of Molecular Structures. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 171-177.
(7) Wipke, W. T.; Dyott, T. M. Stereochemically Unique Naming Algorithm. *J. Am. Chem. Soc.* **1974**, *96*, 4834-4842.
(8) Jochum, C.; Gasteiger, J. Canonical Numbering and Constitutional Symmetry. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 113-117.
(9) Schubert, W.; Ugi, I. Constitutional Symmetry and Unique Description of Molecules. *J. Am. Chem. Soc.* **1978**, *100*, 37-41.
(10) Masinter, L. M.; Sridharan, N. S.; Carhart, R. E.; Smith, D. H. Applications of Artificial Intelligence for Chemical Inference XIII. Labeling of Objects Having Symmetry. *J. Am. Chem. Soc.* **1974**, *96*, 7714-7723.
(11) Carhart, R. E. Erroneous Claims Concerning the Perception of Topological Symmetry. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 108-110.
(12) Dyott, T. M.; Hove, W. J. Canonical Numbering. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 187-187.
(13) Shelley, C. A.; Munk, M. J. An Approach to the Assignment of Canonical Connection Tables and Topological Symetry Perceptions. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 247-250.
(14) Bersohn, M. A Sum Algorithm for Numbering the Atoms of a Molecule. *Comput. Chem.* **1978**, *3*, 113-116.
(15) Heap, B. R. The Production of Graphs by Computer. In *Graph Theory and Computing*; Academic Press: New York, 1972; pp 47-62.
(16) Baker, H. H.; Dewdney, A. K.; Szilard, A. L. Generation the Nine-Point Graphs. *Math. Comp.* **1974**, *127*, 833-838.
(17) Bussemaker, F. S.; Cobejlič, S.; Cvetkovič, L. M.; Seidel, J. J. Computing Investigation of Cubic Graphs; Technical Report 76-WSK-01; Technical University Eindhowen: Eindhowen, 1976.
(18) Arlazarov, V. L.; Zuev, I. I.; Uskov, A. V.; Faradzhev, I. A. Algorithm for Transformation of Finite Nonoriented Graphs to Canonical Form. *Zn. Vychisl. Mat. Mat. Fiz.* **1974**, *14*, 737-743 (in Russian).
(19) Lederberg, J. *Computation of Molecular Formulas for Mass Spectroscopy*; Holden-Day: San Francisco, 1964.
(20) Read, R. C. The Enumeration of Acyclic Chemical Compounds. In *Chemical Application of Graph Theory*; Balaban, A. T., Ed.; Academic Press: London, 1976; pp 25-61.
(21) Trinajstič, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1983; Vol. II.
(22) Harary, F. *Graph Theory*; Addison-Wesley: Reading, MA, 1969.
(23) Lindsay, R. K.; Buchanan, B. G.; Feigenbaum, E. A.; Lederberg, J. *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project*; McGraw-Hill: New York, 1980.
(24) Balaban, A. T. Enumeration of Cyclic Graphs. In *Chemical Application of Graph Theory*; Balaban, A. T., Ed.; Academic Press: London, 1976; pp 63-105.
(25) Gray, N. A. B. *Computer Assisted Structure Elucidation*; Wiley: New York, 1986.
(26) Faradzhev, I. A. *Algorithmic Investigations in Combinatorics*; Nauka: Moscow, 1978 (in Russian).
(27) Essam, J. W.; Fisher, M. E. Supplement: Some Basic Definitions in Graph Theory. *Rev. Mod. Phys.* **1970**, *42*, 271-288.
(28) Lawler, E. L.; Wood, D. E. Branch and Bound Methods. *J. Oper. Res. Soc. Am.* **1966**, *14*, 217-245.
(29) Harary, F.; Palmer, E. M. *Graphical Enumeration*; Academic Press: New York, 1973.
(30) Rouvray, D. H. Topological Indices as Chemical Behaviour Descriptors. *Congr. Numerantium* **1985**, *49*, 161-179.
(31) Rouvray, D. H. The Modeling of Chemical Phenomena Using Topological Indices. *J. Comput. Chem.* **1987**, *8*, 470-480.
(32) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17-20.
(33) Randič, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609-6615.
(34) Skvorcova, M. I.; Baskin, I. I.; Devdariani, R. O.; Zefirov, N. S. On the Problem of Generation of Structures of Organic Compounds with Prescribed Properties. *Proceedings of 8th All-Union Conference on Application of Computers in Molecular Spectroscopy and Chemical Research*; Novosibirsk Institute of Organic Chemistry, Academy of Sciences of USSR: Novosibirsk, USSR, 1989; pp 250-251 (in Russian).
(35) Zefirov, N. S.; Skvorcova, M. I.; Stankevitch, I. V. Generation of Structures of Polycondensed Benzenoid Hydrocarbons with Given Randič Index. *Proceedings of 8th All-Union Conference on Application of Computers in Molecular Spectroscopy and Chemical Research*; Novosibirsk Institute of Organic Chemistry, Academy of Sciences of USSR: Novosibirsk, USSR, 1989; pp 252-253 (in Russian).
(36) Gordeeva, E. V.; Zefirov, N. S. Solution of Inverse Problem for Wiener and Randič Topological Indices. Programs RING and WING. *Proceedings of 8th All-Union Conference on Application of Computers in Molecular Spectroscopy and Chemical Research*; Novosibirsk Institute of Organic Chemistry, Academy of Sciences of USSR: Novosibirsk, Novosibirsk, USSR, 1989; pp 254-255 (in Russian).

