Computer Storage and Retrieval of Generic Chemical Structures in Patents. 9. An Algorithm To Find the Extended Set of Smallest Rings in Structurally Explicit Generics

GEOFFREY M. DOWNS, VALERIE J. GILLET, JOHN D. HOLLIDAY, and MICHAEL F. LYNCH*
Department of Information Studies, University of Sheffield, Sheffield S10 2TN, U.K.

Received October 18, 1988

This paper reports how Zamora's smallest set of smallest rings algorithm has been modified and extended to provide an algorithm that will find the extended set of smallest rings (ESSR) for specific and structurally explicit generic structures. Modifications are necessary to find the ESSR rings within a partial structure connection table, while extensions are required to perceive the ESSR rings that span several such connection tables. Particular care is involved when structures containing primary cut faces or alternative embedments are processed.

INTRODUCTION

Two previous papers review ring perception algorithms for specific structures[1] and discuss the theoretical considerations of ring perception leading to the development of the extended set of smallest rings (ESSR) concept.[2] They consider aspects concerning specific structures and general theory and hence are not included in this series of papers on generic structure handling; however, they do form a foundation for this paper. Consequently, the reader is referred to them for the definition and description of many of the terms and concepts used here. A more comprehensive treatment of the algorithm described in this paper is to be found in reference 3. The next paper in this series[4] shows how the ring information produced by this algorithm is represented and accumulated.

The previous parts of this series[5-12] outline the storage and retrieval environment in which ring perception is required and into which this algorithm has been integrated.

In summary, the storage aspect of this environment consists of input of the generic structure information via the GENSAL language,[6] which is then interpreted[10] to produce an extended connection table representation (ECTR).[8] Structure diagrams and specific nomenclatural terms are converted directly to the partial connection table format, while generic nomenclatural terms are converted to lists of structural parameters.[9] Expressions that cannot be converted to parameters lists, such as "electron withdrawing group", are left as text.

The ECTR is a logical framework linking the various partial structures described in GENSAL in a way that allows automatic derivation of search representations. It has a treelike arrangement in which the internal nodes represent the logical relationship (AND or OR) between the branches, and the leaf nodes contain the partial structure connection tables, parameter lists, or text. The main structural diagram found in most chemical patents forms the root partial structure of the tree. Connections to other partial structures are designated as parental if they go up the tree toward the root or child if they go down the tree away from the root (graph theoretical trees are upside down!).

From a processing point of view, the three levels of GENSAL description can be seen to translate into the following: full generic structures have ECTRs containing both logical relationships (AND, OR) and all three partial structure types (connection table, parameters, text); structurally explicit structures have ECTRs with both relationships but only connection table partial structures; and specific structures have ECTRs with only the AND relationship and connection table partial structures.

The retrieval aspect derives search representations from the ECTR in a variety of ways. Fragment screens are generated automatically from the ECTR by applying the FRAGGEN[9] path-tracing procedures to the partial structure connection tables and by invoking the topological grammar based procedures of TOPOGRAM[7] to process parameter lists. The resultant fragments are posted in one or both of the two bitstrings (POSS and MUST) used to represent the structure.

In place of the atom-by-atom search, a relaxation algorithm search strategy[13] is used and has been further evaluated in a parallel processing implementation.[11] In addition, there is the technique of simplifying structures to the reduced-graph form.[12]

This paper reports a ring perception algorithm developed to find the ESSR in structurally explicit parts of generic structures, i.e., all ECTR leaf nodes containing connection tables. Consequently, the algorithm also processes specific structures, but it will not process the structurally implicit parameter lists (or text) of full generics; work on these will be reported in the future. The algorithm described here is incorporated into the RINGGEN program that also produces ring screens from the lists of perceived rings, as reported in the next paper in this series.[4]

For specific and structurally explicit parts of generic structures it is necessary to find the ESSR rings contained wholly within a partial structure connection table of the ECTR, i.e., the **intra-PS** rings, and also those rings that span two or more partial structure connection tables, i.e., the **inter-PS** rings. The ring perception must also be consistent no matter how the structure is split up (**partitioned**) over partial structures or ordered in terms of the partial structure relationships (**oriented**) and irrespective of whether the inter-PS rings are represented by general, combined, or single substitution. In the discussion on inter-PS rings, there are simple examples to show why the ESSR is the most appropriate ring set to use to maintain a consistent ring analysis in such a complex environment.

The generation, handling, and storage of rings can be accomplished by either vector or path-tracing and array techniques. The nature of the ECTR and its splitting into partial structures precludes the use of a vector technique since the vectors would have to contain at least the number of vertices of the maximum number of partial structures that can contribute to a ring system. In the worst case all partial structures might contribute to a ring system. There is no limit to the number of partial structures possible in an ECTR, and so the dimension of the required vectors is potentially infinite. The use of vectors and sets of unknown, and possibly infinite, dimension would cause severe computational problems!

The alternative is to use a path-tracing technique with storage of the ring sequences in arrays of predetermined length. For ring perception in the ECTR, the array has an arbitrary 32 positions, the same as the arbitrary limit on the number of rows in a partial structure connection table. Each vertex in the ECTR (whether it represents an atom or a variable node) can be given a unique number by using the formula

$$U = N_i + ((\mathrm{PSNO} - 1) \times 32)$$

where U is the unique number, N_i is the partial structure row

0095-2338/89/1629-0207$01.50/0

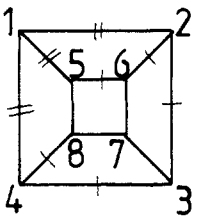


Figure 1.

number for vertex i, and PSNO is the partial structure number. Hence, rings can be stored in an array of integers of length equal to the ring size (up to 32). The use of an array also retains the order of the ring sequence, which a vector does not. This is a very useful attribute for testing the ring perception procedures, although not necessary for the subsequent analysis of the rings to produce suitable descriptors for the search representations, such as ring screens.

