Automatic Extraction of Ring Substructures from a Chemical Structure

Yoshimasa Takahashi

Department of Knowledge-based Information Engineering, Toyohashi University of Technology, Tempaku, Toyohashi 441, Japan

Received June 29, 1993*

In this paper, a computer algorithm is investigated for the systematic perception and extraction of possible ring substructures: smallest set of smallest rings (SSSR), SSSR-dependent rings, fused rings, spiro ring substructures, and so on. Here, a simple ring with no transannular bonds in the ring system is defined as an elementary ring. The set of elementary rings (SER) consists of a SSSR and the SSSR-dependent simple rings. Systematic extraction of possible ring substructures within a chemical structure can be done using a ring adjacency graph based on the SER. The algorithm has been implemented as a computer protram, RSSGEN. The details of the algorithm will be discussed with a couple of illustrative examples.

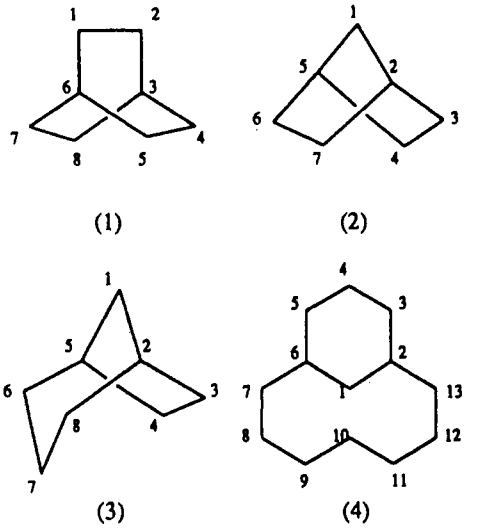
INTRODUCTION

Ring perception is a very important technique for the automatic interpretation of chemical structures. Many algorithms for ring perception within a chemical structure exist.[1] Most of them, however, are just for finding circuits in terms of graph theory. Downs et al. present an exhaustive review of ring perception algorithms.[2] In the review, they categorize the ring perception techniques into three categories, with a single exception, according to their initial approaches as follows: (i) find all possible cycles and then select these required ones, (ii) generate a fundamental basis of cycles from which all other cycles can be derived as necessary, or (iii) determine directly the smallest fundamental basis, known as the smallest set of smallest rings, SSSR. The details are given in their paper. They conclude that each of the algorithms has its disadvantages or failures in terms of efficiency of the algorithm or the completeness of the ring set obtained. They also present an algorithm to find the extended SSSR and its application to generate ring screens for sustructure search in large databases.[3]

In the present work, aiming at the automatic knowledge acquisition for computerized structure design[4] and related areas, an algorithm has been devised for the systematic perception and extraction of possible ring substructures: SSSR, SSSR-dependent rings, bicyclics, tricyclics, and so on.

ALGORITHM

A Set of Elementary Rings. Here, a simple ring with no direct transannular bonds in the ring system is defined as an elementary ring. Therefore, all components of a smallest set of smallest rings (SSSR) are elementary rings. It is known that a SSSR is not an invariant set; that is, a chemical structure often contains additional elementary rings which are not in the SSSR of the structure. The alternative six-membered ring of bicyclo[2,2,2]octane (structure **1**) exemplifies this. The additional elementary rings are referred to as SSSR-dependent elementary rings (SSSR-DER) in the present work. Furthermore, in some cases, additional elementary rings exist, which are larger than any of the rings contained in a SSSR. They are also referred to as SSSR-DER because they are associated with component rings of the SSSR or SSSR-DER(s). Therefore, a SSSR can be regarded as a fundamental basis from which SSSR-DER can be derived. Thus, a set of elementary rings (SER) is defined with a SSSR and the associated SSSR-DER(s).