Enhanced Algorithm for Finding the Smallest Set of Smallest Rings

CHENG QIAN,* WILLIAM FISANICK, DALE E. HARTZLER, and STEVEN W. CHAPMAN
Chemical Abstracts Service, P.O. Box 3012, Columbus, Ohio 43210

Received August 21, 1989

The search algorithm for the smallest set of smallest rings (SSSR) has long been an important basic algorithm for processing chemical information. This paper describes the merits and limitations of one of the SSSR search algorithms used at Chemical Abstracts Service (CAS), provides the mathematical basis for the general approach, presents enhancements to this algorithm, and includes a new, more rigorous approach that extends the scope of the original algorithm while reducing its limitations.

I. INTRODUCTION

Isolated rings and isolated joined rings in a chemical structure are referred to as ring systems—an important part of structural topology used to identify and characterize structures. Because of their importance, the ring systems have been reported in *Chemical Abstracts* since 1907. In 1940, *The Ring Index*, a catalog of all known ring systems, was published by the American Chemical Society; a second edition was published in 1960.[1] The current catalog is the Chemical Abstracts Service (CAS) *Ring Systems Handbook*.[2]

In a ring system, all the possible rings, including the envelope rings, form the all-ring set of the system.[3] Although it is the most complete set and so provides exhaustive information about the ring system, the number of rings in a complex system usually makes the all-ring set unsuitable for practical use. Therefore, the topological features of the ring system are typically characterized by a subset of the all-ring set. However, the subset adopted to describe the ring system is not always the same under different implementations. The smallest set of smallest rings (SSSR)[4] is the ring set most commonly used, although some authors recommend SSSR+, i.e., the SSSR plus other rings.[5,6] [For example, Fujita has suggested a set of rings called the essential set of essential rings (ESER).[7]] All of these ring sets, however, are related: a ring in the all-ring set must be one of the SSSR rings or a linear combination of SSSR rings; SSSR+ sets are subsets of the all-ring set and supersets of an SSSR ring. Therefore, an SSSR or an SSSR+ can be generated by filtering unqualified members from the all-ring set (reduction strategy), or the all-ring set and SSSR+ can be generated by combining SSSR members with joint edges (expansion strategy).

Balaban presented an algorithm to generate the all-ring set of a ring system that uses a homomorphically reduced graph (HRG).[3] The use of the HRG greatly simplifies ring system graphs and makes the reduction strategy feasible even when limited computing resources are available. However, selecting

0095-2338/90/1630-0105$02.50/0

an SSSR or SSSR+ from all rings in a ring system is still very time consuming because candidate rings must be checked for their mutual independency; i.e., none of the rings can be a linear combination of any others. In contrast, the expansion strategy based on an SSSR search algorithm is likely to be less time consuming because of the efficiency of existing SSSR algorithms and because combining SSSR rings is basically straightforward. Unfortunately, most existing SSSR search algorithms are not rigorous.[8-10]

This paper describes the enhancements of an SSSR search algorithm developed at Chemical Abstracts Service (CAS).[10] The enhancements include (1) a modification to provide for extended SSSR rings, (2) modifications to the phase 3 portion of the algorithm to reduce the failure rate, and (3) an improvement of the ring search procedure using breadth-first search. As a comprehensive enhancement, a new, more rigorous algorithm that extends the general approach while removing its remaining limitations is presented along with the mathematical basis that supports the approach and the enhancements.

II. MATHEMATICAL BASIS

In this paper, only simple graphs (i.e., nonoriented connected linear graphs without self-loops and two-node rings) are discussed because simple graphs are used to represent component chemical structures and because an SSSR of a multicomponent structure is a combination of the SSSRs of all the component structures. Discussions are further restricted to ring systems—graphs with nodes and edges that belong to at least one ring. Chain nodes and chain edges of a graph are not essential to a ring identification problem and can be easily removed from a general linear graph.

In this paper, the following terminology applies: "ring" refers to *cycle* or *circuit*, which are often used in graph theory; "N_e" and "N_v" are used to designate number of edges and number of nodes, respectively.

The following definitions are according to graph theory:[11]

1. A ring is a connected linear graph in which every node is of degree 2 (i.e., connectivity 2).

2. The dimension of the ring space of a graph is $N = N_e - N_v + p$, where p is the number of maximum connected subgraphs of the graph. For a connected graph, $p = 1$ and so $N = N_e - N_v + 1$.