PERCEPTION OF INTRA-PS RINGS

In light of the review of published ring perception algorithms,[1] the smallest set of smallest rings (SSSR) algorithm by Zamora[14] seemed the most appropriate for extension to find the ESSR and to find it in a generic environment.

In summary, to find the three classes of rings described by Zamora, the three stages of processing are as follows:

Phase 1, To Include All Vertices in the Ring Set. Initialize all **atom-used** and **bond-used** values associated with the vertices and edges to 0, to denote them as unused. Starting from an unused vertex of highest ring connectivity, the **start atom**, trace the smallest ring associated with that vertex. Set an upper limit to the length of the path trace equal to the size of the smallest ring found so far for that vertex (initially set to the number of vertices in the connection table). For all vertices and edges in the smallest ring, increment their atom-used and bond-used values, respectively, and store the ring. If there is a choice between smallest rings, then apply heuristics based on the numbers of used vertices, used edges, heteroatoms, and the connectivity sum to select just one of them. Continue until all vertices have nonzero atom-used values. Check the number of cycles found; if it equals the **nullity** (Frerejacque number), then terminate; otherwise, continue.

Phase 2, To Include All Edges in the Ring Set. If there are any unused edges left (i.e., with bond-used values of 0), then similarly trace the smallest ring associated with each of these edges and terminate if the nullity is reached.

Phase 3, To Include All Faces in the Ring Set. If there are any unfound faces, then trace these faces by following paths with all bond-used values equal to 1 (the bond-used limit), or at most one bond-used value of more than 1, and with all vertices with a ring connectivity greater than 2.

For manual analysis, bond-used values are marked on diagrams by means of the appropriate number of slashes on the edge concerned.

The selection criteria used by Zamora to differentiate between symmetrically equivalent smallest rings about the same start atom are unnecessary when the ESSR rather than an SSSR is traced. If more than one smallest ring is found from a start atom, then all are included in the ring set. The result in most cases is the generation of all symmetrical equivalents by an algorithm that is faster and uses fewer start atoms than the original SSSR version. For instance, in cubane (Figure 1) three four-edged faces are traced from the first start atom (vertex 1), all are assigned to the ESSR, and the bond-used values are set as shown. There is now only one unused atom left (vertex 7), from which the other three four-edged faces are found.

With the omission of the selection criteria the nullity becomes the *minimum* number of rings that must be found, and so the ring count check can be left until the end of the algorithm to act solely as an error check.

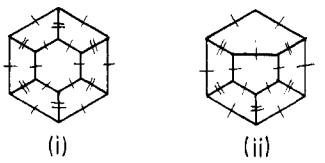


Figure 2.

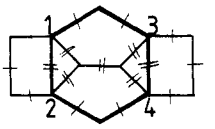


Figure 3.

Another modification concerns the connectivity sum equation used to determine the order of start-atom selection. Zamora's original equation is

$$C_i = 64(K_i) + L_i$$

where C_i is the ring connectivity sum of vertex i, K_i is 1, 8, or 64 depending on whether vertex i has 2, 3, or 4 incident ring edges, respectively (4 being an arbitrary limit), and L_i is the sum of the K values of the congeners to i. Powers of 2 are used presumably in the belief that it enhances the differentiation between various vertex connectivity and congener connectivity combinations to ensure a unique connectivity sum for each. Further investigation shows that it is not necessary to use powers of 2 and that equally good results are obtained by using the simpler equation

$$C_i = (ND_i) + \sum D_j$$

where N is the maximum ring connectivity (6 for the ECTR), D_i is the ring connectivity of vertex i, and D_j is the ring connectivity of vertex j, a congener of i. The main consideration is that the contribution from i should not be less than the sum of the contributions from its congeners. Since a simple unitary difference will achieve this, the use of a power series is redundant.

With these minor modifications, Zamora's phases 1 and 2 have become phases 1 and 2 of the ESSR algorithm used in RINGGEN. The major modifications have occurred with respect to phase 3 and the addition of further stages to cater for inter-PS rings.

In Zamora's phase 3, an unfound face is defined as one in which all vertices have a ring connectivity greater than 2 and only one edge at most with a bond-used value of more than 1. These are very limiting conditions and lead to the failures of the algorithm instanced by Zamora. For ESSR tracing, the first condition will result in the infinite region[2] being detected in certain cases, such as the structure in Figure 2i, but not in others, such as in Figure 2ii. This problem is eliminated by relaxing the condition so that only the start atom needs to have a ring connectivity of more than 2. The second condition, that of limiting the bond-used values, causes failure in cases such as in Figure 3, where the six-edged simple face (marked in bold) is not found (a simple face such as this will be referred to as a maximal infinite region, since we are interested only in simple cycles, and the true maximal infinite region contains the Nachbarpunkte 1,2 and 3,4[2]). This failure is eliminated by extending Zamora's ideas and implementing a two-part phase 3.

The bond-used value that can be exceeded only in certain circumstances will be referred to as the **bond-used limit**. Where this may be exceeded by one edge in the ring only, it will be referred to as a **single** bond-used limit. In Zamora's phase 3,

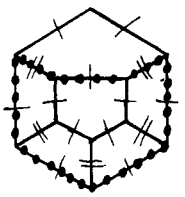


Figure 4.

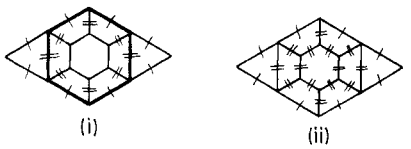


Figure 5.

the second condition controlling path tracing is a single bond-used limit of 1. This has been retained in phase 3-1. For phase 3-2 the new concept of a **rolling** bond-used limit is used. This signifies that more than one edge with a bond-used value greater than the limit is allowed, but no two such edges can be adjacent in the ring sequence.