Finding SSSR-Dependent Elementary Rings. The SSSR is not always unique. For example, there are three possible SSSRs for structure **1**. On the other hand, in some cases, alternative larger rings are found within a structure (e.g., structures **2**–**4**). Structures **2**–**4** yield a unique SSSR, and each of them contains another elementary ring which is not a component of its SSSR: R(2,3,4,5,6,7) for structure **2**, (R(2,3,4,5,6,7,8) for structure **3**, and R(2,3,4,5,6,7,8,9,10,-11,12,13) for structure **4**. These additional elementary rings are referred to as SSSR-DER as mentioned above.

The SSSR-DER can be found easily by introducing the concept of the θ-graph.[5] A θ-graph is defined as a graph in which only two of the vertices, they are not adjacent, have a vertex valence of 3 and all of others have the vertex valence of 2 (Figure 1).

In Figure 1, the θ-graph has three different rings. These are the $(l+m+2)$-membered ring, the $(l+n+2)$-membered ring, and the $(m+n+2)$-membered ring. If $l=m=n$, then two of them arbitrarily can be selected as the components of the SSSR for the graph because they are equivalent rings. For the other case the SSSR of a θ-graph is unique. Now suppose that the SSSR consists of rings A and B which are the components of a SSSR of the graph in Figure 1. In both cases, we can get the third ring with the size of $m+n+2$ drawn in Figure 1. This third ring which is a SSSR-DER can

* Abstract published in *Advance ACS Abstracts,* January 15, 1994.

0095-2338/94/1634-0167$04.50/0

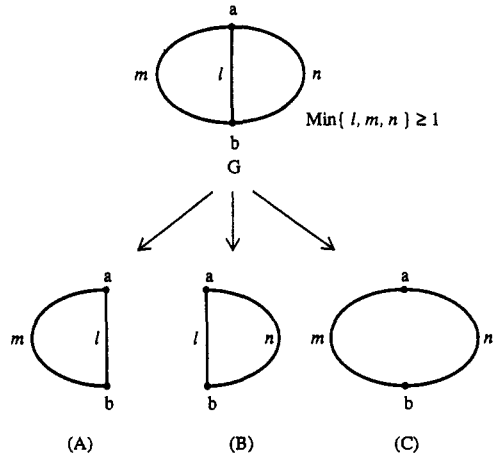


Figure 1. General expression of the θ-graph (G) and cyclic subgraphs (A–C) derived from the θ-graph. l, m, and n express the number of vertices between the vertices a and b, respectively.

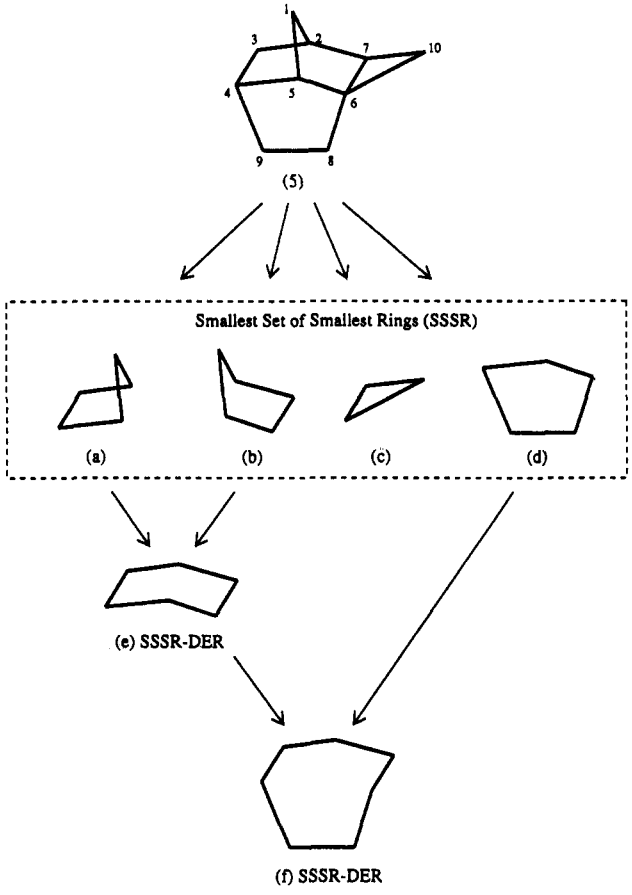


Figure 2. Finding the set of elementary rings, SER. The SER consists of a SSSR and SSSR-dependent elementary rings SSSR-DER(s).

be identified by taking an exclusive sum of the edges of rings A and B. Thus, we can extract the additional elementary rings by finding the θ-graph embedded in an original structure.

