying Theorem 1. If the algorithm were constructed so as to select one of the other 10 bonds as B, its deletion would open the gates for repeated applications of Theorem 1, but the structure remaining after the deletion of B would be found to have four 8-atom rings, all longer than the shortest ring through B, and the algorithm would therefore fail anyway.

FINDING THE RINGS OF K

Theorem 2. Given an SSSR for a structure, if R_m is a longest ring of the SSSR and if P is an unforked path that is part of R_m but not of any other ring of the SSSR, then the rings of the class K that pass through P are the rings of length $|R_m|$ that pass through P. These are the shortest rings through P.

Proof. Since P does not occur in any other ring of the SSSR, any ring R through P is linearly independent of the remaining—i.e., other than R_m—rings of the SSSR. If $|R| = |R_m|$, then the ordering-by-length that produces the given SSSR need only be rearranged to the extent that the ring R come first among the rings of length $|R_m|$, and R will be in the SSSR corresponding to the new ordering-by-length. Therefore R is in K. If on the other hand R is a ring through P and $|R| > |R_m|$, R cannot be in any SSSR, because no SSSR contains a ring longer than R_m. The third case, $|R| < |R_m|$, is impossible because if it were so, R_m could not be in the given SSSR. The proof is complete.

The algorithm selects a longest unforked path P, consisting of a single bond if necessary, through a longest ring R_m in the SSSR which was found earlier. By a procedure by now familiar, it finds all the rings of length $|R_m|$ through P, deletes P, and repeats. It is not necessary to find a new SSSR for the reduced structure—it is enough to delete R_m from the original SSSR. After $d(C)$ iterations there are no rings left and all the rings of the class K have been found.

EFFICACY

As was explained above under Examples, the algorithm for finding a SSSR requires certain conditions, not always fulfilled. If the programs produce a set of rings, that set of rings is the class K, but for certain compounds the program will produce a statement of inability instead. Only one such compound, exhibited above under Examples, is known; none was encountered in the present CIDS files of 50,000 compounds.

The programs that find the rings of class K and all other rings of eight or fewer atoms consist of approximately 1000 FORTRAN statements and ordinarily take between ½ and ¾ seconds per compound on the IBM 7040 computer, starting from an arbitrary connection table.

ACKNOWLEDGMENTS

The author is indebted to Norman London and Ali Semsarzadeh who in writing the programs made many of the algorithm design decisions and to C. T. Van Meter for guidance in chemistry.

REFERENCES

(1) Welch, J. T., Jr., "A Mechanical Analysis of the Cyclic Structure of Undirected Linear Graphs," *Journal for the Association for Computing Machinery*, **13**, 205–10 (1966).
(2) Gibbs, N. E., "A Cycle Generation Algorithm for Finite Undirected Linear Graphs," *ibid.*, **16**, 564–8 (1969).
(3) Gotlieb, C. C., and D. S. Corneil, "Algorithms for Finding a Fundamental Set of Cycles for an Undirected Linear Graph," *Communications of the Association for Computing Machinery*, **10**, 780–3 (1967).
(4) Paton, K., "An Algorithm for Finding a Fundamental Set of Cycles of a Graph," *ibid.*, **12**, 514–8 (1969).
(5) Fugmann, R., U. Dölling, and H. Nickelsen, "A Topological Approach to the Problem of Ring Structures," *Angewandte Chemie*, **6**, 723–33 (1967), in English.
(6) Halmos, P. R., "Finite Dimensional Vector Spaces," Princeton University Press, Princeton, N. J., pp 10–11 (1948).
(7) Harary, F., "Graph Theory," Addison-Wesley, p 64, Theorem 7.1, 1969.
(8) Patterson, A. M., L. T. Capell, and D. F. Walker, "The Ring Index," 2nd ed., ACS 1960.