Phase 3-1 traces unfound faces in a similar manner to Zamora's phase 3. The single bond-used limit is used, but only the start atom must have a ring connectivity of more than 2. In most circumstances an initial single bond-used limit of 1 is adequate. However, under certain circumstances where alternative embedment regions[2] are involved, this phase needs to be repeated with a bond-used limit incremented above 1, as will be shown later. The atom-used values are incremented as each ring is found, but the bond-used values are incremented at the end of phase 3-1 to enable tracing of all symmetrical equivalents. Retention of the single bond-used limit, initially set to 1, is necessary to restrict tracing to simple faces; otherwise, all secondary cut faces[2] will also be traced. This can be shown by consideration of the example in Figure 4, in which the dotted edges highlight a secondary cut face. After phases 1 and 2, the bond-used values are as shown, and the outer six-edged and inner five-edged regions still have to be found. A bond-used limit of 1 will prevent jumping from the inner five-edged region to the outer six-edged region and back to form a simple cycle corresponding to a secondary cut face, such as that emphasized by the dots.

Phase 3-2 is a repetition of phase 3-1, but with the single bond-used limit altered to a rolling bond-used limit, so that more than one edge can have been used more than the limit, but no two such edges should be adjacent on the path. Such a rolling criterion effectively prevents the crossing of bridges. As with phase 3-1, the limit is initially set to 1, but may be progressively incremented above 1 if phase 3 needs to be repeated for alternative embedment regions. This phase will trace those simple faces missed by phase 3-1 due to the use of the single bond-used limit. Their effects are complementary, but phase 3-1 must be conducted first to prevent the rolling bond-used path trace jumping from inner to outer regions and back again, as shown in Figure 4. This complementary effect is illustrated by the structure in Figure 5i. The bold six-edged region will not be found by phase 3-1 or Zamora's phase 3 due to the presence of two edges with a bond-used value of 2. Phase 3-1 will find the inner six-edged region and increment the bond-used values accordingly, as shown in Figure 5ii. The rolling bond-used limit of phase 3-2 will now allow the tracing of the bold six-edged region.

These basic modifications will enable the intra-PS ESSR rings to be found for most structures. The three-phase processing outlined so far defaults to finding all regions of the maximal Schlegel projection of one embedment, and so for structures with unique embedments without primary cut faces[2] the ESSR is found correctly. However, problems can arise

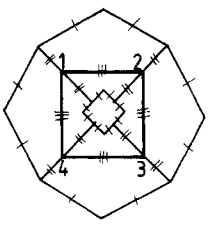


Figure 6.

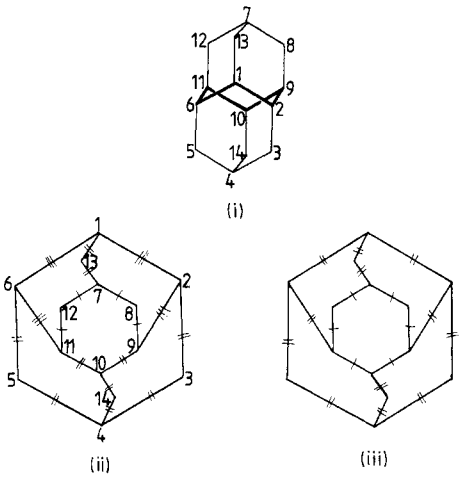


Figure 7.

where structures contain class II primary cut faces (i.e., those that are the same size as one or more adjacent regions), alternative embedment regions, or both. Let us first consider the effects of primary cut faces and then all alternative regions.

Class I primary cut faces (i.e., those that are smaller than all adjacent regions) present no problems for processing. Since they are smaller than the surrounding simple faces, they will be found first, but not to the detriment of finding these simple faces. The structure in Figure 6 shows processing at the end of phase 1. The class I cut face (1,2,3,4) has not hindered the tracing of all adjacent simple faces and has left two simple-faces for phase 3-1 to trace.

Class II primary cut faces can cause difficulties. It is necessary to detect them and decrement their bond-used values before phase 3 so that tracing the remaining simple faces can proceed normally in phase 3. This can be achieved by checking the ring sequences in the ring list after phase 1. By this stage, one characteristic of these cut faces is that their start atom has incident edges with bond-used values greater than or equal to its ring connectivity and other incident edges, not in the the ring sequence, with bond-used values less than its ring connectivity. If, however, all bond-used values are less than the start-atom ring connectivity, then no class II cut faces can have been traced from that start atom. Similarly, if all incident edges have the same bond-used values, then only simple faces can have been traced.

A further characteristic that needs to be used where there are also symmetrically equivalent simple faces from the same start atom is that the cut face has the highest bond-used value sum across the sequence when compared with the simple faces.

For example, consider the situation for the structure in Figure 7i. Phase 1 finds the rings (1,2,3,4,5,6), (1,2,9,8,7,13), (1,2,9,10,11,16), (1,13,7,12,11,6), (14,10,9,2,3,4), and (14,10,11,6,5,4), to leave the bond-used values as shown in Figure 7ii. The class II cut face (1,2,9,10,11,16) has been found to the detriment of the simple face (7,8,9,10,11,12), which now has two adjacent edges with bond-used values of 2 and so cannot be traced by further processing. However, vertex 1 has a ring connectivity of 3, two incident edges, (1,6) and (1,2), with bond-used values of 3, and one incident edge,

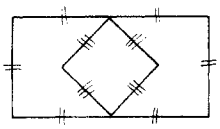


Figure 8.

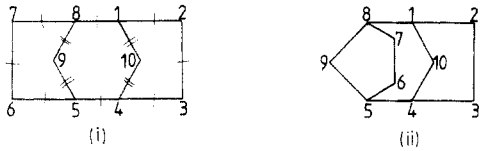


Figure 9.