We now consider structure **5** in Figure 2. We can get an unique SSSR. The SSSR consists of a three-membered ring and three five-membered rings, as shown in Figure 2.

Besides, the structure contains two different additional elementary rings, SSSR-DERs. One is R(2,3,4,5,6,7) which can be found with R(1,2,3,4,5) and R(1,2,7,6,5) on the basis of the θ-graph mentioned above. The other one R-(2,3,4,9,8,6,7), is derived with R(4,5,6,8,9) and the additional elementary ring R(2,3,4,5,6,7) found in the preceding step. No other different θ-graphs can be generated for any pair of

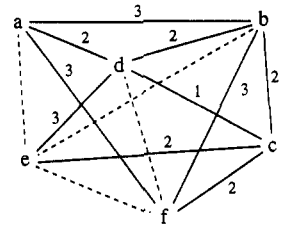


Figure 3. Ring adjacency graph for the SER of structure **5**. The vertices express the elementary rings labeled in Figure 2. A dotted line shows not an adjacency relationship but a dependency relationship between the vertexes (i.e. elementary rings).

elementary rings obtained. Thus the SER of structure **5** consists of six elementary rings. In Figure 2, it is also true that ring e can be derived by exclusive summing of rings a and b without considering the θ-graph. However, such a manner often yields some extra ring like R(1,2,7,10,6,5) from rings b and c, which is not an elementary ring of the original structure. The θ-graph check makes it possible to avoid such a thing.

Ring Adjacency Graph and Its Subgraph Generation. Once the SER is obtained we can generate possible polycyclic ring substructures for a chemical structure. First of all a ring adjacency graph whose vertices correspond to the elementary rings and edges expressing the ring adjacency relations is generated. Every edge is weighted by the number of fused points (atoms) between two rings which correspond to a pair of vertexes on the edge. A ring adjacency graph of structure **5** can be drawn as in Figure 3.

It is clear that each vertex of a ring adjacency graph expresses a simple ring fragment. Here a figure on each edge shows the weighted value mentioned above. Once we get the ring adjacency graph, it is possible to generate larger ring substructures which have two or more elementary rings by subgraph generation for the ring adjacency graph.

It is known that finding all possible rings within a chemical structure is a time consuming task. Not all of them are always needed. The set of rings of our interest often depends on the problem to be approached. The aim of the present paper is not to find circuits of a graph in terms of graph theory but to find possible ring substructures (or ring systems) in a chemical structure.

Supposed that the size of a subgraph is defined by the number of vertices contained in it, a subgraph with the size of 2 corresponds to a bicyclic ring substructure, the size of 3 for a tricyclic ring substructure, and so on because each vertex of the ring adjacency graph corresponds to an elementary ring. Thus, a subgraph with the size of q would correspond to a q-cyclic ring substructure. A single edge subgraph with the edge weighted value of 1 expresses a spiro ring system with two elementary rings. If we focus on the edges which are weighted with a larger value than 1, then only ring substructures without a spiro ring system are generated. In other words, in such a case, the ring adjacency graph is revised. It is clear that the revised ring adjacency graph can become much simpler than the original one. By means of subgraph generation for the ring adjacency graph, we can generate and extract possible ring substructures which have the specified number of rings.