3. $N = N_e - N_v + 1$ rings of a connected graph form the basis of the ring space of the graph if none of the rings can be represented as a linear combination of the others. The basis of a graph's ring space is obviously not unique.

4. If a ring contains one edge that is not contained by any member of a linearly independent set of rings, this ring and all rings of the set still form a linearly independent set.

5. Any ring of the graph can be represented as a linear combination of some or all N rings of a basis S, such as $R = S_1 + ... + S_k$, where $k \leq N$. Replacing R for any one of the k basis rings results in another basis.

The following definitions are used in this paper:

Definition 1. Ring size is the number of nodes contained in the ring. The size of ring R is denoted by size (R). A self-loop is a ring of size 1; a back-and-forth ring is a ring of size 2. All other rings must have sizes larger than 2. (Rings of size 1 or 2 probably have little use in chemical structure representations because interactions between two atoms are generally described by a chemical bond. A bond between an atom and itself is meaningless. Two or more bonds between a pair of atoms are sometimes used to represent special interactions but never considered structurally cyclic.)

Definition 2. The size sequence of a ring set is the sequence of sizes for all rings forming the set in ascending order. A set of three rings of sizes 5, 7, and 4 has the size sequence (4,5,7).

Definition 3. Size sequence A is smaller than B if (1) A and B have the same number of elements, (2) the first $K - 1$ elements of A and B form identical sequences, and (3) the Kth element of A is smaller than the Kth element of B.

Definition 4. A set of rings of a ring system graph G is a smallest set of smallest rings (SSSR) if it is a basis of the ring space of G and its size sequence is the smallest among all bases of the ring space. The SSSR of a ring system must contain $N_e - N_v + 1$ rings, and it may not be unique; however, all the SSSRs must have the same ring size sequence. Although Plotkin has given a definition of an SSSR,[12] the definition in this paper is slightly different in expression.

Definition 5. A ring is the smallest ring at edge E of a graph if it has the smallest ring size among all rings containing E in the graph.

With definitions 1–5 and general graph theory principles previously summarized, we have proven the following theorems:

Theorem 1. Suppose that G' is a graph union of a subset S of an SSSR of a graph G, where

$$S = \{S_i | i = 1, 2, ..., K < N_e - N_v + 1\}$$

The members of S and ring $R \notin S$ form a subset of an SSSR of G if R is the smallest ring at an edge $E \subset G\text{–}G'$.

Proof. Since $E \subset G\text{–}G'$ is contained by R but not by any members of S, R is linearly independent of S. However, R can be represented as a linear combination of members of an SSSR $S' = S$; this combination must contain at least one ring, say C, that belongs to S' (but not to S) and contain edge E. So, R and all S' rings except C form a basis, i.e., an SSSR, because S' is an SSSR and R has size (R) $\leq$ size (C).

In a ring system, we can look for the smallest ring at each unused edge until all edges are covered by at least one local smallest ring. According to theorem 1, the local smallest rings form an SSSR, so we have proven that Zamora's phase 1 and 2 algorithms are rigorous.

Theorem 2. Suppose that S is a subset of an SSSR of a graph G, where

$$S = \{S', S''\} = \{S'_1, ..., S'_k, S''_1, ..., S''_h\}$$

and the graph union of S' rings is graph G', a subgraph of G. All S rings and ring $R \notin S$ form a subset of an SSSR of G, if $R \in \{R_i | i = 1, ..., M\}$ has size (R) = min (size (R_1), ..., size (R_M)), where R_i ($i = 1, 2, ..., M$) are rings satisfying $[R_i \cap (G\text{–}G')] \neq 0$ and $[R_i \cap (G\text{–}G')] \neq [(a_1 \cdot S''_1 \oplus \cdots \oplus a_h \cdot S''_h) \cap (G\text{—}G')]$ with a_j ($j = 1, ..., h$) being coefficients of any nontrivial linear combination.

Proof. From the theorem's conditions regarding R_i, ring $R \in \{R_i | i = 1, ..., M\}$ cannot be represented as a linear combination of S rings; in other words, $\{R,S\}$ is a linearly independent set. However, S is a subset of an SSSR $SS = \{S, SS'\}$, so ring R must be a linear combination of SS rings. The combination must include at least one SS' ring, say C, with $C \cap (G\text{–}G') \neq 0$ and $C \cap (G\text{–}G') \neq (a_1 \cdot S''_1 \oplus \cdots \oplus a_h \cdot S''_h) \cap (G\text{–}G')$ for any $a_1, ..., a_h$ because, otherwise, $(R \oplus b_1 \cdot S''_1 \oplus \cdots \oplus b_h \cdot S''_h) \cap (G\text{–}G') = 0$ for certain b_i, which is contradictory to $R \in \{R_i | i = 1, ..., M\}$. Therefore, ring R, having size (R) $\leq$ size (C), can be substituted for C in SS to form another SSSR that has a subset consisting of R and all S rings.