(1,13), with a bond-used value less than 3. Both (1,2,3,4,5,6) and (1,2,9,10,11,6) exhibit the first class II cut cut-face characteristic, but the former ring is a simple face. However, (1,2,9,10,11,6) has a higher bond-used value sum and so can be flagged as a cut face and the bond-used values decremented, as shown in Figure 7iii. Phase 2 has no unused edges to trace from, but phase 3-1 can now trace (7,8,9,10,11,12), which has bond-used values of 1 on all edges.

Alternative embedment regions can also be difficult to trace correctly due to two general situations. For the first of these situations, consider the structure in Figure 8. The bond-used values are shown after phase 1. The presence of the five-edged alternative embedment regions has incremented all bond-used values to more than 1. Phase 3 cannot now operate with a bond-used limit of 1. However, repeating phase 3 with an incremented bond-used limit (up to 1 less than the maximum bond-used value present after phase 1) overcomes the effect of the alternative embedment regions on the tracing, and the remaining regions can be traced. In the example given, the bond-used limit is incremented to 2, and phase 3-1 will trace the six-edged infinite region. No further repeats are necessary since a further increment gives a bond-used limit of 3.

For the second situation, consider the structure in Figure 9. The embedment given in Figure 9i is the maximal Schlegel projection and is the default that the processing assumes when tracing and subsequently incrementing the bond-used values. The situation is shown after phases 1 and 2; it can readily be seen that phase 3-1 will trace the eight-edged maximal infinite region to leave all bond-used values at 2. The two seven-edged alternative embedment regions (revealed in Figure 9ii) have not been traced. However, a characteristic of structures of this kind is that after phases 1 and 2, one or more rings in the set will have one sequence of high bond-used values and one sequence of lower bond-used values. In Figure 9i the two five-edged regions have this characteristic. By fusion to the other rings in the set that have two or more edges in common with them, the missing alternative embedment regions are produced. The example given is structure number 23 in the test database that was used, and so they are referred to as type 23 or **T23** rings. In terms of the theory presented in the previous paper,[2] this situation arises where two or more cut-vertex pairs are incident to a common region and where there are two or more unlinked paths between the cut vertices of each cut-vertex pair. Where more than two such cut-vertex pairs are involved around the same region, all combinations of fusion of the T23 rings with that region must be taken.

As with the theory, more complex structures are combinations of the general cases discussed here. The phased structure of the processing tackles them as a series of subproblems of increasing complexity. Repetition of certain phases, with incremented parameters, and the incorporation of procedures to identify particular situations are appropriate in practical terms for the majority of ring systems. Consideration has also been given to exceptionally complex theoretical examples in which problems might occur. These will be mentioned in the discussion at the end of this paper.

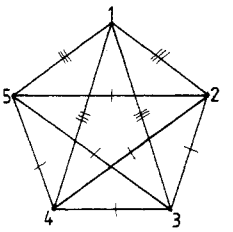


Figure 10.

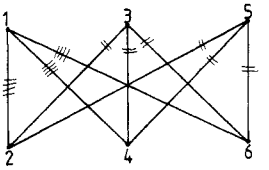


Figure 11.

Although the ESSR is based upon the concepts of finite and infinite simple faces and cut faces, which are not valid in a nonplanar environment, the algorithm will find an analogous ring set. This can be shown by consideration of the two key nonplanar structures K_5 and $K_{3,3}$. The importance of these is that all nonplanar structures must contain a substructure homeomorphic to at least one of them.

K_5 is the quartic graph shown in Figure 10 and has a nullity of 6 (although the Cauchy formula is not intended for use with nonplanar structures). From vertex 1, phase 1 finds the six three-edged rings, (1,2,3), (1,3,4), (1,4,5), (1,2,4), (1,3,5), and (1,2,5), to leave the bond-used values as shown, while phase 3-1 finds the other four three-edged rings, (2,3,5), (2,3,4), (4,5,2), and (4,5,3), from vertices 2 and 4.

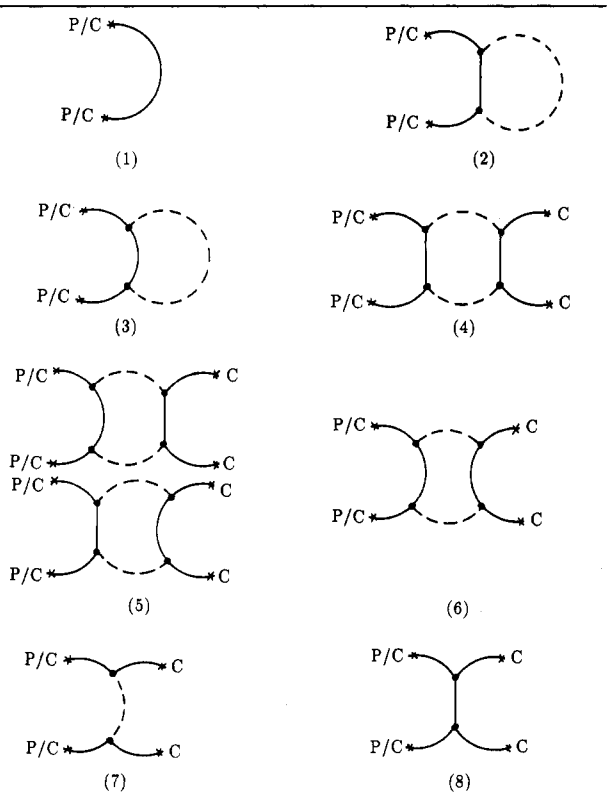
$K_{3,3}$ is the cubic graph given in Figure 11 and requires the repetition of phase 3. Phase 1 traces the six four-edged rings (1,2,3,4), (1,2,3,6), (1,2,5,6), (1,2,5,4), (1,4,5,6), and (1,4,3,6). As shown in the figure, after phase 1, there are no edges with a bond-used value of 0 or 1, and so phases 2 and 3 can trace no further rings on a first iteration. However, the maximum bond-used value is 4, and there are edges present with bond-used values of 2. Phase 3 is repeated with a bond-used limit of 2 to find the two four-edged rings (2,3,4,5) and (2,3,6,5) and then with a bond-used limit of 3 to find the remaining ring (3,4,5,6).