IMPLEMENTATION AND RESULTS

The algorithm described above was implemented as a computer program RSSGEN written in FORTRAN77 on a Data General AV-6000 UNIX workstation. For finding the SSSR within a chemical structure we employ the procedure

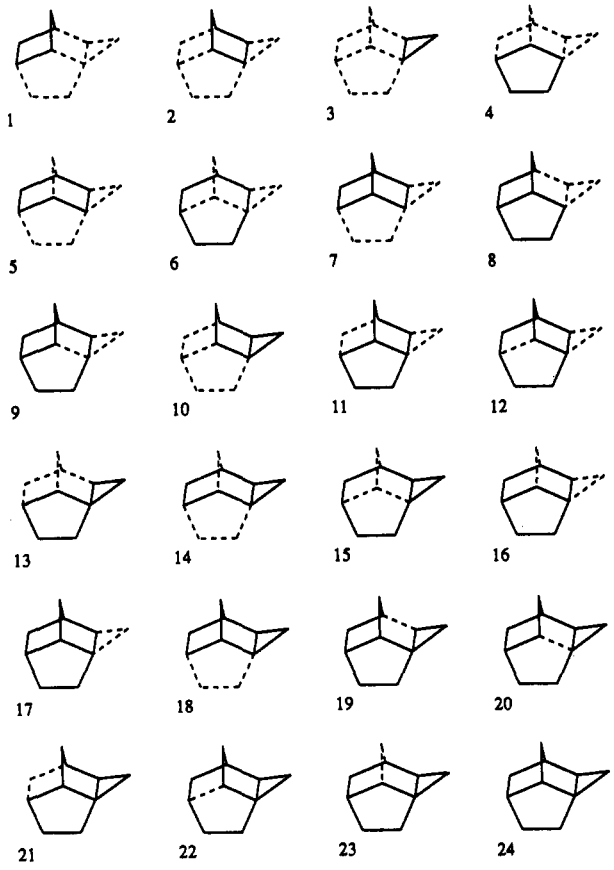


Figure 4. All possible ring substructures of structure **5** derived from the SER. The first four simple rings (1–4) are part of the SSSR, and 5 and 6 are SSSR-DER. These are monocyclic ring substructures. Rings 7–16 are bicyclics. Rings 17–23 are tricyclics. Ring 24 is tetracyclic, which is the full structure in this case.

reported by Schmidt and Fleischhauser[6] with some modifications. Detected components of the SSSR are stored as elementary rings for the structure. Finding additional elementary rings is based on the search for θ-graphs which are constructed from a pair of elementary rings already found. From the SER a ring adjacency graph is generated. A ring connectivity matrix is used as the internal representation of the ring adjacency graph. The algorithm for the subgraph generation of a ring adjacency graph is based on the procedures of GROWSUB,[7] a program which generates all possible substructures from a given chemical structure. RSSGEN has two different modes for finding possible ring substructures. A run-time switch controls whether spiro ring systems are ignored in the substructure generation process or not.

As a first example, the results of the ring substructure analysis of structure **5** are shown in Figure 4. RSSGEN finds twenty-four ring substructures for this compound. The result shows that they consist of six elementary rings (four SSSR members and two SSS-DERs), ten bicyclic ring substructures, seven tricyclic ring substructures, and a tetracyclic one that is the maximal one and the full structure in this case.

A second example was generated from structure **7** (in Figure 5) which is a natural alkaloid compound, gelcemine. It was tested in the two different modes (spiro systems ignored/included). The results are summarized in Table I. In the case of included spiro rings, RSSGEN found eighty-nine ring substructures ranging from monocyclics to hexacyclics. When spiro ring systems are excluded, thirty-seven ring substructures (up to four rings in a system) are found. Four other structures (**6** and **8–10**) were also analyzed by RSSGEN. Their results without spiro ring systems are summarized in Table II together