Theorem 2 has shown that the ring R of theorem 2 must be an additional SSSR ring if it exists. However, it may not exist for certain fragmentations of S into S' and S''. Therefore, an appropriate subset S'' from S needs to be selected so that theorem 2 can be employed. It can be proven that for the ring R to exist, the necessary and sufficient condition of S'' is that the difference of ring space dimensions of G and G' must be greater than the number of member rings in the set S''. Later in this paper, an algorithm for finding additional SSSR rings that makes use of this condition will be discussed.

Lemma 1. Suppose S is a subset of an SSSR of graph G and ring $S1$ belongs to S. All S rings and the smallest one among the rings containing at least one, but not exactly all, edge(s) covered only by ring $S1$ form a subset of an SSSR of G. [This is a special case of theorem 2 when subset S'' has only one member ($S1$).]

The nodes and edges for graphs that represent chemical structures usually have attributes (atom types, bond types, etc.). Therefore, the idea of the SSSR is sometimes further extended to include this additional node/edge information.

Definition 6. The most preferred ring at edge E of a graph G is the ring of the smallest ring size among all rings containing E in G or, if there are more than one of the same size, the one having the most preferred attributes according to a set of preference rules.

Definition 7. The preferred ring sequence is a set of rings sorted in ascending order by ring size and in descending order by the preference of their attributes according to a set of preference rules.

Definition 8. Ring set A is preferred to ring set B if (1) all the members of A and B are sorted in a preferred sequence, (2) the first $K-1$ members of A have the same ring sizes and attributes as their counterparts in B, and (3) the Kth member of A has a smaller size than the Kth member of B or when the sizes are the same preferred attributes.

Definition 9. A basis of a ring system graph is an SSSR under the extended meaning if it is the most preferred one to all other bases of the graph.

A set of preference rules is essential for defining an SSSR under the extended meaning. An example of a preference rule is the ring with the preferred ring elemental formula, defined by (1) the largest number of heteroatoms, (2) the largest variety of heteroatoms, and (3) the largest number of preferred heteroatoms in the descending order of O, S, Se, Te, N, P, As, Sb, Bi, Si, Ge, Sn, Pb, and B.

It is not difficult to extend theorems 1 and 2 and lemma 1 to cover the extended SSSR if the condition of the most preferred ring is added to the condition of the smallest ring size when the size of two rings is the same. However, to guarantee a valid extension, the preference rules must not imply that the choice of the most preferred ring at a certain edge depends on the selection of other SSSR rings identified before or after the choice is made. This is because both the SSSR and the extended SSSR definitions given here are "state functions" (by this definition, one of the bases can be determined as an SSSR or extended SSSR only according to the ring sizes and attributes of the rings in the basis, regardless of the order in which the SSSR rings are found). Any rules related to the order by which the basis rings are found will change the definition to a "function of processes" (the selection of a basis as an extended SSSR is also accounted for by the order in which basis rings are found). For the extended SSSR defined as a state function, all processes following the preference rules must generate SSSRs that have the same smallest ring size sequence and ring attribute sequence (such as ring formula sequence). In contrast, for the SSSR defined as a function of the processes, more than one process may exist—each generating legitimate SSSRs that may have different ring size sequences or different ring attribute sequences. Consequently, such a process-dependent extended SSSR is not well-defined unless rules to select a preferred process are also given.

III. CAS SSSR SEARCH PROGRAM

In the 1970s, CAS developed an SSSR search program that uses the algorithm for SSSR search described by Zamora.[10] Zamora's algorithm for SSSR search consists of three phases. Phases 1 and 2 looked for local smallest rings at each unused node and bond (see discussion of section II). In phase 3, the "unused faces" were identified and selected as SSSR ring candidates. However, it was inappropriately assumed that SSSR rings that remained to be found in phase 3 must be unused faces. In Figure 1, both real and hypothetical structures are presented to demonstrate the possibility that none of the unused faces of a structure is an SSSR ring. It is also possible that an SSSR ring that remains to be found has all edges with a usage count larger than one.

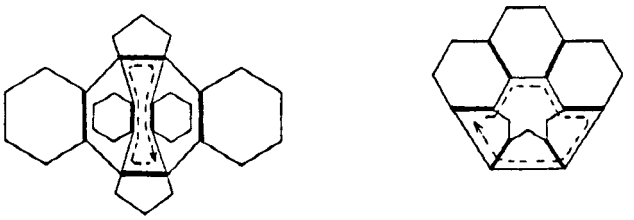

Figure 1. Real (left) and hypothetical structures in which none of the unused faces are SSSR rings: (—) edges of usage count larger than 1; (---) SSSR rings to be found.