PERCEPTION OF INTER-PS RINGS

This section considers how the ESSR algorithm for determining intra-PS rings can be mimicked to find the inter-PS rings in a consistent way. The basic extension is known as phase 4 and is an integral part of the depth-first trace through the ECTR. In each partial structure the intra-PS rings found by phases 1–3 and inter-PS part-ring paths found by phase 4 are traced on the way down the ECTR. Thus, tracing the intra-PS rings and inter-PS part-ring paths is a top-down operation. However, the inter-PS rings can only be assembled at each level on the way back up, and so construction of the inter-PS rings, by linking the relevant part-ring paths, is a bottom-up operation. These operations are incorporated within the screen assignment and bubble-up procedures to be described in the next paper in this series.[4]

A major concern is that the same structure may be divided into identical sets of partial structures, but these may be given in GENSAL in a different order or **orientation**. Similarly the same structure may be split into different sets of partial structures; i.e., it may be **partitioned** in different ways. Ring perception must be both orientation and partition independent so that no matter how identical structures are input they will always yield the same ring set.

Table I. Possible Occurrences of Inter-PS Part-Ring Paths[a]



[a] P = Parental attachment; C = child attachment.

As the name suggests, phase 4 was originally implemented after phases 1–3; however, later developments to accommodate primary cut faces and alternative embedment regions have led to its insertion between phases 2 and 3.

The aim of phase 4 is to trace the part-ring paths that pass through each partial structure. The part-ring paths can either pass up into the parent or down into the children and will be referred to as **PARLINK** and **CHILINK** paths, respectively. These are then joined appropriately to form the inter-PS equivalents of the intra-PS rings found by phases 1–3. The processing within phase 4 reflects the processing of the other three phases and occurs in two stages, phases 4-1 and 4-2. Phases 1 and 2 are designed to find the smallest intra-PS rings associated with each vertex and edge, respectively, and are mimicked by phase 4-1, which finds the shortest paths between each pair of PARLINK or CHILINK connections. These inter-PS part-ring paths are labeled and stored as the **SHORTEST** paths for the associated pair of connections. Phase 3 is designed to find the maximal infinite-region simple cycles and their symmetrical equivalents and is mimicked by phase 4-2, which finds those paths between each pair of PARLINK or CHILINK connections that obey the single and rolling bond-used limits. These inter-PS part-ring paths are labeled and stored as the **MAXIMAL** paths for the appropriate pair of attachments.

The possible occurrences of part-ring paths within any one partial structure can be generalized to those given in Table I. In these diagrams the asterisks represent the points of attachment to the parent (P) or child (C), the dotted lines indicate the possible presence of intra-PS rings, and the straight edges connect Nachbarpunkte vertices.[2] The diagrams are thus simplified and generalized cut-vertex graphs in which only the shortest paths between the parent or child connections are fully represented since there must always be at least one path between each pair.

Diagram 1 is representative of situations where the partial structure has no intra-PS rings but has one pair of either PARLINK or CHILINK connections. There can only be one SHORTEST part-ring path and no MAXIMAL paths through this partial structure.

Diagram 2 has one pair of iether PARLINK or CHILINK connections, but the SHORTEST part-ring path has incident intra-PS rings; however, the fusion points are Nachbarpunkte, and so there is no possibility of a simple cycle being part of both the part-ring path and the intra-PS rings. In effect, the situation reduces to that of diagram 1 with the addition of intra-PS rings being traced separately by phases 1–3.

Diagram 3 shows a similar situation, but there are either more than two intra-PS ring to part-ring path fusion points, or if there are just two fusion points, then they are not Nachbarpunkte. The PARLINK or CHILINK connections will thus have at least one SHORTEST path and the possibility of one or more MAXIMAL paths.

Diagram 4 and subsequent diagrams have more than one pair of double connections; there can be either one or no PARLINK pair, while there can be any number of CHILINK connections due to the nature of the ECTR. In this case there are two part-ring paths, both of which have Nachbarpunkte fusion points with the intra-PS ring system, thus effectively reducing the problem to two separately processable occurrences of the type in diagram 2.

Diagram 5 combines the occurrence of part-ring paths with and without Nachbarpunkte fusion points. A distinction is made between the two possibilities for the PARLINK connections, if present.

Where the PARLINK path has Nachbarpunkte fusion points with the intra-PS ring system, or there is no PARLINK double connection, any inter-PS rings spanning this partial structure and any of its children will terminate in this partial structure. These inter-PS rings can be completed by combining the various CHILINK and child part-ring paths in the knowledge that they cannot affect any inter-PS rings with the other connections; i.e., no MAXIMAL inter-PS rings can pass through this partial structure and up into the parent or down into the other child (the situation of more than one non-Nachbarpunkte fused CHILINK path is given in diagram 6).

Where the PARLINK path does not have a Nachbarpunkte fusion, the intra-PS infinite region will be part of the PARLINK:ROLLING path and must therefore be fused with it. In this situation, the CHILINK path has Nachbarpunkte, and so the PARLINK and CHILINK paths are effectively separated again so that no MAXIMAL inter-PS rings can pass through this partial structure.

Diagram 6 has PARLINK and CHILINK connections or several CHILINK connections, all of which have non-Nachbarpunkte fusions to the intra-PS ring system. The SHORTEST paths of each PARLINK and CHILINK connection can encompass parts of the intra-PS ring system, while the MAXIMAL paths, by incorporating most of the intra-PS infinite-region simple cycle, will also incorporate part of the other connections' SHORTEST paths. Hence, the MAXIMAL inter-PS rings will pass through this partial structure, up into the parent and down into the children.