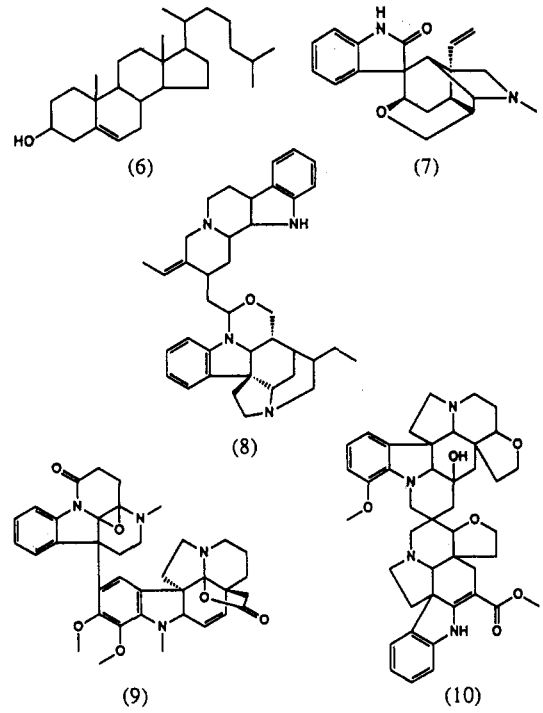


Figure 5. Polycyclic structures tested by RSSGEN.

Table I. Results of Ring (R) Substructure Analysis of Gelcemine (Structure **7**)

spiro ring	SSSR	DER	2R	3R	4R	5R	6R	all
not included	6	4	17	9	1	0	0	37
included	6	4	21	26	22	9	1	89[a]

[a] All of these graphical drawing are available as supplementary materials.

Table II. Results of Ring (R) Substructure Analysis of Polycyclic Structures Listed in Figure 5

structure	SSSR	DER	2R	3R	4R	5R	6R	7R	all
6	4	0	3	2	1				10
7	6	4	17	9	1				37
8	10	2	17	20	16	7	1		73
10	13	0	14	16	15	11	5	1	75

with that of gelcemine obtained above. The required total processing time is less than 1.0 s in every case presented here.

Limitations of RSSGEN. The current RSSGEN program uses Schmidt and Fleischhauser's procedure for finding a SSSR which is used as an initial set of simple rings to get the SER. They described in their paper that if a cyclic subgraph G of length l is surrounded by cyclic subgraph G' of length l' ($l' < l$), then the cyclic subgraph G will not be found. RSSGEN still contains this problem. The procedure can be replaced with the other sophisticated approach by Downs et al. or by Qian et al., whose approach is described in their recent work.[8] It will be done in a future work. In ring substructure generation from a given adjacency graph, if an embedded ring is formed due to being surrounded by the elementary rings RSSGEN, then RSSGEN gives us the number of rings lower by one than that of the real ring system. These are checked by Euler's formula on cyclic graphs.

CONCLUSION

The set of elementary rings, SER, which consists of a SSSR and the SSSR-dependent elementary rings (SSSR-DERs) is defined. The SSSR-DER is defined as an additional simple

ring which can be derived from a pair of components of the SSSR or from a pair of one of the components and the detected SSSR-DER. An algorithm for finding the SSSR-DER based on the θ-graph is discussed. Procedures using the SER to generate and extract possible ring substructures for a chemical structure are described with a couple of illustrative examples. The implemented computer program, RSSGEN, is useful for the systematic ring analysis of polycyclic chemical structures. The presented algorithm is for planar chemical graphs. Therefore it should be noted that some problems concerning polyhedral structures remain to be solved.

ACKNOWLEDGMENT

This work was supported by Special Coordination Funds for Promoting Science and Technology, Science and Technology Agency of Japan.

Supplementary Material Available: Figures showing possible ring structures and text describing the various drawings (6 pages). Ordering information is given on any current masthead page.

REFERENCES AND NOTES

(1) See ref 2 and its cited references.
(2) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 172–187.
(3) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 187–206. Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 207–214. Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 215–224.
(4) Takahashi, Y.; Hosokawa, K.; Yoshida, F.; Ozaki, M.; Sasaki, S. *Anal. Chim. Acta* **1989**, *217*, 61–83.
(5) Harary, F. Graph Theory (Japanese Translation). Kyoritsu Shuppan, Tokyo, 1966.
(6) Schmidt, B.; Fleischhauer, J. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 204–206.
(7) Takahashi, Y.; Satoh, Y.; Suzuki, H.; Sasaki, S. *Anal. Sci.* **1986**, *2*, 321–323.
(8) Qian, C.; Fisanick, W.; Hartzler, D. E.; Chapman, S. W. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 105–110.