Problems with SSSR search results in phase 3 have been detected, some of which are described in ref 10. The problems included incorrect rings being found and failures to find all SSSR rings. Although the percentage rate for these problems was very low, it was obvious that the limitations of the algorithm needed to be removed. Also, even though the algorithm was very efficient for most structures in the CAS Registry File, the CPU time used to locate the SSSR of particular structures was less than optimal. Additionally, the ring elemental analysis data were not available for the initial implementation. Recent enhancement efforts address all of these issues.

IV. ENHANCEMENTS TO THE CAS SSSR SEARCH PROGRAM

Three major enhancements have been made to the existing CAS SSSR search programs.

1. Use of Ring Elemental Analysis Data as Preference Rules of the Extended SSSR. Although Zamora included in his paper a general statement for searching the extended SSSR,[10] this function was not implemented in the original CAS SSSR search program. (The original search program finds an SSSR only with regard to ring sizes.) To enhance the program so that it can search for an extended SSSR carrying ring elemental composition features, we have taken a set of preference rules similar to Dialog's SSSR rules.[13] The following rules apply: (a) Nonoverlapping rings are preferred to overlapping rings, where overlapping is measured by the number of new ring nodes that have been included in the rings already found. (b) Rings with the largest number of heteroatoms are preferred. (c) Rings with the largest variety of heteroatoms are preferred. (d) Rings with the largest number of preferred heteroatoms in the descending order O, S, Se, Te, N, P, As, Sb, Bi, Si, Ge, Sn, Pb, and B are preferred. (e) Rings with the largest number of used atoms are preferred. (f) Rings with the largest number of used edges are preferred. (g) Rings with the greatest connectivity sum are preferred.

These rules diverge from the rules for SSSR used in the CAS *Ring Systems Handbook*[2] in one major way: the *Ring Systems Handbook* takes a nomenclature-dependent approach to determine the overlapping of rings in a bridged ring system.[14] The current SSSR rules are based on a topological interpretation of overlapping and so need less chemistry intelligence.

(We should note here that these rules imply that the extended SSSR is defined as a function of processes. The impact of this implication for finding valid SSSR is described under Discussion.)

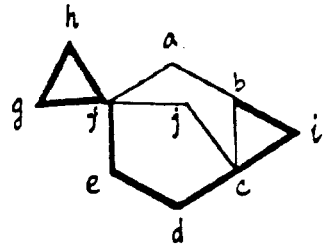

Figure 2. Ring paths can be removed to simplify the graph and avoid phase 3 failure. Isolated spiro ring f–g–h–f: remove bonds g–f, g–h, f–h and nodes g, h. Simple bridge ring c–d–e–f–j–c: remove bonds c–d, d–e, e–f and nodes d, e. Appendage ring b–i–c–b: remove bonds b–i, i–c and node i.

2. Reduction of the Failure Rate of Phase 3. Since phase 3 imposes major limitations on the search program's function, changes have been made to reduce the number of structures that need phase 3 processing and to reduce the failure rate of phase 3.

Only complex structures with rings surrounded by other rings are processed by phase 3. Consequently, if the paths of special rings can be identified and subsequently removed after the rings are discovered in phases 1 and 2, the ring graph is simplified and has less of a chance of being passed to phase 3.

We have found that paths can be removed for at least three types of rings: isolated spiro rings, appendage rings, and simple bridge rings. An isolated spiro ring is a ring having only one node of connectivity greater than two; a simple bridge ring has only two nodes of connectivity greater than two. An appendage ring is a simple bridge ring in which the two nodes of connectivity greater than two are adjacent.

It can be proven that some paths of the three types of rings are never involved in more than one SSSR ring, so these paths can be removed after the SSSR rings involving them are found. For an isolated spiro ring, all bonds and nodes except the one of connectivity greater than two can be removed. For an appendage ring, all nodes and bonds except the two adjacent nodes of connectivity greater than two and the bond between them can be removed. For a simple bridge ring, the less preferred path connecting the two nodes of connectivity greater than two can be removed. The preference can be determined by rules similar to those for SSSR rings. In the current program, in addition to the preference of the shorter length of paths, (b), (c), and (d) of the SSSR preference rules described in section IV.1 are used. Removing the paths simplifies the graph on which remaining SSSR rings are searched for and therefore may avoid failure of phase 3, as shown in Figure 2. The simplification also improves the performance of the program.

Another modification was made to change the definition of an unused face to any rings containing a bond of usage count one. This change substantially reduces the failure rate in phase 3, although it still cannot guarantee that all SSSR rings generated in phase 3 are correct.

3. Improvement of the Ring-Finding Algorithm. The performance of the SSSR search program depends heavily on the performance of the basic ring-finding algorithm. The initial implementation used a depth-first approach that, in general, worked well. However, we have detected cases where many deep paths had been searched before a preferred ring was found. Figure 3 shows one of these cases. When this happened, the CPU time of the search procedure was relatively long—sometimes over 30 min on an IBM 3090. To solve this problem, the depth-first approach was replaced by a breadth-first approach.