Diagram 7 depicts those situations in which there are no intra-PS rings, and the PARLINK or CHILINK SHORTEST paths have at least three vertices in common. Thus, any MAXIMAL inter-PS rings will pass through this partial structure without any PARLINK or CHILINK MAXIMAL paths being involved.

Diagram 8 shows a similar situation, but any MAXIMAL inter-PS rings are prevented from crossing through the partial structure by the Nachbarpunkte fusion of the PARLINK or CHILINK SHORTEST paths.

More complex situations of inter-PS part-ring paths are combinations of the generalizations given above. The main problem to be tackled is that the inter-PS rings can span many

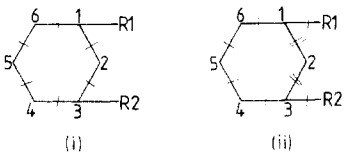


Figure 12.

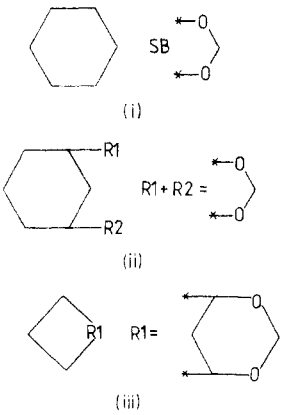


Figure 13.

levels of partial structures depending on the way the structure has been split. This is particularly true of the MAXIMAL inter-PS rings.

The SHORTEST and MAXIMAL part-ring paths associated with each PARLINK and CHILINK connection are found by the two-stage phase 4 as follows:

Phase 4-1 traces the SHORTEST paths between a pair of inter-PS connections (in a breadth-first manner) and increments the bond-used value of each edge contained in these shortest paths by 1. For instance, for PS1 (partial structure 1) in Figure 12i, the bond-used values after phases 1 and 2 are as shown. After phase 4-1, they are as given in Figure 12ii, and the path (R1,1,2,3,R2) is stored as the SHORTEST path associated with R1 + R2.

Phase 4-2 traces all paths between a pair of inter-PS connections that fulfill first the single and then the rolling bond-used limit of phase 3. In Figure 12ii the path (R1,1,6,5,4,3,R2) is traced by using the single bond-used limit to leave all bond-used values at two, thus terminating the tracing (the rolling bond-used limit is not used in this case). This path will be stored as the MAXIMAL path associated with R1 + R2.

It is not necessary to repeat phase 4-1 since the first iteration will guarantee finding all of the SHORTEST paths. However, the MAXIMAL paths may not have been found due to the presence of symmetric SHORTEST paths incrementing the bond-used values to more than the bond-used limit of 1. In such cases, the bond-used limit is incremented in the same way as for phase 3.

After phase 4 is complete, the bond-used values are returned to their pre-phase 4 values to enable phase 3 to proceed correctly.

Double connections to the parent do not cause any problems since they are always represented in the same way; however, double connections to children may be represented in three different ways:

- by general substitution, SB, as in Figure 13i
- by combined substituents, Rx + Ry, as in Figure 13ii
- by a single substituent, Rx, as in Figure 13iii

Doubly connected single substituents do not require phase 4 processing to trace their SHORTEST and MAXIMAL paths since they are found by phases 1–3 and appear as variable groups within the intra-PS rings.

The order in which the depth-first trace proceeds from a partial structure is governed by the nature of its children.

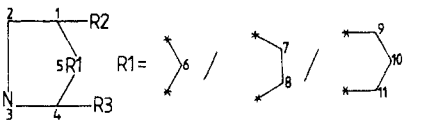


Figure 14.

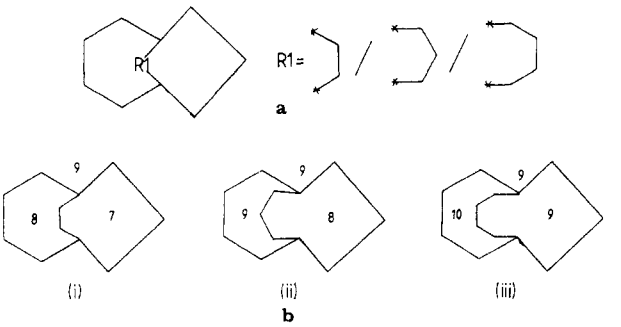


Figure 15.

Consider the simple example in Figure 14, in which R1 lies on a path between R2 + R3. Due to the alternative expansions for R1, the status of the R2 + R3 CHILINK SHORTEST and MAXIMAL paths will change. For the first alternative, (1,6,4) is designated the R2 + R3 CHILINK:SHORTEST path and (1,2,3,4) the CHILINK:MAXIMAL. However, for the second alternative, the paths (1,7,8,4) and (1,2,3,4) are of equal length, and both must be designated CHILINK: SHORTEST; there is no CHILINK:MAXIMAL for this expansion. Furthermore, for the third alternative, the classification is reversed; the path containing R1 is now longer than the other path, and so (1,9,10,11,4) is designated CHILINK:MAXIMAL, while (1,2,3,4) is the CHILINK: SHORTEST. Overall, the CHILINK:SHORTEST for R2 + R3 will contain a 3-vertex all-carbon, a 4-vertex all-carbon, and a 4-vertex heteroatom path, while the PARLINK: ROLLING list will contain a 4-vertex heteroatom and a 5-vertex all-carbon path. All of these will be flagged as alternative to each other.

This trivial example shows the necessity of processing doubly connected single substituents before assigning the SHORTEST and MAXIMAL paths to other doubly connected substituents. The order of processing from a particular partial structure is (1) process doubly connected single substituents, (2) apply phases 1–4 to find all intra-PS rings and all PARLINK and CHILINK SHORTEST and MAXIMAL paths for this partial structure, and (3) process all remaining substituents, inserting the expansion of the previously obtained doubly connected single substituent as appropriate.