The breadth-first algorithm is fairly straightforward. In a search of the most preferred ring containing one particular node, a spanning tree rooted at the node is grown concentrically. If ring closure bonds cannot be found after a new layer of nodes is added, the tree continues to grow. Otherwise, the growth stops and a back-tracing procedure locates all rings containing both the root and a ring closure bond. The most preferred among the rings thus located is then chosen as the search result. In a search of the most preferred ring containing a particular bond, one node connected by the bond is used as the root of the spanning tree, and the back-tracing procedure has to find all rings that contain both the bond and a ring closure bond. The performance of the new ring-finding algorithm was very satisfactory. The new approach achieved improvement of almost 1 order of magnitude in the worst cases and provided about 200% improvement on average. Another advantage of the breadth-first approach is that the new search program works more consistently, regardless of the complexity of the ring system. The CPU time for some of the sample tests is listed in Table I.

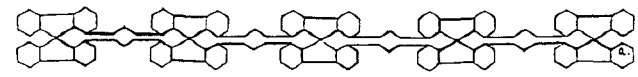

Figure 3. Structure that needs extensive search time to find a smallest ring using depth-first searching. A trace the program searches through, starting from node a, is highlighted.

Table I. CPU Time of SSSR Ring Searching Test

	depth-first		breadth-first	
file	CPU for the job, s	CPU per rings, s	CPU for the job, s	CPU per ring, s
sample 1	33.16	0.066	11.97	0.024
sample 2	139.98	0.279	15.36	0.031
sample 3	17.94	0.036	9.92	0.020

Each of the three files in Table I contains 501 ring systems. Sample 1 has systems of intermediate complexity. File sample 2 contains more complex ring systems. And file sample 3 contains simpler ring systems. It is not surprising to find performance difference for the two algorithms when the sample files are processed. Most complex ring systems contain many nodes, and some have very large envelope rings. The depth-first algorithm has a good chance of finding large-size rings before rings of the smallest size can be found. On the contrary, the breadth-first algorithm always finds the rings of the smallest size first—sometimes along with the smallest size plus one. For simple systems, the performance of the two algorithms becomes closer because there will be fewer paths for the depth-first algorithm to try, while the breadth-first algorithm will still need to spend time building tree branches that do not lead to the smallest rings.

The enhanced SSSR search program was first written in C language on a UNIX operating system and later ported to an IBM mainframe environment. The data in Table I were collected on an IBM 3090.

The enhanced SSSR search program has been used to process the CAS Ring File that contains the unique ring systems in approximately 10 million substances that comprise the CAS Registry File. Processing results show a very low failure rate. Of 166 397 ring structures, only 39 failed phase 3, while 4316 structures (less than 2.6%) entered phase 3. However, the number of structures with incorrect data may still be significant because it is possible that phase 3 was not failed but incorrect SSSR rings were generated. To thoroughly resolve the problem, a mathematically rigorous algorithm was developed. The proposed algorithm is described in the next section.

V. A RIGOROUS ALGORITHM OF SSSR SEARCH

This is a rigorous SSSR search algorithm combining phases 1 and 2 of Zamora's algorithm and a new phase 3 based on

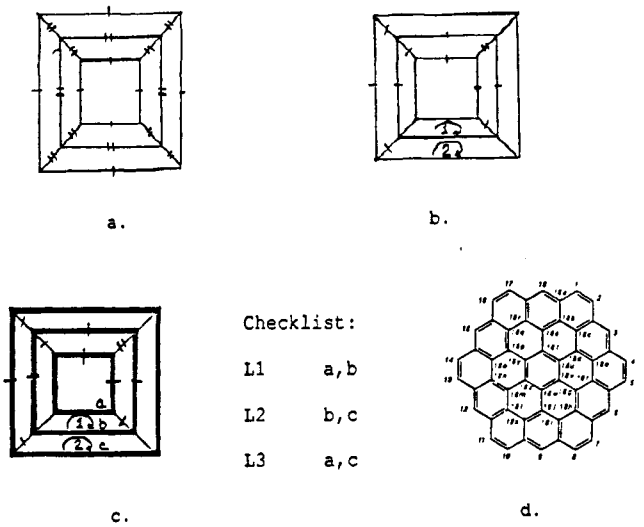

Figure 4. Phase 3 searching: (a) Hypothetical structure after phase 1–2 (usage counts of edges are shown). (b) Beginning of step 1c; rings 1 and 2 have been recorded in the check-ring list. (c) Most preferred ring found at bond a,b,c (highlighted) with checklist shown on the right. (d) Structure of a ring system in the CAS Registry File similar to the hypothetical one.

theorem 2. This algorithm is applicable to a simple graph that represents a ring system in a chemical structure.

Algorithm. Step 1. Calculate the number of SSSR rings of the ring system:

$$N = N_e - N_v + 1$$

Step 2. Perform phases 1 and 2 of Zamora's algorithm to discover as many SSSR rings as possible, that is, until all nodes and edges are covered by at least one ring. Set N_REMAINING to the difference of N and the number of rings discovered. If N_REMAINING is zero, go to step 4.

Step 3. Perform phase 3 to find all remaining SSSR rings.

Step 4. SSSR ring searching completes.

The suggested algorithm for phase 3 is completely different from Zamora's phase 3. The key step of the new phase 3 is to form set S'' (defined in theorem 2) with a minimum number of SSSR rings. To do that, the algorithm first checks the SSSR rings that can be excluded from S'', using the necessary and sufficient condition described in section II (refer to the text following the proof of theorem 2). It then selects the remaining SSSR rings as the S''.

Phase 3 of the SSSR Search Algorithm. Step 1. Find an S'' set to which theorem 2 can be applied: (a) Select the first ring in the SSSR list; let $D = 1$. Mark this ring and its edges and nodes as "used". Also mark the ring as "tried". (b) Let N' equal the number of SSSR rings in the SSSR list minus one. (c) In the SSSR list, select the next untried ring that has at least one node or edge shared with other used rings; calculate $D' = N_e' - N_v'$, where N_e' and N_v' are the numbers of unused edges and nodes of this ring. If all unused rings have been tried, go to step e. (d) If $D + D' < N - (N' - 1)$, mark this ring, as well as its edges and nodes as used, let $D = D + D'$ and $N' = N' - 1$, remove the tried mark from all unused rings in the SSSR list, and go to the step c; otherwise, mark this ring as tried and go to step c. (e) Put all unused SSSR rings in the S'' set.

Step 2. Mark all edges that only belong to the S'' rings identified in step 1.

Step 3. Generate all combinations of the S'' rings by permutation. For each combination, calculate the usage count of every marked edge of the rings involved in the combination and record those of usage count one in one line of a checklist.

Step 4. At each marked edge, search for the most preferred ring from rings of which marked edges do not exactly match those in any lines of the checklist.

Starting from node	Use preference rule (a) of elementary analysis	Without using rule (a)
a	C2N, C6, C5N	C2N, C5N, C5N
b	C5N, C2N, C5N	C5N, C2N, C5N

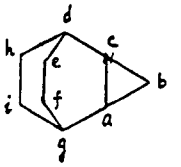

Figure 5. Example of SSSR search results that depends on the first node chosen to start preferred ring search.

Step 5. Among rings found in step 4, choose the most preferred one as an SSSR ring.

Step 6. Set N_REMAINING = N_REMAINING − 1. If N_REMAINING is greater than zero, go to step 1; otherwise, return.

An example of phase 3 searching is given in Figure 4.

The advantage of this algorithm is that a correct SSSR can always be found, regardless of the complexity of structures. This is achieved at the expense of more CPU processing time, especially when structures are so complex that the set S'' becomes large. Fortunately, for ring systems of real chemical structures that need phase 3, the set S'' usually has only one member. Under this situation, theorem 2 is reduced to lemma 1, so the complexity of phase 3 is then largely reduced. (The percentage of ring systems that complete phase 3 with an S'' set of more than two rings is so low that its impact on the total processing performance is negligible.)

VI. DISCUSSION

1. Although the SSSR of a ring system is obviously not unique, each SSSR has the same ring size sequence and ring attribute sequence that is generally implementation independent. However, an extended SSSR can be implementation dependent and even structure representation dependent if some preference rules imply that results are dependent on the order of the SSSR rings being found (refer to section II). (The preference rules a, e, and f described in section IV.1 are of this type.) Therefore, the SSSR that is found may depend on the way the program selects nodes and edges to perform the calculations for phases 1–3. Different implementations of the same set of preference rules may find different SSSRs for the same structure (implementation dependent). On the other hand, even if the same program is used, two identical structures with different atom numbering systems could lead to different results (structures representation dependent), as shown in Figure 5.

This effect has no impact on the uniqueness of the SSSR generated for each structure of the CAS Registry File, because every structure in the CAS Registry File has canonical numbering. However, there may be more than one base set of rings that meet all the rules. The one that the program finds is determined by the process selection rules implied by the program and is not fully defined by the rules so far provided. The solution to this problem could be to include additional rules for process preference or to abandon those rules that cause the problem. (The latter solution might be desirable because the information lost is minor. Sometimes, it may result in a slightly less specific extended SSSR, but that will have little or no effect on most practical uses.)

2. Further improvement of the ring search procedure can be achieved if a ring system is split into subsystems at gen-

eralized spiro nodes. Generalized spiro nodes are ring nodes at which a ring system can be split into separate ring systems without breaking rings. The identification of these nodes is very simple: for each node of connectivity four or greater, temporarily remove all bonds incident to it; if the ring system becomes two or more separate subsystems, the node is a generalized spiro node.