Another interesting aspect of doubly connected single substituents is that their expansion can alter the status of the intra-PS rings. This has immediate consequences for the ring set and reveals one reason for using the ESSR in preference to an SSSR or the set of $\mathcal{K}$-rings (see reference 1). Consider the structure in Figure 15a, for instance. The variable group, R1, has alternative expansions of two, three, and four vertices. The specifics covered by this generic are given in Figure 15b, each of which has a nullity of 2. Figure 15bi shows the first alternative, in which the maximal infinite region is the nine-edged cycle. The seven- and eight-edged rings constitute the only SSSR and are hence the $\mathcal{K}$-rings; the nine-edged infinite region is a non-$\mathcal{K}$-ring. For the second alternative, as given in Figure 15bii, the original nine-edged region is symmetrically equivalent to one of the finite regions. Hence, both nine-edged rings and the eight-edged ring are $\mathcal{K}$-rings. Figure 15biii gives the third alternative, in which the original nine-edged region is symmetrically equivalent to the smallest ring and hence is a $\mathcal{K}$-ring, but the other region is now the maximal infinite region and is a non-$\mathcal{K}$-ring.

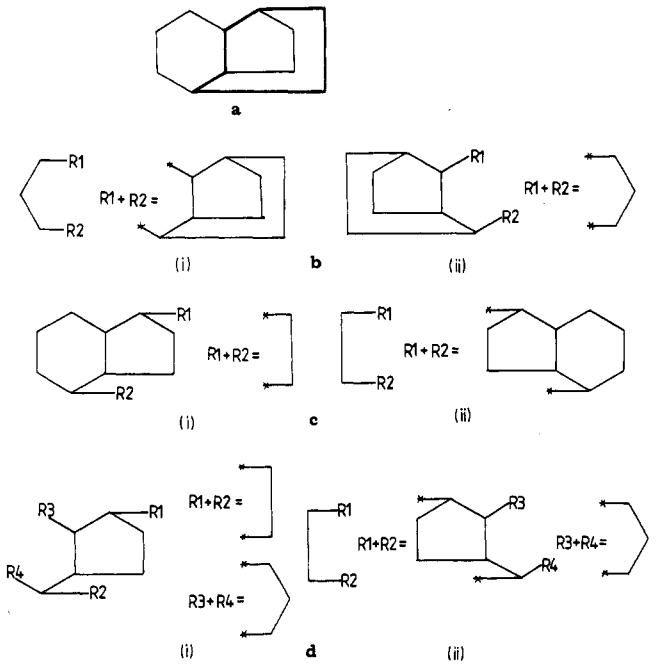


Figure 16.

In this example, the only constant part of the generic form of the structure is the nine-edged maximal infinite region of Figure 15a. If an SSSR or the set of $\mathcal{K}$-rings were used instead of the ESSR, then the nine-edged ring would not be common to the ring analyses for the three possible expansions of R1. Extension of ring perception into structurally explicit generics requires the emphasis of such areas of commonality to achieve better screening performance. Inclusion of the infinite region in the ESSR requires more complex processing but leads to a more accurate representation of such generic structures.

As mentioned earlier, another problem that needs to be addressed when generics are processed is that of orientation and partitioning. This becomes apparent only when complex structures are considered since most structures are simple enough to have unique embedments or easily processable alternative embedments. In terms of finding the ESSR particular problems may be caused where primary cut faces are split over several partial structures. For instance, the structure in Figure 16a contains a class II primary cut face (marked in bold). Three different partitionings of this structure are given in Figure 16b–d, each with two different orientations, i and ii. Examination of examples such as these shows that the consistent detection of primary cut faces in structurally explicit generic structures requires additional procedures:

- Increment all SHORTEST paths after phase 4 to enable the primary cut face detection procedure to operate correctly, with the repetition of phase 4-2 if a cut face is found.
- Check the PARLINK:SHORTEST paths, after all children have been processed, and fuse them to the intra-PS infinite region rings with three or more vertices in common with these paths to form new, replacement or additional PARLINK:MAXIMAL paths for passing back up to the parent.
- Progressively fuse the intra-PS infinite region rings to any MAXIMAL inter-PS rings crossing into the children to form the MAXIMAL inter-PS ring incorporating this partial structure or a larger infinite region ready for fusion to the PARLINK:SHORTEST path for passing up to the parent.
- Fuse the child MAXIMAL inter-PS rings to each other, if they have three or more vertices in common with each other but not with the intra-PS infinite region, to create any MAXIMAL inter-PS rings that pass through this partial

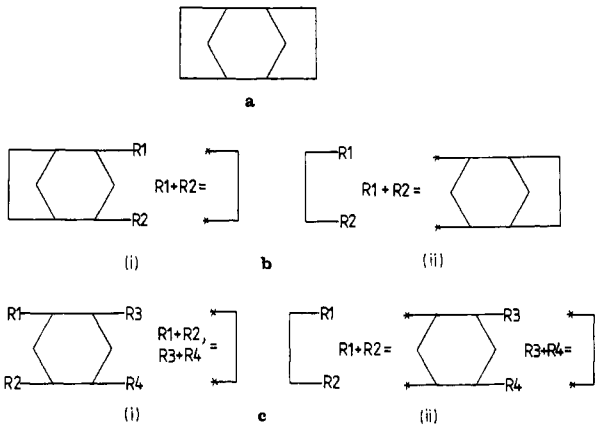


Figure 17.

structure but do not have three or more vertices in common with its intra-PS rings.

The consistent detection of all alternative embedment region has been examined by using different orientations and partitionings of structures such as that given in Figure 17a (the T23 example used earlier). Parts b and c of Figure 17 are two different partitionings of Figure 17a, each with two different orientations, i and ii. To cater for alternative embedment region (T23) structures such as these requires additional procedures:

- Increment all SHORTEST paths after phase 4 to enable the correct operation of the T23 detection procedure.
- Successively fuse the T23 ring with other intra-PS rings and with MAXIMAL inter-PS rings that have three or more vertices in common, with retention of the old rings and flagging of all fused rings as T23.
- Check the PARLINK:SHORTEST paths for fusion with the T23 rings to form new PARLINK:MAXIMAL paths, once again retaining any old paths and flagging the new as T23.

DISCUSSION AND CONCLUSIONS

The intra-PS ring detection of the ESSR algorithm presented here has been developed with the aid of a database of 40 selected specific structures (see reference 2 and its Appendix of the DBR database structures) and has been tested against the Project's database of 1200 specific structures derived from patents and against 49 000 specific structures from the *Fine Chemicals Directory*. The inter-PS ring detection has been developed with the aid of a database of 32 constructed structurally explicit generics (referred to as the DBX database) and has been tested against the Project's database of 77 structurally explicit generics derived from patents. The DBX database consists of a variety of structures that have been constructed to test for consistency in processing alternative embedments and primary cut faces and different partitionings, orientations, sizes, and variable group composition. The development databases include structures of unusual complexity, while the test databases are generally much simpler and are more representative of "real life". All the development and test structures are processed correctly, and the algorithm is currently being evaluated by Internationale Dokumentationsgesellschaft für Chemic mbH (IDC).

The ESSR algorithm as presented works on the full graph of a structure. In the theoretical considerations of ring perception[2] the importance of the interplay between areas exhibiting 2- and 3-polytopal characteristics is revealed. It can be shown that the intra-PS part of the ESSR algorithm (i.e., phases 1–3) can run into problems with complex theoretical examples in which there is a multiple occurrence of different-length unlinked paths (see reference 2) through 2-polytopal areas of the graph. This problem is removed by simplifying

the full graph to the cut-vertex graph form introduced in reference 2. Working on the cut-vertex graph has the advantage of separating the perception of simple faces from the perception of primary cut faces and greatly simplifies the perception in both cases. With minor modifications to the path tracing to prevent secondary cut-face tracing through 3-polytopal areas (linked path regions), the ESSR algorithm can be applied directly to the cut-vertex graph to find all alternative embedment regions, i.e., all possible simple faces. The primary cut faces can be found by using each simple-face start atom to trace any rings of smaller or equal size in the full graph. If they are not already in the ring list, then these are the primary cut faces.

It has been verified manually that all cut-vertex graphs of specific structures of the DBR database will process correctly and in a manner simpler than that of the full graphs. Similarly the cut-vertex graphs used to develop the theory (of which a selection is given in an appendix to reference 2) give no problems.

Thus, the ESSR algorithm presented here will work on the full graph of the vast majority of specific and structurally explicit generic structures to find the correct ESSR, but additional procedures are necessary, and particular care is required, to cater for alternative embedment regions and primary cut faces. If exceptionally complex structures are to be processed, then the same basic algorithm, with a few minor modifications, can be used on the cut-vertex graph to yield the correct ESSR in all cases.

ACKNOWLEDGMENT

This research was made possible through funding provided by IDC (Internationale Dokumentationsgesellschaft für Chemie mbH), whose staff also gave valuable support through discussions. In addition, we are grateful to Fraser Williams, Scientific Systems Ltd., for providing the *Fine Chemicals Directory* and the associated connection table software.

BIBLIOGRAPHY

(1) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Review of Ring Perception Algorithms for Chemical Graphs. *J. Chem. Inf. Comput. Sci.* (first of four papers in this issue).
(2) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Theoretical Aspects of Ring Perception and Development of the Extended Set of Smallest Rings Concept. *J. Chem. Inf. Comput. Sci.* (second of four papers in this issue).
(3) Downs, G. M. Computer storage and retrieval of generic structures in patents: ring perception and screening to extend the search capabilities. Ph.D. Thesis, University of Sheffield, March 1988; Chapters 6 (RINGGEN - a program for finding the ESSR in an ECTR) and 7 (Perception of rings crossing between partial structures (inter-PS rings)).
(4) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 10. Assignment and Logical Bubble-Up of Ring Screens for Structurally Explicit Generics. *J. Chem. Inf. Comput. Sci.* (fourth of four papers in this issue).
(5) Lynch, M. F.; Barnard, J. M.; Welford, S. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 1. Introduction and General Strategy. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 148–150.
(6) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 2. GENSAL, a Formal Language for the Description of Generic Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 151–161.
(7) Welford, S. M.; Lynch, M. F.; Barnard, J. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 3. Chemical Grammars and Their Role in the Manipulation of Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 161–168.
(8) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 4. An Extended Connection Table Representation for Generic Structures. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 160–164.
(9) Welford, S. M.; Lynch, M. F.; Barnard, J. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 5. Algorithmic Generation of Fragment Descriptors for Generic Structure Screening. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 57–66.
(10) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 6. An Interpreter Program for the Generic Structure Description Language GENSAL. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 66–71.
(11) Gillet, V.; Welford, S. M.; Lynch, M. F.; Willet, P.; Barnard, J. M.; Downs, G. M.; Manson, G.; Thompson, J. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 7. Parallel Simulation of a Relaxation Algorithm for Chemical Substructure Search. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 118–126.
(12) Gillet, V. J.; Downs, G. M.; Ling, A.; Lynch, M. F.; Venkataram, P.; Wood, J. V.; Dethlefsen, W. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 8. Reduced Chemical Graphs and Their Applications in Generic Chemical Structure Retrieval. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 126–137.
(13) von Scholley, A. A Relaxation Algorithm for Generic Chemical Structure Screening. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 235–241.
(14) Zamora, A. An Algorithm for Finding the Smallest Set of Smallest Rings. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 40–43.