CONCLUSION

The enhanced SSSR search program described here works efficiently and with a low failure rate. This has made it possible to process the entire ring file in the CAS Registry system. Further improvement can be made by the implementation of the proposed new phase 3 in the SSSR search algorithm. With the new phase 3, the whole algorithm becomes mathematically rigorous and program failure can be eliminated.

ACKNOWLEDGMENT

We thank G. G. Vander Stouw for encouragement and constructive discussions and M. E. Olowinski for providing us with CAS Ring File processing data. We also thank the CAS Documentation Turnover Group for editorial support.

REFERENCES

(1) Patterson, A. M.; Capell, L. T.; Walker, D. F. *The Ring Index*, 2nd ed.; American Chemical Society: Washington, DC, 1960.
(2) *Ring Systems Handbook*; Chemical Abstracts Service, American Chemical Society: Washington, DC, 1988.
(3) Balaban, A. T.; Filip, P.; Balaban, T. S. *J. Comput. Chem.* **1985**, *6*, 316-329.
(4) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93-102.
(5) Corey, E. J.; Petersson, G. A. *J. Am. Chem. Soc.* **1972**, *94*, 460-465.
(6) Wipke, W. T.; Dyott, T. M. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 140-147.
(7) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 78-82.
(8) Gasteiger, J.; Jochum, C. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 43-48.
(9) Schmidt, B.; Fleischhauer, J. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 204-206.
(10) Zamora, A. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 40-43.
(11) Mayeda, W. *Graph Theory*; Wiley-Interscience: New York, 1972.
(12) Plotkin, M. *J. Chem. Doc.* **1971**, *11*, 60-63.
(13) File 301 328-331 300 (*Chemname*, *Chemsis*, *Chemzero*), Dialog Bluesheet, revision April 1984, Dialog Information Retrieval Service, 1984.
(14) *Index Guide*, Appendix IV; Chemical Abstracts Service, American Chemical Society: Washington, DC, 1987.

Topological Structural Information in the CAS File: Statistical Occurrences of DARC Concentric Fragments. 1. Basic Carbon Substructures

GERARD CARRIER, ANNICK PANAYE, and JACQUES-ÉMILE DUBOIS*

Institut de Topologie et de Dynamique des Systèmes, associé au CNRS, Université de Paris VII, 75005 Paris, France

Received October 10, 1989

The Chemical Abstracts Service (CAS) Chemical File of some 7.5 million structures, structured in DARC topological space, was analyzed statistically for the frequency of occurrence of different topological substructures. These topological DARC fragments, known as FREL-Bs, are described concentrically around a single carbon atom with an environment limited to the first (A) and second (B) neighbor atoms and treated as a spanning graph tree. Occurrence searches were carried out for the 70 primitive carbon FRELs based on σ bonding. These represent potential acyclic or cyclic substructures with an ordered environment and are called FREL-AC or FREL-CY, respectively. Such ordered environments provide more than mere statistics of local and global descriptions; they lead to trend analyses. Certain broad correlations, linked to the idea of structural neighbors, can thus be observed. Analysis of the statistics associated with cyclic structures revealed some counterintuitive exceptions to the general rule that occurrence should be inversely related to complexity. As an example, the presence of various familiar chemical families such as steroids and terpenes can be detected within the context of the file and can be tagged by means of specific FREL associations.

INTRODUCTION

Many activities in chemical information, computer-assisted documentation (CADoc),[1-4] or computer-assisted design (CAD)[5-10] are based upon the exploitation of sets of substructures selected within a given context. Selection criteria vary with the application, but a search for the basic organizational parameters should help to formalize and elucidate original sets of substructures for new applications. Furthermore, an understanding of these parameters will permit informed improvement of the existing systems.

There is only one global statistical study that is frequently updated with regard to all known structural data.[11,12] This study deals with the familiar molecular fragments such as rings, chains, substituents, and functional groups. The significance of the statistical occurrence of such substructural fragments is complex. The selection of compounds into the database reflects human interest in the structures both for their properties and for the challenge they present to the synthetic chemist. It is because of this study that we can claim a global view of this large database and can use it indirectly to evaluate chemical knowledge. The database and its analysis provide data pertaining to the existence or nonexistence of compounds, and these can be used to draw conclusions as to the ease or lack of ease of access to structural families and even to specific structural entities. This is the basis of present-day chemical taxonomy, which can be exploited with the help of systematic nomenclature in an effort to understand familial relationships that may exist between various groups of chemical compounds.

In this paper, we outline another approach to this global view. This uses atoms, together with their bonds and local topologies, as structural primitives. The aim is a quantitative and qualitative analysis of the local and global structural environment of an atom.

A minimal structural fragment, such as an atom and its neighbors, constitutes the simplest basic fragment, but it is often an inadequate representation of the whole, particularly

0095-2338/90/1630-0110$02.50/0